Lidar Engineering

Introduction to Basic Principles

GARY G. GIMMESTAD
Georgia Tech Research Institute

DAVID W. ROBERTS
MicroDynamics LLC

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9780521198516

DOI: 10.1017/9781139014106

First published 2023

A catalogue record for this publication is available from the British Library

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-0-521-19851-6 Hardback

Additional resources for this publication at www.cambridge.org/9780521198516

Lidar Engineering

Explore the spectrum of lidar engineering in this one-of-a-kind introduction. For the first time, this multidisciplinary resource covers all the scientific and engineering aspects of atmospheric lidar – including atmospheric science, spectroscopy, lasers and eye safety, classical optics and electro-optics, electrical and mechanical engineering, and software algorithms – in a single comprehensive and authoritative undergraduate textbook. Discover up-to-date material not included in any other book, including simple treatments of the lidar crossover range and depolarization in lidar signals, an improved explanation of lidar data inversion algorithms, digital signal processing applications in lidar, and statistical limitations of lidar signal-to-noise ratios. This is an ideal stand-alone text for students seeking a thorough grounding in lidar, whether through a taught course or self-study.

Gary G. Gimmestad is an instructor in professional education at the Georgia Institute of Technology, Atlanta, USA. He is a fellow of the Institute of Electrical and Electronics Engineers (IEEE), the Optical Society of America (OSA), the American Association for the Advancement of Science (AAAS), and the Society of Photo-Optical Instrumentation Engineers (SPIE) and a Fulbright scholar. He has twice received Order of Merit awards for service to the lidar community for organizing and presenting Lidar Tutorials at International Laser Radar Conferences.

David W. Roberts is Chief Engineer at MicroDynamics LLC, Woodstock, Georgia, USA. He worked for the Georgia Tech Research Institute's (GTRI's) Electro-Optical Systems Laboratory for 30 years where he developed innovative atmospheric lidar systems for measuring aerosols, ozone, water vapor, and optical turbulence.

Lidar Engineering

Introduction to Basic Principles

GARY G. GIMMESTAD
Georgia Tech Research Institute

DAVID W. ROBERTS
MicroDynamics LLC

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9780521198516

DOI: 10.1017/9781139014106

First published 2023

A catalogue record for this publication is available from the British Library

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-0-521-19851-6 Hardback

Additional resources for this publication at www.cambridge.org/9780521198516

For Abbey, Allison, Lauren, Natasa, and Tiff

Contents

Preface

The topic of this textbook is the engineering of lidars that are optical atmospheric remote sensing systems based on pulsed lasers. The technique was originally called laser radar, but as more applications of laser remote sensing arose, other names came into use. Broadly, the acronyms LIDAR and LADAR are now used for military surveillance and targeting and for automotive collision avoidance, and the term LiDAR has more recently been adopted by the airborne topographic and bathymetric mapping communities. Here we are concerned only with the atmosphere, and the name lidar is taken to be an acronym that has passed into the English language in the same way that RADAR became radar. This choice is a personal preference; there is unfortunately no standard naming convention. There is no standard notation for atmospheric lidar either, but the symbols used in this book are at least commonplace in the literature.

The book has its roots in a series of lectures that the authors developed for students at Agnes Scott College in Decatur, Georgia, in 2001–2003. Leanne West, John Stewart, and Jack Wood also contributed to developing and presenting the initial set of lectures. Our goal was to provide a comprehensive introduction to basic atmospheric lidar technology that was appropriate for advanced undergraduates, in preparation for building and operating their own lidar system. The lectures were later adapted into a three-and-a-half-day short course that is offered annually at Georgia Tech in Atlanta, Georgia. Technical material for this book also came from Lidar Tutorials that Dr. Gimmestad presented at each of the International Laser Radar Conferences starting in 2010 and from graduate-level courses on atmospheric lidar that he taught at Georgia Tech and at the National University of Ireland Galway. Developing a specialized set of lectures was necessary because the lidar technique is inherently multidisciplinary, requiring a basic understanding of laser technology, geometric optics, atmospheric optics, optomechanics, photodetectors, statistics, analog and digital electronics, and signal processing. The chapters in this book are expanded versions of the lectures, and because of the need to cover such a broad range of topics, the chapters are not definitive treatments; they are basic introductions to the topics at the advanced undergraduate level. References and suggestions for further reading are provided at the end of each chapter so that students can easily pursue a deeper and more detailed understanding of any of the topics.

Several other books on atmospheric lidar have been written since the first demonstration of the technique in 1963, but this is the only textbook on the topic. The chapters are written and arranged to be studied sequentially, and they include worked

examples and homework problems. Working the problems is important – they include many key derivations. The chapters will serve as handy references on various lidar engineering topics for the student who has worked through them. When technical terms are first introduced, some are italicized to indicate that they deserve special attention, and a glossary of all symbols and a list of abbreviations are provided as well. This book is intended for all newcomers to lidar, including software engineers and atmospheric science researchers who use lidars. Thorough explanations of how lidar techniques are used to characterize different atmospheric parameters and why specific wavelengths must sometimes be used are included. Trade-offs among the engineering parameters of a lidar system are discussed, in terms of how they affect signal-to-noise ratios (SNRs). Guidance is also provided for identifying and understanding common problems in lidar systems and for evaluating their performance.

This book contains material not generally found in other lidar books, including a simple and thorough treatment of the geometrical (crossover) function, a treatment of depolarization in lidar that is consistent with scattering theory and optical physics, a detailed explanation of the signal inversion algorithm known as the Klett retrieval, information on digital signal processing applied to lidar signals, and several implications of Poisson statistics for noise levels in lidar signals. An elastic backscatter lidar, the most basic type, is emphasized because its engineering principles are common to all types of lidars. The basic principles of coherent detection are covered, but its implementation in lidar systems is outside the scope of this book. Software is a key element of any lidar system, so some general requirements for it are mentioned, but software engineering is also outside our scope.

The SNR in lidar signals is used here as the measure of merit that addresses random errors that can be reduced by signal averaging. SNR is fundamental because, no matter what the application, some minimum SNRs must be achieved in the recorded lidar signal to enable the desired accuracy of a scientific measurement. Systematic errors are more problematic, as they cannot be reduced by averaging. They can arise from both optical and electronic distortions of the signal, so engineering "best practices" for avoiding signal distortions are emphasized throughout the book. The other main lidar challenges addressed here include accommodating the large dynamic range of lidar signals and the problem of having only one signal but several unknowns.

This book was written in the tradition of an older generation of researchers passing along technical expertise, gained through long experience, to younger generations. It is often said that we learn more from our mistakes than our successes, so in the hope that such knowledge is transferable, we have included examples of oversights, insufficient attention to detail, and outright blunders that we committed over a period of three decades. Most of the material is based on our direct experience, but for completeness, some sections had to be largely summarized from other books, which are cited. The chapters on atmospheric optics and on optical detection are examples of this practice. We apologize for any errors that we may have introduced in the summaries.

The lectures for the Agnes Scott College students were developed as part of their joint project with the GTRI to develop a lidar system for teaching and research in the undergraduate environment. The students chose the name EARL for Eye safe

Atmospheric Research Lidar. EARL is used as an example system throughout the book. ALE (Astronomical Lidar for Extinction) is also used as an example; ALE resulted from a joint project between GTRI and the University of New Mexico.

Several other GTRI researchers contributed technical materials for the Georgia Tech short course over the years, especially Chris Valenta, Nathan Meraz, Ryan James, and Leda Sox. Kristin Youngquist drafted many of the illustrations for it. Their technical materials and illustrations also appear in this book. The chapter on optomechanics began with a guest lecture by David Smith. The projects with Agnes Scott College were funded by the National Science Foundation under the grant numbers 0116039 and DUE-0836997, and additional support was provided by Agnes Scott College and the College of Engineering at Georgia Tech. ALE was funded by the National Science Foundation grant number 0421087. Much of the work of developing this book would not have been possible without the resources of the Glen P. Robinson Chair in Electro-Optics in GTRI, which Dr. Gimmestad held from 2002 to 2015. The authors gratefully acknowledge all of these types of support.

Glossary

The symbols used in this book are listed below, with units where applicable. Because several technical disciplines are covered, many of the symbols have different meanings in different sections. For this reason, the section numbers are listed where the meanings are first used. Some symbols formed by adding subscripts, and others used only in one place, are not listed.

Symbol	First used in section	Units	Meaning
a	4.6.3		Angstrom parameter
a	8.7.1		parameter in model for PMT secondary emission ratio
A	2.2	m^2	receiver area
A	3.3.2	m	amplitude of oscillation
A	4.1.2	m^2	geometrical area of a particle
A	9.1	V	amplitude of electronic waveform
b	9.1		individual digital bit
B	8.1	Hz	electronic bandwidth
B_{opt}	2.3	μm	receiver optical bandpass
B_v	3.3.2	J, cm^{-1}	rotational energy scale
c	1.1	m/s	speed of light
c_n	10.1.1		filter function coefficient
C	2.2	(varies)	lidar calibration constant
C	4.1.2	m^2	Mie cross section of a particle
C	10.1.1		filter coefficient normalizing factor
d	3.3.2	m	distance or length
d	4.4		depolarized fraction of backscattered light
D	2.4	m	diameter
D_1, D_2	3.3.4		designations of sodium spectral lines
e^-	8.2		an electron
E	3.1.1	W/m^2	irradiance
E	5.2.8	V/m	electric field strength
E	8.7.1	V	voltage between dynodes
E_{phot}	2.2	J	energy of a photon

E_{pulse}	2.2	J	energy in a pulse of laser light
E_{rot}	3.3.2	J, cm^{-1}	energy of rotation
E_{vib}	3.3.2	J, cm^{-1}	energy of vibration
E_0	4.6.1	W/m^2	exo-atmospheric solar irradiance
E	7.1	Pa	elastic modulus
E_c	8.7.2	eV	energy at bottom of conduction band
E_v	8.7.2	eV	energy at top of valence band
E_g	8.7.2	eV	band gap energy
E_F	8.7.2	eV	Fermi energy
E_{rms}	9.14	V	error in digital output for sine wave
f	5.1	m	focal length
f_c	8.1	Hz	critical frequency of an electronic filter
f_c	10.1.1	bin^{-1}	cutoff frequency of an FIR filter
$f/$	6.3.2		f-number (focal ratio)
F	7.1	N	force
F	8.7.2		APD noise factor
F	11.2		coefficient of finesse of an etalon
$F_v(J)$	3.3.2	J, cm^{-1}	term value for rotational energy
$F(E)$	8.7.2		occupancy of electronic states
Δf	8.6	Hz	width of electronic bandpass filter
g	11.3	m/s^2	acceleration due to gravity
$g(t)$	8.1		a function in the time domain
G	8.2		electronic gain
G	8.7.1		PMT current gain
G	10.1.1		gain of an FIR filter
G_v	3.3.2	J, cm^{-1}	term value for vibrational energy
$G(R)$	2.2		the geometrical function
$G(f)$	8.1		a function in the frequency domain
h	2.2	Js	Planck constant
h	3.2	m, km	altitude
h	6.3.2	m	image size
$h(n)$	9.14		measured probability of code n
$\hbar$	3.3.2	Js	Planck constant divided by 2π.
H	2.4	km	air density scale height
$\overline{i}$	8.1	A	current
$\overline{i_N^2}$	8.3	A^2	mean-square noise current
I	3.1.1	(W/sr)	radiant intensity
I	3.3.2	(kg·m)	moment of inertia
I	4.4.1	W	power of backscattered light on receiver
J		3.3.2	rotational quantum number
J	8.7.1	$e^-/cm^2 \cdot s$	thermionic current density
k	3.3.1	J/K	Boltzmann constant
k	3.3.2	N/m	Hooke's law spring constant

k_T	2.2		optical efficiency of transmitter
k_R	2.2		optical efficiency of receiver
k_{ion}	8.7.2		hole-to-electron ionization coefficient ratio
$\vec{k}$	5.2.8	m^{-1}	wave vector
L	5.2.3	m	length
L_λ	2.3	$W/m^2 \cdot \mu m \cdot sr$	spectral radiance
m	3.3.1	kg	mass
m	4.6.1		air mass
m	9.1.2		bin number in an FFT
m	11.5		mixing ratio
M	4.4		Mueller matrix
M	6.1.3		magnification
M	8.7.2		APD gain
M	9.1.2		number of points in an FFT
M	9.14		number of samples required
M^2	5.2.5		beam propagation ratio
n	3.2		refractive index
n	10.1.2		number of range bins
n_S	2.1		number of signal (laser) photons detected
n_B	2.1		number of background photons detected
n_D	2.1		number of dark counts
N_B	2.3		the number of background photons received per range bin (for each laser pulse)
N	3.2	m^{-3}	molecular number density
N	8.7.1		number of dynode stages
N	9.1		number of bits
N	10.1.1	range bins	filter width
N_0	2.2		number of photons in each laser pulse
$N_S(R)$	2.2		number of signal (laser) photons received in a bin at range R per laser pulse
$N(J)$	3.3.2		population of state with rotational quantum number J
$O(R)$	6.3.3		overlap function
$p(n)$	9.14		theoretical probability of code n
P	3.2	(kPa)	pressure
P	5.2.8	C/m^2	dielectric polarization
P	8.1	W	electrical power
P_0	2.2	W	power in the laser pulse
P_S	2.2	W	optical signal power
P_B	2.2	W	optical background power
$P(\theta)$	3.1.1		scattering phase function
q	5.2.6		laser cavity mode number
q	8.3	C	charge on the electron

Q	4.1.2		Mie scattering efficiency factor
Q	5.2.4		laser cavity quality factor
r_{avg}	8.3	s^{-1}	average rate
r_0	8.6	cm	Fried parameter
R	1.1	m	range
R	4.6.1	A.U.	Earth–Sun distance
R	6.2		reflectance
R	7.2.1	m	radius of mirror
R	8.1	Ω	resistance
R	9.2	1/s	count rate in photon counter
R	11.3	J/mol·K	gas constant
R_{sca}	4.5.1		scattering ratio
ΔR	2.2	m	range bin length
S	4.4.1		electronic signal from lidar receiver
S	8.3	A/W	detector responsivity
S_m	3.2	sr	molecular lidar ratio
S_a	4.2.4	sr	aerosol lidar ratio
S_k	10.1.1		the kth raw data point
S'_k	10.1.1		the kth filtered data point
$S(R)$	6.3.3	m	edge of cone of light or shadow
t	7.1	m	thickness of mirror
t_a	9.1.3	s	aperture error
t_d	9.2	s	dead time in photon counter
T	3.2	(°C, K)	temperature
$T^2(R)$	2.2		two-way transmittance
v	3.3.1	(m/s)	velocity
V	4.2.4	m	visibility
V	8.1	V	voltage
V_{out}	4.6.2	volts	photometer output signal
V_0	4.6.2	volts	photometer signal for zero air mass
V_{wind}	8.6	ms^{-1}	wind speed
V_Q	9.1	V	quantum voltage level
w_D	8.7.2	cm	width of depletion region
w_n	10.1.1		filter coefficient weight
W	5.2.5	m	laser beam radius
x_i	2.1		the ith outcome of a set of trials
$\overline{x}$	2.1		the mean of a set of trials
$X(R)$	2.2		range-corrected lidar signal
ΔY	7.1	m	mirror sag
z^*	2.1	std. dev.	confidence interval
z_R	5.2.5	m	Rayleigh range
δz	10.1.1	m	sampling interval
Δz	10.1.1	m	range resolution

α	3.1.2		scattering parameter
α	7.1	1/°C	coefficient of thermal expansion (also CTE)
α	8.7.1		parameter in model for PMT secondary emission ratio
α	8.7.2	cm^{-1}	absorption coefficient
α	11.4	m^{-1}	lidar extinction coefficient
β	9.1.4		desired accuracy of measurement
β	3.1.1	m^{-1}	volume total scattering coefficient
$\beta(\theta)$	3.1.1	$m^{-1}sr^{-1}$	volume angular scattering coefficient
$\beta(R)$	2.2	$m^{-1}sr^{-1}$	volume backscatter coefficient
Γ	8.7.1		PMT excess noise factor
δ	4.4.3		lidar depolarization ratio
δ	6.3.3	radians	angle between transmitter and receiver OAs
δ	8.7.1		secondary emission ratio
ε	7.1		strain (normalized length change)
ε	9.1	V	quantization error
ε_0	5.2.8	F/m	dielectric constant
η	8.3		quantum efficiency
θ	2.3	rad	receiver FOV plane angle
θ	3.1.1	degrees, rad	scattering angle
θ	5.2.5	rad	laser beam divergence
λ	2.2	nm, μm, m	wavelength
μ	10.1.2	counts	background in lidar data
ν	2.2	Hz	frequency
ν	3.3.2		vibrational quantum number
ν	7.1		Poisson's ratio
$\bar{\nu}$	3.3.2	cm^{-1}	spatial frequency (wavenumbers)
$\Delta\nu$	8.6	Hz	range of signal frequencies
ρ	10.1.3		scale factor
σ	7.1	Pa	Stress
σ	2.1		standard deviation
σ^2	2.1		Variance
$\sigma(r)$	2.2	m^{-1}	extinction coefficient
$\sigma(\theta)$	4.1.1	m^2/sr	angular scattering cross section
σ_p	4.1.1	m^2/sr	scattering cross section of a particle
σ	11.4	m^2	absorption cross section of molecule
$(d\sigma / d\Omega)_\pi$	11.5.1	m^2/sr	Raman differential backscattering cross section
τ	1.1	s	time interval
τ	4.6.1		optical depth
τ_{pulse}	2.2	s	pulse width
τ_{int}	8.3	s	detector integration time
ϕ	3.2	rad	angle measured from y-axis

ϕ	8.4	Wb	magnetic flux
ϕ	8.6	rad	phase angle
Φ	8.7.1	eV	work function
χ	5.2.8		susceptibility
χ_e	3.3.2		anharmonicity constant
ω	3.3.2	rad/s	angular velocity of rotation
ω	5.2.8	rad/s	angular frequency of light
ω_e	3.3.2	cm^{-1}	harmonic wavenumber
Ω	2.3	sr	receiver FOV solid angle

Abbreviations

AC	alternating current
ACCD	accumulation CCD
ADC	analog-to-digital converter (also A to D, A–D, A/D)
AERONET	aerosol robotic network
ALADIN	atmospheric laser Doppler instrument
ALE	Astronomical Lidar for Extinction
ALOMAR	Arctic Lidar Observatory for Middle Atmosphere Research
AMU	atomic mass unit
AMV	atmospheric motion vector
ANSI	American National Standards Institute
AOD	aerosol optical depth
APD	avalanche photodiode
AR	anti-reflection
AU	arbitrary units (on plot axes)
BBO	β-barium borate
BPP	beam parameter product
CAD	cloud-aerosol discriminator
CALIOP	cloud-aerosol lidar with orthogonal polarization
CALIPSO	cloud-aerosol lidar and infrared pathfinder satellite
CB	citizen's band (radio)
CCFU	cloud climatology field unit
CDRH	Center for Devices and Radiological Health
CH	Chanin–Hauchecorne (algorithm)
CR	color ratio
CTE	coefficient of thermal expansion
CW	continuous wave
DAOD	differential atmospheric optical depth
DC	direct current
DFG	difference frequency generation
DIAL	differential absorption lidar
DNL	differential nonlinearity
EARL	Eye safe Atmospheric Research Lidar
EMI	electromagnetic interference
ENOB	effective number of bits

ERBW	effective resolution bandwidth
FAA	Federal Aviation Administration
FADOF	Faraday anomalous dispersion optical filter
FDA	Food and Drug Administration
FFT	fast Fourier transform
FHWM	full width at half maximum
FIR	far infrared, also finite impulse response
FOV	field of view
FPA	focal plane array
FSR	full-scale range
GHG	greenhouse gas
GLOW	Goddard Lidar Observatory for Winds
GMAO	Global Modeling and Assimilation Office
GTRI	Georgia Tech Research Institute
HERA	hybrid extinction retrieval algorithm
HF	high frequency
HHG	high harmonic generation
HITRAN	high-resolution transmission molecular absorption database
HLOS	horizontally projected line of sight
HSRL	high spectral resolution lidar
IABS	integrated aerosol backscatter
IMD	intermodulation distortion
INL	integral nonlinearity
IPDA	integrated path differential absorption
ISO	International Organization for Standardization
KDP	potassium dihydrogen phosphate
KTP	potassium titanyl phosphate
LBLRTM	line-by-line radiative transfer model
LCVR	liquid crystal variable phase retarder
LHC	left-hand circular polarization
LO	local oscillator
LR	long range (receiver)
LSB	least significant bit
LWIR	long-wave infrared
MERLIN	methane remote sensing lidar mission
MOPA	master oscillator – power amplifier
MPE	maximum permissible exposure
MWIR	mid-wave infrared
NAT	nitric acid trihydrate
NCAR	National Center for Atmospheric Research
ND	neutral density
NDACC	Network for the Detection of Atmospheric Composition Change
NEP	noise equivalent power

NIR	near infrared
NLC	noctilucent cloud (also called polar mesospheric cloud)
NOHD	nominal ocular hazard distance
NWP	numerical weather prediction
OA	optical axis
OD	optical depth
OPA	optical parametric amplification
OPO	optical parametric oscillator
PBL	planetary boundary layer
PDF	probability density function
PDR	particle depolarization ratio
PER	polarization extinction ratio
PHD	pulse height distribution
PIN	P-type, intrinsic, N-type photodiode
PIV	particle imaging velocimetry
PM	10 particulate matter with aerodynamic diameter less than 10 μm
PM	2.5 particulate matter with aerodynamic diameter less than 2.5 μm
PMT	photomultiplier tube
PRF	pulse repetition frequency
PSC	polar stratospheric cloud
QE	quantum efficiency (also called η)
REAL	Raman-shifted Eye safe Aerosol Lidar
RF	radio frequency
RHC	right-hand circular polarization
RISTRA	rotated image singly resonant twisted rectangle
RTV	room temperature vulcanizing
SAE	Society of Automotive Engineers
SFDR	spurious-free dynamic range
SFG	sum frequency generation
SHG	second harmonic generation
SIBYL	selective iterated boundary locator
SNDR	signal-to-noise-and-distortion ratio (also S/N+D, SINAD)
SNR	signal-to-noise ratio
SOP	standard operating procedure
SPAD	single-photon avalanche diode
SR	short range (receiver)
SRS	stimulated Raman scattering
SSW	sudden stratospheric warming
STP	standard temperature and pressure
SWIR	short-wave infrared
TEM	transverse electromagnetic mode
THD	total harmonic distortion
THG	third harmonic generation
TIA	transimpedance amplifier

TIR	Total internal reflection
TROPOS	Leibniz Institute for Tropospheric Research
ULE	ultra-low expansion
UTS	unified thread standard
UV	ultraviolet
VHF	very high frequency
VIL	volume imaging lidar
VIS	visible

1 Introduction

Lidar (light detection and ranging) is an optical remote sensing technique used to measure properties of the atmosphere at long ranges without direct contact. Because a lidar system transmits pulses of laser light, the atmospheric measurements are functions of range, where the range is calculated from the pulse time-of-flight multiplied by the speed of light. Lidar is a very powerful tool for atmospheric remote sensing. No other active remote sensing technique can measure as many parameters as lidar.

1.1 The Atmospheric Lidar Technique

A basic lidar system is illustrated in Figure 1.1, where the laser emits pulses of light in a narrow beam into the atmosphere toward an opaque cloud. The receiver telescope collects backscattered photons and concentrates them onto a photodetector that converts the photons into an electronic signal. If that signal is plotted versus time, it will have the characteristics shown graphically in the box on the right: For short ranges, the signal will be zero, because none of the scattered photons can get through the receiver optics to reach the detector. At intermediate ranges, the signal is caused by backscatter from molecules, and it rises suddenly to a peak, from which it falls rapidly, until a second peak appears due to a strong signal caused by backscatter from the water droplets in the cloud. After the cloud, the signal returns to zero because the laser light cannot penetrate an opaque cloud, so no laser light is received from ranges beyond it.

In practice, the time axis shown in Figure 1.1 is always converted to range, by using the formula distance = velocity × time, where the velocity of light c is 3×10^8 m/s. The light must travel out and back, so the distance $2R$ is equal to ct for a time t, and the lidar range R is therefore $ct/2$. For example, a time of 1 μs corresponds to a lidar range of 150 m. The lidar signal is often referred to as a *transient* waveform because it occurs in such a short time. The lidar signal is discussed in detail in the chapters that follow, but for now, it is sufficient to note three features: (1) a lidar system always has a nonzero minimum range; (2) the signal tends to span a very large dynamic range; and (3) the time duration of the signal is very short, generally less than 1 ms for ground-based lidars.

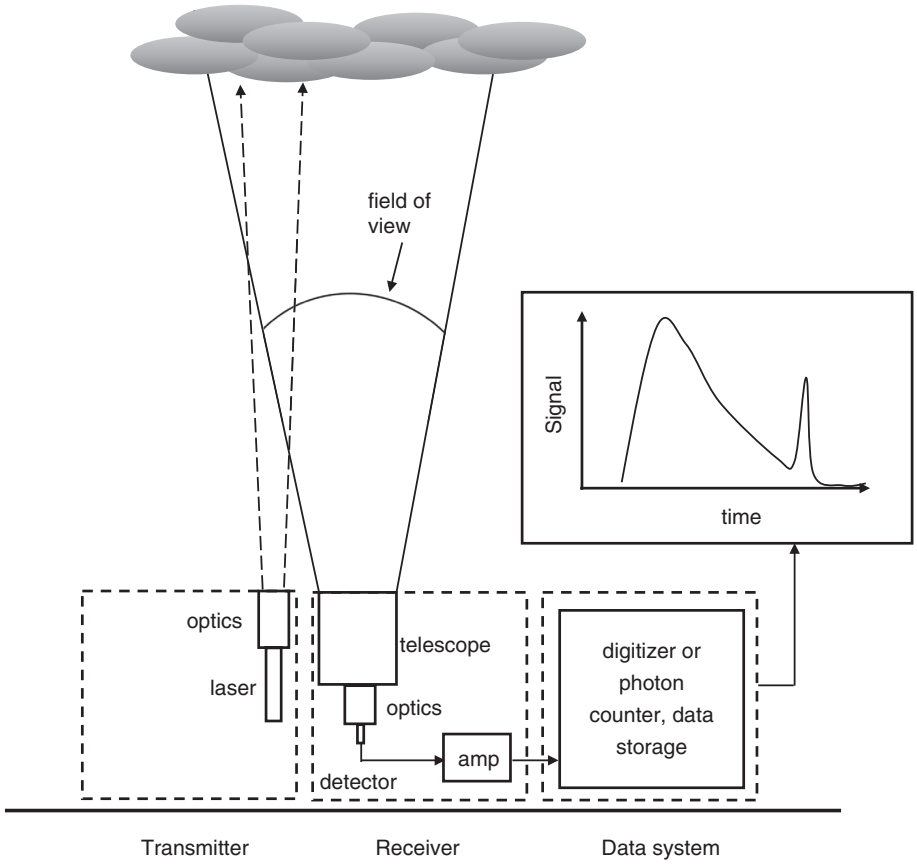


Figure 1.1 A basic lidar system. The lidar has three main components: a transmitter, a receiver, and a data system that acquires a transient electronic signal versus time, for each laser pulse.

1.2 Structure and Composition of the Atmosphere

To appreciate the wide variety and value of lidar measurements, we must understand the structure and composition of Earth's atmosphere. The atmosphere is conventionally described as a set of concentric spherical shells (troposphere, stratosphere, etc.) where the shell boundaries, known as *pauses*, are determined by inflections of the temperature profile, as shown in Figure 1.2. Clouds, convection, and weather phenomena are primarily in the troposphere, whereas there is much less vertical mixing in the stratosphere. The mesosphere is free of aerosols, and it was a poorly known region until remote sensing measurements became available. The thermosphere lies above these layers, from roughly 100 to 500 km, where the temperature rises rapidly with altitude and atmospheric gases may be ionized. From the ground all the way up to the top of the mesosphere at 100 km, the atmospheric gases are well mixed and neutral. Being well mixed means that air is always 78% nitrogen, 21% oxygen, and 1%

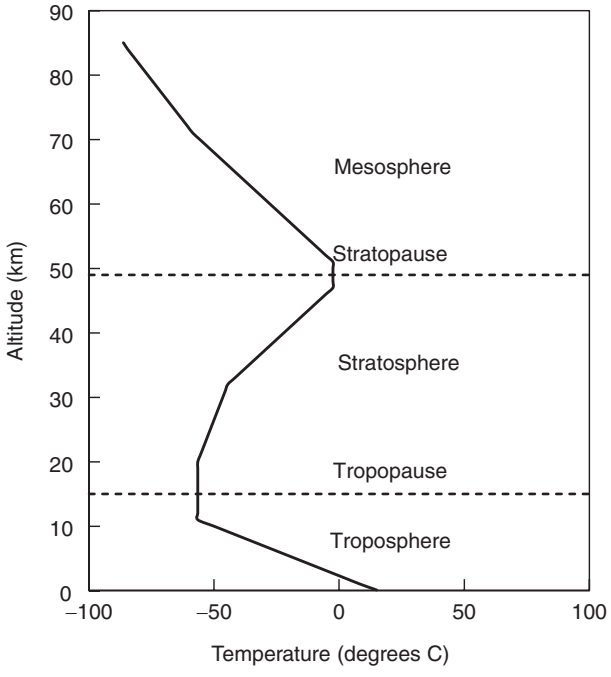


Figure 1.2 The structure of Earth's atmosphere. The conventional description of the atmosphere in terms of spherical shells is illustrated using the temperature profile data in the U.S. Standard Atmosphere, 1976 [1].

argon, with smaller amounts of other gases, including 0.04% carbon dioxide (CO_2) and highly variable amounts of water vapor. Being neutral means that the concentration of ions is negligible in this altitude region.

The decrease in air density with altitude is nearly exponential, and it decreases by six orders of magnitude between the ground and 100 km, as indicated in Figure 1.3, where the density scale is logarithmic. In geoscience, altitude is usually plotted on the vertical axis even though it is an independent variable, so this convention is followed in Figures 1.2 and 1.3. In lidar, signals are often plotted with range on the horizontal axis, as in the Figure 1.1 inset, but reduced data products in the form of altitude profiles are usually plotted using the geoscience convention. Modeled lidar profiles are plotted both ways.

1.3 Atmospheric Lidar Applications

Figure 1.2 describes the atmosphere in terms of the conventional spherical shells of atmospheric science and meteorology, with a typical temperature profile. However, the lidar point of view is much more complex, as illustrated in Figure 1.4, because the atmosphere contains important observable constituents at all levels, including both trace gases and solid and liquid particles of matter. Lidar systems have been developed

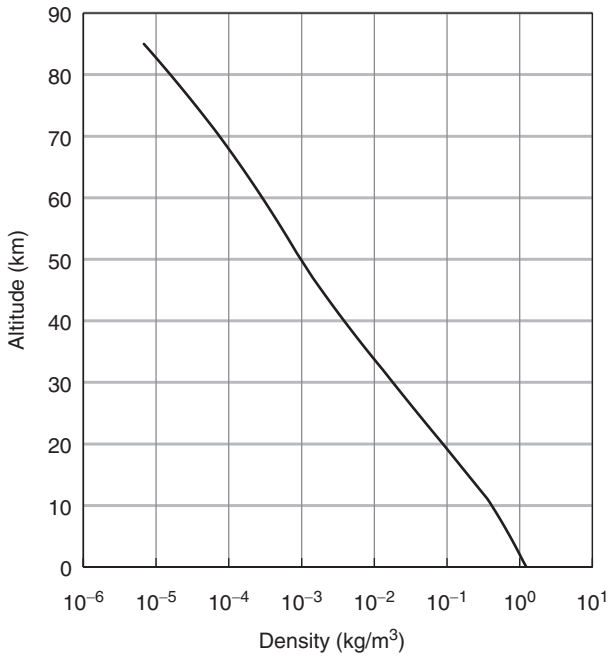


Figure 1.3 Air density versus altitude. The air density profile is from the U.S. Standard Atmosphere, 1976 [1].

to measure all the constituents shown in Figure 1.4 as well as winds and temperatures. Some of those lidar measurements are briefly described in the following sections.

1.3.1 Troposphere

The troposphere is where virtually all human activity takes place. It has two distinct regions: the mixed layer, which is typically 1–2 km thick, and the free troposphere, where the word "free" refers to unimpeded winds. The greatest source of aerosols is at Earth's surface, and those aerosols are carried upward by convection in the mixed layer. Photochemistry related to air pollution also occurs in the mixed layer, producing both ozone (O_3) and additional aerosols. Industrial emissions of pollutants occur at the surface, and those pollutants are diluted by the mixing action in this layer. Ozone and smog are major components of bad air quality, which is a regional hazard to human health often associated with urban areas. Some industrial emissions such as mercury (Hg) find their way into the environment and consequently into the food chain, again presenting a health hazard. Lidar techniques have been developed to provide information for managing all these issues. As shown in Figure 1.5, urban air-quality problems are intrinsically time varying and three-dimensional (3-D), depending on the mixing layer thickness, the emission rates of pollutants, prevailing winds, photochemical reactions, and entrainment of pollutants or their precursors in the free troposphere. Urban areas often have arrays of surface air-quality sensors to monitor the air that people breathe, but a detailed understanding of air quality can

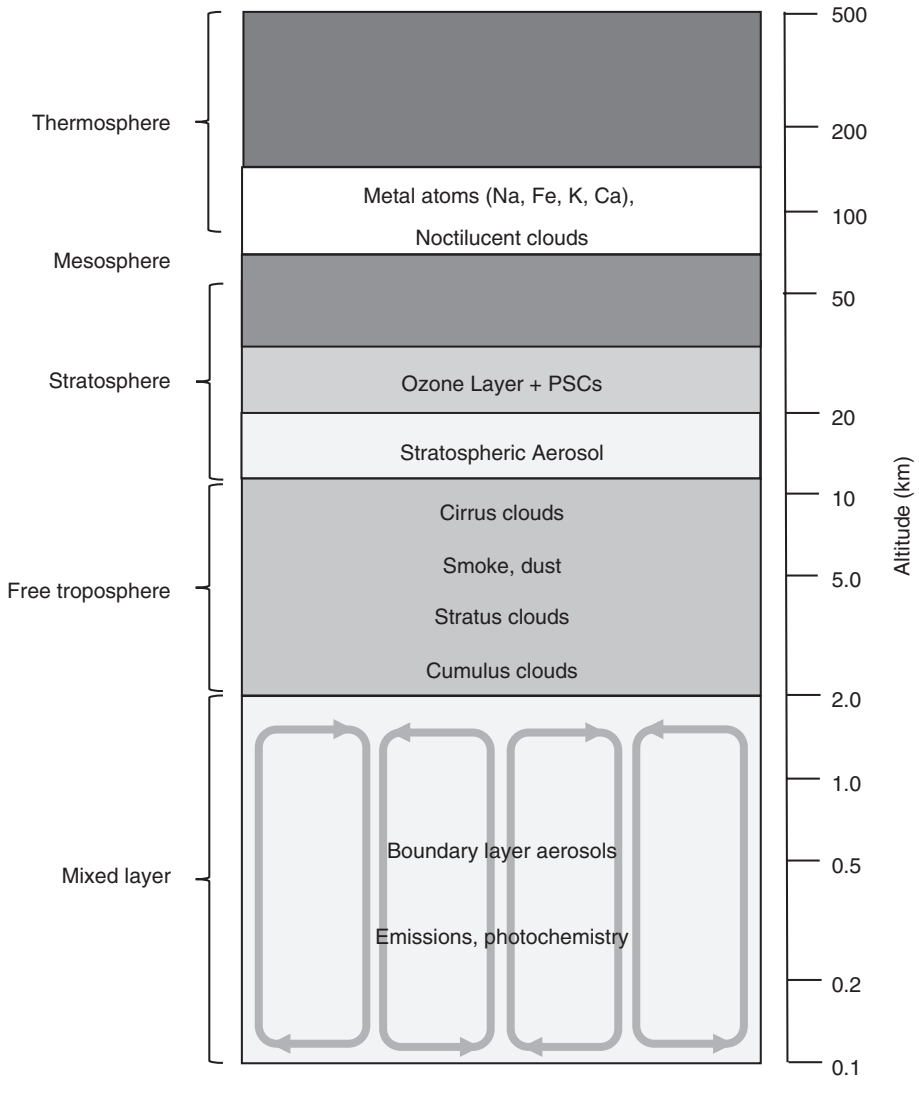


Figure 1.4 Lidar observables in the atmosphere. The main constituents that are measurable with lidar systems are shown for the various regions of the atmosphere. Note that the altitude scale is logarithmic.

only come from measurements throughout the 3-D volume, which are difficult to obtain. For this reason, periodic air-quality measurement campaigns employ suites of ground-based and airborne instruments that usually include lidars because of their long-range measurement capability.

Lidar techniques are also being employed to monitor the related but much larger problem of global climate change. Although Earth's climate does have natural periodic changes on geological timescales, as shown by phenomena such as ice ages, man-made emissions of certain gases and aerosols are introducing unusual changes to Earth's energy balance. Without an atmosphere, the average temperature of our planet

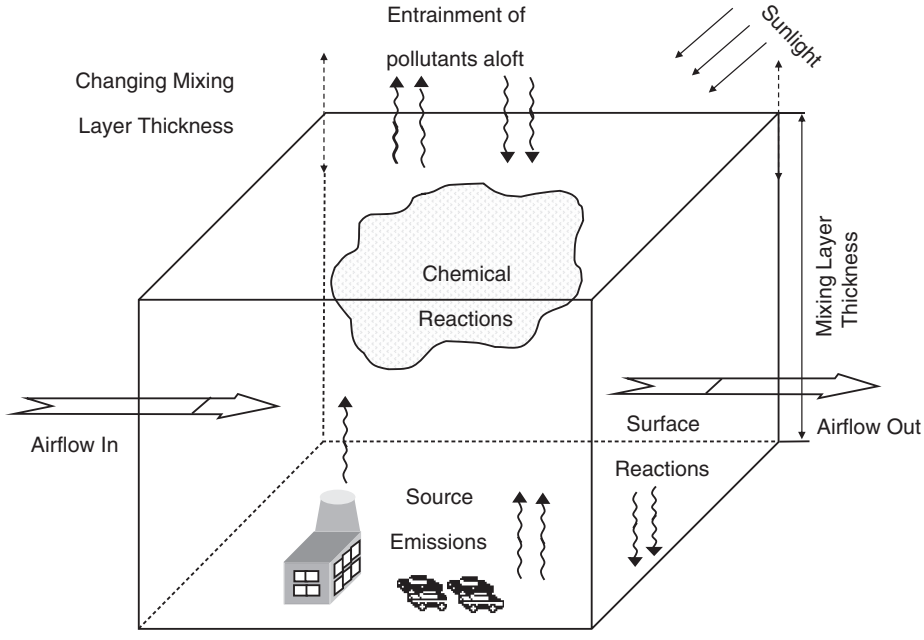


Figure 1.5 The three-dimensional nature of urban air quality. In any volume of the mixing layer, there are many sources and sinks for pollutants. All the factors shown tend to be time varying.

would be about 260 K, which is below the freezing point of water. The actual average temperature is a much more habitable 295 K because of the greenhouse effect, which is caused by atmospheric gases with strong absorption bands in the infrared region of the spectrum. The incoming and outgoing radiations are in a delicate balance: The sun's irradiance outside the atmosphere (known as the solar constant) is about 1,400 W/m^2, and climate-changing imbalances are of the order of 1 W/m^2. Around 30% of the solar radiation (averaged over the globe and all seasons) is scattered to space by the atmosphere, land, and oceans. The remaining 70% is absorbed, and an equal amount of power escapes to space as thermal radiation. Both scattering and absorption of solar radiation in the atmosphere are highly dependent on aerosols and clouds, and small changes in the amounts or the optical properties of aerosols and clouds cause significant changes in the energy balance. In addition, aerosols cause an indirect change by influencing the formation of clouds and their ability to scatter sunlight. Lidar is an essential tool for studying and monitoring many of these competing effects.

Of the absorbed 70% of the incoming sunlight, around 20% is absorbed in the atmosphere and the remaining 50% is absorbed at the surface. Much of the latter is reradiated upward as thermal radiation. However, only 6% is radiated directly to space, and the rest is absorbed by clouds and greenhouse gases (GHGs), of which the most important are water vapor, CO_2, methane (CH_4), and nitrous oxide (N_2O). The concentrations of the last three are increasing rapidly due to human activities. Since the start of the industrial revolution, CO_2 has increased by almost 50%, and its rate

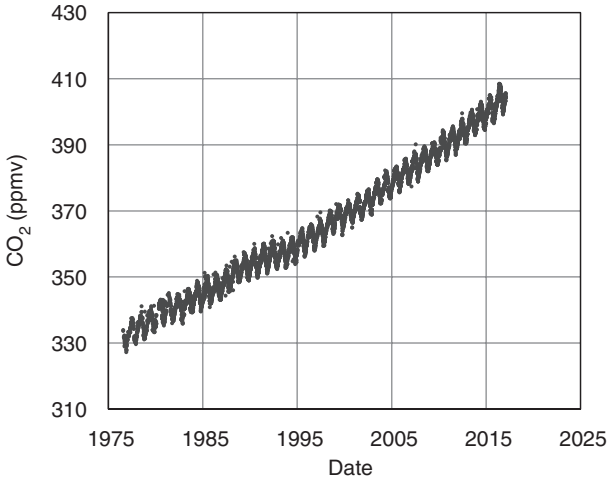


Figure 1.6 Monthly mean CO_2 mixing ratios measured at Mauna Loa, Hawaii, known as the Keeling Curve. The jagged appearance is due to seasonal variations. The CO_2 concentration was in the 260–280 ppm range during the 10,000 years, leading up to the industrial revolution. Data are from [2], used by permission.

of increase is accelerating, as shown in Figure 1.6 [2]. As of this writing, half of the CO_2 humans have ever generated has been put into the atmosphere in the past three decades. CH_4 concentrations have more than doubled, and the greenhouse effect of CH_4 is 25 times that of CO_2. GHGs are already causing measurable climate changes, and the continued rise in their concentrations can lead to environmental catastrophes. The natural "sinks" that remove both natural and man-made gases are not understood, but the effects are already apparent. For example, CO_2 is being absorbed by oceans and land areas much faster than models predict, but there is great concern that the current absorption rate will not be sustained in the future. Carbonic acid (CH_2O_3) produced by CO_2 dissolving into sea water has already increased ocean acidity by 25% over the past two centuries, affecting the survival of many marine animals at the bottom of the food chain. As industrialized nations attempt to grapple with this situation, a need has arisen for monitoring GHG concentrations and emission rates globally. In response to this challenge, the international lidar community is pioneering a suite of techniques to monitor GHGs in Earth's atmosphere, in which ground-based, airborne, and spaceborne measurement platforms all have applications.

In the free troposphere, both aerosols and trace gases are transported over long distances, so they cause environmental harm far from the source of pollutants. For example, nitrogen dioxide (NO_2) is a precursor for ozone, but local regulations to control its concentration are not an effective strategy in the eastern U.S. because so much of it is transported from other regions upwind. The situation with sulfur dioxide (SO_2) is similar – once it gets into the free troposphere, it is transported by the prevailing winds, causing acid rain over wide geographical areas. Smoke from forest fires and ash from volcanic eruptions can also rise into the free troposphere and be transported

to thousands of kilometers, causing widespread air-quality problems as well as hazards to aviation. Ground-based and airborne lidar systems are used to monitor and understand the transport of all these pollutants.

Several lidar techniques have been developed for measuring wind speed, which is a key parameter for weather forecasting and aviation safety. Wind lidars are in use at airfields to monitor local winds near the ground and to detect wake vortices, which are serious hazards to small aircraft. Studies have shown that measurements of the global wind field are the most important missing input for weather forecasting, and for this reason, spaceborne wind-profiling lidar technology is being developed in both Europe and the U.S.

1.3.2 Stratosphere

The stratosphere lies above the troposphere. As the name suggests, this region is stratified, meaning that there is much less vertical mixing than in the troposphere, but the stratosphere contains important layers from the lidar point of view. The stratospheric aerosol layer is mainly tiny drops of sulfuric acid, although it also contains ashes of meteors that burned up much higher in the atmosphere. This aerosol layer is constantly being depleted as the aerosols fall into the troposphere and become entrained in weather, and it is constantly being replenished by SO_2 from volcanoes and other sources. After exceptionally large volcanic eruptions, scattering of sunlight by this aerosol layer significantly alters Earth's energy balance, and this effect persists for years. The stratospheric aerosol layer has been continuously monitored with ground-based lidars for more than three decades, as shown in Figure 1.7 [3].

A thick layer of ozone occurs roughly in the middle of the stratosphere, and this layer helps to make the earth habitable by absorbing harmful ultraviolet light from the sun at wavelengths shorter than about 300 nm. In the 1980s, an alarming discovery was announced: Each year in springtime, the stratospheric ozone layer becomes depleted in a large area over the South Pole. This phenomenon, which became known as the "ozone hole," was traced to chemical reactions involving chlorine and bromine atoms in man-made chlorofluorocarbons, which were manufactured for use as refrigerants and propellants in aerosol cans. This explanation was puzzling because such reactions do not occur in the gas phase; they require a solid surface. Lidar instruments showed that the required surface was provided by polar stratospheric clouds (PSCs) that occur at both poles when the stratospheric temperature drops below about 198 K. Lidar techniques were used to classify these clouds into several distinct types based on their backscattering and depolarization of the laser light [4]. As a result of the solid scientific understanding that lidar helped provide, nearly 200 nations have ratified an international treaty, known as the Montreal Protocol, banning the production of certain chlorofluorocarbons. Permanent lidar facilities have been installed in the polar regions to monitor ozone and PSCs, and the ozone layer is expected to fully recover by the year 2050. The Montreal Protocol is widely heralded as illustrating the level of international cooperation on man-made atmospheric change that can be achieved when the science is convincing enough. The stratospheric ozone layer

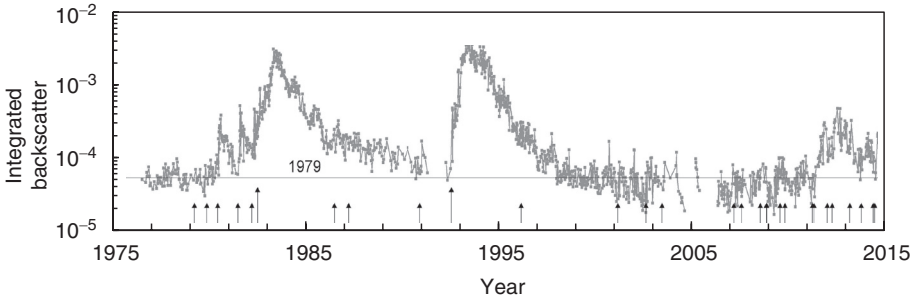


Figure 1.7 Thirty-year lidar record of stratospheric aerosols at Garmisch-Partenkirchen. The layer is characterized by its lidar backscattering signal, which varies by more than two orders of magnitude, depending on volcanic activity. The horizontal line shows the 1979 average value, and the vertical arrows show the dates of eruptions. This figure is adapted from [3] and used by permission.

is now routinely monitored by an international network that includes more than 30 ground-based lidar stations, the Network for the Detection of Atmospheric Composition Change (NDACC), that is providing a long-term record [5].

Much of the upper stratosphere is free of aerosols, and in that region, the lidar signal is caused by Rayleigh scattering from air molecules, which is proportional to the air density. Temperature profiles can be derived from density profiles by using the principle of hydrostatic equilibrium and the ideal gas law. Long-term stratospheric temperature variations have been monitored in this way by researchers at several locations, including a lidar station in France with a continuous record spanning four decades [6].

1.3.3 Mesosphere

The mesosphere (the prefix meso- means "middle") extends from the stratosphere to an altitude of about 100 km. Until the development of lidar techniques, measurements in the mesosphere were generally made with sounding rockets, which provided very sparse data; hence, little was known about this region. In addition, the air in the mesosphere is so rarified that molecular scattering provides a very small signal, which is a serious challenge for ground-based lidars. However, the mesosphere contains a layer of atoms in the 80–100 km range that are deposited by meteors as they burn up, including the elements sodium (Na), potassium (K), iron (Fe), and calcium (Ca), and lidar techniques have been developed to observe all of them. The scattering mechanism is called resonance fluorescence scattering, and it is 12 orders of magnitude stronger than molecular (Rayleigh) scattering. For this reason, the mesospheric elements cause measurable lidar backscatter even though their concentrations are tiny. Both temperatures and winds are measured in the mesosphere with ground-based lidars that exploit resonance fluorescence scattering by the metal atoms [7].

In summary, lidar systems are now deployed on the ground, in the air, and in space to help keep our planet habitable and to protect public health. Lidar remote sensing

is used in atmospheric science to understand and monitor both natural and man-made changes, including bad air quality caused by emissions of pollutants, the ozone hole, and the climate-forcing effects of GHGs and aerosols. Lidars are being used to improve weather forecasting, which will have a huge benefit for both agriculture and mitigating natural disasters. Lidars are used to improve the safety and efficiency of air travel by monitoring hazards such as volcanic ash, terminal-area winds, and wake vortices. For many of these applications, there is no other practical way to obtain the combination of spatial coverage, range resolution, and time resolution that lidar provides, because the atmosphere is large, dynamic, and three dimensional. As shown by these applications, lidar is a powerful and essential remote sensing tool. However, lidar engineering is inherently multidisciplinary, requiring expertise in lasers, geometrical optics, atmospheric optics, optomechanics, photodetectors, statistics, digital electronics, and signal processing. All these topics are introduced in the chapters that follow, as they apply to lidar, and best practices for designing, constructing, and operating lidars are given so that the final data products will be unbiased results with reliable error estimates.

1.4 Book Contents and Structure

This book is an introduction to the engineering and application of atmospheric lidar instruments. The main engineering challenges are described, along with engineering trade-offs and best practices. Many tips are given for avoiding common pitfalls. A thorough understanding of atmospheric lidar systems requires a basic familiarity with a broad range of technical topics, and there is perhaps no ideal order in which to present them. The approach taken in this book is to address a set of sequential questions that arise naturally from the material in each chapter. For example, a statistical model for the lidar signal-to-noise ratio (SNR) in Chapter 2 requires as inputs the numbers of laser and sky photoelectrons in a measurement, and this fact leads to the question of how to find those numbers. That question is answered in the following two sections, with mathematical models for those two parameters. But the models also depend on atmospheric parameters, which leads to the question of how to find them, and that question is addressed in Chapters 3 and 4 on atmospheric optics. The models in Chapter 2 also depend on instrumental parameters, and the question of how to get them is answered in Chapters 5 and 6 on lidar transmitters and receivers. At that point, the reader will also be aware of the necessity to maintain optical alignment between the transmitter and the receiver, and techniques for achieving this are given in Chapter 7. Chapters 3–6 also lead to the question of how to detect photons, so Chapter 8 covers the conversion of photon arrival rates to electronic lidar signals. Methods of recording the signals are addressed in Chapter 9, and finally, the analysis of the resulting digital lidar data is addressed in Chapter 10. This book is heavily slanted toward elastic backscatter lidar, which is the simplest type, but the physical bases for more sophisticated types of lidars are covered in Chapters 3 and 4, and some comments about data analysis for them are provided in Chapter 11. The material is introductory because it

is intended to be accessible to advanced undergraduates, and recommendations for further reading are provided at the end of each chapter, along with references.

1.5 Further Reading

E. D. Hinckley, Ed., *Laser Monitoring of the Atmosphere*. New York: Springer, 1976.

Written when the lidar technique was only 13 years old, this book has seven chapters with a different author (or team of authors) for each chapter. It covers several of the major types of atmospheric lidars, and it remains perhaps a valuable resource for newcomers to lidar because of its clear and very readable expositions on the physical basis of each lidar measurement technique.

R. M. Measures, *Laser Remote Sensing: Fundamentals and Application*. New York: Wiley-Interscience, 1984.

This book, written when lidar was about two decades old, tends to be very theoretical, but it contains photos of early lidar instruments as well as numerous examples of the corresponding lidar data. It has a chapter on atmospheric applications.

V. Kovalev and W. Eichinger, *Elastic Lidar: Theory, Practice, and Analysis Methods*. Hoboken: Wiley-Interscience, 2004.

This book was written as a handbook of elastic backscatter lidar and a guide to the lidar literature. The authors included some information on lidar instrumentation, but their book is more focused on lidar techniques and analysis. It has the most comprehensive treatment of the multi-angle lidar technique in the literature.

C. Weitkamp, Ed., *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere*. New York: Springer, 2005.

This book has 14 chapters covering all major types of atmospheric lidar, with a different author (or team of authors) for each chapter. It is highly recommended as a general resource.

T. Fujii and T. Fukuchi, Eds., *Laser Remote Sensing*. New York: Taylor & Francis, 2005.

This is another book with different authors for each of its nine chapters. It is complementary to the Springer book described above, and its chapters on resonance fluorescence lidar and wind lidar are much more comprehensive.

C.-Y. She and J. S. Friedman, *Atmospheric Lidar Fundamentals: Laser Light Scattering from Atoms and Linear Molecules*. New York: Cambridge University Press, 2022.

This recent book is a rigorous explanation from first-principles physics of many standard types of lidar measurements. As the subtitle implies, atoms and molecules are emphasized, not aerosols. The book includes a few comments on lidar optical systems and detectors, but it is primarily theoretical.

References

[1] National Atmospheric and Oceanic Administration, National Aeronautics and Space Administration, and United States Air Force, *The U.S. Standard Atmosphere*, *1976*. NOAA-S/T 76–1562 (1976). [Online]. Available: https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19770009539.pdf. [Accessed December 7, 2020]. NASA Technical Reports Service.

[2] K. W. Thoning, A. M. Crotwell, and J. W. Mund, *Atmospheric Carbon Dioxide Dry Air Mole Fractions from Continuous Measurements at Mauna Loa, Hawaii, Barrow, Alaska, American Samoa and South Pole. 1973–2019, Version 2020–08*. Boulder, CO: National Oceanic and Atmospheric Administration Global Monitoring Laboratory, 2020. [Online]. Available: ftp://aftp.cmdl.noaa.gov/data/greenhouse_gases/co2/in-situ/surface/. [Accessed December 7, 2020].

[3] T. Trickl, H. Giehl, H. Jaeger, and H. Vogelmann, "35 yr of Stratospheric Aerosol Measurements at Garmisch-Partenkirchen: From Fuego to Eyjafjallajökull, and Beyond," *Atmospheric Chemistry and Physics*, vol. 13, pp. 5205–5225, 2013. [Online serial]. Available: https://acp.copernicus.org/articles/13/5205/2013/acp-13-5205-2013-discussion.html. [Accessed December 8, 2020].

[4] M. A. Felton, Jr., T. A. Kovacs, A. H. Omar, and C. A. Hostetler, "Classification of Polar Stratospheric Clouds Using LIDAR Measurements from the SAGE III Ozone Loss and Validation Experiment," U.S. Army Research Laboratory, Adelphi, MD, Tech. Report. ARL-TR-4154, 2007. [Online]. Available: https://apps.dtic.mil/dtic/tr/fulltext/u2/a469817.pdf. [Accessed December 8, 2020].

[5] "Network for the Detection of Atmospheric Composition Change," *ndaccdemo.org*. Available: https://www.ndaccdemo.org. [Accessed December 8, 2020].

[6] T. Leblanc, I. S. McDermid, P. Keckhut, A. Hauchecorne, C. Y. She, and D. A. Krueger, "Temperature Climatology of the Middle Atmosphere from Long-Term Lidar Measurements at Middle and Low Latitudes," *Journal of Geophysical Research* vol. 103, pp. 17191–17204, 1998. [Online serial]. Available: https://agupubs.onlinelibrary.wiley.com/doi/epdf/10.1029/98JD01347. [Accessed December 8, 2020].

[7] X. Chu and C. G. Papen, "Resonance Fluorescence Lidar for Measurements of the Middle and Upper Atmosphere," in *Laser Remote Sensing*, T. Fujii and T. Fukuchi, Eds. New York: Taylor & Francis, 2005, pp. 179–432.

2 The Basic Lidar Models

During the development of any useful sensor system, a mathematical model describing its inputs and outputs is needed so that the system can be optimized for any given application. Lidar is certainly no exception, and the main model is the lidar equation described in Section 2.2. The lidar equation is a model of the received power, or equivalently the photon arrival rate, as a function of range, based on instrumental and atmospheric parameters. It is used for understanding lidar measurements, for engineering lidar systems, for developing data analysis algorithms, and for developing new lidar techniques. The lidar equation yields the number of laser photons received from each range interval, which can be used to find the signal-to-noise ratio (SNR) of a lidar signal by the methods shown in Section 2.1. The number of background photons received is also required for SNR, and the background model is presented in Section 2.3. The background model also uses both instrumental and atmospheric parameters as inputs.

2.1 Photon Statistics and SNR

Reliable error estimates are required for atmospheric data. A geophysical measurement result without error limits is basically meaningless because it cannot be compared to other data: If the size of measurement errors is unknown, trends over time cannot be established and neither can variations from one location to another. There are two basic types of error, *random* and *systematic*, and in this chapter, we are concerned only with random error. Best practices for avoiding systematic error are described in several of the chapters that follow. SNR is a standard measure of merit; higher values of SNR mean more accurate measurements. In optics, the SNR is usually taken to be the mean value of a series of measurements divided by their standard deviation. Poisson statistics describes small numbers of rare events, such as photons that are detected after their journey (shown in Figure 1.1) from the laser, through the transmitter optics, through the atmosphere to a scatterer, back through the atmosphere, through the receiver optics, and onto the detector. To elucidate Poisson statistics, a few standard definitions are necessary. Statistical results pertain to the outcome of sets of observations, which are also known as *trials*. The variable N is the total number of trials (total photons transmitted, for example), the variable x_i is the outcome of the ith trial (the ith photon is detected or not), and $\overline{x}$ is the *mean* (or average) outcome, defined by

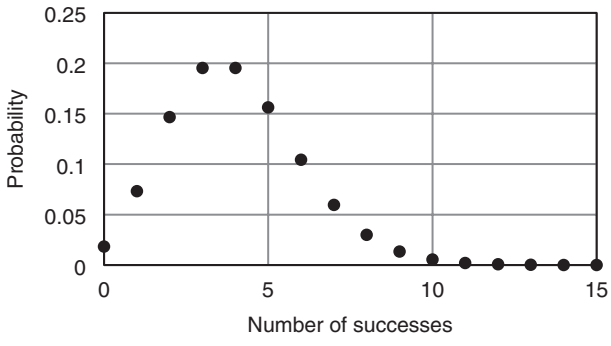


Figure 2.1 The Poisson distribution for mean value a equal to four.

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{2.1}$$

The *variance* σ^2 is defined as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^2, \tag{2.2}$$

and the *standard deviation* σ is defined as

$$\sigma = \sqrt{\frac{1}{N}\sum_{i-1}^{N} (x_i - \overline{x})^2}. \tag{2.3}$$

The standard deviation is also known as the root mean square (r.m.s.) deviation because of the order of its mathematical operations. As mentioned above, the optical SNR is defined as

$$\text{SNR} = \frac{\overline{x}}{\sigma}. \tag{2.4}$$

Note that the parameters defined in Eqs. (2.1)–(2.4) are all calculated from actual measurements. On the other hand, the statistics for differing situations are described by mathematical models known as *frequency distributions*. The Poisson frequency distribution is given by

$$f_a(n) = \frac{a^n \exp(-a)}{n!}, \tag{2.5}$$

where $f_a(n)$ is the probability of n events (such as n photons detected) when the mean value is a. The number n can only have integer values, so the frequency distribution only has discrete values. An example of this distribution is shown in Figure 2.1 for $a = 4$. Note that the distribution is not symmetrical about the mean value for such a small a (the probability of three detections is essentially the same as for four), and that the probability of zero detections is not insignificant in this case. Most importantly, when the mean value is four, individual trials will yield anywhere from zero to about ten successes. This unavoidable spread in values is called *statistical uncertainty*, and it is considered to be a type of noise, most often called *shot noise* or *Poisson noise*.

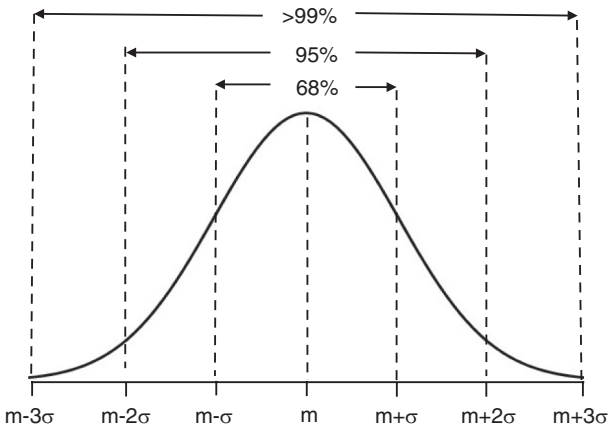


Figure 2.2 The Gauss probability distribution. The percentage of scores (often photon counts, in lidar) is shown in terms of the standard deviation σ.

The definition of SNR in Eq. (2.4) takes a particularly simple form for the Poisson distribution because the mean is the value a and the standard deviation is $\sqrt{a}$, so

$$\text{SNR} = \frac{a}{\sqrt{a}} = \sqrt{a}. \tag{2.6}$$

As the value of the mean increases, the Poisson distribution rapidly becomes indistinguishable from the Gauss distribution, which is defined by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-(x-m)^2 / 2\sigma^2\right], \tag{2.7}$$

where m is the mean and σ is the standard deviation. The Gauss distribution is different from the Poisson distribution in several ways: It is continuous, written in terms of the variable x, where $f(x)dx$ is the probability that a measured value will be in the range $x + dx$; it is symmetrical about the mean; and the mean and standard deviation are independent variables. The Gauss distribution has a convenient and often used relationship between the number of standard deviations about the mean and the number of results that fall within those limits, as illustrated in Figure 2.2. In scientific literature, uncertainties are usually quoted as $\pm\sigma$, but only 68% of measured results fall within that range. However, more than 99% of them fall within $\pm 3\sigma$. For engineering, it is often more convenient to speak of the *confidence intervals* z^*, which are the numbers of standard deviations required to include a certain percentage of events. The value of z^* is 1.0 for 63%, 1.96 for 95%, and 2.58 for 99%.

Mathematical probability distributions such as Eqs. (2.5) and (2.7) are known as *infinite parent distributions*, because they describe results in the mathematical limit where the number of trials is infinite. The practical question is what value to use for a when our experimental sample size is necessarily limited. Fortunately, when the Poisson distribution applies, it can be shown that the best choice for the mean $\overline{x}$ is simply the experimental mean σ, and the best estimate of the standard deviation σ is $\sqrt{\overline{x}}$, so

SNR = $\sqrt{\bar{x}}$, in accordance with Eq. (2.6). For example, if 100 photons are detected from one shot of the laser, the best estimate of the SNR of the measurements is 10. Lidar researchers almost always add the results from many laser shots to increase $\bar{x}$ and hence improve SNR. If the measurement is repeated M times, the SNR becomes $\sqrt{M \cdot \bar{x}}$. The SNR therefore increases, but only as the square root of the number of laser shots. A factor of 10 improvement requires a factor of 100 times more averaging, and the data acquisition time is therefore 100 times longer. In statistics, this reduction of uncertainty by averaging is called reducing the *standard deviation of the mean*.

In lidar systems, the SNR has complicating factors: Referring again to Figure 1.1, the lidar receiver is aimed at the sky, which has significant brightness over a broad spectral range during daytime and hence is radiating photons that will be received and detected along with the backscattered laser photons. In addition, most photodetectors produce apparent detections to some extent even when they are in the dark. The signal is always the number of laser photon detections, but the noise is the square root of the total number of photon detections. For these reasons, the lidar SNR is more generally given by

$$\mathrm{SNR} = \frac{n_\mathrm{S}}{\sqrt{n_\mathrm{S} + n_\mathrm{B} + n_\mathrm{D}}}, \tag{2.8}$$

where n_S is the average number of *signal* photon detections, n_B is the average number of *background* photon detections, and n_D is the number of *dark counts*. Variances have the useful property that they add together, as in the denominator in Eq. (2.8), if the noise sources are statistically independent and their distributions are Gauss. Equation (2.8) has two limiting cases, known as *signal-limited* and *background-limited* detection. If $n_\mathrm{S} \gg n_\mathrm{B}$, and assuming that n_D is zero, Eq. (2.8) reduces to

$$\mathrm{SNR}_\mathrm{SL} = \sqrt{n_\mathrm{S}}, \tag{2.9}$$

which is the signal-limited case. If $n_\mathrm{B} \gg n_\mathrm{S}$, Eq. (2.8) reduces to

$$\mathrm{SNR}_\mathrm{BL} = \frac{n_\mathrm{S}}{\sqrt{n_\mathrm{B}}}, \tag{2.10}$$

which is the background-limited case. As will be shown in the next section, lidar signals decrease very rapidly with range, so lidar SNR may well be signal limited at close ranges, background limited at far ranges, and described by Eq. (2.8) in between. The predictions of Eq. (2.8) are shown as a nomogram in Figure 2.3, which illustrates the decrease in SNR caused by the background. For example, achieving SNR = 10 only requires 100 photon detections when n_B is small, but a factor of 100 more when $n_B = 10^6$. The two basic ways of minimizing background photon detections are discussed in Section 2.3.

The effect of statistical noise and the meaning of optical SNR are illustrated graphically in Figure 2.4. The data points were produced in a spreadsheet random number generator as Gauss distributions with the standard deviation equal to the square root of the mean. Sample numbers 1–200 represent the background, with no signal. The background has a mean value of 1000. The standard deviation is therefore about

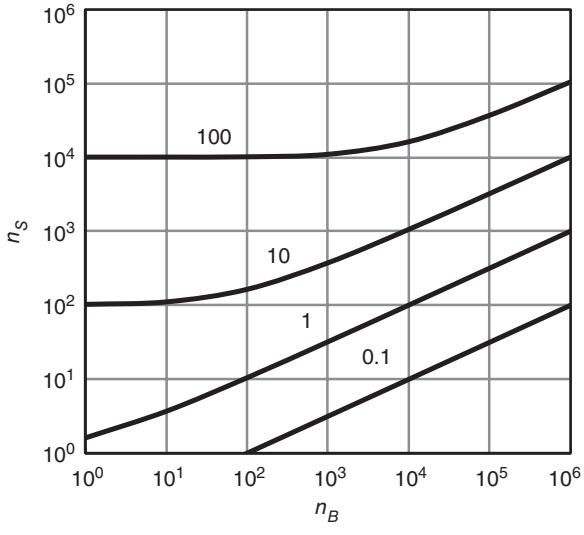


Figure 2.3 An SNR nomogram. The heavy black lines have constant SNR values given by their labels, for n_S and n_B values ranging from 1 to 10^6.

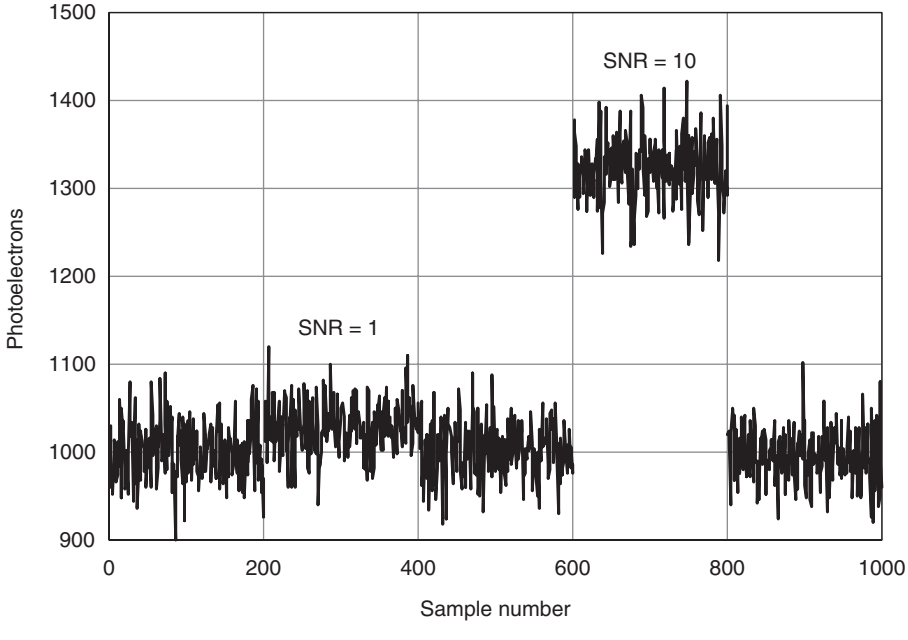


Figure 2.4 SNR illustrations. A signal with SNR = 1 is almost undetectable, whereas SNR = 10 is easy to detect.

32, and almost all the data is within ±96 photoelectrons, in accordance with Figure 2.2. Samples 201–400 represent SNR = 1, where the mean value has been increased by 1 standard deviation, to 1032. Samples 401–600 are the same as samples 1–200.

A signal with SNR = 1 is very difficult to detect. Samples 601–800 represent SNR = 10, where the mean value has been increased by 10 times the standard deviation, to 1320, and samples 801–1000 are the same as samples 1–200. When SNR = 10, the signal is obvious. A common rule of thumb is that a signal (such as lidar backscatter from a thin haze layer) can be detected when the SNR is 3 or greater.

The foregoing discussion of Poisson and Gauss statistics describes an ideal situation in which the "noise" is simply unavoidable statistical fluctuations. The SNR achieved by an actual lidar system may well be lower because of additional electronic noise sources, but it will never be higher. Equation (2.8) therefore gives an upper limit to SNR, which is valuable both for engineering new systems and for understanding the performance of existing systems. The numbers n_S and n_B depend on both instrumental and atmospheric parameters. To calculate them, we need mathematical models. The signal is modeled by the lidar equation described in the following section, and the background model follows in Section 2.3.

2.2 The Lidar Equation

The models presented in this chapter are for the most basic type of lidar, known as *elastic backscatter lidar*. The word "elastic" means that the backscattered photons are due to elastic scattering and hence have the same wavelength as the transmitted photons. The lidar receiver is also assumed to be the most basic type, which is insensitive to polarization. The models predict the numbers of signal and background photons that reach the lidar system's detector in each sampling interval. However, not all those photons are detected. Optical detectors are characterized by a quantum efficiency η, also called QE, which is the fraction of incident photons that are detected, so the signal and background photon numbers must be multiplied by η to find the parameters n_S and n_B that appear in Eq. (2.8).

The models can be used to generate synthetic lidar signals. Although the lidar's received power is a continuous time-varying signal for each laser pulse, lidar signals are always digitized at discrete range intervals to facilitate averaging, storing, and analyzing them. There are two options for creating digital lidar data: photon counting and analog-to-digital conversion. Photon counting is conceptually the simplest. When the laser pulse is transmitted, the data system begins counting the received photons in successive time increments of duration τ, known as *sampling intervals*. As described in Chapter 1, the corresponding range increments are then $\Delta R = c\tau / 2$.

Example. If the data acquisition rate is 10 Msamples/s (ten million samples per second), then τ is 100 ns and ΔR is 15 m ($c = 3 \times 10^8$ m/s, within 0.07%). The range increments are universally known as *range bins*. The lidar equation and the background model describe the numbers of photons received in each of these bins, from some minimum range to a maximum range. A 10 Msamples/s data system with 2000 range bins records lidar signals from 0 to 30 km. In analog-to-digital

conversion data systems, the continuous signal is electronically sampled at time increments τ that correspond to range increments ΔR in the same way. Again, these range increments are universally referred to as bins.

The lidar signal format is shown schematically in Figure 2.5. Thousands of range bins are typically used, and modern digitizers produce tens of thousands of counts (although photon-counting lidars are limited to a few counts per bin, from each laser pulse). A model for a photon-counting lidar is described first in this chapter, for clarity.

The lidar signal is provided by *backscattering*, which is scattering of photons directly back at the lidar system, as illustrated in Figure 2.6. As a pulse of laser light propagates upward, photons (indicated by wavy arrows) are scattered continuously in all directions, and some of them are scattered in the direction of the receiver telescope. The spatial extent of the pulse is assumed to be smaller than the range bin length. The number of photons received from a range bin is proportional to the bin length ΔR, and the laser beam cross-sectional area and the bin length define the *scattering volume* where the scattering occurs. Only a small fraction of the scattered photons is received by the lidar, and that fraction is proportional to A / R^2, which is the solid angle subtended by the receiver. For optical wavelengths, scattering is caused by both gas molecules and particles of matter in the atmosphere, which include both aerosols and the water droplets or ice crystals in clouds. The strength of scattering directly backward toward the receiver is given by the volume backscatter coefficient $\beta(R)$. The origins of the lidar parameter $\beta(R)$ are explained in Chapters 3 and 4.

The number of photons per range bin received at the detector is a product of several terms:

1) The number of photons in the laser pulse,
2) optical efficiency factors that include all losses in the system,
3) a function that describes losses at close ranges due to geometrical effects,
4) the solid angle that the receiver telescope subtends,
5) the backscatter coefficient $\beta(R)$, and
6) the two-way transmittance to and from the scattering volume.

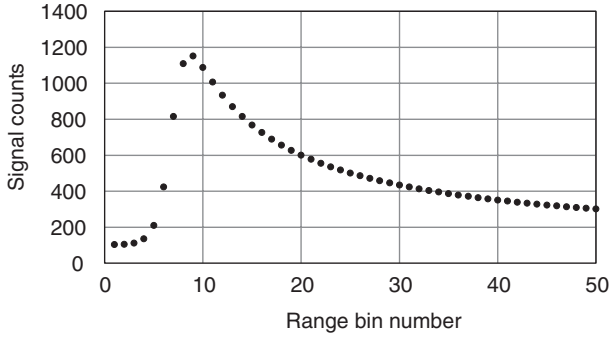


Figure 2.5 The lidar signal. The signal, in either photon counts or digitizer counts, is recorded at discrete ranges known as range bins.

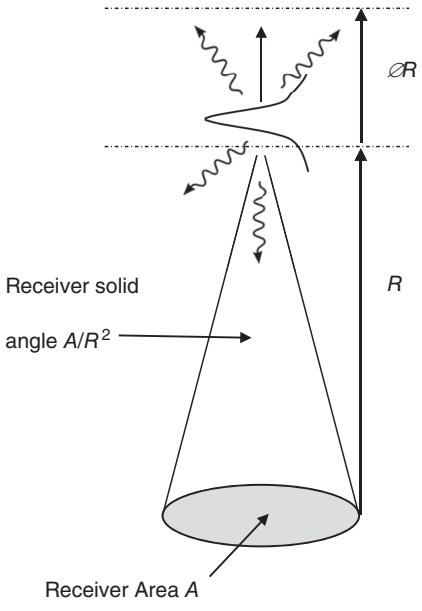


Figure 2.6 Lidar geometry. As a pulse of laser light propagates upward, photons (indicated by wavy arrows) are scattered continuously in all directions and a small fraction of them is scattered into the solid angle subtended by the receiver telescope.

Symbolically, the lidar signal model is written as

$$N_S(R) = N_0 k_T k_R \, G(R) \left(\frac{A}{R^2} \right) (c\tau / 2) \beta(R) \exp\left[-2 \int_0^R \sigma(r) dr \right], \qquad (2.11)$$

where

$N_S(R)$	is the number of laser photons received in a bin at range R per laser pulse,
N_0	is the number of photons in each laser pulse,
k_T	is the optical efficiency of the transmitter,
k_R	is the optical efficiency of the receiver,
$G(R)$	is the geometrical function,
$\left(\frac{A}{R^2} \right)$	is the receiver solid angle (sr),
$c\tau / 2$	is the range bin length (m)
$\beta(R)$	is the volume backscatter coefficient ($m^{-1}sr^{-1}$), and
$\sigma(r)$	is the extinction coefficient (m^{-1}).

Equation (2.11) is known as the *lidar equation* and it rests on several implicit assumptions: The entire laser pulse is assumed to be contained within one range bin; all received photons have experienced only one scattering event during their trip through the atmosphere; and only one laser pulse is in the atmosphere during the time in which the data system is recording a profile. The photon-counting lidar equation is dimensionless on both sides, as it must be. The instrumental parameters are conventionally grouped at the beginning of the equation, and the atmospheric parameters

β and σ appear only in the last two terms. It may seem odd that the lidar equation is written with range as a continuous variable when lidar data are always recorded in discrete range bins. The reason for this convention is that it facilitates the development of various schemes for extracting atmospheric information from lidar data, because functions of continuous variables are easier to deal with. Algorithms for processing lidar data are of course implemented in terms of discrete variables (see Chapter 10). The parameters in the lidar equation are discussed term by term in the following paragraphs.

The number of photons per laser pulse N_0 is found from the energy of each photon, which is given by the relation $E_{\text{phot}} = h\nu$, where h is the Planck constant (6.63×10^{-34} J·s), and ν is the photon frequency (Hz). In the field of optics, wavelength λ is generally used instead of frequency, with the conversion being $\nu = c / \lambda$, so $E_{\text{phot}} = hc / \lambda$. The pulse energy E_{pulse} is equal to $N_0 hc / \lambda$, so $N_0 = E_{\text{pulse}} \lambda / hc$. The numerical value of hc can be approximated as 2.00×10^{-25} with an error of less than 2%.

Example. A laser operating at 532 nm wavelength with a pulse energy of 1.00 J produces $N_0 = (1.00 \times 532 \times 10^{-9}) / (2.00 \times 10^{-25}) = 2.66 \times 10^{18}$ photons per pulse. This is a truly huge number! The energy per photon is very small, so the number of photons per joule is very large, and the arrival rate of the photons can easily exceed the capabilities of photon counting data systems at short ranges. However, other factors in the lidar equation can be very small, especially at long ranges, so it is not unusual for the average number of photons per range bin to be less than 1 per laser pulse in actual measurements. Multi-pulse averaging is necessary in such cases, and high-altitude lidars often average for hours.

The parameters k_T and k_R represent all the losses in the transmitter and receiver optical systems due to mirror reflectances and the transmittances of optical elements that are less than unity. For any one element, such losses may be small, but the compounded loss from several elements can be serious and it is not unusual for k_R to be as small as 0.1 due to the cumulative effect of many optical components. For example, consider a receiver with two mirror surfaces that reflect 88% of incident light, eight antireflection (AR) coated lens/window surfaces that transmit 99%, and one narrowband filter that transmits 45%: $k_R = 0.88^2 \times 0.99^8 \times 0.45 = 0.32$, so two-thirds of the photons that make it back to the receiver aperture are lost. The laser is often the highest-cost component of a lidar, so those photons are expensive. For this reason, it is important to minimize the number of elements in the optical system and maximize their reflectances and transmittances. Reflectance and transmittance may also have a strong dependence on polarization at angles away from normal incidence, potentially making k_T and k_R even smaller (see Chapters 5 and 6).

The geometrical function $G(R)$ has the value zero at the lidar (zero range) and it smoothly rises to unity at the *crossover range*, as illustrated in Figure 2.7. For this reason, lidar systems are generally not capable of measurements starting at zero range but rather have some minimum measurement range. The optical reasons for this behavior, and techniques for engineering $G(R)$, are described in Chapter 6.

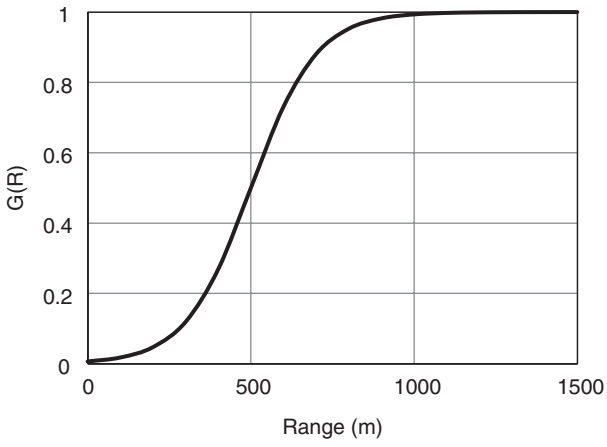


Figure 2.7 The geometrical function $G(R)$. The shape of the function is illustrated for a lidar system with a crossover range of 1 km.

The term A/R^2 is the solid angle subtended by the lidar receiver from the scattering volume. The inverse-range-squared dependence of the lidar signal caused by this term is a purely geometrical effect that has nothing to do with atmospheric parameters, so the first step in lidar data analysis is often to remove it by multiplying the signal at all ranges by R^2 (after subtracting the background), a procedure known as range correction. However, the A/R^2 term nevertheless contributes to a very large dynamic range in lidar signals, and accommodating that dynamic range is one of the challenges of lidar engineering.

Measuring either or both atmospheric parameters β and σ is often the first goal of lidar measurements, but a problem is apparent in the model given by Eq. (2.11): The lidar equation yields only one signal value based on two atmospheric input parameters. This fact means that the mathematical problem of retrieving those parameters from the signal does not have a unique solution. Much of the lidar literature is concerned with ways to circumvent this problem, by using restrictive assumptions or scenarios and/or by constraining the solution with additional information. In practice, the situation is even worse than having one equation with two unknowns, because β and σ are caused by atmospheric gases as well as particulate matter and so they each have two or more independent components. The atmospheric parameters β and σ are discussed in Chapters 3 and 4.

The term $\exp\left[-2\int_0^R \sigma(r)dr\right]$ is the two-way transmittance from the lidar to the scattering volume and back (the dummy integration variable r is the range). This term is often abbreviated as $T^2(R)$ in the lidar equation. With this convention, if all instrumental parameters are lumped into a constant C and both sides are multiplied by R^2, then the range-corrected lidar equation becomes $R^2 N_S(R) = C\beta(R)T^2(R)$ at

ranges where $G(R)$ is equal to unity. The range-corrected signal $R^2 N_S(R)$ is often called $X(R)$ in the lidar literature and that notation is used in this book, that is, $X(R) = C\beta(R)T^2(R)$. Maximizing $X(R)$ is generally desirable because of its effect on SNR. The instrumental parameters N_0, k_T, k_R, and A all appear as factors in the lidar equation, so increasing any of them will increase $X(R)$ (the sampling interval τ could also be increased at the cost of lower range resolution, but this would not represent a real gain in performance). The considerations for engineering the lidar transmitter are covered in Chapter 5.

As mentioned earlier, there are two basic types of detection used in lidar, photon counting and analog. For analog detection systems, the power form of the lidar equation is required, which is

$$P_S(R) = P_0 k_T k_R \, G(R) \left(\frac{A}{R^2} \right) (c\tau_{\text{pulse}} / 2) \beta(R) \exp\left[-2 \int_0^R \sigma(r) dr \right], \qquad (2.12)$$

where τ_{pulse} is the pulse width (s) and $P_0 = E_{\text{pulse}} / \tau_{\text{pulse}}$ is the average power in the laser pulse. The power form of the lidar equation is more common in the lidar literature than Eq. (2.11). Equation (2.12) is a model of received laser power on the detector, in watts, as a function of range.

2.3 The Background Model

As explained in Section 2.1, the background, whether sky or terrain, is a source of photons that may be received along with the backscattered laser photons, decreasing the SNR. This is a serious problem, especially for visible-light lidars operating during daytime. The rate at which background photons arrive at the detector is determined by the background's *spectral radiance* L_λ, along with instrumental parameters. Spectral radiance characterizes the radiative transfer of power, and it has the units of $\text{Wm}^{-2}\mu\text{m}^{-1}\text{sr}^{-1}$. The optical power transferred from the background to a lidar system's detector is found by multiplying the background's spectral radiance by the receiver's area, optical bandpass, and field of view (FOV) expressed as solid angle. The receiver area is the parameter A, which also appears in the lidar equation. An optical filter with a narrow bandpass is necessary in lidar receivers because the daytime background is due to scattered sunlight and it is very broadband, extending from the ultraviolet to the infrared, and a lidar's detector may be sensitive to much of this wavelength range. For this reason, all lidar systems employ a narrow-band optical filter with its peak transmittance centered on the transmitted laser line, to block as much of the background as possible. The spectral width of this filter determines the receiver's optical bandpass B_{opt} (μm). Lidar receivers also employ a small FOV to minimize the arrival rate of background photons. FOV is most often specified as a plane angle θ in radians. However, the arrival rate of photons is proportional to the solid angle Ω, in steradians. For small angles, the relationship between the plane and solid angles

is given by $\Omega = (\pi / 4)\theta^2$. Multiplying these terms together, along with the receiver efficiency, yields the background model for analog detection (the background power P_B on the detector) as

$$P_B = k_R L_\lambda A \Omega B_{opt}. \tag{2.13}$$

The photon-counting background model must give the number of photons per sampling interval, as Eq. (2.11) does. Power equals energy/time, so optical power is proportional to the rate at which photons are received, because each of them carries energy $E_{phot} = h\nu = hc / \lambda$. If N_B background photons arrive in a sampling interval τ, the optical power is then $P_B = (N_B / \tau)(hc / \lambda)$, and by inverting this relation, we find $N_B = P_B \tau (\lambda / hc)$. The photon-counting background model is therefore

$$N_B = k_R L_\lambda A \Omega B_{opt} \tau (\lambda / hc), \tag{2.14}$$

where

- N_B is the number of background photons received in each bin for each laser pulse,
- k_R is the optical efficiency of the receiver,
- L_λ is spectral radiance of the background ($Wm^{-2}\mu m^{-1}sr^{-1}$),
- A is the receiver area (m^2),
- Ω is the receiver FOV as solid angle (sr),
- B_{opt} is the receiver optical bandpass (μm),
- τ is the sampling interval (s), and
- λ is the photon wavelength (m).

The photon-counting background equation is dimensionless on both sides like Eq. (2.11), but N_B is not a function of range because the background radiance is assumed to be constant during the short acquisition time of the lidar profile. For this reason, the signal shown schematically in Figure 2.5 was drawn with 100 counts at zero range and it asymptotically decreases toward 100 counts at long ranges. Those counts were intended to represent the background. Equation (2.14) shows that N_B is proportional to the instrumental parameters k_R, A, Ω, and B_{opt}, so N_B could be reduced by decreasing any one of these instrumental parameters. Reducing k_R or A would reduce the SNR by reducing $N_S(R)$ as shown by Eq. (2.11), so the only useful engineering parameters for minimizing N_B are Ω and B_{opt}. Techniques and trade-offs for minimizing those parameters are discussed in Chapter 6. As mentioned earlier, Eqs. (2.11) and (2.14) predict the numbers of laser and background photons that reach the lidar system's detector in each sampling interval, but not all of those photons will be detected, so $N_S(R)$ and N_B must be multiplied by the quantum efficiency η to find the parameters n_S and n_B that appear in Eq. (2.8).

Some lidars are operated only at night and they may have detectors with low dark count rates. In that case, the SNR depends only on n_S, as shown in Eq. (2.9), and one way to rate the performance of such a lidar system is based on how quickly it can reach a given SNR value by multi-pulse averaging. This idea gave rise to an

often-used measure of merit for night-only systems, the *power-aperture product*, in units of W-m^2. With multi-pulse averaging, the time-averaged transmitted power is $k_T E_{\text{pulse}}$ times the *pulse repetition frequency* (PRF). The power-aperture product is a somewhat crude measure of merit because it does not include η nor k_{R}, but it is commonly used in the lidar literature to compare one lidar system to another. The situation is different during daytime, because a lidar with a high pulse energy and low PRF will yield higher SNR than a low pulse energy and high PRF system of the same average power, in the presence of background light.

2.4 Example Lidar System

The application of the lidar equation and the background model is illustrated with a real lidar system, the educational lidar known as EARL (Eye safe Atmospheric Research Lidar) [1, 2]. EARL was developed by the Georgia Tech Research Institute (GTRI) lidar team in collaboration with Agnes Scott College in Decatur, Georgia. Agnes Scott College is an undergraduate women's college, so EARL was designed for training and research in the undergraduate environment. EARL embodies many of the engineering principles that are discussed in this textbook and it also has some shortcomings, so it will be used as an example lidar in several chapters. Eye safety was deemed to be essential for an undergraduate lidar, so EARL employs the *micro-pulse lidar* technique in which the laser pulse energy is only tens of microjoules. This type of lidar, which was pioneered by Spinhirne in 1993 [3], is made eye safe by combining low pulse energy with a large transmitted laser beam diameter, but it must use extensive multi-pulse averaging to achieve useful SNR, and small values of the FOV and optical bandpass are required for daytime operation. The overall optical configuration of EARL is shown in Figure 2.8 and its main design parameters are listed in the Table 2.1. EARL has two receiver channels, short range and long range. Only 9% of the received light is directed to the short-range receiver channel, which is why its efficiency is so low. EARL's transmitter is discussed in Chapter 5, the receiver is discussed in Chapter 6, and the mechanical structure is discussed in Chapter 7.

Simulations of EARL's received signal and background photons per range bin are shown in Figure 2.9 to illustrate the application of the basic lidar models described in this chapter. Instrumental parameters were taken from Table 2.1, and the measurement scenario was chosen as 10 minutes of data acquisition (1.5 million laser pulses) with a daytime sky spectral radiance of 80 $\text{Wm}^{-2}\mu\text{m}^{-1}\text{sr}^{-1}$. The purpose of the simulation was to investigate EARL's maximum SNR for measurements of clear air backscatter in the free troposphere during both day and night, with an SNR goal of at least 100 at 4 km altitude during daytime. The nighttime sky radiance was assumed to be 6 orders of magnitude smaller than daytime. Aerosol extinction was included in this simulation by assuming that the two-way transmission to the

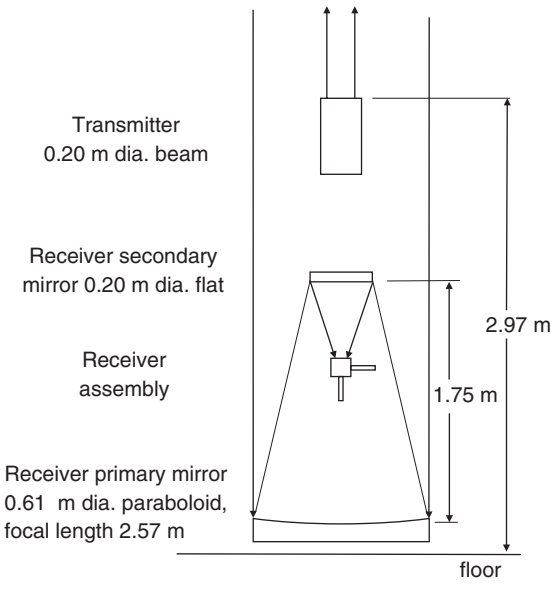


Figure 2.8 EARL overall optical configuration. The lidar is coaxial, with the transmitter centered above the receiver. A flat mirror folds the light path down into the receiver, which has short-range and long-range channels.

Table 2.1 Design parameters of EARL

Transmitter parameter	Symbol	Value	
Wavelength	λ	523.5 nm	
Laser pulse energy	E_{pulse}	50 μJ	
Pulse repetition frequency	PRF	2.5 kHz	
Optical efficiency	k_{T}	0.30	
Receiver parameter		Long range	Short range
Primary mirror diameter	D_{R}	0.61 m	0.61 m
Field of view	FOV	0.389 mrad	1.17 mrad
Optical bandpass	B_{opt}	0.21 nm	1.06 nm
Digitization rate	-	10 MSample/s	10 MSample/s
Detector quantum efficiency	η	0.12	0.12
Optical efficiency	k_{R}	0.183	0.028

tropospheric altitudes was 0.80. The clear air backscatter coefficient was taken to be $\beta = 1.39 \times 10^{-6} \times (550/523.5)^4\,(\text{m}^{-1}\text{sr}^{-1})$ at sea level (see Chapter 3), decreasing with altitude with a scale height H of 7 km (see problem 5). EARL's elevation is 0.300 km. EARL's detectors have a very low dark count rate and they are sensitive to single photons. Although EARL does not quite achieve the statistical limit of SNR,

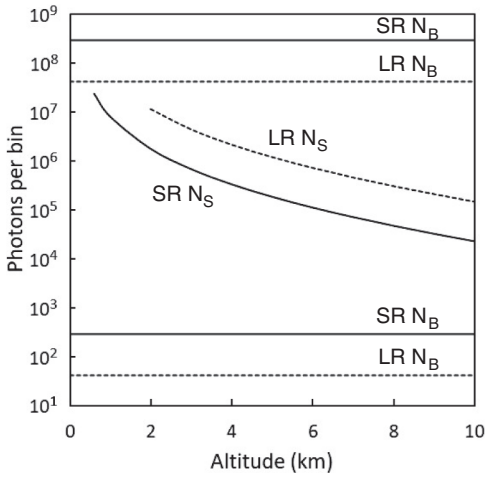


Figure 2.9 EARL simulations. Numbers of received signal and background photons per range bin from 1.5 million laser shots versus altitude are plotted. Results are plotted only for altitudes where $G(R)$ is unity. Solid black lines pertain to the short-range receiver channel (SR) and dashed lines to the long-range receiver channel (LR). Upper background levels are daytime, and lower levels are nighttime.

it does come close to it and the simulations are fairly representative of actual measurements. SNR calculations based on the simulations are left as a problem for the student. Both receiver channels are background limited during day and signal limited at night, which is typical of micro-pulse lidars.

2.5 Further Reading

U. Wandinger, "Introduction to Lidar," in *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere*, C. Weitcamp, Ed. New York: Springer, 2005, pp. 1–18.

See this chapter for the history of lidar, another derivation of the lidar equation, and other insights. The notation used in this textbook is largely consistent with Wandinger's chapter.

H. D. Young, *Statistical Treatment of Experimental Data*. New York: McGraw-Hill, 1962.

This book is a concise presentation of statistics as it relates to measurements. It covers the statistics material in this chapter and much more, and it serves as an excellent resource.

2.6 Problems

2.6.1 The Poisson distribution given in Eq. (2.5) is often said to closely resemble the Gauss distribution given in Eq. (2.7) for mean values of 6 or greater. Is this correct? Plot the two distributions for mean = 6 on the same graph and compare them.

2.6.2 A lidar researcher makes a nighttime measurement at a certain range only once, and he collects 100 lidar photoelectrons. What is his best estimate of the average value he would find if he repeated the measurement many times? What is his best estimate of the standard deviation of each set of measurements?

2.6.3 In Figure 2.3, SNR increases with n_S by a factor of 10 every decade on the right vertical axis (background limited), but only with every two decades on the left axis (signal limited). Does this fact mean that the common lidar technique of multipulse averaging can yield a linear increase in SNR in background-limited conditions? Using the general formula in Eq. (2.8), show that the increase in SNR obtained by summing the photon detections from M pulses is always proportional to $\sqrt{M}$.

2.6.4 Estimated errors are normally expressed as fractions of the mean value, for example, ±10%. How is the fractional uncertainty related to the SNR?

2.6.5 The Standard Atmosphere air density profile plotted in Figure 1.3 is almost a straight line on the semi-logarithmic graph, which means that the decrease with altitude is very nearly exponential. Using the data shown in Figure 1.3, find the *scale height* H of the atmospheric density, in the relation $\rho(h) = \rho_0 \exp(-h / H)$, with h and H in km. Assume that $\rho(1)$ is 1 kg/m^3 and $\rho(100)$ is 1×10^{-6} kg/m^3. How much of the mass of the atmosphere is in the troposphere? Assume the tropopause height is 15 km as shown in Figure 1.2 and use the scale height found above.

2.6.6 Did EARL meet the design goal of 10-minute measurements of clear air backscatter at 4 km altitude during both day and night with an SNR of at least 100? Values of signal and background photons are shown in the table below for both receiver channels during both day and night, for the conditions of Figure 2.9. Fill in the missing entries, recalling that the quantum efficiency is 0.12. Find the SNR values using Eq. (2.8). The dark count rate is effectively zero.

		N_S	n_S	N_B	n_B	SNR
Short range	Day	3.33×10^5		2.94×10^8		
	Night	3.33×10^5		2.94×10^2		
Long range	Day	2.15×10^6		4.14×10^7		
	Night	2.15×10^6		4.14×10^1		

References

[1] L. L. West, G. G. Gimmestad, D. W. Roberts et al., "Atmospheric Laser Radar as an Undergraduate Educational Experience," *American Journal of Physics*, vol. 74, pp. 665–669, 2006.

[2] H. N. Forrister, D. W. Roberts, A. J. Mercer, and G. G. Gimmestad, "Infrared Lidar Measurements of Stratospheric Aerosols," *Applied Optics*, vol. 53, pp. D40–D48, 2014.

[3] J. D. Spinhirne, "Micro Pulse Lidar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 31, pp. 48–55, 1993.

3 The Molecular Atmosphere

The lidar models introduced in Chapter 2 require both instrumental and atmospheric input parameters. The purpose of this and the next chapter (Chapter 4) is to describe the atmospheric parameters and some procedures for determining them. All lidar measurement techniques are enabled by the various ways that the atmosphere affects laser beams, which have several unusual characteristics when compared with other types of light: They are monochromatic, meaning that they are single wavelength; they are highly directional; and they usually have a well-defined polarization state. Our goal is to understand the effects of the atmosphere's constituents and its state (temperature, pressure, and winds) on laser beams. The study of these effects is a specialized subtopic of the subject known as *atmospheric optics*, which has been called "… the province of all alterations of the characteristics of light as it propagates in the atmosphere" [1]. Atmospheric optics is a popular and fascinating subject, because so many interesting and beautiful phenomena can be observed in the open air with the naked eye [2, 3], but this chapter and Chapter 4 have a narrower focus that is specific to lidar. In this chapter, the effects of atmospheric molecules on laser beams are examined, beginning with an overview of optical scattering and absorption followed by a review of atmospheric structure and constituents, and finally, the physical bases of all main types of lidar techniques are elucidated with a discussion of the ways that energy is stored by molecules. Interactions of laser beams with particulate atmospheric matter are described in Chapter 4.

3.1 Overview of Atmospheric Scattering

The lidar transmitter sends a directional beam of pulses of light into the atmosphere, where the photons encounter molecules, aerosols, and water droplets or ice crystals in fog and clouds. Typical order-of-magnitude sizes and concentrations are shown for some representative atmospheric scatterers in Table 3.1, which was adapted from E. J. McCartney's book on atmospheric optics [1], which is an invaluable resource for lidar researchers. Much of this chapter, especially Section 3.1, was adapted from that book. The sizes and concentrations span very large ranges, but in any practical situation, the lidar scattering volumes include huge numbers of scatterers (Table 3.1). For this reason, the optical properties of the atmosphere are always due to averages, rather than the specific encounters of one photon with one particle. The large numbers of particles involved cause good statistical estimates of the average properties.

Table 3.1 Atmospheric light scatterers

Type	Radius (μm)	Concentration (m^{-3})
Air molecule	10^{-4}	10^{25}
Aitken nucleus	10^{-2}–10^{-1}	10^{10}–10^{8}
Haze particle	10^{-2}–1	10^{9}–10^{7}
Cloud droplet	1–100	3×10^{8}–10^{7}

3.1.1 Scattering Formalism

As the laser beam propagates through the atmosphere, several types of interactions with matter may occur, but the dominant phenomenon is scattering. The blue sky is caused by scattered sunlight for example, and lidar signals are caused by scattering of laser light in the backward direction. For this reason, optical scattering is the first phenomenon to consider. In keeping with standard treatments, a radiometric approach is used in which the quantity of interest is power P (watts, W), and beams of light are characterized by irradiance E (W/m^2) and radiant intensity I (W/sr). These quantities generally vary with wavelength, but for clarity, the wavelength dependence is not shown here. The parallels between power levels derived from radiometry and photon arrival rates from pulsed laser light are drawn at the end of this section.

Consider first a collimated beam of light on the x-axis with unit cross-sectional area and irradiance E_0 incident on a unit volume of scatterers, as shown in Figure 3.1. The effect of scattering is to remove optical power from the forward direction and send it into all directions, meaning a solid angle of 4π steradians. Because power has been removed, the exiting irradiance E is less than E_0. The scattering angle θ is measured from the forward direction, and the incident beam direction and the scattering direction define the *scattering plane*. Assuming that the scattered intensity is axially symmetric about the forward direction, the angular distribution of the scattered light can be characterized by $I(\theta)$ from 0 to 180 degrees (0 to π radians). The assumption of axial symmetry is justified by the fact that, in any practical lidar scenario, the pulses of light encounter large numbers of scatterers that are almost always randomly oriented (the main exception is ice crystals that are oriented by aerodynamic forces as they fall).

The radiant intensity per unit scattering volume is related to the incident irradiance by $I(\theta)/m^3 = \beta(\theta)E_0$, where the constant of proportionality $\beta(\theta)$ is known as the *volume angular scattering coefficient*, in units of m^2/m^3-sr or 1/m-sr. The total power removed from the incident beam by scattering in the unit volume is given by the integral of $I(\theta)$ over solid angle as

$$P/m^3 = \int_0^{4\pi} I(\theta)d\omega = E_0 \int_0^{4\pi} \beta(\theta)d\omega = E_0\beta, \tag{3.1}$$

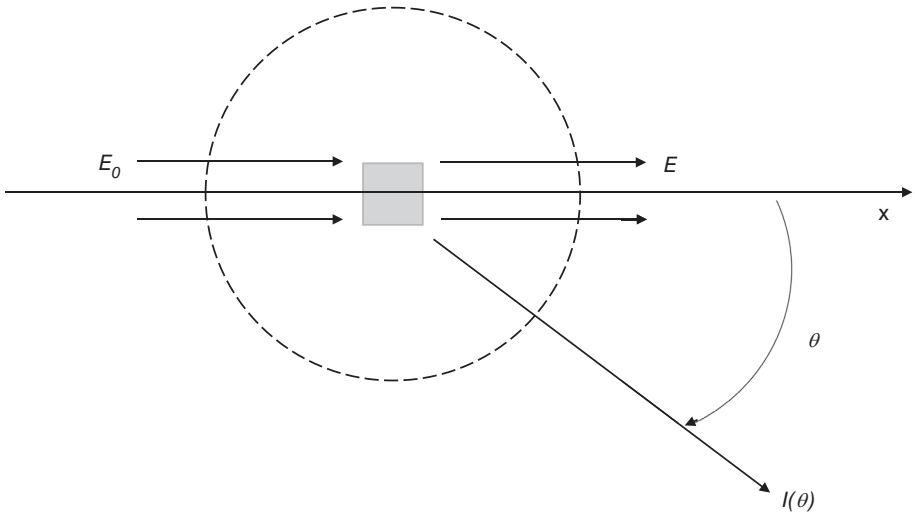


Figure 3.1 Scattering. A light beam incident on a unit volume of scatterers (shown in gray) is scattered into a 4π solid angle represented by the dashed circle. The scattering angle θ is measured from the forward direction, and the radiant intensity in that direction is $I(\theta)$.

where the relationship between β and $\beta(\theta)$ is $\beta = \int_{4\pi} \beta(\theta) d\omega$. The units of β are square meters per cubic meter, or m^{-1}. The decrease in irradiance ΔE after the light travels through 1 m of the scattering medium is the incident irradiance E_0 times the parameter β, which is called the *volume total scattering coefficient*. The proportional loss of irradiance is the same for a differential distance dx through the medium, so we can write $dE / E = -\beta dx$, which is the classic differential equation for an exponential decrease with distance. This equation has the solution

$$E(x) = E_0 \exp(-\beta x). \tag{3.2}$$

Equation (3.2) has many names, including *Beer's Law, Bouguer's Law*, and the *Beer-Lambert Law*. It is incorporated in the lidar equation, Eq. (2.11), in the two-way transmission term. The dimensionless argument βx in the exponential function is called the *optical depth* (abbreviated OD; also called optical thickness), and the ratio $E(x) / E_0$ is called the *transmittance* to the range x. A helpful naming convention, followed here, is that quantities ending in *-ance* have a range of values from 0 to 1. This is to distinguish them from quantities such as *transmission,* which is usually expressed as a percentage, with a range of values from 0% to 100%.

Only the beam power loss due to scattering is included in β as defined above. The other potential loss is caused by *absorption*. If light is absorbed by a particle, its energy is converted to heat, which is quickly exchanged with surrounding air molecules, with the effect of heating the air on the beam path. The heating generally has a negligible effect on lidar measurements. If a gas molecule absorbs light, its internal energy state is raised to a higher level, and it will tend to return to its original state

either by emitting a photon or by losing energy in collisions with other molecules, which again leads to beam path heating. The total of all power losses in a light beam is called *extinction*, and the *extinction coefficient* β_{ext} is defined as

$$\beta_{\text{ext}} = \beta_{\text{sca}} + \beta_{\text{abs}}, \tag{3.3}$$

where the subscripts ext, sca, and abs refer to extinction, scattering, and absorption, respectively. Using β_{ext} in Beer's law and squaring yields the two-way transmission from a lidar system to a scattering volume at range R and back, which is

$$T^2(R) = \exp\left[-2\int_0^R \beta_{\text{ext}}(r)dr\right], \tag{3.4}$$

where we have integrated β_{ext} along the beam path because it is not generally constant with range.

The other term needed in the lidar equation is the backscatter coefficient $\beta(\pi)$, which is the value of the volume angular scattering coefficient $\beta(\theta)$ at the scattering angle π. The angular scattering coefficient is related to the total scattering coefficient by the relation

$$\beta(\theta) = P(\theta)\beta, \tag{3.5}$$

where $P(\theta)$ is the *phase function*, which describes the angular pattern of scattered radiation. The phase function can be defined as the ratio of the power per unit solid angle scattered in the direction θ to the average power per unit solid angle scattered in all directions. With that definition, the normalization condition on $P(\theta)$ is

$$\frac{1}{4\pi}\int_0^{4\pi} P(\theta)d\omega = 1. \tag{3.6}$$

The phase function is quite sensitive to the ratio of the wavelength to the particle size. Atmospheric scattering volumes generally contain a mix of gas molecules (which are very small compared to laser wavelengths) and aerosols (which may be comparable to wavelength), and for this reason, the two constituents generally have very different phase functions, and the total lidar backscatter coefficient is

$$\beta(\pi) = P_{\text{a}}(\pi)\beta_{\text{a}} + P_{\text{m}}(\pi)\beta_{\text{m}}, \tag{3.7}$$

where the subscripts a and m refer to aerosols and molecules, respectively. There are other subscript notations in the lidar literature, but these are used here for simplicity, along with p for particle and atm to mean the total atmospheric effect. The lidar equation is sometimes written using the notation in Eqs. (3.5) and (3.7) as

$$\begin{aligned} N_{\text{L}}(R) = {} & N_0 k_{\text{T}} k_{\text{R}}\, G(R)\left(\frac{A}{R^2}\right)(c\tau/2)[P_{\text{a}}(\pi,R)\beta_{\text{a}}(R) \\ & + P_{\text{m}}(\pi,R)\beta_m(R)\exp\left[-2\int_0^R \beta_{\text{ext}}(r)dr\right] \end{aligned} \tag{3.8}$$

which is a good practice because it harmonizes the lidar notation with standard treatments of light scattering and absorption, and it explicitly separates the molecular and aerosol backscatter effects. However, most of the lidar community long ago merged the products of the phase functions and total scattering coefficients into a specialized

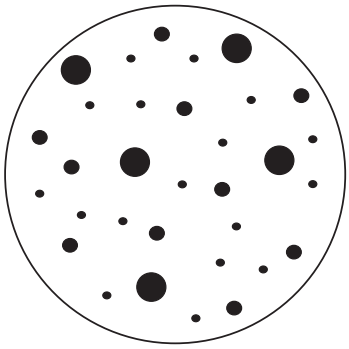


Figure 3.2 A scattering volume. The outer circle represents a light beam with unit area, and the dark-filled circles represent the extinction cross sections of individual scatterers.

lidar volume backscatter coefficient simply called β ($m^{-1}sr^{-1}$), as shown in Eq. (2.11), and the extinction coefficient β_{ext} in Eq. (3.3) is generally replaced with the symbol σ as shown in Eq. (2.11). Nevertheless, the origins and meanings of these specialized parameters, as described above, should be borne in mind.

The discussion of scattering that led to Eqs. (3.1) and (3.5) was in terms of radiometric quantities (P, I, and E) that all involve power. Those quantities usually describe a static situation, but the results also apply to the lidar scenario in which a short pulse of light propagates through the scattering volume losing scattered photons as it goes, as illustrated in Figure 2.6, because Eq. (2.11) yields photons per sampling time, which is proportional to power. The correspondence is perhaps more obvious when the lidar equation is written in terms of power as in Eq. (2.12). Thinking in terms of photons rather than power leads to a simple interpretation of the volume total scattering coefficient β, which is illustrated in a schematic way in Figure 3.2 as a view from the left side of Figure 3.1, along the x-axis. The outer circle represents the boundary of a light beam with unit area, and the dark filled circles represent the *extinction cross sections* of individual scatterers. The depth of the scattering volume is one meter, so its volume is one cubic meter. The probability that an incident photon will be scattered or absorbed when passing through the volume is just the total extinction cross section of the scatterers divided by the beam area (unity), and that ratio is the total scattering coefficient β (square meters per cubic meter). When the incident pulse contains N_0 photons, the number removed by scattering is $N_0\beta_{ext}$, which leads to the Beer's Law relation $N(x) = N_0 \exp(-\beta_{ext}x)$, which is analogous to Eq. (3.2). In this conceptual picture, the extinction coefficient β_{ext} is the probability that an incident photon will be scattered or absorbed as it propagates through 1 m of the scatterers.

The derivations above assumed a scattering volume 1 m long, but practical lidar range bins are from a few meters to hundreds of meters in length. This fact does not present a problem, assuming that the scatterers are distributed homogeneously in space, because β has dimensions of inverse length, meaning per meter of scattering

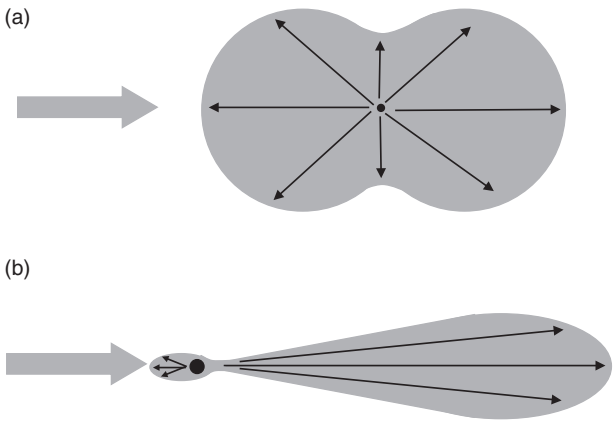


Figure 3.3 Phase functions for small (a) and large (b) particles. The scattering particle is shown by a black dot, and the distribution of scattered light is shown in polar plots by the gray areas and arrows.

volume, and it is multiplied by the bin length in the lidar equation. The lidar equation therefore accommodates any range bin length greater than the laser pulse length. The derivations were also based on a laser beam with unit area, but lidars employ beams that are usually diverging, ranging in size from 0.01 meter to tens of meters. This fact does not affect the validity of the results either, again assuming homogenous scatterers. The reason is that the beam irradiance E scales inversely with area, whereas the total cross section of the scatterers in the beam scales linearly with area, so the two effects exactly cancel, and the results are therefore independent of the laser beam size if the scattering volume is within the receiver's FOV.

3.1.2 Three Types of Scattering

The phase function, which describes the angular pattern of scattered light, changes greatly with particle size, as illustrated with polar plots in Figure 3.3. For small particles, the pattern is symmetrical between the forward and backward directions, whereas for large particles, much more light is scattered in the forward direction. In addition, the wavelength dependence of scattering is quite different for small and large particles. This description leads to the question, what is meant by "small" and "large"? The answer lies in the *scattering parameter*, defined as $\alpha = 2\pi r / \lambda$, where r is a characteristic dimension of the particle and λ is the wavelength. For a sphere, r is the radius and hence α is the ratio of the circumference to the wavelength, which is a convenient way of remembering the definition of the scattering parameter. The radii of scatterers in Table 3.1 span the range from much smaller to much larger than the laser wavelengths used in lidar systems. Conventionally, there are said to be three types of light scattering by particles, delineated by the scattering parameter, as shown

Table 3.2 The three types of scattering

Scattering parameter range	Type of scattering	Wavelength dependence
$\alpha < 0.5$	Rayleigh	λ^{-4}
$0.5 < \alpha < 50$	Mie	Intermediate
$50 < \alpha$	Geometric optics	λ^{0}

in Table 3.2. Rayleigh scattering is due to particles much smaller than the wavelength, in the Mie regime the particle size is comparable to the wavelength, and geometric optics describes scattering by particles much larger than the wavelength. The boundaries between the types are somewhat subjective, and some authors recommend $\alpha < 0.2$ for Rayleigh scattering [1].

3.2 Rayleigh Scattering

The structure of Earth's atmosphere and its constituents were briefly discussed in Chapter 1. Figure 1.2 illustrates the conventional way of looking at the structure in terms of the altitude profile of temperature, while the air density profile is illustrated in Figure 1.3. As noted in Chapter 1, the atmosphere is >99% nitrogen, oxygen, and argon in a constant mix from the ground up to about 100 km. Molecules have dimensions of about 0.3 nm and lidar laser wavelengths are always at least hundreds of nanometers, so the wavelength is always much greater, and the molecular atmosphere is solidly in the Rayleigh scattering regime for lidar. The word "molecular" includes both molecules and atoms in this chapter, as it often does in atmospheric physics.

As shown in Table 3.2, when α is less than about 0.5, the scattering process is called Rayleigh scattering, after British scientist Lord Rayleigh (John W. Strutt, 1842–1919), who first investigated this regime theoretically. Rayleigh scattering has a characteristic wavelength dependence of inverse wavelength to the fourth power. In clear conditions, skylight is due to Rayleigh scattering of sunlight by air molecules. The blue end of the visible spectrum is about 400 nm and the red end is about 700 nm, so the blue wavelengths in sunlight are scattered more strongly than the red by a factor of $(700/400)^4 = 9.4$. The Rayleigh wavelength exponent of −4 is the reason that the sky is perceived as being blue. In addition to the wavelength dependence, the other main features of Rayleigh scattering are that equal amounts of light are scattered into the forward and backward hemispheres, as illustrated by the top phase function in Figure 3.3, and that when the incident light is unpolarized, the light scattered at 90 degrees is almost completely linearly polarized. These features of Rayleigh scattering are predicted by a simple classical model of an atom, in which the electronic charge surrounding the nucleus is displaced by the electric field associated with the incident light. Assuming a Hooke's law restoring force and small damping, the displacement varies sinusoidally in time, synchronously with the incident field. The nucleus, with

its positive charge, essentially remains at rest because it so much more massive than the electrons. The motion of the electronic charge therefore creates an oscillating dipole moment aligned with the incident electric field vector, which causes classical dipole radiation. The geometry is shown in Figure 3.4, where the incident wave travels from the left along the *x*-axis, with its electric field vector in the *x*–*y* plane. The molecule, with its induced oscillating dipole moment, is at the origin. The dipole radiates a secondary wave, which is a function of the angle ϕ from the *y*-axis, but the dipole radiation pattern is axially symmetric about the *y*-axis. The secondary wave radiated by the dipole has maximum power orthogonal to the dipole and no power parallel to the dipole (on the *y*-axis), with linear polarization parallel to the dipole. The radiation pattern is shown in Figure 3.5, as a polar plot in the *x*–*y* plane. To a good approximation, the volume angular scattering coefficient is given by

$$\beta_m(\phi) = \frac{\pi^2 (n^2 - 1)^2}{N\lambda^4} \sin^2 \phi, \tag{3.9}$$

where n is the refractive index of air, N is the molecular number density (m^{-3}), and λ is expressed in m. Because $n^2 - 1$ is closely proportional to N, $\beta_m(\phi)$ is proportional to air density.

Unpolarized incident light (such as sunlight) can be treated as the sum of two orthogonal polarizations. In Figure 3.5, there is no angular dependence in the scattering plane (coming out of the page) and so the pattern is just a circle of unit radius, hence the sum of the intensities for the two incident polarizations is the pattern shown in Figure 3.5 plus a circle, as shown in Figure 3.6. The result is a function of the scattering angle θ given by

$$\beta_{\mathrm{m}}(\theta) = \frac{\pi^2 (n^2 - 1)^2}{2N\lambda^4} (1 + \cos^2 \theta)(\mathrm{m}^{-1}\mathrm{sr}^{-1}), \tag{3.10}$$

which has the familiar dumbbell shape in the *x*–*z* plane shown at the top in Figure 3.3 and by the solid curve in Figure 3.6. This pattern has rotational symmetry about the *x*-axis. The scattered light is strongly polarized at scattering angles near 90 degrees, and this fact is relevant to lidar measurements because many lidar receivers are sensitive to polarization. The amount of sky background light received by such systems is therefore determined not only by the sky radiance, but also by the polarization state of the skylight. For example, a ground-based lidar aimed at the zenith observes sunlight scattered near 90 degrees whenever the Sun is near the horizon. If the receiver has two orthogonal polarization channels, their sky background signals may be very different, and their ratio may vary with time of day.

Clear-sky spectral radiance data measured at 33 degrees north latitude in Atlanta, Georgia are shown in Figure 3.7. These clear-sky spectral curves were recorded at midday in winter, at the zenith and at 3 degrees above the horizon. Because there were no clouds and the aerosol concentration was minimal, the sky radiance was nearly all due to Rayleigh scattering. It is highest in the ultraviolet-visible spectral region, and the peak value near the horizon is roughly twice as high as the zenith value. Scattering from clouds and aerosols may also contribute to the sky radiance, and as a useful rule of thumb, the highest zenith radiance that a mid-visible lidar is likely to see is about

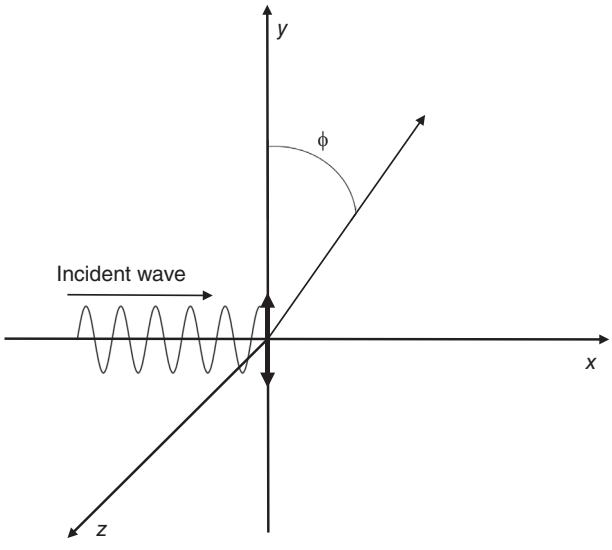


Figure 3.4 Geometry of Rayleigh scattering. The incident wave induces an oscillating dipole moment in the molecule, shown by the double-ended arrow at the origin. The dipole radiates a secondary wave.

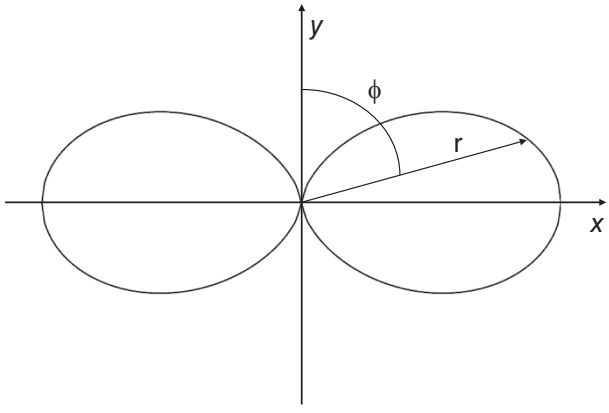


Figure 3.5 Dipole radiation pattern. The radial distance from the origin to the solid curve shows the intensity of radiation as a function of angle from the *y*-axis. The pattern has axial symmetry about the *y*-axis.

100 $Wm^{-2}sr^{-1}\mu m^{-1}$. Because of the λ^{-4} dependence of molecular scattering, the clear-sky radiance decreases very rapidly with wavelength. This effect is shown in Figure 3.8, where the zenith sky radiance data from Figure 3.7 is shown on the same plot as ASTM E490, *Zero Air Mass Solar Spectral Irradiance*, which is a model of the solar irradiance spectrum outside Earth's atmosphere [4]. For clarity, the solar spectrum

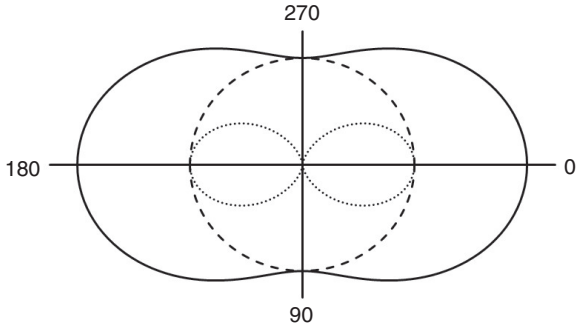


Figure 3.6 Rayleigh scattering intensity for unpolarized light. In this polar plot, the two components for orthogonal incident polarizations are shown by dashed and dotted lines and the solid line shows their sum. Scattering angles are marked in degrees.

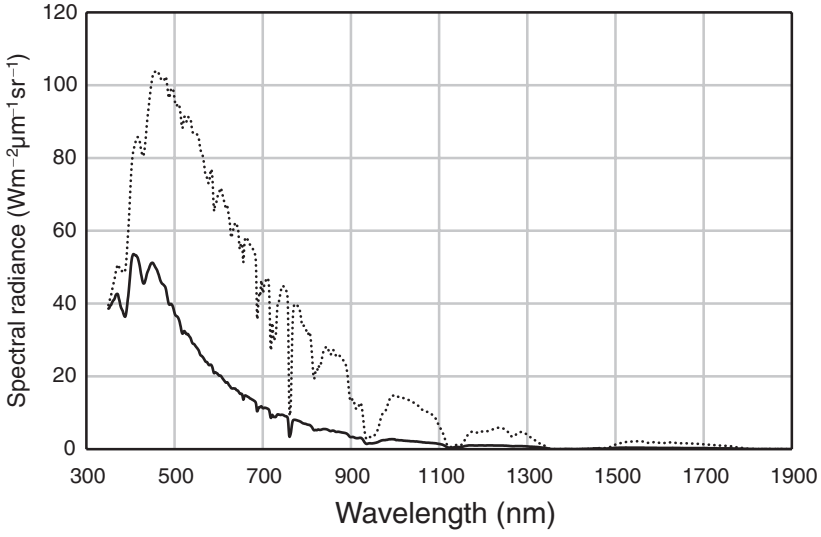


Figure 3.7 Measured spectral sky radiance. Solid line – at the zenith; dotted line – near the horizon (measured by Sarah Lane, GTRI, near noon on November 20, 2008, using an ASD Spectrophotometer).

has been smoothed to a resolution varying from 4.5 to 18 nm, depending on the wavelength region. The peak of the sky radiance is clearly shifted toward the ultraviolet, and its falloff with wavelength is dramatically faster than the solar curve (absorptions due to water vapor and trace gases are also evident in the measured data).

For modeling lidar signals, the important parameter is the molecular backscattering coefficient $\beta_m(\pi)$ (which is usually written as simply β_m in the lidar equation). McCartney [1] gave the value of $\beta_{mol}(\pi)$ at 0.55 μm wavelength as about 1.4×10^{-6} $m^{-1}sr^{-1}$ at Standard Temperature and Pressure (STP), and Collis and Russell [5] gave

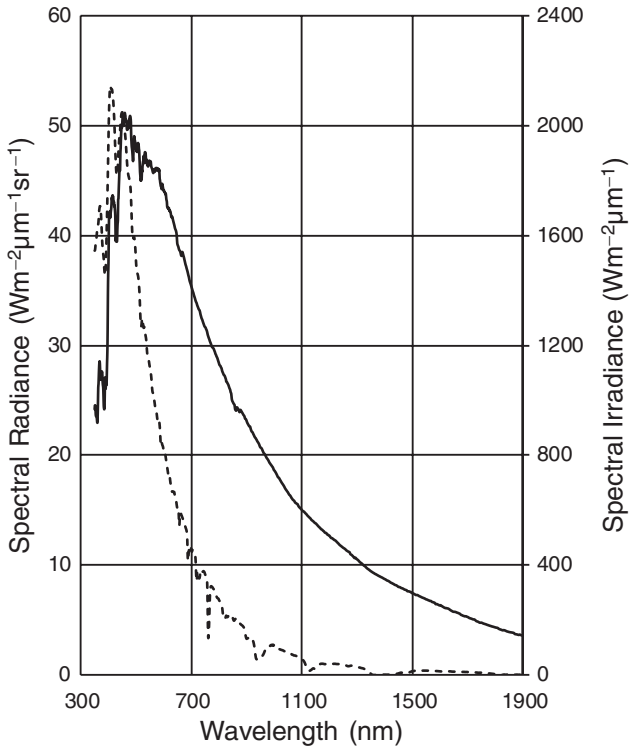


Figure 3.8 Sky radiance and solar irradiance. Zenith spectral sky radiance is shown by the dashed curve and the scale on the left, while the zero air mass spectral solar irradiance is shown by the solid curve and the scale on the right.

the value as 1.47×10^{-6} $m^{-1}sr^{-1}$, when adjusted to STP. However, all these authors use a temperature of 273.15 K in their definitions of STP rather than the International Standard Metric Conditions used here: T_s = 288.15 K and P_s = 101.325 kPa, where the subscript s refers to standard conditions. Collis and Russell's numerical value of β_m for lidar backscatter with this definition of STP is

$$\beta_m = 1.39 \times \left[0.55/\lambda\right]^4 \times 10^{-6} \ (m^{-1}sr^{-1}). \tag{3.11}$$

The lidar backscatter coefficient of the molecular atmosphere is proportional to air density, which decreases close to exponentially with altitude, by a factor of about one million from the surface to 100 km, as indicated in Figure 1.3. The molecular signal can be calculated from an atmospheric model or from measured profiles of temperature and pressure, so it may appear uninteresting as a lidar measurement. However, it has at least two important uses: (1) the molecular atmosphere is implicitly used as a calibration target in some types of lidars (see Chapter 11); and (2) measured lidar backscatter profiles in the aerosol-free upper stratosphere and mesosphere can be inverted to find temperature profiles by invoking hydrostatic equilibrium and the

ideal gas law in a technique known as *Rayleigh lidar*. When the molecular atmosphere is being used for lidar calibration, a density profile specific to the time and place of the measurement is required because the density varies with weather patterns, latitude, and seasons. The variability increases with altitude and reaches its maximum in the 60–80 km region, where the extremes (occurring 1% of the time) are as high as ±80% of the standard value [6]. Simple models are available for extrapolating density profiles from the surface temperature and pressure, but a more accurate approach for ground lidar stations is to use upper air data from the nearest radiosonde launching site. The backscatter coefficient at any altitude $\beta_{\text{mol}}(h)$ can then be calculated from the value at STP using the upper air temperature and pressure measurements with the relation

$$\beta_{\text{m}}(h) = \beta_s \frac{P(h)T_s}{P_s T(h)}, \tag{3.12}$$

which is based on the ideal gas law. This law is not strictly accurate at sea level, but its accuracy rapidly increases with altitude as the air becomes more rarified. Airborne and spaceborne lidars require density profiles over wide areas, so they typically rely on a global model such as the GMAO model provided by NASA's Global Modeling and Assimilation Office [7].

Molecular scattering in the atmosphere causes extinction of a laser beam because power is scattered out of the beam. The lidar extinction coefficient σ_m can be found from Eq. (3.11) and the molecular *lidar ratio* S_m, which is the extinction coefficient divided by the backscatter coefficient. To a close approximation, the ratio is given by

$$S_m = 8\pi / 3(\text{sr}), \tag{3.13}$$

which is numerically equal to 8.38 sr [5]. For example, at 532 nm wavelength and STP, $\beta_{\text{m}} = 1.59 \times 10^{-6}\ \text{m}^{-1}\ \text{sr}^{-1}$ and $\sigma_{\text{m}} = 1.33 \times 10^{-5}\ \text{m}^{-1}$. The molecular extinction by scattering of incoming solar radiation is characterized by the OD of the atmosphere, which is the altitude integral of σ_m from the surface to space. The atmospheric OD as a function of wavelength has been modeled extensively over the years. Bucholtz [8] provided tabulated values of OD for six different atmospheric models, for wavelengths from 0.2 to 4.0 μm. Summaries of those results are shown in Table 3.3, interpolated to several common laser wavelengths. For lidar, the extinction is doubled because of the two-way path. For example, a ground-based high-altitude lidar operating at 355 nm would be hampered by a two-way transmittance of about $\exp[-2(0.596)] = 0.30$. This is one reason that high-altitude lidars, such as Rayleigh lidars, usually operate at 532 nm or longer wavelengths.

The foregoing treatment of Rayleigh scattering is subject to small corrections in the molecular atmosphere, largely because O_2 and N_2 molecules are not exactly spheres and consequently they are not perfectly represented by the simple model described above. Bucholtz provides a concise summary of several of the corrections [8]. For a lidar-specific discussion of Rayleigh scattering and its corrections, including a correction to the value of S_m given in Eq. (3.13), see Adam [9].

Table 3.3 Atmospheric OD at laser wavelengths

Wavelength (nm)	OD range
266	2.060–2.077
355	0.5933–0.5983
532	0.1110–0.1119
1064	0.006869–0.006918

3.3 Molecular Energy Effects

Many of the important features of atmospheric optics for lidar are related to the ways in which energy is stored in the molecules of the atmospheric gases. Molecules have kinetic energy of motion; molecules with two or more atoms store both rotational and vibrational energy; all molecules can store energy in excited states of their electrons; and molecules store energy in molecular bonds. All these energy storage mechanisms cause optical effects that are exploited by the various kinds of lidars that are described in this section.

3.3.1 Kinetic Energy of Motion

In equilibrium, the temperature T of a system is proportional to the energy in the system. According to the classical equipartition of energy theorem, the molecules in a gas have an average energy of $kT/2$ per degree of freedom, where k is Boltzmann's constant, 1.38×10^{-23} J/K, and T is the gas temperature in Kelvins. Molecular masses for the three major gases are quite small, in the range $5–7 \times 10^{-26}$ kg, which means that their average speeds are appreciable, equal to hundreds of meters per second at atmospheric temperatures. For motion, there are three degrees of freedom because of the three orthogonal directions in space, so each particle has an average kinetic energy $3kT/2$. Setting this thermal energy equal to the kinetic energy we have the relation $m\mathrm{v}^2/2 = 3kT/2$, where m is the particle mass and v is its speed. Solving this relation for v and using the mass of nitrogen yields a speed of about 500 m/s when $T = 300\ K$. This speed is high enough to cause a measurable Doppler shift in laser light backscattered by air, with important consequences for lidar. The discussion above refers to an average molecular speed, but in fact, the velocities of the molecules in a gas are randomized by frequent collisions (each molecule experiences about 10^9 collisions per second at surface temperature and pressure), and the speeds follow a statistical distribution known as the *Maxwell–Boltzmann distribution*. This distribution, which gives the probability per unit speed of finding the particle with a speed near v, is given by

$$f(\mathrm{v}) = \left(\frac{m}{2\pi kT}\right)^{3/2} 4\pi \mathrm{v}^2 \exp\left(\frac{-m\mathrm{v}^2}{2kT}\right). \tag{3.14}$$

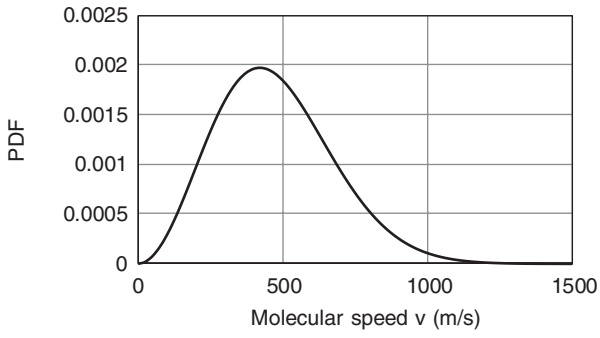


Figure 3.9 The Maxwell–Boltzmann speed distribution. The probability distribution function (per unit speed) is shown for nitrogen gas at 300 K.

The distribution is shown graphically in Figure 3.9, for diatomic molecules of ^{14}N at 300 K. The function is not symmetrical, so the peak probability is not equal to the mean speed, which is 517 m/s in this example.

The lidar Doppler shift is given by $\Delta\lambda/\lambda = 2v/c$, that is, the fractional change in backscattered wavelength is equal to twice the ratio of the molecule's speed (away from the lidar along the lidar line of sight) to the speed of light. Molecules moving toward the laser cause shifts to shorter wavelengths; those moving away cause shifts to longer wavelengths; and those moving transversely to the beam cause no Doppler shifts at all. The net result of these phenomena coupled with the speed distribution shown above yields a lidar backscatter spectrum similar to the curve illustrated in Figure 3.10. The spectrum is the sum of a broad molecular curve that is Gaussian to a close approximation, spanning a few gigahertz, and a narrow peak due to aerosols in the center. Up to this point, wavelengths have been expressed in nanometers or micrometers, but the Doppler shifts shown in Figure 3.10 are conventionally expressed in frequency units (GHz). The reason for this change is that the shifts are quite small. The conversion can be worked out from the equation $c = \nu\lambda$, which pertains to all waves. Rearranging as $\nu = c/\lambda$ and using differentials yields $\Delta\nu = -c\Delta\lambda / \lambda^2$, which can be solved for $\Delta\lambda$ in terms of $\Delta\nu$, λ, and c as $\Delta\lambda = -\Delta\nu\lambda^2/c$.

Example. The Doppler shifts due to molecular motion shown in Figure 3.10 span about ±2 GHz, which corresponds to a wavelength spread of $\pm(2.0 \times 10^9) \times (0.532 \times 10^{-6})^2 / (3 \times 10^8) = 1.9 \times 10^{-12}$ m at 532 nm wavelength, or 1.9 picometers (pm). The aerosol peak in Figure 3.10 is narrow because aerosol masses are huge compared to molecular masses. The density of water is 1×10^3 kg/m^3, so the mass of a 1-micron radius water drop is $(4\pi/3) \times (1 \times 10^{-6})^3 = 4.2 \times 10^{-18}$ kg, which is about 8 orders of magnitude larger than the mass of an air molecule.

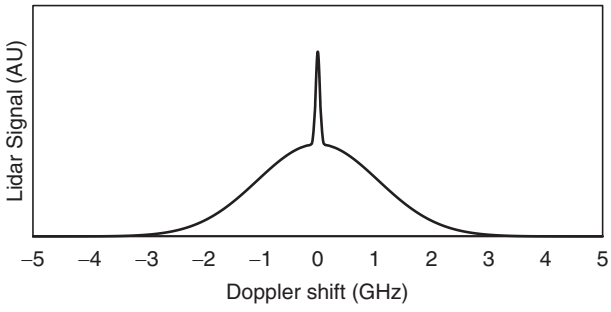


Figure 3.10 The Doppler-broadened lidar signal.

Aerosols experience Brownian motion but only on the order of ~1 mm/s, so their motion in the atmosphere is therefore simply due to wind; they are advected by air motion. For this reason, aerosol motion is used in eddy correlation lidars and coherent wind-sounding lidars as a tracer for wind vectors. The width of the aerosol peak in Figure 3.10 is largely due to the spectral width of the transmitted laser pulse.

Doppler broadening has important consequences for lidar. First, it enables the separation of aerosol and molecular signals, by using lidar receivers with very high spectral resolution. The spectral separation of aerosol and molecular signals enables the measurement of aerosol parameters such as the backscatter coefficient in absolute units, which a simple elastic backscatter lidar cannot do. This type of instrument, known as the high spectral resolution lidar (HSRL), is described in more detail in Chapter 11. Wind is a bulk motion of air, and wind toward or away from a lidar will simply shift the spectrum illustrated in Figure 3.10 to the left or right. Because atmospheric wind speeds are usually much smaller than molecular speeds, the shift is smaller than the ±2 GHz spread, but it can nevertheless be measured optically. For example, the space-based ALADIN lidar (atmospheric laser Doppler instrument) exploited the Rayleigh signal using a dual-filter Fabry–Perot interferometer with a 5.5 GHz spacing [10]. The two filter passbands were centered on the spectrum, so that the two receiver channels received equal signals in the absence of wind. When the Doppler-broadened spectrum shifted due to wind, the signal in one channel decreased while the other channel's signal increased. This measurement method, known as the *double-edge technique* is illustrated in Figure 3.11. It is used at high altitudes where there is negligible aerosol. ALADIN also had a Mie receiver channel for lower altitudes, which exploited shifts in the aerosol peak position measured with a Fizeau interferometer. Its frequency resolution was 100 MHz, which allowed a wind speed resolution of 18 m/s (ALADIN operated at 355 nm). Ground-based Rayleigh lidars also measure high-altitude horizontal winds this way, by operating off zenith and exploiting shifts of the Doppler-broadened spectrum [11].

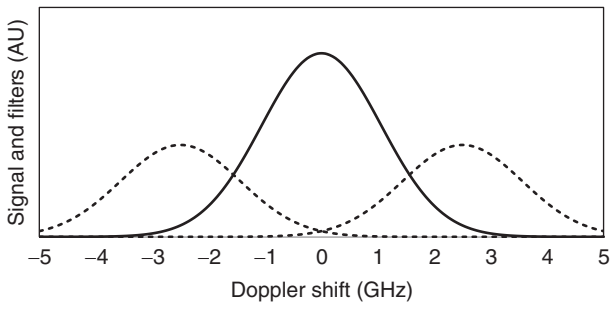


Figure 3.11 The double-edge technique. The filter passbands (dashed lines) are placed symmetrically about zero Doppler shift. If the Rayleigh spectrum (solid line) shifts to the right or left due to wind, the signal through one filter will increase and through the other it will decrease.

3.3.2 Rotational and Vibrational Energy

In addition to their kinetic energy of motion, molecules in the gas phase store energy by rotating and vibrating. These modes are responsible for the infrared absorption spectra of atmospheric gases as well as for Raman spectra, and they are very different from the energy of motion because they are quantized, meaning that only certain discrete energy levels are available. Although the energy levels are a quantum-mechanical phenomenon, a simple classical picture yields some valuable insight. In a diatomic molecule, rotational energy is stored in rotations about the center of mass of the two atoms, and vibrational energy is stored in oscillatory motion of the atoms toward and away from each other, along with stretching and compression of the chemical bond between them. The classical model of such a molecule is a spring-mass oscillator tumbling end over end, with the two atoms orbiting the system's center of mass with two rotational degrees of freedom, as illustrated in Figure 3.12. The rotational energy is given by $E_{\text{rot}} = I\omega^2 / 2$, where I is the moment of inertia about an axis through the center of mass and perpendicular to the line joining the two atoms and ω is the angular rotation frequency. The vibrational energy is $E_{\text{vib}} = kA^2 / 2$, where k is the spring constant and A is the amplitude of the oscillation. In classical mechanics, ω and A are continuous variables that can take on any values, so the rotational and vibrational energies are also continuous. The moment of inertia of the system is given by $I = d(m_A m_B) / (m_A + m_B)$ where d is the mean separation, and the oscillation is modeled as simple harmonic motion because the restoring force is assumed to obey Hooke's law, $F = -kx$, where x is the distance from the equilibrium separation d.

One might anticipate some problems with this model: Increasing rotational speed ω will cause a centrifugal force that will increase the separation d, and there is no *a priori* reason to assume that stretching of a molecular bond will obey Hooke's law, hence the motion is not necessarily harmonic. It turns out that these problems do occur, and they are dealt with in the quantum mechanical model. The rotational and

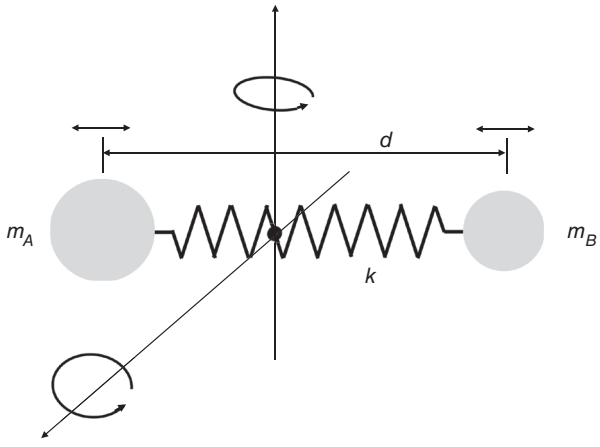


Figure 3.12 Classical model of a diatomic molecule. The atoms are represented by the masses m_A and m_B, which are separated by a mean distance d and joined together by a spring with a Hooke's law constant k. The two atoms rotate about their common center of mass (black dot) with two rotational degrees of freedom and oscillate toward and away from each other.

vibrational energies in gas phase molecules are both quantized. Angular momentum is quantized as integral multiples of $\hbar$ (Planck's constant h divided by 2π), which leads to quantized rotational energies specified by a *term value* $F_v(J)$, as

$$F_v(J) = B_v J(J+1) + DJ^2(J+1)^2, \tag{3.15}$$

where J is the rotational quantum number (an integer value starting at zero), $B_v = h/8\pi^2 cI_v$, and D is a centrifugal distortion constant that corrects for centrifugal stretching. The moment of inertia is the same as the classical term, except that the interatomic distance is a function of the vibrational quantum number v: $I_v = d_v(m_A m_B)/(m_A + m_B)$. Term values are in the spectroscopic units of waves per centimeter (cm^{-1}) which are universally called *wavenumbers,* with the symbol $\bar{\nu}$. Wavenumbers are proportional to energy – they are energies in Joules divided by hc, so energy levels are obtained by multiplying the term values by hc, with c in units of cm/s (not m/s). To obtain wavelengths in μm, divide wavenumbers into 10,000. For example, 2,000 cm^{-1} corresponds to 10,000/2,000 = 5 μm. Vibrational energies are quantized with term values given by

$$G(v) = \omega_e(v + 1/2) - \omega_e\chi_e(v + 1/2)^2, \tag{3.16}$$

where v is the vibrational quantum number, ω_e is the harmonic wavenumber, and χ_e is the anharmonicity constant. The second term on the right is a correction for vibrational motion that is not simple harmonic. As before, the units are wavenumbers, and energies are obtained by multiplying the term values by hc. Note that the vibrational energy is not zero in the *ground state* $v = 0$. This phenomenon is known as *zero-point motion*, which contrasts with the classical view that all molecular motion ceases at a temperature of absolute zero. The vibrational energy quantum steps are generally

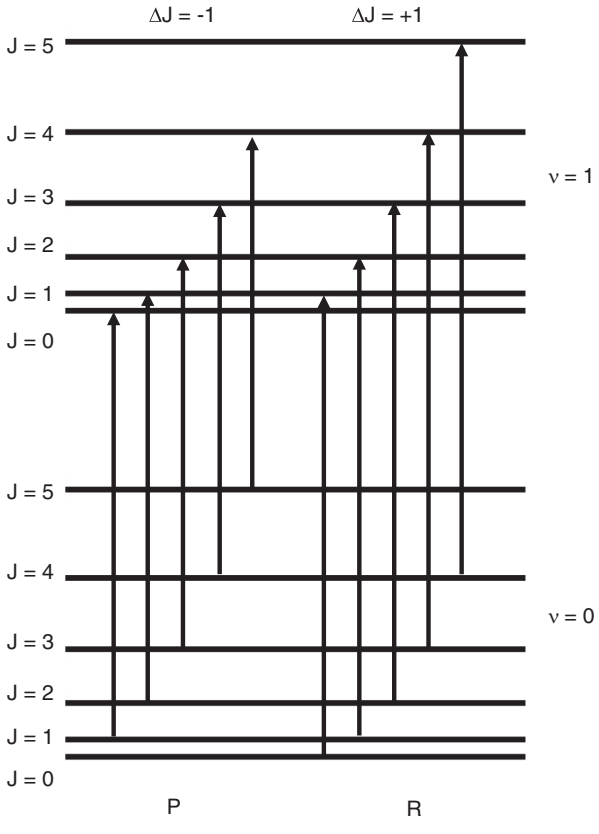


Figure 3.13 Rotation and vibration energy levels and transitions. Energy levels in a diatomic molecule are shown for the first two vibrational states and the first five rotational states. Allowed transitions between levels due to photon absorptions are shown by arrows.

larger than rotational steps, which gives rise to the type of energy level diagram shown in Figure 3.13, which is illustrated for two vibrational levels (the ground state and the first *excited state*) and five rotational levels.

When a molecule absorbs energy, for example by absorbing a photon, a *transition* occurs in which the molecule goes from a lower energy state to a higher state. Radiative transitions are governed by *selection rules*: For diatomics, radiative transitions are only allowed for $\Delta v = \pm 1, \pm 2, \pm 3$ etc., and $\Delta J = \pm 1$. Transitions obeying these rules are illustrated in Figure 3.13. For atmospheric temperatures, these molecular absorptions are dominated by transitions from $v = 0$ to $v = 1$. The radiative transitions result in *spectral lines*, which occur as *bands* of lines known as *branches*, due to the multiple rotational levels and the selection rules. The bands are labeled according to ΔJ as O, P, Q, R, S for $\Delta J = -2, -1, 0, +1, +2$. For a diatomic, there is no Q-branch because $\Delta J = 0$ transitions are not allowed by the selection rules. The change in the molecule's rotational energy can be either negative or positive, giving rise to two

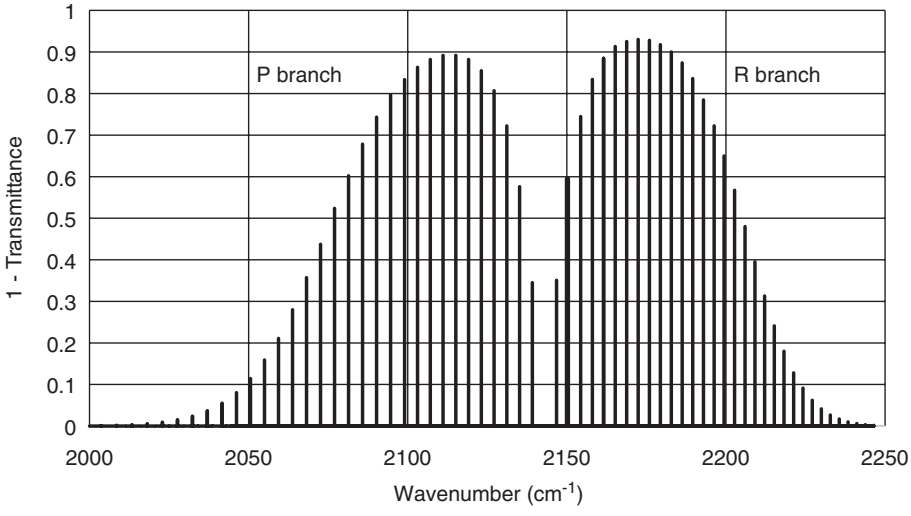


Figure 3.14 The CO absorption spectrum.

branches of the spectrum. The line spacing is $2B_v$. As an example, the calculated spectrum of CO near 5 μm wavelength is illustrated in Figure 3.14, which shows absorptance (1 – transmittance) through a gas cell 0.5 cm long, filled with $^{12}C^{16}O$ at 100 mb pressure at 296 K, with a volume mixing ratio of 0.1. The spectral branches shown in Figure 3.14 both have a minimum in the center, then they rise to a maximum away from the center, and then asymptotically approach zero. The quantum number J is increasing to the left in the P-branch, and to the right in the R-branch. The band shapes reflect the relative probabilities of the various transitions, which are proportional to the *populations* of the energy states, that is, the number of molecules that are in each state in thermal equilibrium. The probability of a transition increases with the population of the initial state. The populations are given by the *Boltzmann distribution*, which describes an exponential decrease in population with energy:

$$\frac{N(J)}{N(J=0)} = \exp\left(\frac{-hcF(J)}{kT}\right), \tag{3.17}$$

where the N values are the populations. The other factor in the transition probability is the *degeneracy* of the state. This factor is required because the quantum number J does not completely specify the state. An additional quantum number m is needed, with integer values from $+J$ to $-J$, which adds an additional factor of $2J+1$. Note that the degeneracy factor increases with J, whereas $N(J)$ decreases with J. The product of the two factors therefore goes thru a maximum, which explains the spectral shape of the branches in Figure 3.14. The selection rules also allow transitions from higher vibrational states, for example from $v=1$ to $v=2$, but the Boltzmann distribution also explains why the resulting bands of absorption lines are very weak: Only a very tiny fraction of the molecules is in the $v=1$ state at atmospheric temperatures. Spectroscopists refer to such spectra as *hot bands*, because they are only

observed in heated gas cells. The consequence for lidar is that the infrared (IR) spectral properties of the atmosphere are mainly determined by $v = 0$ to $v = 1$ transitions with many J values.

To absorb optical energy in a transition such as those shown in Figure 3.13, a molecule must be able to couple to the electromagnetic field, which requires that it have a dipole moment. CO has an electric dipole moment, hence the spectrum shown in Figure 3.14 can be observed in the atmosphere. However, the two main atmospheric gases, N_2 and O_2, do not have dipole moments, so they do not have infrared (IR) absorption spectra. Diatomic molecules are the simplest absorbers, with only one vibrational mode and two equal principal moments of inertia. Other molecules are more complicated, with both stretching and bending vibrational modes and up to three different principal axes of rotation and associated moments of inertia, so their spectra are correspondingly more complex than those of the diatomics. The H_2O molecule has all these complications plus a very large dipole moment, so its rotation-vibration spectrum is extensive, with strongly absorbing bands throughout the infrared spectral region. CO_2 does not have a dipole moment when undergoing symmetric stretching, but asymmetric stretching creates a dipole moment and so does bending, so again the absorption spectrum is extensive in the IR region. These two gases account for the general shape of the 1–15 μm atmospheric absorption spectrum. The other main gases that contribute to it are CO, CH_4, HDO, N_2O, and O_3. Procedures for modeling such spectra are extensions of the models presented here for diatomics, and they are covered in many standard books on molecular spectroscopy. Low-resolution zenith spectra of atmospheric gases from 1 to 15 μm are illustrated in Figure 3.15 for the gases individually and for the typical atmospheric mix.

The spectra shown in Figure 3.15 are plotted at low spectral resolution. There are hundreds of thousands of molecular absorption lines in the 1–15 μm region, and laser lines are generally very narrow, so the atmospheric spectrum must be modeled at an appropriately high resolution for each lidar application. Fortunately, high-fidelity models are readily available. The most common database is HITRAN, which is an acronym from the name high-resolution transmission molecular absorption database. HITRAN is a compilation of spectroscopic parameters that a variety of computer codes use to predict and simulate the transmission and emission of light in the atmosphere. The database is a long-running project started by the Air Force Cambridge Research Laboratories (AFCRL) in the late 1960s in response to the need for detailed knowledge of the infrared properties of the atmosphere for military sensing systems, and it can be accessed online [12]. The Ontar Corporation markets and maintains modeling software for personal computers called HITRAN-PC [13]. The spaceborne remote sensing community tends to use the Line-By-Line Radiative Transfer Model (LBLRTM), which is said to be an accurate, efficient, and highly flexible model for calculating spectral transmittance and radiance. LBLRTM is available from Atmospheric and Experimental Research (AER) in Lexington, Massachusetts [14]. The spectrum plotted in Figure 3.14 was generated with SpectralCalc, which is a commercial web-based spectral calculator that is especially easy to use [15]. The models for molecular spectra are sufficiently refined that the main

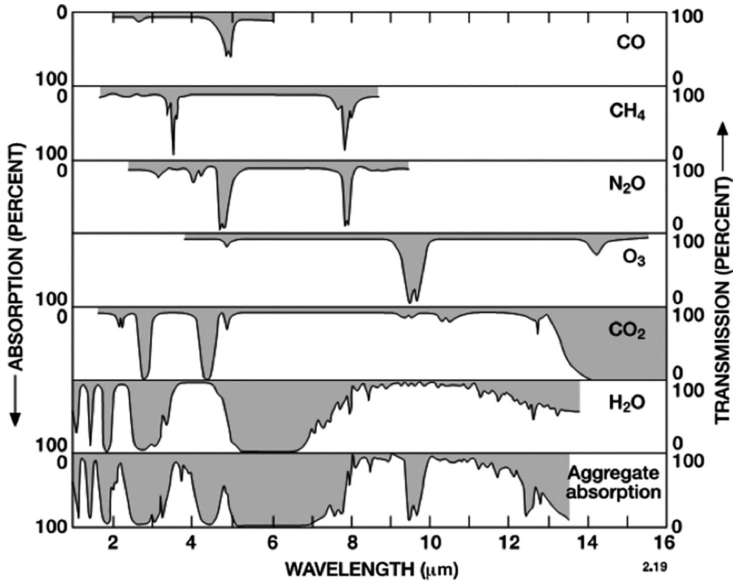


Figure 3.15 Low-resolution spectra of atmospheric gases. Transmission and absorption are shown in the zenith direction. From *Modeling of Atmospheric Chemistry*, G. P. Brasseur and D. J. Jacob, used by permission.

source of uncertainty in atmospheric spectra is in the inputs to the models (such as gas concentration, temperature, and pressure).

The rotation-vibration absorption spectra of atmospheric gases have two implications for lidar: First, they determine the *windows*, meaning semitransparent wavelength regions where optical remote sensing techniques can be used, and second, they enable the lidar technique known as differential absorption lidar (DIAL). The IR spectrum is commonly subdivided into five regions according to the scheme shown in Table 3.4, and the windows are named according to the subdivisions they are in. For example, the semitransparent region shown from 8 to 12 μm in Figure 3.15 is called the LWIR window. This naming convention is not universal, and it is not an official standard, but it is commonly used. Definitions for the ultraviolet and visible regions used in this book have been added to the IR nomenclature in Table 3.4 for convenience, because these regions are so often used in lidar. With those definitions, there are seven spectral regions.

The DIAL technique uses two or more wavelengths to exploit the absorption spectrum of a gas to measure the range profile of its concentration. The basic technique is illustrated in Figure 3.16. The lidar transmits two wavelengths, one on an absorption line and the other off the line. The two lidar signals have different shapes because the "on" wavelength is attenuated by the gas of interest, and so its signal decreases faster with range than the "off" wavelength signal. A simple algorithm, derived in Chapter 11, is used to find the gas concentration profile. DIAL is used for profiling water vapor and GHGs including CH_4 and CO_2, and industrial emissions that have suitable

Table 3.4 Subdivisions of the optical spectrum

Name	Abbreviation	Wavelength range
Ultraviolet	UV	0.25–0.4 μm
Visible	VIS	0.4–0.75 μm
Near infrared	NIR	0.75–1.4 μm
Short-wave infrared	SWIR	1.4–3μm
Mid-wave infrared	MWIR	3–8 μm
Long-wave infrared	LWIR	8–15 μm
Far infrared	FIR	15–1000 μm

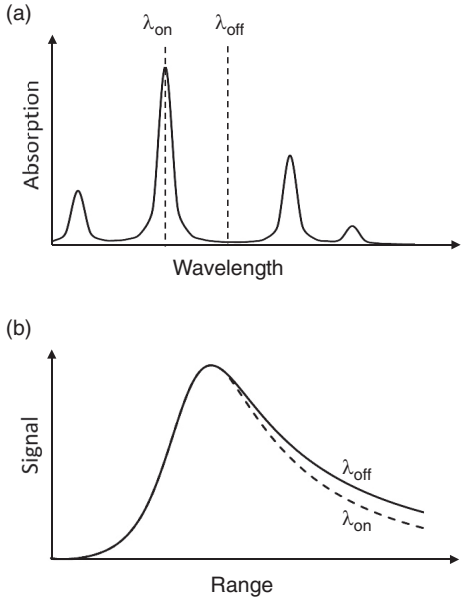


Figure 3.16 The DIAL Technique. (a) Two wavelengths are chosen to be on and off an absorption line. (b) The lidar signal for the "on" wavelength decreases faster with range than the "off" wavelength.

absorption lines in window regions [16]. Finding optimal wavelength pairs for DIAL is complicated because there are usually spectra of other interfering gases in any window region; because the absorption strength must not be too strong nor too weak; and because the absorption line should not be temperature dependent. In addition, the two lines should be close together, so that the atmospheric backscatter coefficient is the same for both wavelengths. Ozone DIAL does not exploit rotation-vibration spectra; it is a special case described in Section 3.3.5.

3.3.3 Raman Scattering

In addition to absorption, laser light may interact with a molecule through *Raman scattering*. This is *inelastic scattering*, in which the molecule undergoes a transition from one rotation-vibrational energy level to another and a photon emerges with a wavelength shorter or longer than the incident photon's wavelength, thereby conserving energy. Raman scattering is illustrated in Figure 3.17. Absorption of the incident photon elevates the molecule to a *virtual* energy level from which it decays very rapidly by emitting a photon, ending up in a rotation-vibrational level other than the initial one. In this example, $\Delta J = +2$ and $\Delta v = +1$. For diatomic molecules such as N_2 and O_2, the selection rules for Raman transitions are $\Delta v = 0, \pm 1$ and $\Delta J = 0, \pm 2$, which leads to O, Q, and S spectral branches. If ΔJ is positive the transition is said to be *Stokes*, and if it is negative, it is *Anti-Stokes*.

The Raman wavelength shifts are characteristic of the scattering molecule. This means that the quantized energy levels of N_2 and O_2 can be exploited by lidars even though those molecules have no dipole moments. Raman cross sections are orders of magnitude weaker than Rayleigh, but they still cause measurable lidar signals, and Raman lidar has become an invaluable tool in atmospheric research. Raman cross sections are also proportional to $1/\lambda^4$, so Raman lidars operate in the visible and UV spectral regions. Note that Rayleigh scattering corresponds to the process shown in Figure 3.17 when $\Delta v = 0$ and $\Delta J = 0$. One upshot of the Raman scattering phenomenon for lidar is that atmospheric molecular backscatter is not just a single spectral line at the laser wavelength. When $\Delta v = 0$, it is a strong central line due to Rayleigh scattering with bands of much weaker closely spaced rotational Raman lines. The width of the Raman spectrum is similar to that of commonly used lidar receiver filters, so some or all of the Raman lines contribute to the received signal in simple receivers. The Rayleigh line, known as the *Cabannes line*, is due to elastic scattering, while the Raman lines are not, but the lidar literature is inconsistent on this point and the surrounding rotational Raman lines are sometimes lumped in with the Cabannes line and referred to as Rayleigh scattering. The rotational Raman lines surrounding the Cabannes line have one other consequence for lidar: They have varying degrees of depolarization, so the measured lidar depolarization by clear air depends on the optical bandwidth of the receiver. This fact was not appreciated in the early days, so there are conflicting results in the lidar literature for depolarization by clear air. The values of the depolarization ratio in dry, clear air are small, ranging from 0.004 for the Cabannes line to 0.014 for all the rotational lines [17]. The Raman spectra of N_2, O_2, and H_2O are shown in Figure 3.18 for a lidar transmitting at 532 nm [18]. Relative intensities are shown for the Q branches ($\Delta J = 0$) and for the sums of the rotation bands. The Raman spectra shown in Figure 3.18 are utilized by water vapor Raman lidars, which transmit at 532 or 355 nm and receive the N_2 and H_2O signals at two longer wavelengths. Profiles of the water vapor mixing ratio are proportional to the ratio of the two Raman signals, as described in Chapter 11.

The N_2 Raman spectrum also enables aerosol extinction coefficient profiling. This technique arose from the realization that Raman scattering creates a new light source

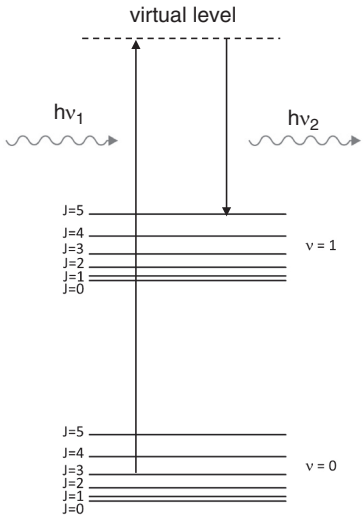


Figure 3.17 Raman scattering. The process is inelastic because the final state of the molecule is different from the initial, and the scattered photon has a frequency different from the incident photon.

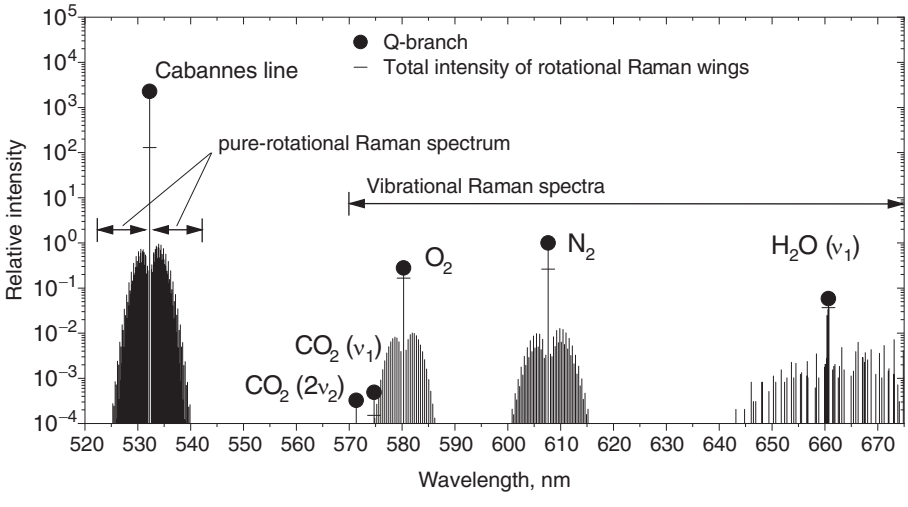


Figure 3.18 Atmospheric Raman spectra. Reprinted with permission from [18] © The Optical Society.

in the scattering volume, so for example 532 nm has two-way path through the aerosols, whereas 608 nm only propagates one way [19]. This fact can be exploited to find aerosol extinction coefficient profiles. The data analysis algorithm requires the molecular backscatter and extinction as inputs, so the molecular atmosphere effectively serves as a calibration target. Raman aerosol lidar is one of only two widely

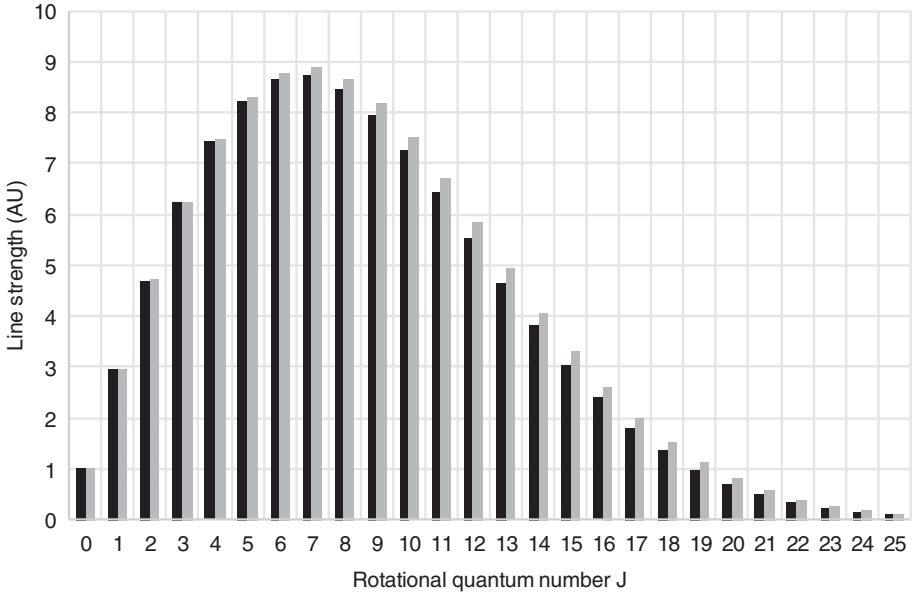


Figure 3.19 Temperature dependence of populations. The populations of CO rotational levels for $v = 0$ are shown for temperatures of 288.15 K (black) and 298.15 K (gray).

accepted techniques for quantitative aerosol extinction coefficient profiling (the other is HSRL), so it is widely deployed. Researchers at the Leibniz Institute for Tropospheric Research, where the Raman aerosol technique was developed, have developed a series of multi-wavelength Raman-elastic lidars over the years that they characterize by the numbers of different wavelengths at which they generate backscatter and extinction profiles, for example 3β, 2σ. This notation was later shortened to simply 3 + 2, and if a depolarization channel is added, it becomes 3 + 2 + 1.

The third lidar technique enabled by Raman scattering is called *rotational Raman*, and it exploits the rotational Raman spectrum to derive temperature profiles. This technique can be used in the presence of aerosols, so it is used in the troposphere, where Rayleigh lidar cannot operate. The populations of the rotational states are given by the Boltzmann distribution, Eq. (3.15), so they are temperature dependent. This fact means that the line strengths are temperature dependent, so measurements on pairs of rotational Raman lines can be analyzed to find temperature. To illustrate this principle, the populations of CO rotational levels for $v = 0$ are shown in Figure 3.19 for temperatures of 288.15 K and 298.15 K. The populations were calculated using Eq. (3.15), Eq. (3.17), and the degeneracy factor $2J + 1$, along with CO molecular constants. The rotational populations shown in Figure 3.19 give rise to the absorptance spectrum shown in Figure 3.14, so the ratio of the absorptions due to an appropriately chosen pair of lines will be a function of temperature. This is the basis of the technique employed with the nitrogen Raman rotational lines to sense atmospheric temperatures. This type of temperature sensing is known as a Boltzmann technique because

it is based on the temperature dependence of the Boltzmann distribution. Raman lidar techniques are described in more detail in Chapter 11.

3.3.4 Electronic Excitations

The energy levels of electrons in atoms and molecules are also quantized, and electronic excitations have higher energies than those associated with rotation-vibration. For this reason, the spectral lines from electronic transitions are in the UV-VIS region. Electronic transitions are used by *resonance fluorescence* lidars, which exploit the process illustrated in Figure 3.20, where $h\nu_1 = E_2 - E_1$, which is the condition of resonance. If the laser is precisely tuned to a transition, the scattering cross section is increased by orders of magnitude, because the photons are strongly absorbed and very quickly re-emitted. Resonance fluorescence lidars probe the mesosphere and lower thermosphere (the MLT region) at altitudes from 75 to 170 km, using the metal atoms sodium (Na), potassium (K), lithium (Li), iron (Fe), and calcium (Ca), which are deposited high in the atmosphere by meteor ablation. The number density of those atoms is typically smaller than the air density by a factor of 10^{-10}, but the resonance cross section is about 10^{14} times larger than the Rayleigh cross section, so the lidar signals are large enough for useful measurements.

The other scattering processes described in this chapter are essentially instantaneous, but resonance fluorescence is not. The term fluorescence means that the excited state has a nonzero lifetime. Fortunately, it is on the order of 10 ns, so it does not cause significant range errors in lidar. It can however cause *saturation*, which means that the lidar signal is no longer proportional to laser power because a significant fraction of the atoms has already been pumped up to the excited state. The atoms are not quickly de-excited through collisions (that process is called *quenching*) because the atmosphere is so rarified at high altitudes that the collision rate is small. The first resonance fluorescence lidar measurements on the sodium layer at about 80 km altitude were reported in 1969 [20]. Early researchers were able to measure sodium atom density and column abundance with what came to be called *broadband* lidars, meaning that the laser line was spectrally broader than the resonance lines. Broadband lidars were also capable of observing gravity waves and tidal waves. Over the years, several generations of custom-built lasers have improved resonance fluorescence lidar capabilities to the point where they now have excellent tuning accuracy, pulse-to-pulse stability, and high peak power, and these ground-based instruments now do precision atomic laser spectroscopic measurements, monitoring both temperatures and wind speeds in the MLT region in addition to abundances and wave phenomena.

Atomic energy levels are much more complicated than rotation-vibration levels because more bodies are involved (in the case of sodium, the nucleus plus eleven electrons) and the electrons have angular momentum associated with their motion plus intrinsic angular momentum known as spin. The nucleus can also have spin. The energy levels are specified with a set of quantum numbers, and selection rules are specified in terms of those numbers, as they are for rotation-vibration levels,

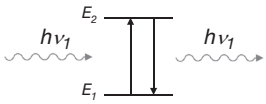


Figure 3.20 Resonance scattering. When the photon energy matches the energy level difference, the scattering cross section is greatly enhanced.

but with more complexity. The partial energy level diagram of sodium shown in Figure 3.21 results in a doublet (two closely spaced spectral lines) at 588.9950 and 589.5924 nm. The orange color of the lines is familiar to the public because these two lines are the dominant output of the sodium vapor lamps commonly used for outdoor lighting. Briefly, the *spectroscopic notation* in Figure 3.21 has the following meanings: the first numeral (3) is the principal quantum number n, which has integer values 1, 2, 3, …; the lower-case s and p mean 0 and 1, respectively, and they indicate the azimuthal quantum number, with integer values 0, 1, …, n–1. These first two quantum numbers date from the semiclassical model known as the Bohr atom and its extension by Sommerfeld. The capital P means that the total orbital angular momentum is 1 (in units of $\hbar$) and the capital S means 0. The raised prefix 2 is equal to 2S + 1, meaning that the total electron spin is ½; and the lowered suffixes ½ and ⅔ specify the total angular momentum, which is a vector sum of the spin and the orbital angular momentum. Transitions between the P and S states cause two spectral lines because the $^2P_{3/2}$ and $^2P_{1/2}$ states have different energies; this feature of the spectrum is known as *fine structure*. These two sodium lines were first observed as dark lines in the solar spectrum by Joseph von Fraunhofer, who labeled them with the letter D, and they are conventionally labeled D_1 and D_2. Sodium is hydrogenic, meaning that its electron configuration is a set of closed shells of electrons with paired spins (which sum to zero) plus one lone electron, so the sodium atom's total electronic spin is ½. If the nucleus has angular momentum (it is ⅔ in sodium), the energy levels are further split by interactions between the electron spin and the nuclear angular momentum. This additional splitting causes the D_2 line to be two groups of three transitions, all with slightly different wavelengths, which results in a Doppler-broadened spectrum spanning about 5 GHz and having two peaks known as D_{2a} and D_{2b}, as shown in Figure 3.22. The features in this spectrum are known as *hyperfine structure*. Classic references on atomic spectra are listed in Section 3.5.

Exploiting the sodium spectrum requires a tunable narrow-band laser with a linewidth of about 100 MHz (laser linewidths are discussed in Chapter 5). The shape of the whole spectrum can be determined by a series of many measurements at discrete wavelengths, but much faster techniques have been developed for measuring temperatures, for example the three-wavelength technique illustrated in Figure 3.22, which uses two wavelengths where the temperature effect is maximal and one wavelength that is not temperature sensitive. The first temperature measurements using the sodium D_2 line were reported around 1980 and the first routine temperature profile

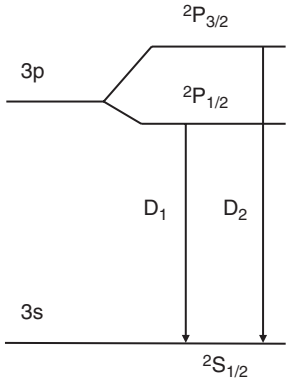


Figure 3.21 Partial energy level diagram for sodium atoms. The radiative transitions of the D lines are shown by arrows.

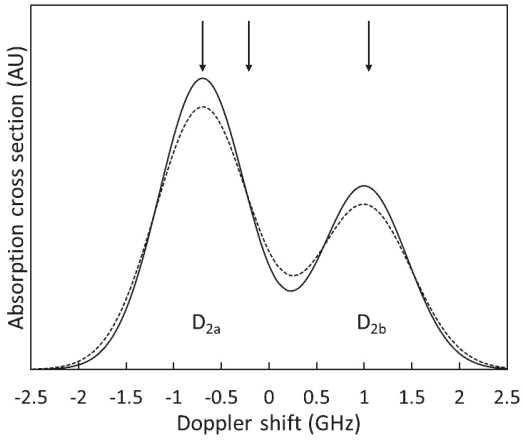


Figure 3.22 Hyperfine spectrum of sodium D_2 line. Solid line – 200 K; dashed line – 250 K. The arrows show three wavelengths used for temperature measurements. Adapted with permission from [21].

measurements started in the mid-1980s. By the early 1990s, uncertainty estimates of ±3 K were reported with 1-km vertical resolution and 5-minute time resolution. The width of the D_2 spectrum is also temperature dependent because it is Doppler broadened. The sodium D_2 lines are also used for wind measurements by measuring Doppler shifts in the whole spectrum, which is analogous to the Doppler shifts in the Rayleigh spectrum shown in Figure 3.10 that were exploited by the ALADIN wind lidar. Again, narrowband lasers are required for these measurements. The UV atomic iron spectrum includes two separated lines at 372 nm and 374 nm that are more convenient for Boltzmann temperature measurements in the sense that they can be done

with broadband lidars with linewidths of about 1 GHz. Chu and Papen [21] presented a comprehensive history of resonance fluorescence lidar up to 2003 that included instruments exploiting the electronic energy levels in Na, K, Li, Ca, and Fe. Doppler techniques for winds in the MLT region were routinely used by that time, along with both Doppler and Boltzmann techniques for temperatures. Resonance fluorescence lidar researchers continued improving their instrumentation over the ensuing years and they also developed transportable systems for use at any latitude, including polar regions. Most ground-based lidars operate at altitudes below about 100 km where the atmosphere is well mixed and neutral, but resonance fluorescence lidar measurements now extend to 170 km in the lower thermosphere, which is in a transition region between the atmosphere and space known as *geospace*. In the polar regions, the theory of such lidar measurements involves electrodynamics, neutral dynamics, chemistry, composition, and energetics, along with neutral-plasma coupling and the influences of polar electric fields, vertical winds, and aurora activity [22]. The applications of resonance fluorescence lidars in polar regions have opened a challenging new area for lidar.

3.3.5 Chemical Bonds

Absorption of UV light by O_3 in the Hartley band, which extends from 245 to 340 nm, is used extensively for ozone DIAL measurements, as shown in Figure 3.23. That absorption is different from the foregoing discussions because it does not result from either rotation-vibration transitions or electronic transitions. Rather, it is caused by photodissociation, meaning that each time an O_3 molecule absorbs a photon, it splits into O_2 and O, so the photon's energy is used to break a chemical bond. The result is a broad, continuous spectrum, unlike the line spectra resulting from quantized energy levels. Figure 3.23 also shows several specific wavelengths that are commonly used to profile tropospheric O_3 concentrations [23]. Most of those wavelengths are in the *solar blind* region, generally taken to be 240–280 nm, where there is negligible sky background because wavelengths in that range are absorbed by the stratospheric ozone layer. Stratospheric O_3 DIAL systems use wavelengths in the 308–353 nm range because stratospheric concentrations of O_3 are much higher than those in the troposphere, so lower cross sections are required to avoid completely absorbing the laser light. There are four overlapping O_3 absorption bands spanning the UV–NIR spectral region: Hartley, 245–340 nm; Huggins, 325–340 nm; Chappuis, 410–690 nm; and Wulf, 633–1000 nm. High-resolution spectra have been measured at several temperatures for the entire wavelength region [25] and they are available online [26]. Those absorption bands contribute to the overall atmospheric optical depth, in addition to the Rayleigh ODs listed in Table 3.3. For example, in GTRI stratospheric measurements with EARL, the OD due to O_3 was about 0.02 at 523.5 nm, which is comparable to the Rayleigh scattering OD for the atmosphere above 25 km [27]. Absorption by O_3 must be included in UV-VIS stratospheric lidar models or data analyses when accurate ODs are required.

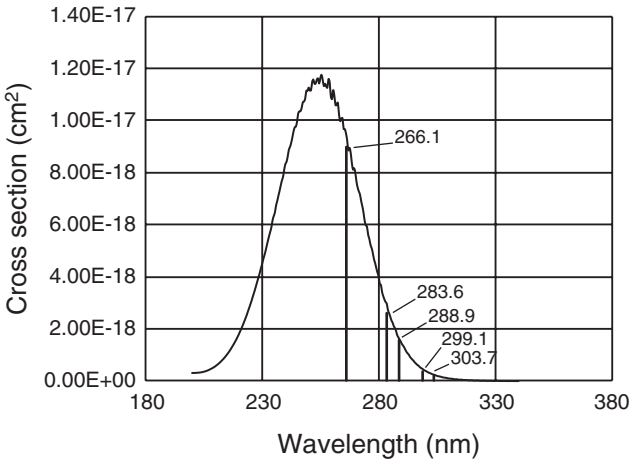


Figure 3.23 The UV absorption spectrum of ozone. The ozone spectrum at 298 K is shown by the curve, and the vertical lines show some wavelengths often used by ozone DIAL systems. The spectral data are from [24].

3.4 Summary

In this chapter, the phenomena in the molecular atmosphere that enable most of the major types of lidar are explained in terms of the underlying physics. The theory of Rayleigh scattering shows that the lidar backscatter coefficient is proportional to the molecular number density and inversely proportional to wavelength to the fourth power. Those facts, along with the constant mixing ratio of atmospheric gases up to at least 100 km, enable UV and VIS Rayleigh lidars, which measure density profiles that are inverted to find temperature profiles in the aerosol-free part of the atmosphere above 30 km. Rayleigh theory also shows that the molecular lidar ratio is approximately $8\pi/3$, and it explains the strong spectral dependence of the solar sky background. The kinetic theory of gases shows that atmospheric molecules have mean speeds on the order of hundreds of meters per second. These speeds, combined with the Doppler shift, cause significant broadening of the Rayleigh backscatter spectrum. Aerosols, whose masses are orders of magnitude larger than molecular masses, have negligible speeds of kinetic energy, but they are advected by wind and serve as tracers for it. The spectral width difference in the backscatter from molecules and aerosols enables the technique known as HSRL to separate the aerosol signal from the molecular signal, and Doppler shifts due to wind-driven motion enable both coherent and incoherent wind-sounding lidars. The coherent lidars depend on aerosol backscatter for their signals, so they tend to have short ranges. Fortunately, aerosol backscatter coefficients usually have a much smaller dependence on wavelength than Rayleigh, which enables coherent wind sounders to operate in several infrared window regions. The theory of quantized rotation-vibration energy levels in multiatomic molecules, along with the Boltzmann distribution of their populations, explains the absorption spectra of atmospheric gases. Those spectra

determine the window regions where lidars can operate, and they also enable the DIAL technique, which is used to measure profiles of water vapor and trace gases, including GHGs. Their spectra are at NIR and longer wavelengths, so these lidars also depend on aerosol backscatter for their signals. Raman scattering theory predicts the spectra of backscattered light as discrete spectral lines that are shifted in wavelength by amounts characteristic of the scattering molecules. This effect occurs even for molecules such as nitrogen that have no dipole moment, and it enables water vapor Raman lidar, which exploits water vapor and nitrogen Raman signals to find mixing ratio profiles of atmospheric water vapor. Tropospheric temperature-profiling lidars exploit the rotational Raman spectrum, in which ratios of individual line strengths have a temperature dependence due to the Boltzmann distribution. The nitrogen Raman backscatter, measured along with Rayleigh backscatter, is widely used to measure aerosol extinction. Raman theory also shows that Rayleigh molecular backscatter is a special case of Raman scattering, and it includes a band of rotational lines around the central Cabannes line. The electronic energy levels of metal atoms in the mesosphere and lower thermosphere are also Boltzmann-distributed, and lidar techniques have been developed for exploiting atomic spectra in these high-altitude regions. Those techniques include Doppler wind sounding as well as temperature measurements using both Boltzmann and Doppler broadening phenomena. Finally, the Hartley absorption band of ozone has a very broad and smooth UV spectrum because it is not due to transitions between discrete energy levels but rather to photodissociation of ozone molecules. That band enables ozone lidar and it also accounts for the solar-blind wavelength region, in which sunlight cannot penetrate the atmosphere.

The wavelength regions where these various types of lidar operate are shown in Figure 3.24. The logarithmic horizontal scale is in wavenumbers, which are proportional to energy. Atmospheric window boundaries are shown by gray vertical lines, and they are labeled according to the scheme in Table 3.4. Lidars must of course operate in windows. The small value of kT at 300 K (207 cm^{-1}) shown at the lower left, along with the Boltzmann distribution, explains why many rotational levels are populated but molecules are essentially all in the vibrational ground state at atmospheric temperatures. Two smooth curves are shown in the figure: the dashed curve, proportional to frequency (cm^{-1}), is meant to illustrate a typical spectral dependence of aerosol backscatter coefficients, and the solid curve, proportional to frequency to the fourth power, illustrates the spectral dependence of Rayleigh and Raman backscatter coefficients. The solid curve shows why Rayleigh, Raman, HSRL, and incoherent Doppler wind lidars all operate in the UV-VIS region: There is not enough backscatter at longer wavelengths to make such systems practical. Resonance fluorescence lidars also operate in this region because they exploit electronic energy level transitions, which have higher energies than the vibration-rotation transitions. Ozone DIAL exploits the Hartley absorption band, which is in the UV. Water vapor DIAL and GHG DIAL systems require isolated absorption lines in window regions, which occur in the NIR, SWIR, and MWIR. These systems rely on aerosol backscatter, and so do coherent wind lidars, which operate in the SWIR and LWIR. Figure 3.24 shows why many of the most capable and valuable types of lidar must

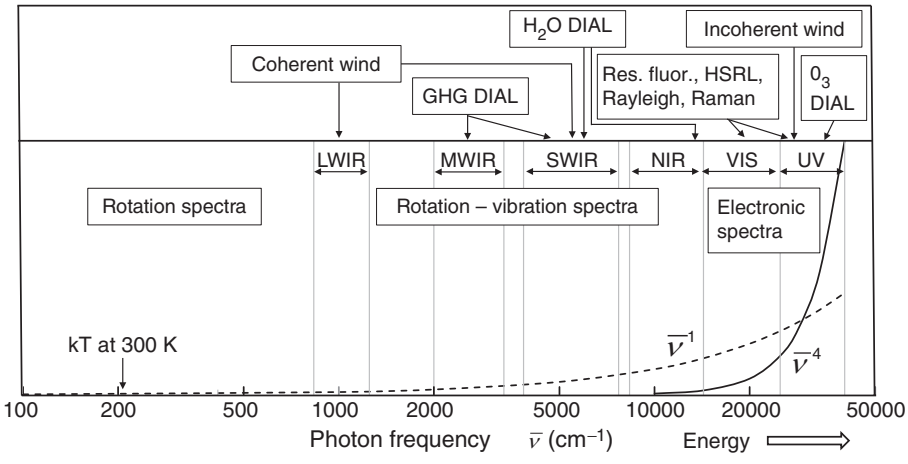


Figure 3.24 Lidar implications of the molecular atmosphere. Types of lidars used in all optical windows are shown vs. wavenumber. Lidar techniques are driven by the physics of optical scattering and radiative transitions between energy levels in molecules.

operate in the UV-VIS region, which has the worst problems with background light, eye hazards, and optical damage by laser photons. These are some of the factors that make lidar engineering challenging.

3.5 Further Reading

Much of the material in this chapter and Chapter 4 was adapted from McCartney's book [1], which is a very readable and comprehensive treatment of atmospheric optics. It is well worth reading for lidar researchers.

U. Wandinger, "Raman Lidar", in *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere*, C. Weitcamp, Ed. New York: Springer, 2005, pp. 241–271.

This chapter includes a very concise treatment of Raman spectroscopy with sufficient detail to calculate spectra such as the Raman spectrum of N_2 shown in Figure 3.18.

G. Herzberg, *Molecular Spectra and Molecular Structure: I. Spectra of Diatomic Molecules*. Malabar, FL: Krieger, 1989.

This is the first of a series of classic texts on molecular spectra by Gerhard Herzberg.

G. Herzberg, *Atomic Spectra and Atomic Structure*. Garden City, NY: Dover, 2010.

This book is another classic by Gerhard Herzberg. First published in 1937, it is often called the spectroscopist's Bible.

C.-Y. She and J. S. Friedman, *Atmospheric Lidar Fundamentals: Laser Light Scattering from Atoms and Linear Molecules*. New York: Cambridge University Press, 2022.

The physical bases of the various types of lidar described in this chapter are rigorously explained from first-principles physics in this book. It is the best resource for the interested reader.

3.6 Problems

3.6.1 The Weather Surveillance Radar 1988 Doppler (WSR-88D) detects radio waves scattered by raindrops. The radar has an operating wavelength of 3 cm and raindrops are on the order of 1 mm in diameter. Which of the three types of scattering does the radar detect?

3.6.2 Is extinction due to molecular scattering an important effect in tropospheric lidar? Consider a 10 km horizontal path with the backscatter coefficient at STP given by Eq. (3.11) and the extinction coefficient calculated using Eq. (3.13): what is the one-way OD at 355, 532, and 1064 nm? How do these OD values compare to the total atmospheric (zenith) ODs in Table 3.3?

3.6.3 Consider the molecules in nitrogen gas at temperature $T = 300\ K$. The molecules have an average kinetic energy of $3kT/2$, where k is Boltzmann's constant (1.38×10^{-23} J/K), and T is the gas temperature in kelvins. The kinetic energy is $m\mathrm{v}^2/2$, where m is the particle mass and v is its speed.

(a) What is the average speed v of the molecules? The mass of the most common isotope of nitrogen, ^{14}N, is 14 AMU, so assume that the N_2 molecular mass is 28 AMU. One AMU = 1.66×10^{-27} kg.

(b) What is the Doppler shift of backscattered 532-nm laser light (in nm) caused by this speed, for molecules moving directly away from the laser? The Doppler shift can be found from the relation $\Delta\lambda/\lambda = 2\mathrm{v}/c$. What is this Doppler shift expressed in GHz?

3.6.4 In atmospheric molecules, many rotational energy levels are populated but almost all the molecules are in the lowest vibrational state. The Boltzmann distribution shows that the ratio of the populations in two states is $\exp(-\Delta E/kT)$, where ΔE is the energy difference of the two states. Calculate that ratio for the two lowest rotational and vibrational levels in N_2, for which $B = 1.99\ \mathrm{cm}^{-1}$ (when the molecule is in its lowest vibrational state) and $\omega_e = 2331\ \mathrm{cm}^{-1}$. Note that the value of kT at 300 K is $207\ \mathrm{cm}^{-1}$, as shown in Figure 3.24, and ignore centrifugal stretching and anharmonicity.

References

[1] E. McCartney, *Optics of the Atmosphere*. New York: Wiley, 1976.

[2] M. Minnaert, *Light and Color in the Outdoors* (originally published as *Light and Color in the Open Air*). New York: Springer-Verlag, 1993.

[3] R. Greenler, *Rainbows, Halos, and Glories*. New York: Cambridge University Press, 1980.
[4] ASTM International, "Standard Solar Constant and Zero Air Mass Solar Spectral Irradiance Tables," ASTM E490-00a, 2014.
[5] R. T. H. Collis and P. B. Russell, "Lidar Measurement of Particles and Gases by Elastic Backscattering and Differential Absorption," in *Laser Monitoring of the Atmosphere*, E. D. Hinckley, Ed. New York: Springer, 1976, pp. 71–151.
[6] National Atmospheric and Oceanic Administration, National Aeronautics and Space Administration, and United States Air Force, The U.S. Standard Atmosphere, 1976. NOAA-S/T 76–1562 (1976). [Online]. Available: https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19770009539.pdf. [Accessed January 15, 2021].
[7] Global Modelling and Assimilation Office. [Online]. Available: https://gmao.gsfc.nasa.gov. [Accessed January 15, 2021].
[8] A. Bucholtz, "Rayleigh-Scattering Calculations for the Terrestrial Atmosphere," *Applied Optics*, vol. 34, pp. 2765–2773, 1995.
[9] M. Adam, "Notes on Rayleigh Scattering in Lidar Signals," *Applied Optics*, vol. 51, pp. 2135–2149, 2012.
[10] European Space Agency (ESA), "ADM-Aeolus Science Report," ESA SP-1311, 2008. [Online]. Available: https://earth.esa.int/documents/10174/1590943/AEOL002.pdf. [Accessed January 17, 2021].
[11] G. Baumgarten, "Doppler Rayleigh/Mie/Raman lidar for Wind and Temperature Measurements in the Middle Atmosphere up to 80 km," *Atmospheric Measurement Techniques*, vol. 3, pp. 1509–1518, 2010.
[12] HITRAN Online. [Online]. Available: https://hitran.org. [Accessed January 14, 2021].
[13] HITRAN PC. [Online]. Available: https://ontar.com/software. [Accessed January 14, 2021].
[14] AER's Radiative Transfer Working Group – Main Window. [Online]. Available: rtweb.aer.com. [Accessed January 14, 2021].
[15] SpectralCalc Hi-resolution spectral modeling. [Online]. Available: www.spectralcalc.com. [Accessed January 14, 2021].
[16] G. G. Gimmestad, "Differential-Absorption Lidar for Ozone and Industrial Emissions," in *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere*, C. Weitkamp, Ed. New York: Springer, 2005, pp. 187–212.
[17] A. Behrendt and T. Nakamura, "Calculation of the Calibration Constant of Polarization Lidar and Its Dependency on Atmospheric temperature," *Optics Express*, vol. 10, no. 16, pp. 805–817, 2005.
[18] A. Behrendt, T. Nakamura, M. Onishi, R. Baumgart, and T. Tsuda, " Combined Raman Lidar for the Measurement of Atmospheric Temperature, Water Vapor, Particle Extinction Coefficient, and Particle Backscatter Coefficient," *Applied Optics*, vol. 41, pp. 7657–7666, 2002.
[19] Ansmann, A. et al., "Combined Raman Elastic-Backscatter LIDAR for Vertical Profiling of Moisture, Aerosol Extinction, Backscatter, and LIDAR Ratio," *Applied Physics B*, vol. 55, pp. 18–28, 1990.
[20] M. R. Bowman, A. J. Gibson, and M. C. W. Sandford, "Atmospheric Sodium Measured by a Tuned Laser Radar," *Nature*, vol. 221, pp. 456–457, 1969.
[21] X. Chu and G. C. Papen, "Resonance Fluorescence Lidar for Measurements of the Middle and Upper Atmosphere," in *Laser Remote Sensing*, T. Fujii and T. Fukuchi, Eds. New York: Taylor and Francis, 2005, pp. 179–432.

[22] X. Chu et al., "First simultaneous lidar observations of thermosphere-ionosphere Fe and Na (TIFe and TINa) layers at McMurdo (77.84°S, 166.67°E), Antarctica with concurrent measurements of aurora activity, enhanced ionization layers, and converging electric field," *Geophysical Research Letters*, vol. 47, e2020GL090181 (2020). [Online serial]. Available: https://doi.org/10.1029/2020GL090181. [Accessed 25 February 2021].

[23] B. Calpini and V. Simeonov, "Trace Gas Species Detection in the Lower Atmosphere by Lidar: From Remote Sensing of Atmospheric Pollutants to Possible Air Pollution Abatement Strategies," in *Laser Remote Sensing*, T. Fujii and T. Fukuchi, Eds. New York: Taylor and Francis 2005, pp. 123–177.

[24] L. T. Molina and M. J. Molina, "Absolute Absorption Cross Sections of Ozone in the 185- to 350-nm Wavelength Range," *Journal of Geophysical Research Atmospheres*, vol. 91, pp. 14501–14508, 1986.

[25] V. Gorshelev, A. Serdyuchenko, M. Weber, et al., "High Spectral Resolution Ozone Absorption Cross-Sections – Part 1: Measurements, Data Analysis and Comparison with Previous Measurements around 293 K," *Atmospheric Measurement Techniques*, vol. 7, pp. 609–624, 2014.

[26] O3 Spectra, [Online]. Available: O3 Spectra (uni-bremen.de). [Accessed February 7, 2022].

[27] H. N. Forrister, D. W. Roberts, A. J. Mercer, and G. G. Gimmestad, "Infrared Lidar Measurements of Stratospheric Aerosols," *Applied Optics*, vol. 53, pp. D40–D48, 2014.

4 Particles in the Atmosphere

As illustrated in Figure 1.4, the atmosphere contains many different types of matter in the form of particles, from the surface to at least 25 km. The particles include *aerosols*, which are dispersed systems of small particles suspended in the air; liquid drops in mist, fog, and low clouds; dust and smoke; and ice crystals in cirrus clouds, noctilucent clouds (NLCs), and PSCs. The size ranges of both molecules and particles are illustrated in Figure 4.1, which spans eight orders of magnitude. *Hydrometeors* (falling drops of water or ice crystals) are not considered here because lidar is rarely used during precipitation. Atmospheric particles span a size range from much smaller than lidar wavelengths to much larger. Condensation nuclei, smoke particles, and dust tend to be small, whereas the droplets and crystals in clouds tend to be large, up to several mm in cirrus. Moreover, within any of these major types, the particles are never a single size; they have a *size distribution*, which may be changing with time. This is just one complication of particulate matter compared to molecules. In Chapter 3, accurate models were described for Rayleigh scattering, rotation-vibration absorption spectra, Raman spectra, and electronic spectra, and those models are used for both predicting and understanding lidar signals. Particles are much harder to model because of their variable size distributions, shapes, concentrations, and compositions, so lidar signals from particulates are consequently harder to interpret.

Remote sensing information on atmospheric particles is needed for a wide range of disciplines. For example, researchers concerned with Earth's energy balance need the OD of aerosol layers and their *single-scattering albedo*, which is the ratio of scattering efficiency to total extinction efficiency. Single-scattering albedo is unitless; a value of unity implies that all particle extinction is due to scattering and a single-scattering albedo of zero implies that all extinction is due to absorption. Unfortunately, lidar does not measure this quantity directly because it only receives light in the backscattered direction, but lidar is used to characterize and classify aerosol layers, and classification can constrain models. Aerosols also have an *indirect effect* on energy balance because they influence cloud formation through nucleation, thereby affecting a cloud's optical properties. Lidar is often used to estimate aerosol concentrations at cloud bases to support research on the indirect effect. Climate modelers also need to know the spatial extent and the altitudes of clouds and aerosol layers, and spaceborne lidars monitor those parameters globally.

Air quality researchers need to know the sizes of pollutant particles, or more precisely, their *aerodynamic diameter*, which for an irregular particle is defined as the

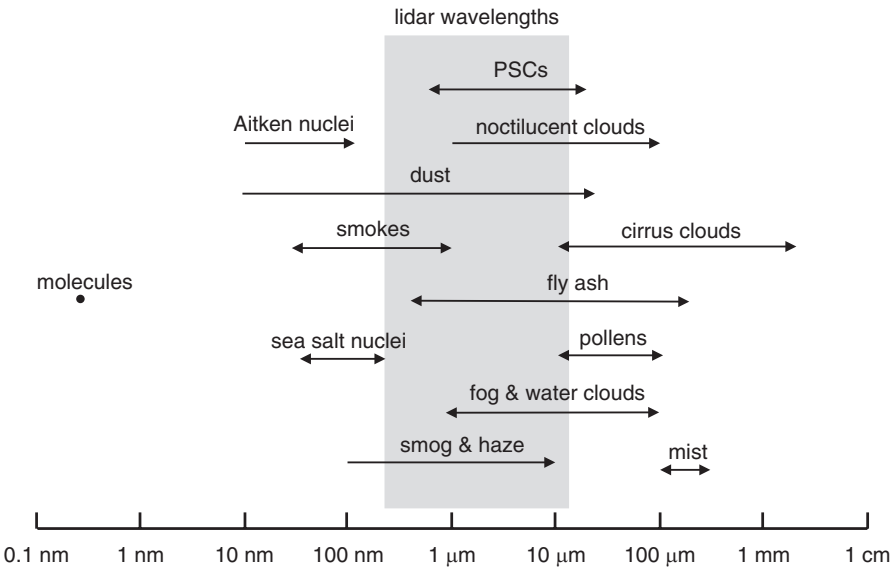


Figure 4.1 Atmospheric scatterer sizes. Size ranges are indicated by the black dot for molecules and by arrows for particles, and the range of wavelengths used by lidars (0.266–10.6 μm) is shown in gray. Adapted with permission from [1] with additional data from [2].

diameter of a spherical particle with the density of water and the same settling velocity as the irregular particle. Aerodynamic diameter determines how far into a person's lungs airborne particles are respired, and air quality is partially characterized by the values of PM 10 (particulate matter 10 μm or less in diameter, generally described as inhalable particles) and PM 2.5 (particulate matter 2.5 μm or less in diameter, generally described as fine particles). Air quality standards are set in terms of micrograms per cubic meter ($\mu g/m^3$) for both long-term and daily averages. Measuring particle size is a challenge for lidar, for reasons that are elucidated in this chapter, but a great deal of effort has been expended on such techniques. Mass per unit volume, often called *mass loading*, is not directly measured by lidar but can be estimated from lidar data if the aerosol optical properties can be established well enough by means of auxiliary data and models. Mass loading is needed for several applications. In aviation, safety for example, the mass loading in volcanic ash clouds is required for assessing the hazard they present to aircraft.

For the molecular atmosphere, universal models are used for all applications including lidar remote sensing. For aerosols, a universal model might be imagined that includes a size distribution (including aerodynamic diameters) and some description of shape, density, chemical composition, and complex refractive index as a function of wavelength. Along with scattering models for both spherical and non-spherical particles, such a model could be used to find the lidar extinction and backscatter coefficients at any wavelength, in a manner analogous to the

molecular modeling described in Chapter 3. Currently however, the research communities are a long way from having such models (except for the spherical droplets in haze, fog, and water clouds) and the underpinnings for them, such as the theory of scattering by nonspherical particles, are still active areas of research. For these reasons, models tend to be application specific, and lidar data are more often used to classify layers of particles by their types, using phenomenology rather than first principles modeling. Scattering theories for spherical particles are covered in the next section, which is followed by discussions of aerosols and clouds at all altitudes, lidar depolarization, and classifiers. Finally, sun photometers are described because they provide a wealth of auxiliary data to support lidar measurements at low cost. As in Chapter 3, the scattering theory and much of the information on the optical properties of atmospheric particles is adapted from McCartney [2].

4.1 Scattering Regimes

As illustrated in Figure 4.1, at lidar wavelengths, atmospheric particles span all three scattering regimes defined in Table 3.2: Rayleigh, Mie, and geometric optics. The boundaries of those regimes in terms of particle radius are shown in Figure 4.2 for two common lidar wavelengths, 0.532 μm and 1.064 μm. Radii from 0.042 μm to 4.2 μm are in the Mie regime for the shorter wavelength and from 0.085 to 8.5 μm at the longer wavelength. Those sizes span the particulate scatterers shown in Figure 4.1, except for pollens, mist, and some of the larger cloud particles, so Mie scattering dominates VIS-NIR lidar measurements in the presence of particles, although the other regimes are also important to lidar in general.

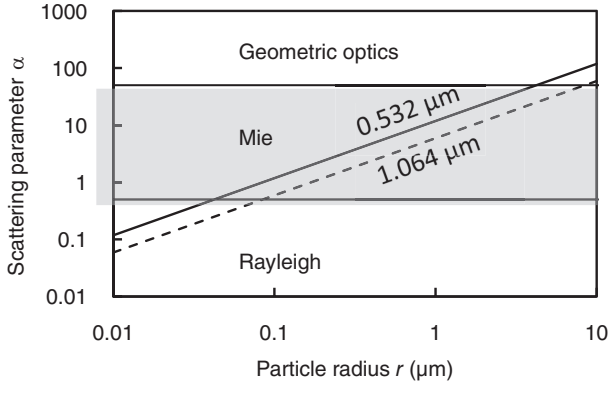


Figure 4.2 Scattering regimes. The horizontal lines mark the boundaries of the three regimes defined in Table 3.2. The solid diagonal line is the scattering parameter α at 0.532 μm versus particle radius, and the dashed line is α at 1.064 μm.

4.1.1 Rayleigh Scattering

Rayleigh scattering by molecules was covered in Chapter 3 in terms of the values of β_m and σ_m in the lidar equation, along with a description of how the angular distribution of scattered sunlight intensity and its polarization influence the sky background level in lidar receivers. To discuss scattering by particles in all three regimes, it is useful to introduce the concept of scattering cross section. The *angular scattering cross section* $\sigma(\theta)$ of a particle is defined as that cross section of the incident wave, acted upon by the scatterer, having an area such that the power flowing across it is equal to the power scattered per steradian at angle θ. Referring to Figure 3.1, this relation implies that

$$\sigma(\theta)E_0 = I(\theta) \tag{4.1}$$

when the scattering volume contains only one scatterer. Using this definition, the Rayleigh cross section of an air molecule is

$$\sigma_m(\phi) = \frac{9\pi^2}{N^2\lambda^4}\left(\frac{n^2-1}{n^2+2}\right)^2 \sin^2(\phi), \tag{4.2}$$

where ϕ is the angle from the polarization axis as shown in Figure 3.4 (the Rayleigh scattering intensity is independent of the scattering angle θ) [2]. The quantity n is the refractive index of air, which is very close to unity, and (n^2-1) is proportional to the number of particles per unit volume N, so the cross section is independent of the number density, as it must be. For small particles, the Rayleigh cross section is given by

$$\sigma_p(\phi) = \frac{16\pi^4 r^6}{\lambda^4}\left(\frac{n^2-1}{n^2+2}\right)^2 \sin^2(\phi), \tag{4.3}$$

where n is now the refractive index of the particle, which is not close to unity. Small particles are usually salts, which are hygroscopic. Their dry refractive indices span a range of roughly 1.5 to 1.6, but at high humidity, they accumulate so much water that their n values approach that of water, which is 1.33. For the extinction term in the lidar equation, we need the particle's total cross scattering section, which is given by integrating Eq. (4.3) over all solid angle to obtain [2]

$$\sigma_p = \frac{128\pi^5 r^6}{3\lambda^4}\left(\frac{n^2-1}{n^2+2}\right)^2. \tag{4.4}$$

The lidar extinction coefficient σ due to Rayleigh scattering by small particles is then given by

$$\sigma = N\sigma_p, \tag{4.5}$$

where N is the number of such particles per unit volume. Note that σ has dimensions of inverse length, or 1/m in the SI units used throughout this book. In practice, the calculation is not this simple because atmospheric particles are never all one size; they have size distributions, which are described in Section 4.2.

4.1.2 Mie Scattering

Much of the lidar scattering by atmospheric particles lies in the Mie scattering regime, as shown in Figures 4.1 and 4.2. Mie scattering theory was developed by German physicist Gustav Mie (1868–1857). It applies to dielectric spheres of any size: When the scattering parameter is small, Mie theory reduces to the Rayleigh result, and when it is large, it reduces to the geometric optics result. Although many particles are non-spherical, it is still instructive to examine the predictions of Mie theory, and in lidar, the term "Mie" has come to mean "not molecular," especially when it is used as a subscript. For example, the atmospheric backscatter coefficient is often said to have two components, β_{mol} and β_{Mie}. This usage is technically incorrect, but it has become standard in the atmospheric lidar literature.

The reason that the mathematical formulation of Rayleigh scattering is straightforward is that the phase of the incident electromagnetic wave is constant across the scattering particle, because of the constraint that the particle is much smaller than the wavelength. For this reason, all the electronic charges oscillate together, and the angular intensity distribution of the radiated wave is a simple dipole pattern. When the particle size and the wavelength are comparable, the electronic charges oscillate with different phases and the radiated wave is due to the interference of many waves. Consequently, the radiation pattern is more complicated and so is the theory. Mie scattering is formulated as a series expansion in multipoles (dipole, quadrupole, etc.) and the series converges slowly, so calculations were laborious in the days before electronic computers. For this reason, results were published as tabulations, and scattering in the forward direction was sometimes calculated as a diffraction pattern for larger particles because that calculation was easier. Most researchers made use of the published tables rather than attempting their own calculations. Nowadays, Mie scattering codes are available in several different computer languages and online, so calculations are straightforward.

In Mie theory, the *scattering efficiency factor* Q_{sca} is defined as the ratio of the total scattering cross section to the geometric cross section A of the particle, which is πr^2. Dividing the right-hand side of Eq. (4.4) by πr^2, we find that

$$Q_{\text{sca}} = \frac{128\pi^4 r^4}{3\lambda^4}\left(\frac{n^2-1}{n^2+2}\right)^2, \tag{4.6}$$

when the scattering by a sphere is in the Rayleigh regime. Remembering that the scattering parameter α is defined as $2\pi r/\lambda$, we see that Q_{sca} is proportional to α^4 for small particles, which is consistent with the Rayleigh result. Mie theory is formulated to yield the efficiency factor Q for given values of α and refractive index n. A set of tabulations is plotted in Figure 4.3 for α values from 0.1 to 30 and two different values of n. The curves have several notable characteristics; they rise rapidly from zero with the α^4 dependence mentioned above to peak values around four; then they resemble damped oscillations with a period of about 10 in scattering parameter; and the oscillations asymptotically approach a scattering efficiency of 2. In addition to the long-period oscillations, additional structure appears on a much finer scale. That finer

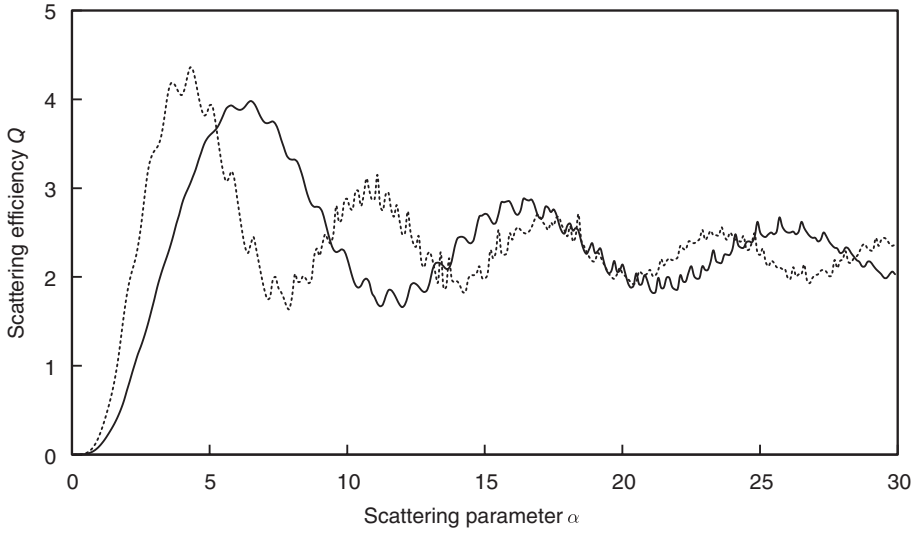


Figure 4.3 Mie scattering efficiencies vs. α. Solid line: $n = 1.33$, dashed line: $n = 1.50$. The data are from Appendix J in [2].

structure is usually smoothed out in published plots like Figure 4.3, but it is shown here in full detail to reveal the complexity of Mie scattering.

It may seem odd that a large non-absorbing sphere would have a cross section that is twice its geometrical area. The reason is that it causes both refraction (for rays passing through the sphere) as well as diffraction (for rays going around the sphere) in equal amounts, hence the factor of two. An absorbing sphere absorbs the incident light on it and diffracts light around it, so either way, the extinction efficiency is 2. The efficiencies are related by $Q_{\text{ext}} = Q_{\text{sca}} + Q_{\text{abs}}$, and Q_{ext} approaches 2.0 as α increases. Mie cross sections are often denoted by the letter C, as

$$\begin{aligned} C_{\text{ext}} &= Q_{\text{ext}} \times A \\ C_{\text{sca}} &= Q_{\text{sca}} \times A \\ C_{\text{abs}} &= Q_{\text{abs}} \times A. \end{aligned} \tag{4.7}$$

When there are N particles per unit volume, the extinction coefficient for the lidar equation σ is calculated as

$$\sigma = NC_{\text{ext}}, \tag{4.8}$$

which is analogous to Eq. (4.5) for the Rayleigh case and could also be written as $\sigma = N\pi r^2 Q_{\text{ext}}$. The lidar backscatter coefficient β is defined in a similar way, as the backscatter efficiency times the geometrical area times the number of particles per unit volume. With these relations, Mie theory calculations can be used to find the lidar parameters β and σ for any spherical aerosols such as those in haze, fog, and clouds, provided their refractive index is known. Such aerosols never occur in nature with just one size; rather they have size distributions, which are illustrated in the next section.

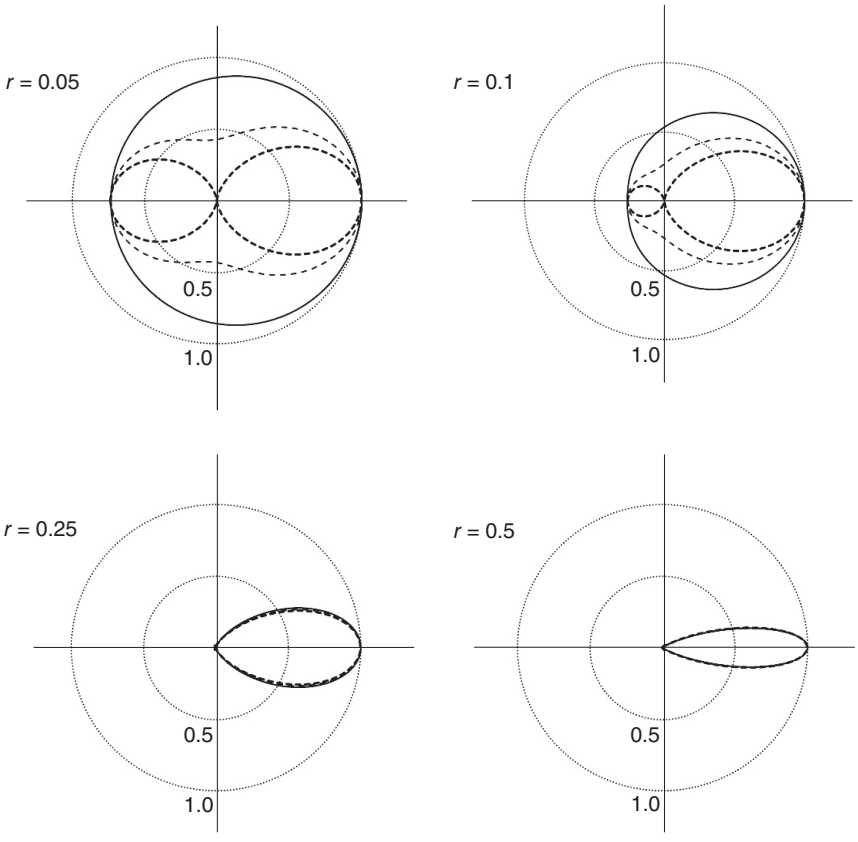


Figure 4.4 Mie scattering polar plots. All plots are normalized to unity in the forward direction. The refractive index is 1.33, the wavelength is 0.532 μm, and the droplet radii are as shown. Scattered natural light is shown by the light dashed line, perpendicular polarization by the solid line, and parallel by the heavy dashed line.

Models of size distributions for various water droplet phenomena are available in the literature [3], and methods for calculating optical properties from size distributions are described in [2]. The single-scattering albedo for spherical aerosols is also calculated by Mie theory.

Mie theory calculations can be accomplished by implementing any of the various Mie scattering codes that are available, or by using an online calculator [4] that produces results including the total scattering efficiency, the backscattering efficiency, and the angular scattering intensity. The online calculator was used to make the polar plots of Mie phase functions for water droplets shown in Figure 4.4, which are analogous to Figure 3.6 for Rayleigh scattering. The results are shown for natural (unpolarized) light and for linear polarization parallel and perpendicular to the *y*-axis. Figure 4.4 shows the changes in the phase function shape as the drops grow from near-Rayleigh through the Mie regime and approaching geometric optics, where the

scattering pattern is dominated by a forward diffraction peak. For the smallest radius shown, 0.05 μm, the scattering parameter α is 0.59 and the phase function resembles the Rayleigh pattern shown in Figure 3.6. Figure 4.4 may be misleading, because the curves have all been normalized to unity in the forward direction to illustrate the change in the shape of the phase function with droplet radius, and the normalization makes the backscattering appear to decrease as droplet radius increases. In fact, the total scattering and backscattering efficiencies increase by several orders of magnitude as the water droplet radius increases from 0.05 to 0.5 μm, as do the corresponding lidar equation parameters σ and β. The demonstration of this fact is left as a problem for the student.

Spheres are the simplest particle geometry for calculations, and they cause no depolarization in singly scattered lidar signals because of their symmetry. Many atmospheric particles, such as pollens and grains of salt or dust, have irregular shapes, and Mie theory results are just a rough approximation for their scattering properties. More sophisticated scattering theories have been developed for non-spherical particles, and the development of tractable calculations is still an active area of research [5].

4.1.3 Geometric Optics

When the scattering parameter α is greater than 50, the scattering regime is said to be geometric optics. In this limit, the extinction efficiency is 2 and the cross section of a particle for extinction is just twice its area. Geometric optics is diffraction, reflection, and refraction, calculated by ray optics. The angular scattering pattern for large spheres is mostly a forward diffraction peak, as shown in Figure 4.4.

4.2 Aerosols

Many of the particles shown in Figure 4.1 are aerosols, that is, small particles suspended in the air. Although atmospheric scientists often use the phrase "the aerosol" to mean a layer of aerosols, the word aerosol as used in this chapter refers to an individual particle. The main sources of aerosols are at the surface, so they tend to be mainly in the *mixed layer*, which typically extends from the surface to a height of 1–2 km. The mixed layer is also called the *mixing layer* and the *planetary boundary layer* (PBL). The other major layers of aerosols are smoke and dust that have been lofted into the free troposphere, and the stratospheric aerosols, which have a different source, described in Section 4.2.2.

4.2.1 Tropospheric Aerosols

The mixed layer is very dynamic, and it goes through a diurnal cycle as illustrated in Figure 4.5. At night, the layer is stratified and almost no vertical mixing occurs. A stable surface layer forms that can be as thin as 50 m. When sunlight heats the

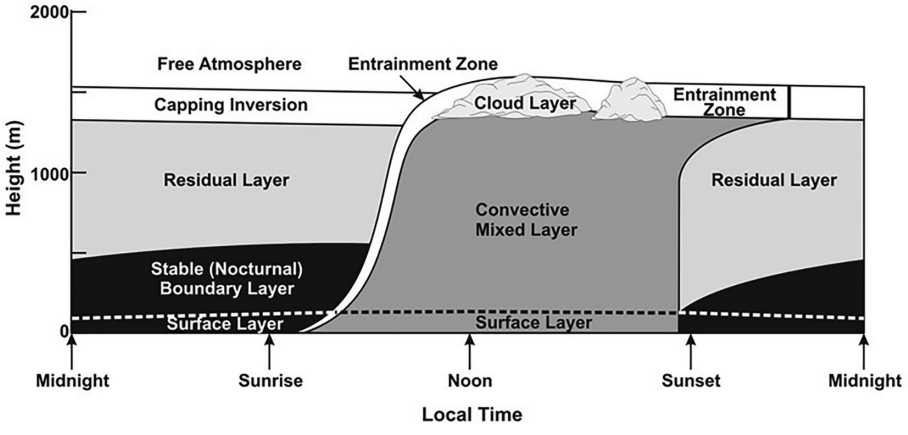


Figure 4.5 Mixed layer dynamics. The mixed layer goes through a diurnal cycle, driven by solar heating of the surface. Aerosols from the surface and from photochemistry are well mixed in the layer during daytime, and they remain in a residual layer at night. Reprinted by permission from Springer: *An Introduction to Boundary Layer Meteorology* by R. B. Stull. ©1988 [6].

ground in the morning, convective cells form and mixing begins. The cells are circulation patterns with updrafts and downdrafts, as illustrated in Figure 1.4. As the day wears on, the cells grow higher and higher to a limit that depends on location, time of year, and weather patterns. The width of the cells is about 1.4 times their height. Aerosols from the surface are distributed throughout the mixed layer by this convection, and aerosols generated in the layer by photochemistry (such as the production of ozone) are also well mixed. Any pollutant chemicals or smoke that may be in the free troposphere above the mixed layer will be entrained when the convective cells reach them and then mixed all the way down to the surface. The puffy white cumulus clouds frequently seen in summertime form at the top of the mixed layer. Later in the day when the sunlight is no longer heating the surface, the mixing stops. However, the aerosols remain because their fall velocity is low, forming the *residual layer*.

Because of boundary layer dynamics and the fact that most aerosols originate at the surface, aerosols give rise to signals from ground-based lidars that typically look like the data shown in Figure 4.6, which were recorded with EARL. This type of figure is known as a *time-height plot*, in which the horizontal axis is time, the vertical axis is height, and the quantity of interest (attenuated backscatter in this case) is displayed on a gray scale or color scale. A similar format is often used for airborne or spaceborne data displays, with latitude–longitude on the horizontal axis. A general term for all such displays is *curtain plot*. The geometric function $G(R)$ reaches unity at 550 m altitude in EARL's short-range receiver, so data are not plotted at altitudes below that level. Figure 4.6 illustrates two aspects of boundary layer dynamics revealed by lidar that are not obvious from the standard meteorological diagram shown in Figure 4.5: First, the residual aerosol layer becomes stratified during the

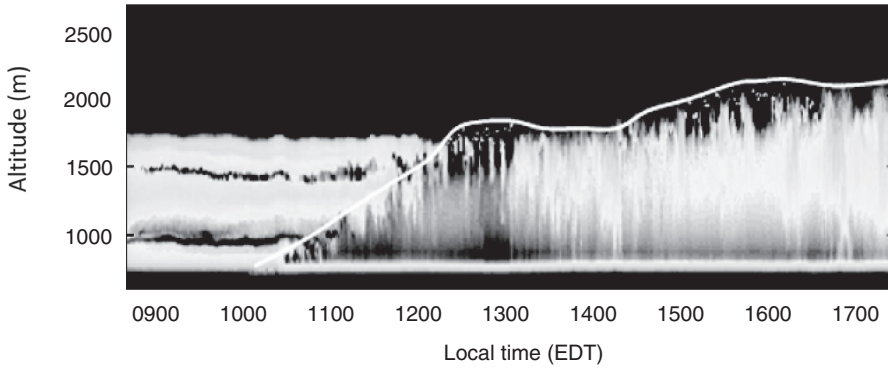


Figure 4.6 Mixed layer evolution. The white line was added to show how the mixed layer height evolves with time of day. The data were recorded with EARL in Atlanta, Georgia, on September 30, 2004.

night due to air motions that form thin layers with high backscatter. The signal from a ground-based lidar will usually show several individual layers of varying backscatter intensity. Second, the mixed layer is not uniform; it builds up as individual convective cells, with updrafts in the centers and downdrafts at their edges. As the cells pass through a vertical laser beam due to prevailing winds, a lidar traces out their structure using backscatter from the aerosols that are entrained in them. Aerosols are said to be *tracers* for air motion.

Mixed-layer aerosols are always present to some extent, and they are usually visible as the phenomenon of haze. There are many sources of haze aerosols. Soil dust becomes airborne at wind speeds of just a few m/s in dry conditions, and major deserts such as the Gobi and the Sahara experience dust storms that loft dust into the free troposphere, where it is transported for many km. Dust from the Gobi Desert is observed by lidars in the U.S., for example, and Saharan dust is observed by lidars all over Europe. Industrial operations such as steel mills, smelters, and cement plants are major sources of aerosols, as is combustion of any type, whether in industry, transportation, or as grass fires and forest fires. These aerosols are mostly non-hygroscopic, meaning that they do not collect water molecules from the water vapor in the air. Plants and trees are the sources of another type of aerosol. They exude aromatic hydrocarbons (terpenes), which, in the presence of sunlight and ozone, form particles. These particles are hygroscopic, so they act as condensation nuclei. Other aerosols are sulfates from several sources including photochemistry, volcanic ash, and bits of sea salt from bursting bubbles in ocean whitecaps. All these types are commonly lumped into three major categories: (1) natural inorganic aerosols, including fine dust, volcanic ash, sea salt, and water drops; (2) natural organic aerosols, including smoke, plant and tree aerosols, and pollen spores; and (3) anthropogenic, including smoke, ash, dust, and aerosols due to ozone production in polluted urban environments.

Haze and fog particles span a size range of 0.01–10 μm, as shown in Figure 4.1. They are said to be in three modes: *Aitken nuclei* (also called the *nuclei mode*) have radius less than 0.1 μm (by definition); *accumulation mode* aerosols span 0.5–2.5 μm and have a long lifetime in the atmosphere; and *coarse mode* aerosols have radii greater than 2.5 μm and they settle out of the air in a few hours. The first two modes taken together are called *fine aerosols*. The *nuclei mode* aerosols consist primarily of combustion products and particles formed in the atmosphere by photochemistry. Because of their high number concentration, especially near their source, these small particles coagulate rapidly. Consequently, nuclei particles have relatively short lifetimes in the atmosphere and end up in the accumulation mode. The *accumulation mode* includes combustion particles, smog particles, and coagulated nuclei mode particles. The smog particles are formed in the atmosphere by photochemical reactions. Particles in this mode are small and they coagulate too slowly to reach the coarse mode. Hence, they have a relatively long lifetime in the atmosphere, and they account for most of the visibility effects of atmospheric aerosols. The *coarse mode* consists of windblown dust, large salt particles from sea spray, and mechanically generated anthropogenic particles such as those from agriculture and surface mining. Because of their large size, the coarse particles readily settle out or impact on surfaces, so their lifetime in the atmosphere is only a few hours.

Aerosols experience two kinds of growth: When two aerosols collide, they may coalesce to form one larger aerosol. This process is called *coagulation*, and it grows larger particles at the expense of the smaller ones, shifting the distribution of sizes toward the larger. A more dramatic phenomenon is the rapid growth of hygroscopic aerosols at relative humidity above about 70%, in which the aerosols collect water molecules from the vapor phase, causing a haze to become fog, for example. A dramatic change in optical properties occurs with a growth of radius which is typically only a factor of 10.

Aerosols are removed from the air by several mechanisms. One is coagulation, described above. Another is *fallout* due to gravity. The fall speed depends on both drop size and altitude. Near the surface, an increase of radius by a factor of 10 causes an increase of fall speed by a factor of about 100, so fallout depends strongly on radius. Finally, *washout* by rain or snow is very effective at removing aerosols. Through these mechanisms, aerosols are continuously being removed and hence there is not an unlimited buildup of aerosols over time.

Because of the mechanisms of aerosol creation, growth, and removal described above, aerosols are not all one size, but rather have size distributions. This fact complicates the modeling of aerosol optical properties, but standard methods have been developed that use measured or modeled distributions [2]. Examples of modeled distributions are shown in Figure 4.7 for number, surface area, and volume versus diameter. Mass is related to volume through a particle's density. The same hypothetical log-normal distribution is used in all three plots: Note that the horizontal axis is logarithmic, and the curves have the familiar shape of the normal distribution. Each distribution is normalized so that the total area is 1000.

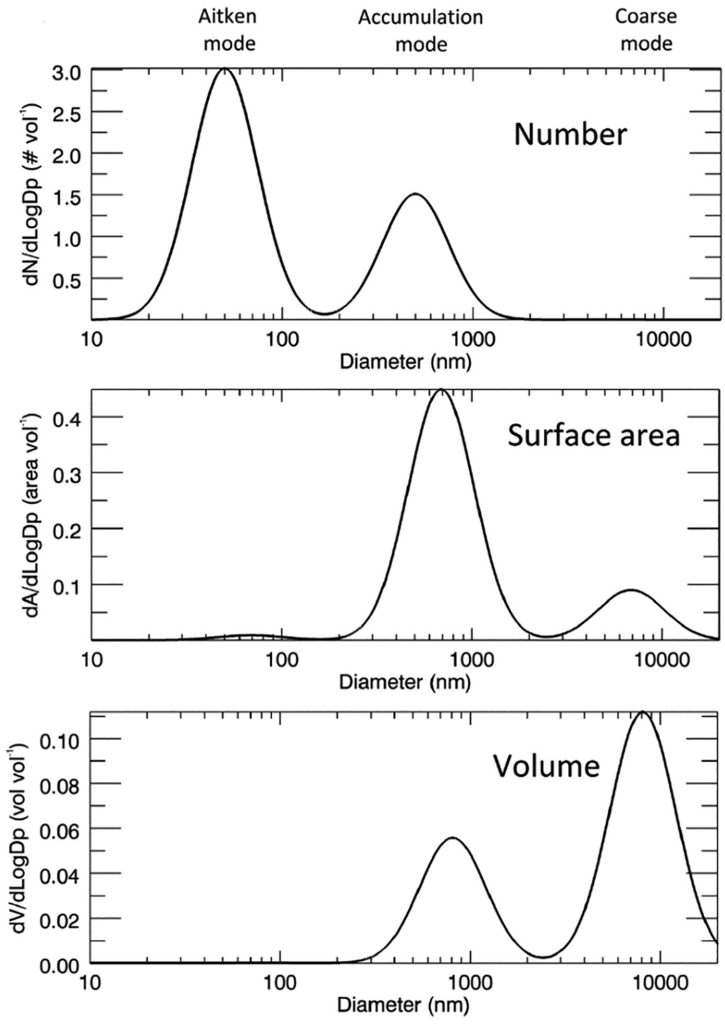


Figure 4.7 Aerosol size distributions. The same hypothetical log-normal aerosol distribution is plotted, from top to bottom, as a number vs. diameter distribution, a surface area vs. diameter distribution, and a volume vs. diameter distribution. Typical mode names are shown at the top. "Synthetic aerosol distribution in number area and volume space" by Niall Robinson is licensed under the Creative Commons CC0 1.0 Universal Public Domain Dedication [7].

4.2.2 Stratospheric Aerosols

The stratospheric aerosol layer, sometimes called the Junge layer, was determined to be a persistent global layer of particles in the stratosphere that contained sulfur by Junge and Manson in 1961 [8]. The layer is now known to be predominantly composed of sulfuric acid droplets. It generally starts at, or slightly below, the tropopause and it extends up into the stratosphere. The aerosols slowly sink and become entrained

in thunderstorms and wash out, and they are replenished by SO_2 that diffuses into the stratosphere from the troposphere. Volcanic eruptions can inject large amounts of SO_2, causing the layer backscatter to increase by orders of magnitude. As shown in Figure 1.7, the stratospheric aerosol layer has been monitored with lidars for more than 30 years. Those measurements began at the ruby laser wavelength of 694 nm but later transitioned to the Nd:YAG laser wavelengths of 355, 532, and 1064 nm. The most common way of characterizing the stratospheric aerosol layer with lidar is to integrate the aerosol backscatter coefficient from some lower bounding altitude to an upper bound, to obtain the integrated aerosol backscatter (IABS) in units of sr^{-1}. This procedure ignores extinction in the layer, but during times of low volcanic activity, the extinction is negligible even when integrating through layers that are several km thick. Jaeger and Hofmann [9] showed that the visible-light extinction-to-backscatter ratio for the aerosols is about 40 sr, so an IABS value of 1×10^{-4} corresponds to an OD of 4×10^{-3} and a two-way transmission greater than 99%. The wavelength dependence of the backscatter coefficient was found to be "almost lambda to the −2" [10]. GTRI researchers extended stratospheric aerosol backscatter measurements to 1570 nm in 2011 by using a photon-counting lidar, and obtained the exponent 1.9 ± 0.1, which agrees well with the UV-NIR value [11].

4.2.3 Aerosol Sizes and Lidar Measurements

As mentioned in the introduction to this chapter, there are many applications for information about atmospheric particle size. However, size information is hard to obtain with lidar. Figure 4.3 shows that the scattering efficiency Q has a strong dependence on the scattering parameter when α is less than four or five, which suggests that multi-wavelength lidar might be sensitive to size. However, Q approaches zero as α does, so there is a limit to how small α can be for useful measurements. An appropriate range for lidar size measurements might be taken as $\alpha = 1 - 5$. The choice of lidar wavelengths depends on the particle sizes of interest, which are shown in Table 4.1 for particles as small as nuclei up to the droplets in fog and clouds. Table 4.1 shows that wavelengths for both the smallest and largest particles are outside the lidar range that would be sensitive to size, and wavelengths for particles with radius 0.1 μm to 1 μm span the UV through MWIR range. Unfortunately, current MWIR lidar technology is not sensitive enough for such aerosol measurements and even SWIR lidar aerosol data is quite limited. In addition, the two accepted lidar types for aerosol extinction operate in UV-VIS region. A pair of wavelengths in the UV-VIS would be sensitive to size for aerosols around 0.1 μm radius, but otherwise, lidars are generally measured in the region of Figure 4.3, where α is large enough that the scattering efficiency oscillates about the value 2, and particles with a range of sizes, as shown in Figure 4.7, tend to average out those oscillations. For these reasons, atmospheric particle size information is difficult to obtain with multi-wavelength lidar, although considerable effort has been made and some progress has been reported in the literature [12].

Table 4.1 Wavelengths required for scattering parameter in the range 1–5

Aerosol radius (μm)	Aerosol types	λ_{min} (μm)	λ_{max} (μm)
0.01	Aitken nuclei	0.01	0.06
0.1	Smoke, dust sea salt nuclei	0.13	0.63
1	Haze, fog, clouds, dust	1.26	6.28
10	Clouds, pollen, ash	12.6	62.8

4.2.4 Aerosol Lidar Signal Modeling

To model the lidar signal from an atmospheric aerosol layer, the lidar equation requires the parameters σ and β, which are proportional to the aerosol number density N. For spherical particles, these parameters can be calculated for any wavelength using the Mie theory, provided the refractive index and the size distribution are known. In the more general case, typical values are taken from summaries of measured data or from auxiliary data. A common way of finding σ at the surface is to estimate the *visibility* V (also called the meteorological range) in a layer of haze or a fog. The visibility is the distance at which a trained observer can just distinguish the contrast between a dark object and the horizon sky. The extinction coefficient σ can be found from the visibility by using the equation

$$\sigma = 3.912 / V. \tag{4.9}$$

Equation 4.9 is attributed to Koschmieder [13], who made many observations at Danzig using large targets mounted on barges that were towed out to sea. Observations of visibility are subjective, and they only pertain to the mid-visible where the eye is most sensitive, but they are simple. An alternative is to use a nephelometer, which is a common meteorological instrument that measures visibility by means of forward scattering in a small volume defined by the intersection of a transmitted beam of light with a receiver FOV. After σ has been determined from the visibility, the next question is how to find β. The aerosol lidar ratio S_a is defined as

$$S_{\mathrm{a}} = \sigma_{\mathrm{a}} / \beta_{\mathrm{a}}, \tag{4.10}$$

so $\beta_{\mathrm{a}} = \sigma_{\mathrm{a}} / S_{\mathrm{a}}$. Many other symbols are used for the lidar ratio, including simply LR, and some authors use the inverse of this ratio instead. Unfortunately, S_{a} values from 20 to 120 have been measured, which is a rather large range, in contrast with the molecular case where $S_m = 8.38$. The lower values generally correspond to clean atmospheres and the higher ones to polluted atmospheres, which is of some help, but often the value of S_a is not accurately known and it must be taken from a table of typical values for various types of aerosols, at the wavelength of interest. Lidar parameters at mid-visible are often extrapolated to other wavelengths by assuming or measuring a *wavelength exponent*, which is –4 for Rayleigh scattering but anywhere from –4 to 0 for particles, as shown in Table 3.2. Measurement of the wavelength exponent for aerosol layer ODs by sun photometry is described in Section 4.6.

4.3 Clouds

Clouds are observed at altitudes from the surface to 80 km. They are all at least partly composed of water in its liquid or solid phases. In the troposphere, clouds are all composed of water droplets or ice crystals. In the stratosphere, several types of PSC have been observed, and NLCs occur in the mesosphere, around 80 km.

4.3.1 Troposphere

Meteorologists have categorized and named many different types of clouds based on their morphologies and altitudes, but three main types are cumulus, stratus, and cirrus. Cumulus has a characteristic puffy appearance, and it is formed as individual clouds when water vapor condenses into liquid drops at the tops of the updrafts in the mixed layer. Stratus clouds have a layered appearance, and they tend to be at higher altitudes than cumulus. Cirrus forms above 5.5 km in temperate regions, and it always includes ice crystals, which give it a wispy appearance. Mixed phase (ice and water) clouds sometimes occur. Cirrus ice crystals have dimensions of 0.01 mm to several mm, making them the largest particles in the atmosphere other than hailstones. Tropospheric clouds have huge backscatter coefficients: At visible wavelengths, cumulus and stratus clouds have β values in the range $1–6 \times 10^{-3}$ $m^{-1}sr^{-1}$ [2], which is three orders of magnitude greater than the backscatter due to clear air, so a sensitive lidar may need optical attenuation in its receiver for cloud observations to avoid exceeding the dynamic range of its detection circuitry. The cumulus clouds at the top of the boundary layer generally cause the largest signal that a ground-based lidar will ever see. The large β values of clouds enable the commercial instruments called ceilometers, which are commonly used at airports to measure cloud base height, known as the *cloud ceiling* in aviation.

The cloud extinction-to-backscatter ratio is on the order of 20 sr, so extinction coefficients can be as high as about 0.1 m^{-1}. Assuming $\sigma = 2 \times 10^{-2}$, a 100-m path in a cloud would have a two-way transmittance of 0.02, so lidar cannot probe very deeply into water clouds. In addition to extinction, clouds present another problem: The assumption in the lidar equation of single scattering is often violated. However, lidar researchers have developed a technique to exploit multiple scattering in clouds. The amount of multiply scattered light in a lidar signal depends on the lidar's FOV and the drop size distribution. By operating lidars with multiple FOVs, information about the drop size distribution can be acquired [14].

4.3.2 Polar Stratospheric Clouds

The stratosphere is very dry, so water clouds rarely form there. However, in the extreme cold of polar winters, when temperatures drop below about 198 K, stratospheric clouds of different types may form at altitudes of 15–25 km. They are more common in the Antarctic than the Arctic because the Antarctic atmosphere is colder. These polar stratospheric clouds are classified according to their physical state and chemical composition. There are two main types: Type I PSCs are mostly supercooled

droplets of water and nitric acid, and they are involved in ozone depletion; and Type II PSCs consist only of frozen ice crystals and they are not harmful. Type I PSCs have three subtypes with varying sizes, shapes, and compositions. The important role of lidar measurements in characterizing these PSCs is described in Section 4.5, along with information on their compositions and particle size ranges.

4.3.3 Noctilucent Clouds

Noctilucent clouds (the name means night shining clouds) are tenuous layers of ice crystals that are only visible at twilight, when they are illuminated by the sun's rays, but the sun has set at the observer's location. They are most often observed during summer months from latitudes between ±50° and ±70°. Noctilucent clouds are the highest clouds in the atmosphere, occurring near the mesopause at altitudes of 76–85 km. As shown in Figure 1.2, this is the coldest region in the neutral atmosphere. Noctilucent clouds are also known as *polar mesospheric clouds*. They are composed of tiny crystals of water ice with sizes <100 nm. They were first seen by lidar at Andenes, Norway in 1989 [15] and have since been studied with lidars in both hemispheres and at mid-latitudes. The RMR (Rayleigh/Mie/Raman) lidar at ALOMAR (Arctic Lidar Observatory for Middle Atmosphere Research) in Northern Norway was used from 2011 to 2018 to acquire a large data set on NLCs with 1 s time resolution and 25 m spatial resolution, and a comprehensive analysis of that data was reported in 2020 [16]. The authors called NLCs a "unique tracer to visualize middle atmosphere dynamics," including wave phenomena and dissipation. Clouds were observed with thicknesses from <100 m to a few km, with β values in the range $2–6 \times 10^{-9}$ $m^{-1}sr^{-1}$.

4.4 Depolarization in Lidar Signals

Lidar systems typically transmit polarized light, and analysis of the polarization state of the received light is a powerful remote sensing tool. In the case of clouds, lidar depolarization provides unambiguous ice/water phase discrimination as well as identification of some crystal types and orientations. Lidar depolarization is also used to identify desert dust in the free troposphere and to study its role in ice nucleation. Lidar depolarization studies made a major contribution to the identification and classification of polar stratospheric clouds, and space-based lidars such as CALIOP (cloud-aerosol lidar with orthogonal polarization) include a cross-polarization receiver channel because of the valuable information that it provides for classifying clouds and aerosols. The theory of depolarization of laser light by atmospheric particles is presented here, and instrumentation for depolarization lidar is described in Chapter 6.

Light is a transverse electromagnetic wave, meaning that the electric and magnetic fields are perpendicular to each other and to the direction of propagation. The *polarization state* of a beam of light describes the motion of the tip of the electric field vector in a fixed plane perpendicular to the direction of propagation. Three examples are shown in Figure 4.8.

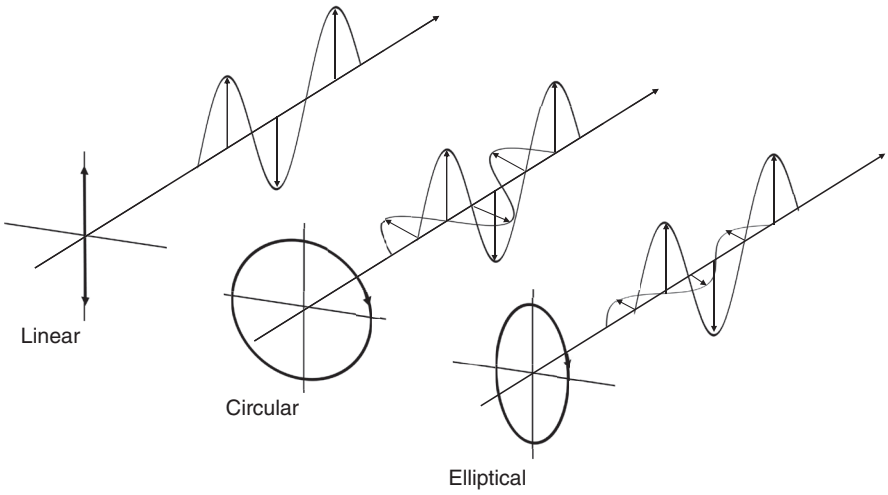


Figure 4.8 Polarization states. The polarization state of a beam of light describes the motion of the tip of the electric field vector in a fixed plane perpendicular to the direction of propagation.

If the tip of the electric field vector moves back and forth in a straight line, the light is said to be linearly polarized. If the beam of light is composed of two perpendicular linearly polarized beams of equal amplitude and ¼ wave out of phase, the electric field vector will rotate, and its tip will describe a circle. Such a beam is said to have circular polarization. In the most general case, elliptical polarization, the tip of the electric field vector moves in an ellipse, as shown in Figure 4.9. One way to describe the polarization state is by specifying the semimajor and semiminor axes A and B of the polarization ellipse, its orientation θ, and the sense of rotation for light coming at the observer, which is counterclockwise in the figure. For this reason, the historical name for polarization measurements was *ellipsometry*. Note that four parameters are required to completely describe the polarization state of a beam of light, and that the coordinate system (x and y axes) must be defined first. The rotation of a visible-light electric field vector is not directly observable because it occurs at hundreds of THz. Optical physicists in the nineteenth century needed a description in terms of measurable quantities, and for this purpose George G. Stokes (1819–1903), an Anglo-Irish physicist and mathematician, defined a four-component vector in 1852 that completely describes the polarization state of a beam of light in terms of intensities, which are time-averaged quantities. The Stokes vector is conventionally written as S = [I, Q, U, V] where the four elements describe the total intensity, the intensity on the x and y axes, the intensity on +45- and –45-degree axes, and the intensity that is left-hand circular (LHC) or right-hand circular (RHC). The four components of the Stokes vector, known as the *Stokes parameters*, provide a description of the polarization state that is experimentally convenient because each parameter corresponds to a sum or difference of measurable intensities. Stokes vectors are often normalized such that the first element I is equal to unity. As examples,

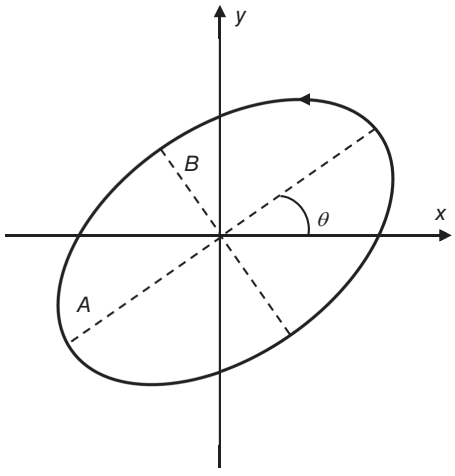


Figure 4.9 The polarization ellipse. The four parameters A, B, θ and the direction of rotation completely specify the polarization state.

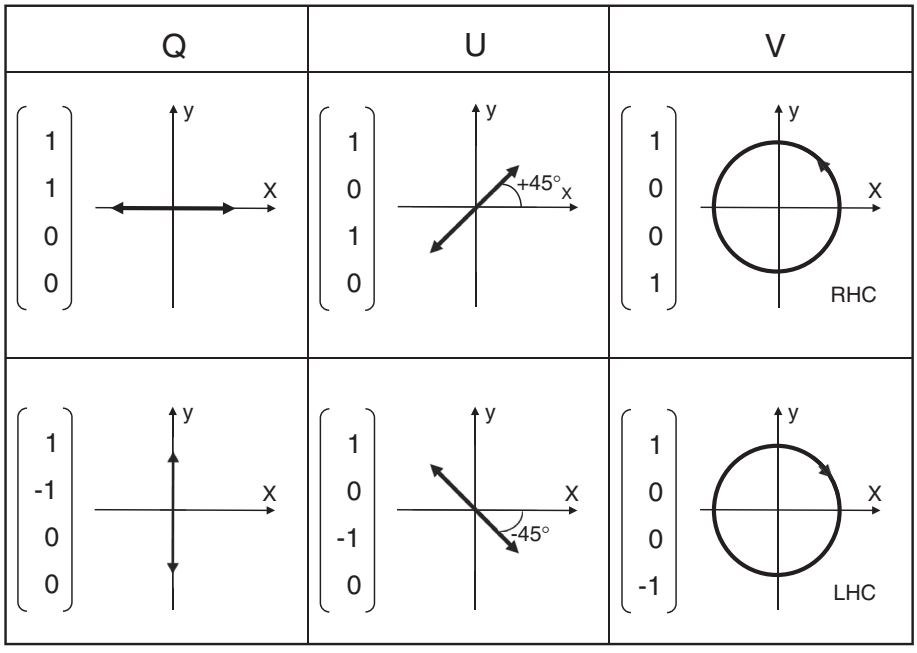


Figure 4.10 The polarization basis states. The path followed by the tip of the electric field vector is shown on the axes in each box, with the corresponding normalized Stokes vector to the left.

the vector S = [1, 1, 0, 0] describes a beam of light linearly polarized along the x-axis, and the vector S = [1, –1, 0, 0] describes a beam of light linearly polarized along the y-axis. The vector S = [1, 0, 0, 0] describes light that is completely unpolarized. The polarization basis states are illustrated in Figure 4.10. Note that they are

Table 4.2 Mueller matrices.

Reflection	*x*-axis analyzer	*y*-axis analyzer
$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$	$\frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$	$\frac{1}{2}\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
$\lambda/4$ plate, fast axis vertical	+45° analyzer	–45° analyzer
$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$	$\frac{1}{2}\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

defined for light coming at the observer, so positive circular polarization refers to counterclockwise rotation, and +45 degree refers to the electric field vector rotated from the positive *x*-axis toward the positive *y*-axis. Positive circular polarization is also called RHC polarization, because it corresponds to the right-hand rule, where the thumb of the right hand points in the direction of propagation while the fingers curl from the *x*-axis to the *y*-axis. Light in any of the basis states is often referred to as being in a pure polarization state, and lidar transmitters are most often intended to transmit laser beams with either linear or circular polarization.

As light interacts with the atmosphere and with the optical elements in a lidar system, its polarization state may be changed. Such changes are conveniently described by the 4 × 4 Mueller matrices, which multiply the Stokes vectors using the standard rules of matrix multiplication. The six Mueller matrices necessary for the analysis presented in this section are listed in Table 4.2. Many others are available to represent the interaction of light with al most any sort of optical element [17, 18].

Example. The Mueller matrix for reflection, as listed in Table 4.2, describes scattering in the backward direction by a spherical particle (or incoherent scattering from an ensemble of spherical particles) and is given by

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}. \tag{4.11}$$

Consider a laser beam linearly polarized along the *x*-axis. The polarization state of the light backscattered from such particles is found from

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \tag{4.12}$$

which shows that the polarization state is unchanged. Note that the matrix operates on the initial Stokes vector, which is left of the equals sign, and the final Stokes vector is to the right of the equals sign. Each element of a diagonal Mueller matrix only modifies the corresponding polarization parameter in the Stokes vector, and the matrix defined in Eq. (4.11) describes a polarization-preserving process. This latter fact may not be immediately apparent, because the [3,3] and [4,4] elements are –1, which is to say, they reverse ±45° polarization and exchange LHC and RHC polarization states. The reason for these reversals is that backscattering reverses the direction of propagation, and polarization states are defined for light coming toward the observer.

In lidar measurements, we are more interested in *depolarization,* which refers to atmospheric scattering phenomena (other than the reversal of propagation direction) that change the polarization state of the light received by a lidar system relative to the state of the transmitted light. An instrument that can determine all four components of the Stokes vector is called *fully polarimetric*. Lidar systems with that capability have been constructed, but a much more common approach is to transmit polarized light and employ two receiver channels with analyzers corresponding to the transmitted polarization and the cross-polarization. The analyses presented below pertain to such a lidar receiving light due to single scattering by randomly oriented scatterers, which corresponds to the vast majority of lidar depolarization measurements. The normalized Mueller matrix for randomly oriented scatterers is

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-d & 0 & 0 \\ 0 & 0 & d-1 & 0 \\ 0 & 0 & 0 & 2d-1 \end{bmatrix}, \tag{4.13}$$

where the parameter d, which has a range 0–1, is a measure of the propensity of the scattering medium to depolarize the incident polarization, as may be seen by setting d equal to zero, in which case Eq. (4.13) reduces to Eq. (4.11). The matrix defined in Eq. (4.13) was not published until 1995 [19], it was not noticed and used by a lidar researcher until 2007 [20], and the wider lidar community was not made aware of it until 2008 [21]. For this reason, the standard treatment of polarization with Stokes vectors and Mueller matrices was not used during much of the history of the lidar technique, and the pre-2008 lidar literature employs misleading notation and concepts that should be avoided. In the limit where d equals 1, the [2,2] element of the matrix in Eq. (4.13) becomes zero, which means that a linearly polarized incident beam becomes completely depolarized by the scattering process (S = [1, 0, 0, 0]). Because Eq. (4.13) defines a matrix with only one parameter, d, the obvious task of analyzing lidar depolarization data is to retrieve that parameter. All the elements of

the scattering matrix are then determined. In the next sections, algorithms for finding d from the lidar signals are developed for the two most common polarization-sensitive lidars, which transmit laser beams with linear or circular polarization.

4.4.1 Linear Polarization

A linear polarization lidar is illustrated schematically in Figure 4.11. The analysis starts with the initial polarization state, modifies it with the scattering process, and determines the received intensities after the two orthogonal polarizers in the receiver. The Stokes vector describing the transmitted laser beam is S = [1, 1, 0, 0]. For clarity, all vectors and matrices are normalized. The relative intensities that emerge from the parallel and perpendicular analyzers are calculated below.

Parallel

$$\frac{1}{2}\begin{bmatrix}1&1&0&0\\1&1&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1-d&0&0\\0&0&d-1&0\\0&0&0&2d-1\end{bmatrix}\begin{bmatrix}1\\1\\0\\0\end{bmatrix}=\begin{bmatrix}1-d/2\\1-d/2\\0\\0\end{bmatrix}. \tag{4.14}$$

Intensities add incoherently, so the Stokes vectors can be decomposed as $I_{\text{total}} = I_{\text{pol}} + I_{\text{unpol}}$. The result described in Eq. (4.14) can therefore be decomposed as follows:

$$\begin{bmatrix}1-d/2\\1-d/2\\0\\0\end{bmatrix}=(1-d)\begin{bmatrix}1\\1\\0\\0\end{bmatrix}+(d/2)\begin{bmatrix}1\\1\\0\\0\end{bmatrix}. \tag{4.15}$$

Equation (4.15) shows that all of the polarized fraction I_{pol} plus one-half of the unpolarized fraction I_{unpol} of the received light passes through the parallel analyzer.

Perpendicular

$$\frac{1}{2}\begin{bmatrix}1&-1&0&0\\-1&1&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1-d&0&0\\0&0&d-1&0\\0&0&0&2d-1\end{bmatrix}\begin{bmatrix}1\\1\\0\\0\end{bmatrix}=(d/2)\begin{bmatrix}1\\-1\\0\\0\end{bmatrix}. \tag{4.16}$$

Equation (4.16) shows that one-half of the unpolarized fraction I_{unpol} of the received light passes through the perpendicular analyzer. The signals S from the detectors are therefore

$$S_{\parallel} = C_{\parallel}(I_{\text{pol}} + \frac{1}{2} I_{\text{unpol}}) \text{ and} \tag{4.17}$$

$$S_{\perp} = C_{\perp} \frac{1}{2} I_{\text{unpol}}, \tag{4.18}$$

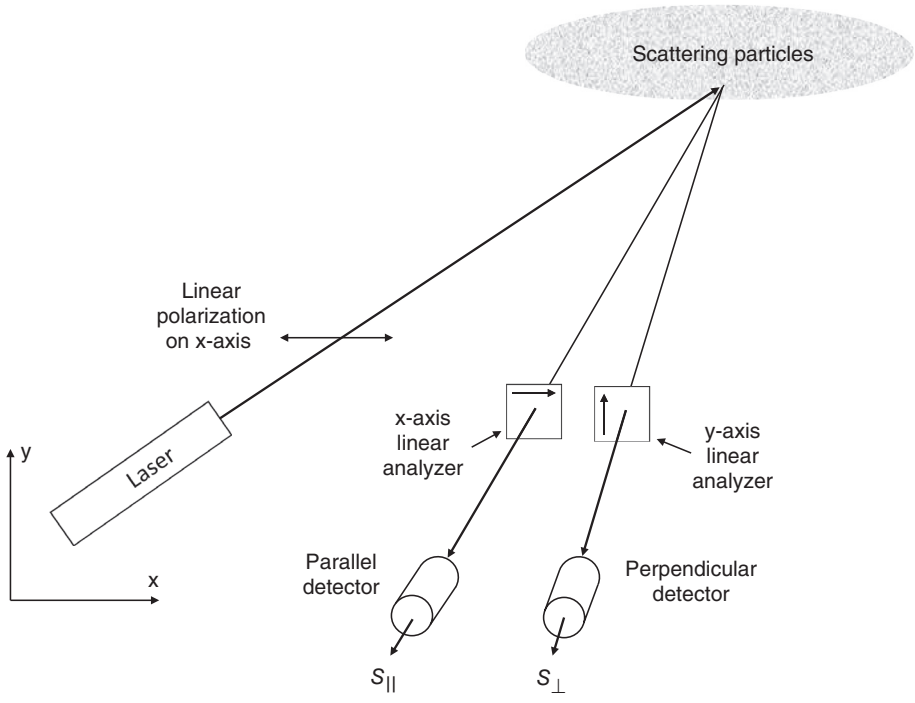


Figure 4.11 Schematic illustration of a linear lidar depolarization measurement. The transmitted polarization is linear along the *x*-axis (horizontal in the figure). The received power is analyzed in two receiver channels with linear polarization analyzers oriented horizontally (parallel) and vertically (perpendicular) to the transmitted polarization state. The signal powers emerging from the analyzers are converted to electronic signals $S_{||}$ and $S_{\perp}$ by the detectors. For single scattering by spheres, all the received power is in the parallel receiver channel.

where $C_{||}$ and $C_{\perp}$ are the calibration factors for the two receiver channels. Note that $(½)I_{\text{unpol}}$ appears in both receiver channels because a polarization analyzer at any orientation passes one-half of the unpolarized incident intensity. This is the experimental definition of unpolarized light, and the fact that $(½)I_{\text{unpol}}$ appears in both receiver channels is a key to understanding linear depolarization lidar signals. From Eqs. (4.17) and (4.18), we find the depolarization parameter d from the receiver signals S as

$$d = \frac{2S_{\perp}}{S_{||} + S_{\perp}}. \tag{4.19}$$

Note also that the [2,2] element of the matrix defined in Eq. (4.13) can be found from the lidar signals as

$$1 - d = \frac{S_{||} - S_{\perp}}{S_{||} + S_{\perp}}. \tag{4.20}$$

The signals S require the subscripts $\parallel$ and $\perp$ because they result from the action of the receiver's polarization analyzers on the received light, and they depend on the relative orientations of the receiver analyzers to the transmitted beam polarization. The intensity components I have subscripts referring to their polarization states (*pol*, *unpol*, and *total*).

4.4.2 Circular Polarization

Circular lidar depolarization is quite different from linear depolarization, and it does not lend itself to a simple physical description such as that following Eq. (4.15) for the linear case. As shown by the factor of two in the [4,4] element in Eq. (4.13), circular depolarization measurements are twice as sensitive to the parameter d, compared to linear depolarization. A schematic diagram for a circular depolarization lidar is shown in Figure 4.12. A configuration often used in common-optics systems uses the same ¼-wave plate at 45 degrees in both transmit and receive paths to separate the transmitted and received light. This optical technique is known as a transmit/receive (T/R) switch, by analogy with a waveguide device used in radar systems. The lidar conceptually illustrated in Figure 4.12 passes linear laser polarization at +45 degrees through a ¼-wave plate with the fast axis vertical, converting the transmitted laser beam to RHC, so that the normalized Stokes vector describing it is S = [1, 0, 0, 1]. The received light passes through another ¼-wave plate at the same orientation followed by linear analyzers at +45 degrees and –45 degrees. In circular depolarization, the term *co-polar* refers to the light in the state expected when d is equal to zero (LHC in the case illustrated), whereas *cross-polar* refers to the opposite handedness. These terms can also be substituted for parallel and perpendicular in the case of linear depolarization. The normalized intensities of the light that emerges from the analyzers are calculated below.

+45 Degrees (Co-polar)

$$\frac{1}{2}\begin{bmatrix}1&0&1&0\\0&0&0&0\\1&0&1&0\\0&0&0&0\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&0&-1\\0&0&1&0\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1-d&0&0\\0&0&d-1&0\\0&0&0&2d-1\end{bmatrix}\begin{bmatrix}1\\0\\0\\1\end{bmatrix}=(1-d)\begin{bmatrix}1\\0\\1\\0\end{bmatrix} \quad (4.21)$$

Equation (4.18) shows that all of the co-polar light passes through the +45-degree analyzer.

–45 Degrees (Cross-Polar)

$$\frac{1}{2}\begin{bmatrix}1&0&-1&0\\0&0&0&0\\-1&0&1&0\\0&0&0&0\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&0&-1\\0&0&1&0\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1-d&0&0\\0&0&d-1&0\\0&0&0&2d-1\end{bmatrix}\begin{bmatrix}1\\0\\0\\1\end{bmatrix}=d\begin{bmatrix}1\\0\\-1\\0\end{bmatrix} \quad (4.22)$$

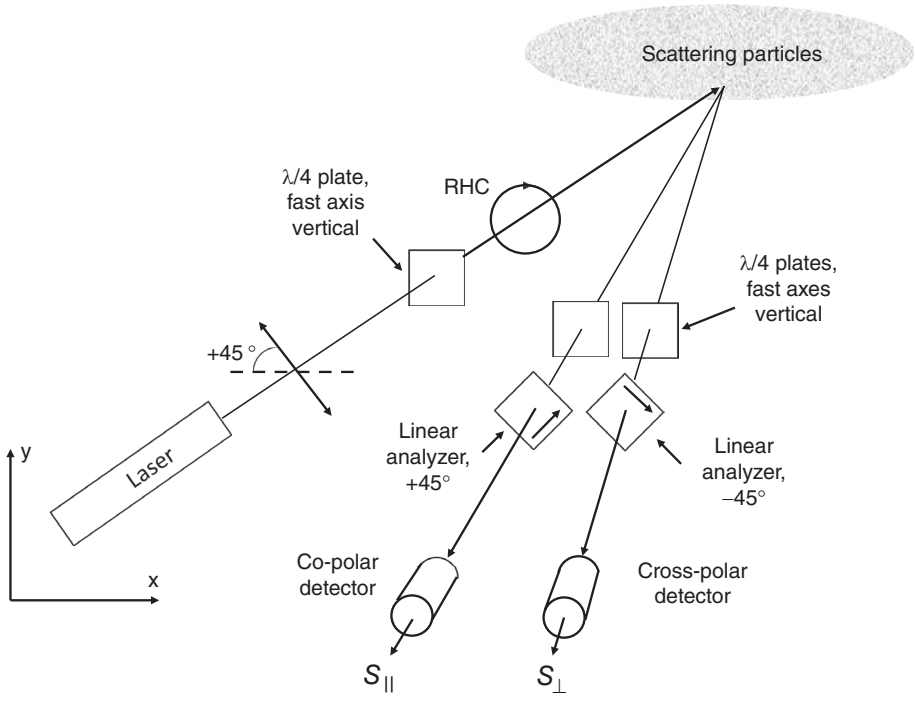


Figure 4.12 Schematic illustration of a circular lidar depolarization measurement. The transmitted polarization is linear at +45 degrees and it passes through a ¼-wave plate with the fast axis vertical to form RHC light. The received power passes through ¼-wave plates with the fast axes vertical and linear analyzers at +45 degrees and –45 degrees. The optical signal powers emerging from the analyzers are converted to electronic signals by detectors.

Equation (4.19) shows that all of the cross-polar light passes through the –45-degree analyzer.

An analysis like that shown for Eq. (4.19) shows that the parameter d is related to the signals from the two receiver channels by the equation

$$d = \frac{S_{\perp}}{S_{||} + S_{\perp}}, \tag{4.23}$$

where the subscripts ‖ and refer ⊥ to the co-polar and cross-polar receiver channels, respectively. Note that the [4,4] element in the scattering matrix becomes zero when d reaches ½, and at that value, the signals in the co-polar and cross-polar receiver channels are equal. If d becomes greater than ½, the [4,4] element becomes positive and the signal in the cross-polar channel becomes greater than the signal in the co-polar channel. In the mathematical limit where d is equal to unity, all the received signal would be in the cross-polar channel. This behavior is quite different from linear depolarization, where complete depolarization corresponds to equal signals in the two receiver channels.

4.4.3 Experimental Verification

Because the matrix given in Eq. (4.13) was not available for much of the history of lidar and depolarization by particle backscattering was not well understood, there was a persistent rumor that linear and circular depolarization lidar measurements yielded different sorts of information. The preceding analysis shows that this is not true for randomly oriented scatterers, because the value of d is the only information that can be obtained from any type of lidar depolarization measurement. The matrix given in Eq. (4.13) is a theoretical result, but the theory has been confirmed by an extensive set of measurements. The data were acquired with a lidar capable of making co-polar and cross-polar measurements with both linearly and circularly polarized light, at four wavelengths. The scattering aerosols, which were confined to a test chamber, included 25 different types of pollen and 4 types of dust (116 measurements) [22]. All the data were analyzed to retrieve the parameter d, using Eqs. (4.19) and (4.23). The results are shown in Figure 4.13 as a plot of d obtained from circular depolarization lidar measurements versus the same parameter from linear depolarization lidar measurements. Within the limits of experimental error shown by the scatter in the data points, the slope of the fitted line is consistent with unity, and the offset is consistent with zero, which shows that the information contained in the circular depolarization data is equivalent to the information in the linear depolarization data. This result is predicted by the matrix in Eq. (4.13).

The analyses in this section should be used with some caution because they only pertain to randomly oriented particles. Certain ice crystals in cirrus, hexagonal plates for example, are oriented by aerodynamic forces as they fall. They act as small mirrors, providing a large backscatter but no depolarization, even though they are crystalline. Some lidars are pointed a few degrees off the zenith to avoid artifacts in their data due to this phenomenon.

In the pre-2007 lidar literature, the degree of depolarization was always measured in terms of the *depolarization ratio* δ, which is simply

$$\delta = \frac{S_{\perp}}{S_{\parallel}}. \tag{4.24}$$

The parameter δ is not linear in the amount of depolarized light, but it has a simple definition, and it is also used in radar. For those reasons, it will most likely always be used. When the lidar transmits linear polarization and has parallel and crossed analyzers as in Figure 4.11, δ is called the *linear depolarization ratio.* The unpolarized and polarized fractions of the received light can be found from δ by the relations

$$d = \frac{2\delta}{1+\delta} \text{ and} \tag{4.25}$$

$$1-d = \frac{1-\delta}{1+\delta}. \tag{4.26}$$

The foregoing treatment of lidar depolarization measurements is idealized because the polarization states of the transmitted laser beams are assumed to be purely linear or purely circular; the polarization analyzers are assumed to be perfect; and the

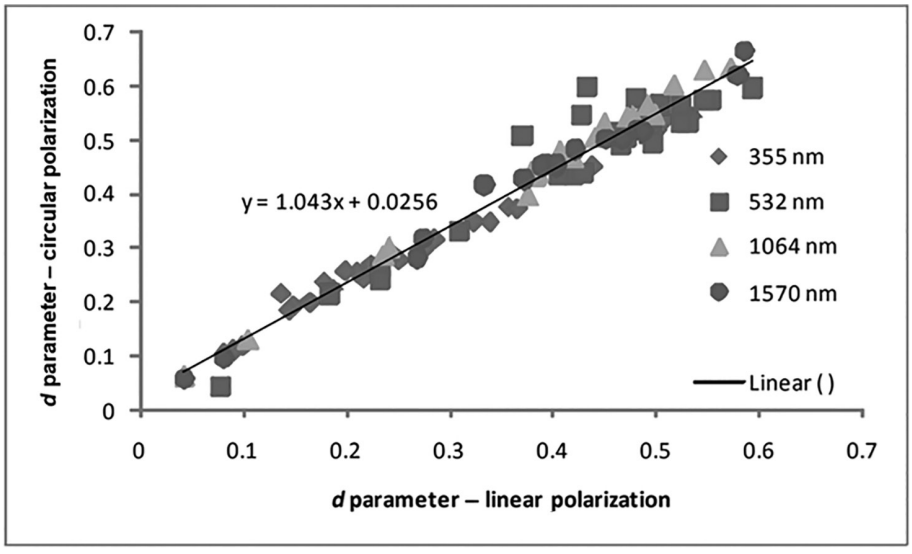


Figure 4.13 Depolarization measurements. The parameter d retrieved from linear and circular depolarization lidar measurements using 29 types of pollen and dust, at four wavelengths. Reprinted with permission from G. Roy, X. Cao, and R. Bernier, "On the information content of linear and circular depolarization signatures of bioaerosols," In Chemical, Biological, Radiological, Nuclear, and Explosives (CBRNE) Sensing XII, Proceedings of SPIE, 2011, vol. 8018 [22].

relative orientation of the transmitted polarization and the analyzers is exactly parallel or orthogonal. In addition, depolarizing affects in the transmitter and receiver optics are ignored. Polarization analyzers are described in Chapter 6, and recommended techniques for building, calibrating, and operating depolarization lidars using commercially available components are described in [23]. Those techniques enable lidar researchers to optimize the accuracy of depolarization measurements and also establish firm error bounds.

An example of depolarization lidar data is shown in Figure 4.14 as a time–height plot. The data were recorded by EARL on November 3, 2005 from 8:10 pm to 9:10 pm EST. About 22 minutes of data is plotted, when a cirrus layer at 9–11 km moved through the lidar beam. Each complete data set (2000 laser pulses parallel; 2000 crossed) required ~5 seconds for acquisition. The lidar data are plotted as range-corrected signal and as depolarization ratio δ. Both plots have 267 range bins vertically and 260 profiles horizontally (about 70k pixels). The gray scale for the signal is in arbitrary units. Upper air data from a National Weather Service balloon launched at Peachtree City at 19:00 EST showed temperatures of –40°C at 9 km and –50°C at 11 km. Supercooled water does not exist at temperatures below –40°C in the atmosphere, so the entire cloud must have been ice crystals. The upper air data also showed wind shear, with wind speeds of 12 m/s at the bottom of the layer and 16 m/s at the top.

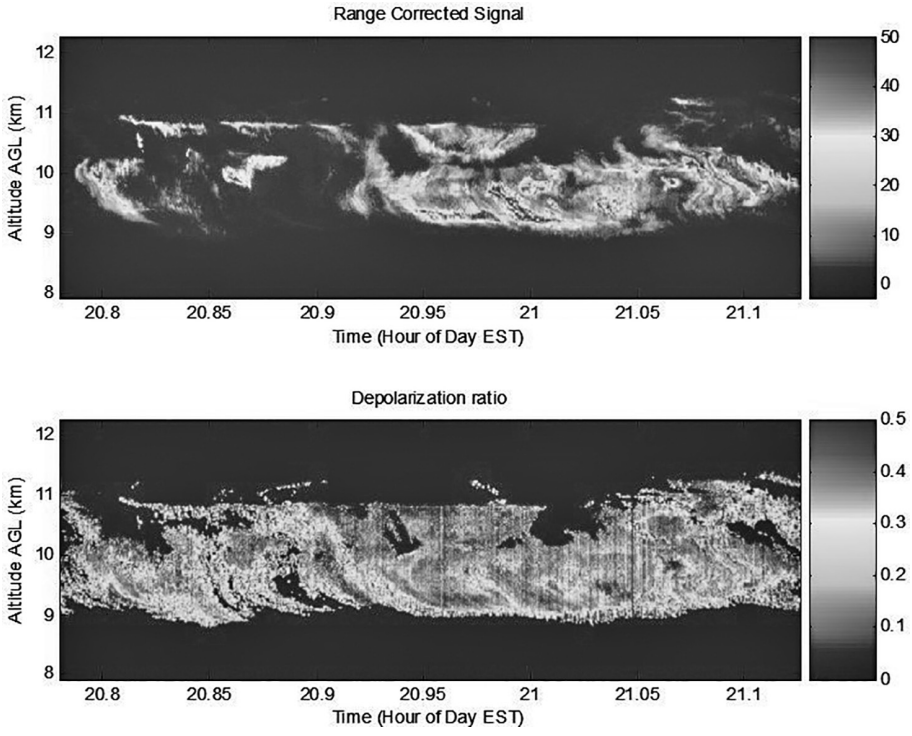


Figure 4.14 Cirrus depolarization. The data were obtained in Atlanta, Georgia by EARL on November 3, 2005.

The wind shear caused the curvature in the fall streaks, which is a common phenomenon in cirrus. Each 5-second data set corresponds to 60 m on the cloud bottom, and the total extent of cloud plotted is 15.8 km. Although the entire cloud was ice crystals, cirrus clouds tend to be inhomogeneous, as shown by their streaky appearance. In addition to variable number densities of crystals, the *crystal habit* (plates, columns, etc.) was probably varying as well as the size distribution. For these reasons, high depolarization regions do not always correspond to high backscatter regions. For example, from 20.82 to 20.92 hours, the lower part of the cloud shows very high depolarization but low backscatter. The opposite situation occurs in some regions after 21.00 hours, with high backscatter and low depolarization. This latter phenomenon may be due to hexagonal plates, which are horizontally oriented by aerodynamic forces as they fall, acting like tiny mirrors – highly reflective and maintaining polarization, as mentioned above. Oriented hexagonal plates are commonplace; they are the cause of the visual phenomena known as sun pillars.

The lower panel in Figure 4.14 shows *atmospheric* depolarization (calculated from the total signal). In the cirrus layer, where the signal is almost solely due to particulate matter, the depolarization ratio is presumably representative of the ice crystals. Atmospheric depolarization is often reported in the lidar literature, but it is

generally a bad practice because it mixes intrinsic and extrinsic properties. In an aerosol layer, for example, atmospheric depolarization depends on the extrinsic aerosol number density as well as the intrinsic tendency of the aerosols to depolarize backscattered light. The particle depolarization ratio (PDR) should always be calculated instead of atmospheric depolarization. The symbol δ_p is sometimes used instead of the acronym PDR.

4.5 Classifiers

Because first-principles models based on electromagnetic scattering theory are generally not available for atmospheric particles due to their variable size distributions, shapes, concentrations, and compositions, the lidar community has had to develop other approaches for analyzing lidar data from layers of particles. One approach, for particles that are stable and commonly used as a standard, is to make relevant laboratory measurements of their optical properties. Such measurements have been reported for Arizona road dust, for example [24]. More generally, lidar researchers have developed methods to characterize particles as they occur in the atmosphere. These *classifiers* are used to identify the types of particles in aerosol layers or in clouds based on lidar signal phenomenology. Classification schemes are based on lidar measurements that usually include δ_p because the particles of interest are all non-spherical, except for water droplets. Multiple wavelengths are often used by classifiers because measured backscatter coefficients and δ_p values at a single wavelength are rarely adequate for identifying a particle type. The previous sections of this chapter showed how scattering phenomena depend on particle size, and that depolarization in backscattered light provides information on particle shape. However, the spectral dependence of the depolarization ratio does not have an obvious relationship to particle size; it may be higher at 532 nm than at either 355 nm or 1064 nm, for example.

4.5.1 Polar Stratospheric Clouds

The studies of polar stratospheric clouds in the late 1980s provide good examples of the success of classifiers using multi-wavelength lidars with depolarization. In the extreme cold of the polar winter, several different types of PSC form at temperatures below −78°C, at altitudes in the 15–25 km range, as mentioned in Section 4.3.2. To understand the ozone hole, large campaigns were conducted that included airborne lidars along with *in situ* sensors that were both airborne and balloon-borne. One campaign included an ozone DIAL with depolarization at two wavelengths, 0.603 μm and 1.064 μm [25]. At the outset, the lidar researchers knew little about the particles they might be observing, but they identified five different types. Their reported results are summarized in Table 4.3. Airborne lidar led to classifications that were called Type 1a, 1b, 1c, Mixed, and Type 2 PSCs, which was an important step in understanding ozone destruction in arctic regions.

Table 4.3 Five types of PSC identified by lidar [25]

Type	Scattering ratio	δ_p		Composition	Size
		532 nm	1064 nm		
1a	Low	High	High	Nitric acid trihydrate (NAT)	≥1 μm
1b	Low	Low	Low	$H_2SO_4/HNO_3/H_2O$	
1c		High	Low	Small solid particles of water-rich HNO_3	A few × 0.1 μm
Mixed		Low	High	Mix of 1a & 1b	
2	High	High	High	Ice	

Type 1 clouds contain water, nitric acid, and/or sulfuric acid and are a cause of polar ozone depletion because they support chemical reactions that produce active chlorine that catalyzes ozone destruction, and because they remove gaseous nitric acid, perturbing nitrogen and chlorine cycles in a way that also increases ozone depletion. Type 2 PSCs are water ice, and they are sometimes observed as *nacreous clouds* with bright iridescent colors. The World Meteorological Organization no longer uses the alpha-numeric nomenclature shown in Table 4.3, distinguishing only between super-cooled stratiform acid-water PSCs and cirrus PSCs, but lidar nevertheless played a key role in understanding the destruction of ozone in polar winters. This example shows the power of lidar classifiers, especially considering that the researchers did not initially know what kind of particles they were observing.

The classifications were based on the results in columns 2, 3, and 4 in Table 4.3. The *scattering ratio* R_{sca} (second column) is defined as the total backscatter coefficient divided by the molecular backscatter coefficient, which can be found from measured quantities if a clear air signal is apparent above and/or below the cloud. Writing the lidar equation as $X(R) = C\beta(R)T^2(R)$ and forming the ratio in the aerosol layer,

$$R_{\text{sca}} = \frac{C(\beta_{\text{mol}} + \beta_{\text{aer}})T^2(R)}{C\beta_{\text{mol}}T^2(R)} = \frac{\beta_{\text{mol}} + \beta_{\text{aer}}}{\beta_{\text{mol}}}, \tag{4.27}$$

where C is the lidar's calibration constant. Most lidars are operated in the uncalibrated mode, so it is convenient that the calibration constant is eliminated by the ratio, along with the generally unknown transmittance to/from the layer. However, the scattering ratio not only served as a discriminant, but it also enabled the calculation of the δ_p values in the table from the measured atmospheric depolarization ratios. The process of finding scattering ratios from lidar data and the algorithm used for finding δ_p are described in [26] and in Chapter 10.

In a later study of PSCs and ozone loss, an airborne lidar had two wavelengths, 532 and 1064 nm, but depolarization capability only on the 532 nm channel [27]. That study used a 532/1064 *color ratio* (CR) as a discriminant, in which the signal at 532 nm was ratioed against the signal at 1064 nm. The authors also showed that

simultaneous curtain plots of 532 nm depolarization, 532/1064 CR, and 1064 nm signal were quite useful for identifying and classifying PSCs.

4.5.2 Aerosols

Aerosols are more varied than PSCs and they have a short lifetime, so aerosol layers are highly variable in both time and space. They usually have localized sources, they may be transported long distances by winds, and they are eventually removed by fallout and/or washout. Spaceborne lidars monitor cloud and aerosol layers globally, but their raw data provide limited information on the optical properties that climate modelers require. For this reason, lidar researchers have made considerable efforts to develop automated aerosol classifiers so that layers can at least be identified by their generic types in the large data sets from spaceborne lidars. As with PSCs, curtain plots are used to detect the presence of the layers, aerosol types can sometimes be discriminated by simply plotting one measured parameter against another, and sometimes color ratios are a powerful discriminant. Auxiliary information is often crucial for spaceborne lidar data, such as the altitude or geographical locations where an aerosol layer is observed. When developing a classifier, scatter plots are often used for identifying useful discriminants. When data points from different aerosol types occur in isolated clusters, the scatter plots can be used to establish the ranges of parameters that bound the clusters. Examples of scatter plots are schematically illustrated in Figure 4.15, which has been adapted from a study reported by Gross et al., in 2012 [28] to support future HSRLs in space (the original figure shows actual data points in color and it is much easier to interpret). The study was conducted as four missions with an HSRL on an aircraft that also carried *in situ* sensors, and the flight paths included lidar data acquisition as well as flights through the layers to measure aerosol microphysical and optical properties. Back trajectory analyses were used to identify the sources of the aerosol layers. The HSRL 532-nm data included σ, β, S_a, δ_p, and layer OD. The value of β at 1064 nm was also measured, and the color ratio CR was calculated as β_{532}/β_{1064}.

The first step in developing the classifier was to make scatter plots of one measured parameter vs. another, including CR, to see whether simple rules could be used to separate the different aerosol types. In Figure 4.15(a) for example, some of the types are in isolated areas but the aerosol from Canadian biomass burning badly overlaps anthropogenic pollution, and in (b) and (c), they are better separated from each other by CR values. The next step in automating classification is to develop a flowchart. The phenomenology shown in Figure 4.15 was implemented in the flowchart shown in Figure 4.16, where the parameter values in the decision nodes were taken from the scatter plots. In the example discussed above, Canadian biomass burning and anthropogenic pollution both have S > 30 sr and $\delta_p < 10\%$, but they are identified by their CR values, where CR < 3 means Canadian biomass burning and CR > 3 means anthropogenic pollution. The other four types are classified in similar ways. Ice and volcanic ash were added by using values taken from the literature.

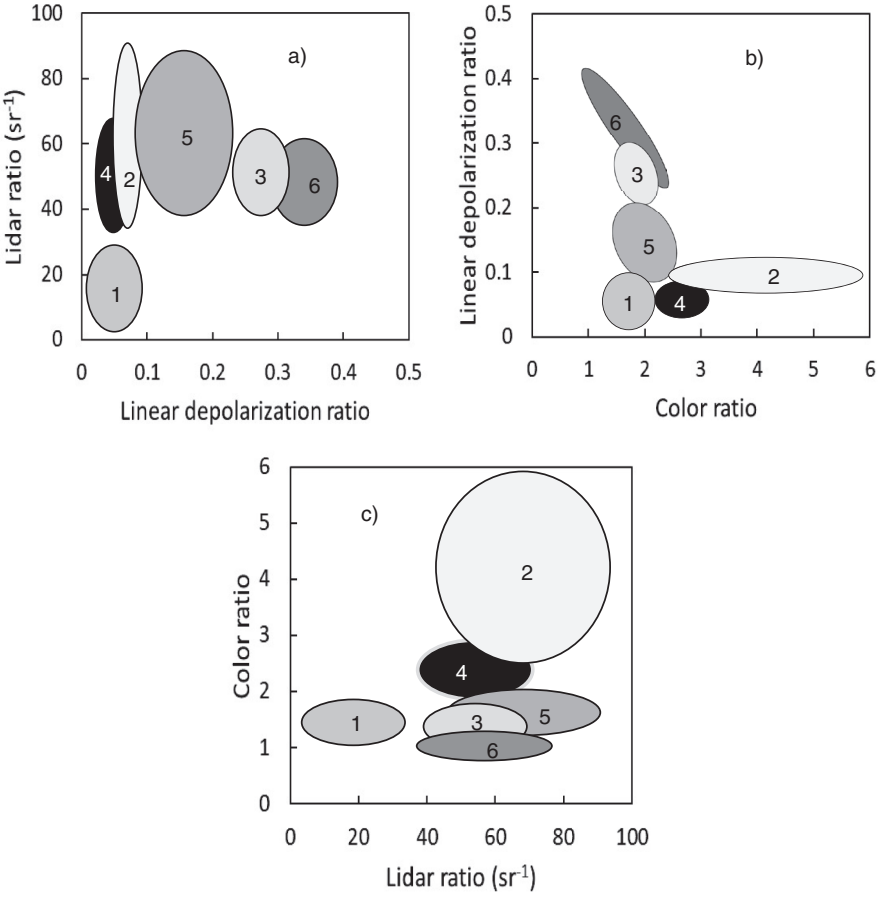


Figure 4.15 Scatter plots. In (a), the lidar ratio S_a at 532 nm is plotted vs. the depolarization ratio δ_p, in (b), δ_p is plotted vs. CR, and in (c), CR is plotted vs. S_a at 532 nm. All three plots include the same six classes of aerosols: 1 – Marine aerosol; 2 – Canadian biomass burning; 3 – Mixed Saharan dust; 4 – Anthropogenic pollution; 5 – African biomass burning mixture; 6 – Saharan dust. Adapted with permission from Figure 5 in [28].

4.5.3 Cloud-Aerosol Lidar with Orthogonal Polarization

Because of the difficulty of putting a reliable lidar in orbit, early spaceborne lidars were all the simplest type, elastic backscatter. The CALIOP lidar, launched in 2006 aboard CALIPSO (cloud-aerosol lidar and infrared pathfinder satellite), was an elastic backscatter lidar with two wavelengths, 532 nm and 1064 nm, and depolarization capability at 532 nm only. CALIOP was unusual in that it was calibrated. The 532 nm channel was calibrated by averaging the signals from high in the atmosphere (above all the aerosols) during half-orbits on Earth's dark side and comparing them to

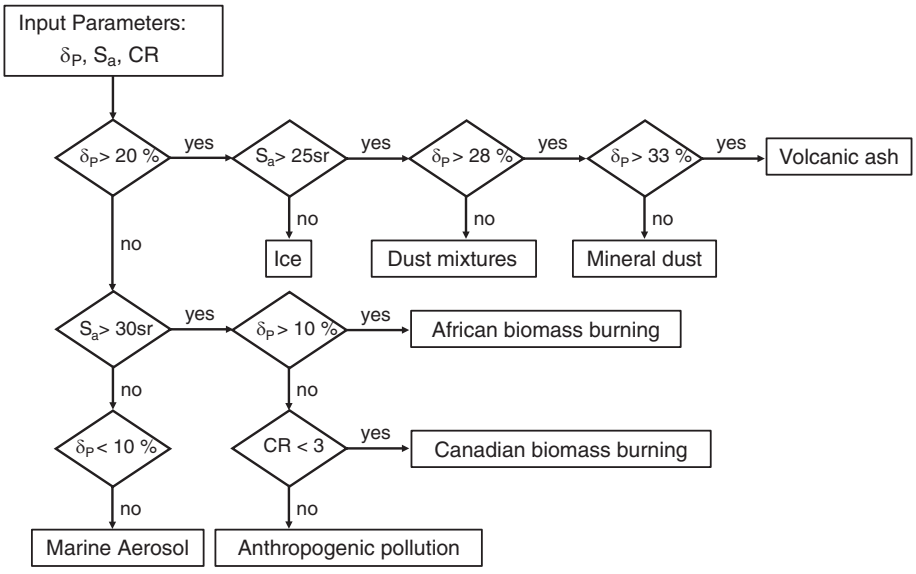


Figure 4.16 A classifier flowchart. The parameter values in the decision nodes were derived from the scatterplots shown in Figure 4.15. Adapted with permission from Figure 8 in [28].

predictions using the GMAO model for atmospheric density. The calibration was then transferred to the 1064 nm channel by using selected cirrus clouds as calibration targets. Because cirrus particles are large, scattering is in the geometric optics regime and the wavelength exponent of backscatter was therefore expected to be zero. Although CALIOP had limited measurement capabilities (calibrated backscatter at 532 nm and 1064 nm plus the depolarization ratio at 532 nm), it provided global coverage for more than a decade, and its cloud and aerosol classifiers were highly developed and refined several times during its long lifetime. Version 4, which is described here, was released in 2016 [29]. The classification proceeded in steps, starting with a curtain plot of attenuated backscatter at 532 nm on a color scale, with altitude vertically and latitude–longitude horizontally, as shown in Figure 4.17(a). Using the backscatter data, *features* are located and classified as types, as shown in Figure 4.17(b), which is known as a *feature mask*. Features are areas where the backscatter is greater than that of clear air. The data were then processed with the cloud-aerosol discriminator (CAD) at all altitudes from the surface to 30 km. The CAD algorithm uses multidimensional probability density functions (PDFs) derived from an extensive training set of CALIOP measurements to accurately distinguish clouds from aerosol layers [30]. The aerosol features were then classified by subtype, as shown in Figure 4.17(c). See [29] for the color version of Figure 4.17.

Two flowcharts are used in the CALIOP classifier, one for the troposphere (below the tropopause) and one for the stratosphere. They are like Figure 4.16 but more complicated. The inputs to the troposphere flowchart are the attenuated 532 nm backscatter integrated through the layer, the PDR estimated from 532 nm cross-polarization data using scattering ratios, the top and base altitudes of the feature, and

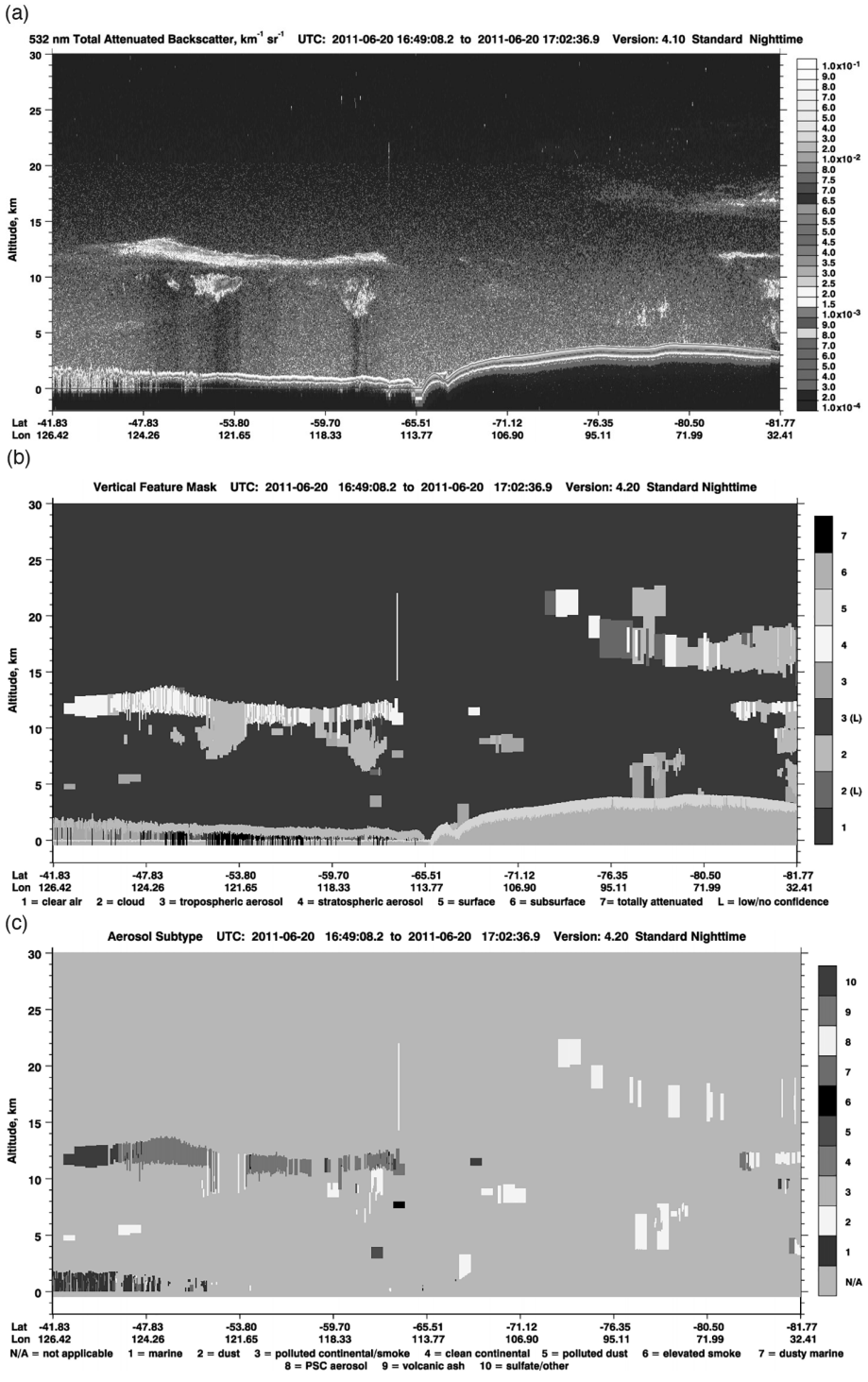


Figure 4.17 The CALIOP classifier. (a) Curtain plot of 532 nm total attenuated backscatter; (b) the resulting feature mask; (c) aerosol type classifications of the features. The date and time range are noted above the panels. This figure is Figure 7 in [29], reprinted by permission.

Table 4.4 CALIOP V4 aerosol subtypes

CALIOP class number	Aerosol subtype
1	Marine
2	Dust
3	Polluted continental/smoke
4	Clean continental
5	Polluted dust
6	Elevated smoke
7	Dusty marine
8	PSC aerosol
9	Volcanic ash
10	Sulfate/other

the surface type (land/ocean). The stratospheric flowchart again uses the integrated 532 backscatter and the PDR estimate, but also the latitude, month of the year, and the color ratio, defined as the ratio of the 1064 nm integrated backscatter to the 532 nm integrated backscatter. The result is classification into the 10 aerosol subtypes listed in Table 4.4.

4.6 Sun Photometry

Sun photometers are passive instruments, so they have lower costs than lidars. They are described here because they provide a wealth of auxiliary aerosol data that can be used to support lidar measurements. Sun photometer data include aerosol optical depth (AOD), the wavelength dependence of aerosol extinction, and even information on aerosol size distributions. They do have some drawbacks: they are ground-based instruments, and their operation and data analysis schemes assume that there is primarily one layer of aerosols (in the PBL) between them and the sun. If an elevated layer is also present, such as a dust layer in the free troposphere, its AOD will be included in the data, and information such as wavelength dependencies will not be correct for either layer. Sun photometers cannot provide AOD values at night, and clouds are a major problem for their operation. Despite these drawbacks, sun photometer data are often a very valuable resource for lidar researchers. By measuring the solar irradiance at ground level on a cloud-free line of sight, a sun photometer can provide the following data products:

(1) The aerosol optical depth, normalized to the vertical direction,
(2) the aerosol extinction spectral dependence,
(3) the column concentration of water vapor, and
(4) aerosol size and shape information (by using angle scans of a clear sky).

These auxiliary data products provide constraints for lidar data analysis; for example, the aerosol extinction coefficient profile, when integrated vertically, should match

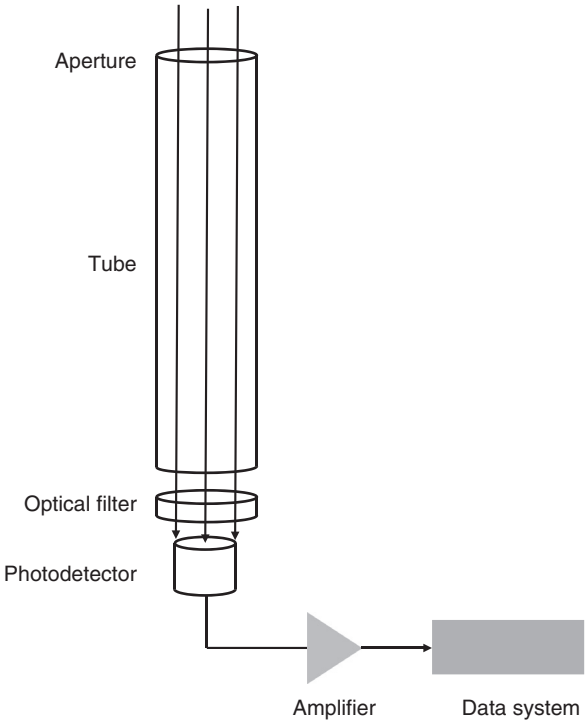


Figure 4.18 Sun photometer basic configuration. The downward vertical arrows represent rays of light from the sun.

the aerosol OD measured by the sun photometer at the same wavelength. The sun photometer's aerosol extinction spectral behavior enables extrapolation of lidar data to other wavelengths, at a much lower cost than additional lidar wavelengths. While sun photometry is simple in principle, it does have difficulties in practice. A photometer must be calibrated, and the calibration may drift with temperature and time, or with dust on its optical elements. The sun must be tracked accurately, and cloud-contaminated data must be detected and discarded. Automated data collection is needed for regular measurements.

4.6.1 Instrumentation

Sun photometers are conceptually simple, as illustrated in Figure 4.18, and they can even be homebuilt [31]. Many types are commercially available, including hand-held versions, and a commonly used configuration consists of a tube aimed at the sun, an optical filter, a photodetector and amplifier, and a data system. If the electro-optical components have a linear response, the photometer records signals that are proportional to the solar irradiance at ground level in a spectral region determined by the filter and detector.

The aperture, tube length, and detector diameter determine the instrument's field of view as shown in Figure 4.19. The FOV half-angle is given by

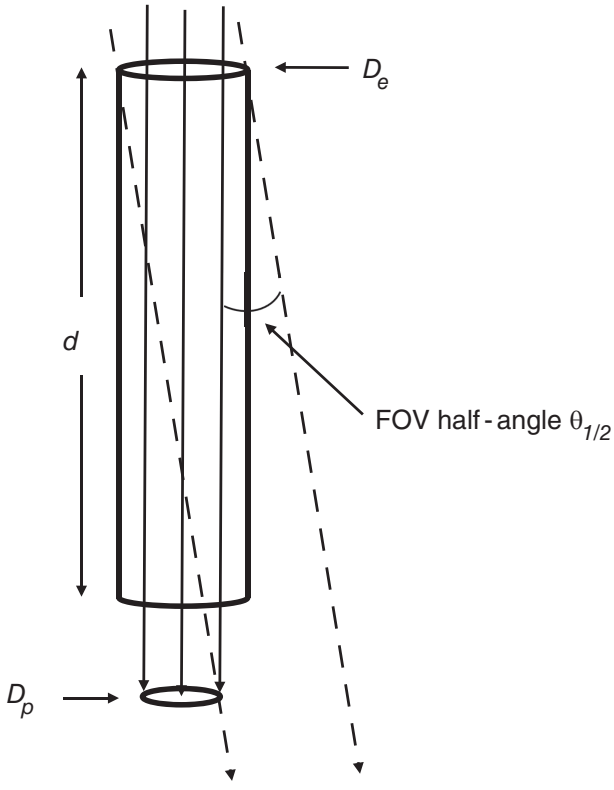


Figure 4.19 FOV geometry. The field of view of the photometer is limited by the tube length d and the diameters of the entrance aperture D_e and the photo detector D_p.

$$\theta_{1/2} = \tan^{-1}\left(\frac{D_e + D_p}{2d}\right), \tag{4.28}$$

where the variables are defined in Figure 4.19. Photometer FOV full angles are generally 1 degree or more (the angular size of the Sun is approximately one-half degree).

The Sun is a very stable light source, providing a total exo-atmospheric irradiance E_0 of 1373 ± 3 Wm^{-2} in the wavelength region from 0.18 to 4.0 μm. However, Earth's orbit is slightly elliptical, and the Earth–Sun distance R is 1 ± 0.017 Astronomical Units. The spectral irradiance at the Earth's surface is given by

$$E_e(\lambda) = \frac{E_0(\lambda)}{R^2}\exp[-m\tau(\lambda)], \tag{4.29}$$

where $E_e(\lambda)$ is the irradiance at the Earth's surface at wavelength λ, $E_0(\lambda)$ is the exo-atmospheric spectral irradiance, m is the air mass, and $\tau(\lambda)$ is the optical depth of the atmosphere at wavelength λ. The air mass m is equal to unity at the zenith and is equal to the secant of the zenith angle for angles less than about 75 degrees. At larger

zenith angles, atmospheric refraction and the curvature of Earth require a more accurate expression for air mass [32].

The solid curve in Figure 3.8 is the exo-atmospheric solar spectral irradiance, which is attenuated in several ways as the sunlight travels through the atmosphere: The molecular atmosphere scatters the sunlight in the UV-VIS-NIR region, causing the ODs in Table 3.3; some atmospheric gases have spectral absorption lines, as shown in Figure 3.15; and aerosols scatter and absorb light. A main goal of sun photometry is to characterize aerosols, and for that purpose, other sources of extinction must be removed. As the optical depth has contributions from Rayleigh scattering by molecules, molecular absorption, and aerosols, we can write

$$\tau(\lambda) = \tau_m(\lambda) + \tau_a(\lambda) + \tau_{\text{gases}}(\lambda), \tag{4.30}$$

where τ_a is the AOD and the molecular optical depth τ_m was discussed in Section 3.2. Sun photometers also derive column concentrations of water vapor by using a method similar to DIAL, but with bands of lines rather than single lines. Optical depths due to absorptions by other gases are estimated by various means.

4.6.2 Calibration

Calibration of sun photometers is accomplished by measuring the detector-amplifier output voltage at many zenith angles in very clear conditions when the atmosphere is considered to be homogeneous and static [33]. That voltage is given by

$$V_{\text{out}}(\lambda) = C(\lambda)E(\lambda)\Delta\lambda, \tag{4.31}$$

which can be inverted and combined with Eq. (4.29) to derive the relation

$$E(\lambda) = \frac{V_{\text{out}}}{C(\lambda)\Delta\lambda} = \frac{E_0(\lambda)}{R^2}\exp[-m\tau(\lambda)], \tag{4.32}$$

which can be solved for τ to yield

$$\tau(\lambda) = \frac{1}{m}\ln\left(\frac{E_0(\lambda)}{R^2}\frac{C(\lambda)\Delta\lambda}{V_{\text{out}}(\lambda)}\right). \tag{4.33}$$

The problem then is to find the calibration constant $C(\lambda)$. Accurate radiometric calibrations are notoriously difficult, but this problem can be circumvented by noting that

$$V_{\text{out}}(\lambda) = V_0\exp[-m\tau(\lambda)]. \tag{4.34}$$

Taking the natural logarithm of both sides yields

$$\ln V_{\text{out}}(\lambda) = \ln V_0(\lambda) - m\tau(\lambda), \tag{4.35}$$

and solving for the total optical depth $\tau(\lambda)$ yields the result

$$\tau(\lambda) = \frac{1}{m}[\ln V_0(\lambda) - \ln V_{\text{out}}(\lambda)], \tag{4.36}$$

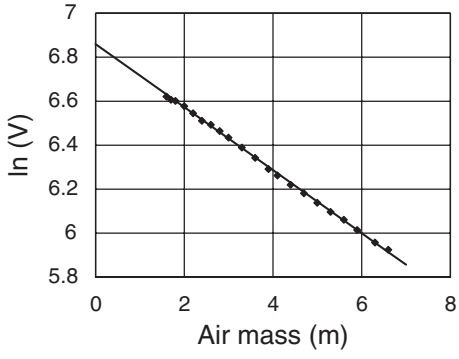


Figure 4.20 A Langley plot. The calibration plot for Microtops photometer no. 03769, described in [34].

which means that the calibration constant is not required if we can find the logarithm of the exo-atmospheric voltage $V_0(\lambda)$. The standard method of accomplishing this is known as the *Langley plot*, named for Samuel Pierpont Langley (1834–1906), who was an American astronomer and physicist. The procedure is to record the output voltage at many different air mass values as the sun travels across the sky, and then plot $\ln V_{\text{out}}$ versus air mass m, as shown in Figure 4.20. A straight line fitted to the data points is then extrapolated back to air mass zero to find $\ln V_0$, enabling the use of Eq. (4.36) for finding τ as the negative slope of the line. Assuming no gas absorptions, the aerosol optical depth is then found as

$$\tau_a(\lambda) = \tau(\lambda) - \tau_m(\lambda), \tag{4.37}$$

where the symbol τ_a is used interchangeably with AOD in this section. The Langley plot method requires accurate air mass values, so the time of the observation as well as the latitude and longitude are required to calculate the solar zenith angle. Accurate time is particularly important when m is large because it is changing quickly when the sun is near the horizon. A static atmosphere during the time of the observation is required, and this type of calibration is generally done only at specially selected sites. Accurate values of the molecular optical depth can be calculated from atmospheric density data, as described in Chapter 3. Sun photometer AOD values are normalized to the zenith, but their measurements are of course on slant paths to the Sun.

4.6.3 Aerosol Robotic Network

AERONET (aerosol robotic network) is a network of ground-based sun photometers that measure atmospheric aerosol properties [35]. The measurement system is a Cimel CE-318 spectral radiometer, shown in Figure 4.21, that measures sun and sky radiances at eight or more wavelengths in the UV-VIS-NIR range. AERONET provides continuous observations of spectral AOD, precipitable water, and inversion

Figure 4.21 An AERONET instrument. Graduate students are shown installing a Cimel CE-318 sun photometer on the roof of the Baker Research Building at Georgia Tech in 2011.

aerosol products, which include aerosol volume size distribution, complex refractive index, single scattering albedo, and the aerosol scattering phase function. AERONET was originally developed by NASA under the leadership of Dr. Brent Holben, and by 2020, it had expanded to more than 1500 stations in a federated global network with national agencies, institutes, universities, individual scientists, and partners as collaborators. The network imposes standardization of instruments, calibration, processing, and distribution. Data are relayed regularly via geostationary communication satellites, making it possible to continuously monitor atmospheric aerosols with this ground network over approximately 90% of the Earth's surface. The resultant database is completely open access and users can directly download data from the AERONET network with a user-friendly graphical user interface [36]. It requires only typing in a site name, selecting a data type (AOD, for example), and selecting a date from the calendar. Aerosol optical depth data are computed for three data quality levels: Level 1.0 (unscreened), Level 1.5 (cloud-screened), and Level 2.0 (cloud-screened and quality-assured).

Examples of several standard AERONET data products are shown in the figures that follow. All the data were acquired at the Georgia Tech site on September 23, 2014. This day was chosen because it was typical in being partly cloudy. *Cloud screening* means deleting data during periods when clouds are present, and it is the reason for the data gaps in the figures. The first example, Figure 4.22, is a plot of AOD vs. time for all eight wavelengths of the Georgia Tech instrument (1640 nm was added because the GTRI lidar team has operated several lidars in the 1500–1600 nm wavelength range). Data gaps are evident during two cloudy periods. AOD always increased monotonically as wavelength decreased, but the spread of AOD values grew over time. Figures 4.22, 4.24, 4.25, and 4.27 were plotted with downloaded data, but the AERONET web site generates such figures automatically, with slightly different formatting.

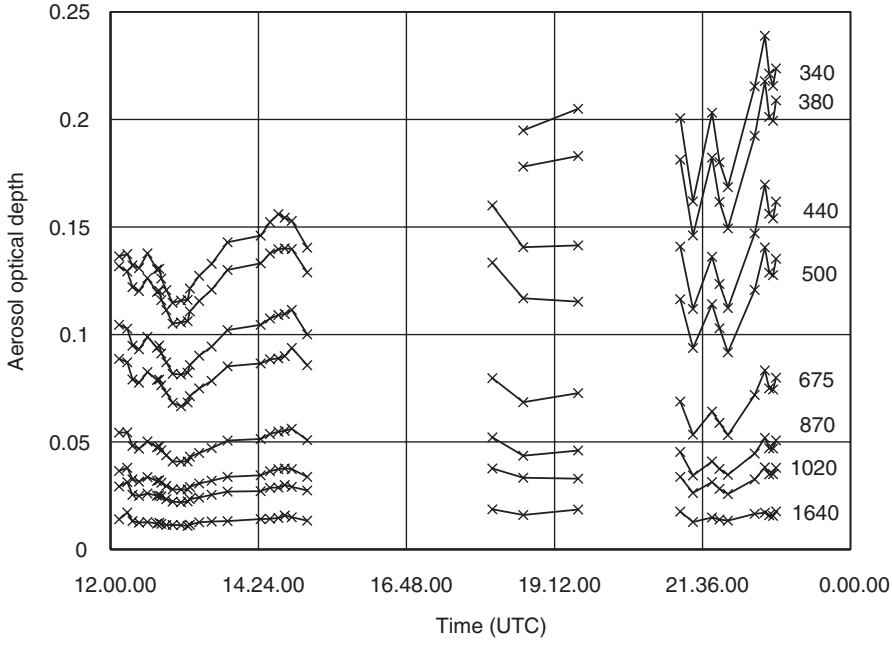


Figure 4.22 AERONET AODs versus time. The data were acquired at the Georgia Tech site on September 23, 2014.

The wavelength dependence of aerosol AOD is characterized by the *Angstrom parameter* (also called the Angstrom exponent), named for Anders Jonas Ångström (1814–1874) who was a Swedish physicist and one of the founders of the science of spectroscopy. The Angstrom parameter is a one-parameter fit to multi-wavelength AOD data. If we write

$$\left(\frac{\lambda_2}{\lambda_1}\right)^{-a} = \left(\frac{\tau(\lambda_2)}{\tau(\lambda_1)}\right), \tag{4.38}$$

then solving for the Angstrom parameter a gives the solution

$$a = -\frac{\ln\tau(\lambda_2) - \ln\tau(\lambda_1)}{\ln\lambda_2 - \ln\lambda_1}. \tag{4.39}$$

The parameter a is therefore the negative of the slope of a line fitted through data points on a plot of $\ln\tau$ vs. $\ln\lambda$. Such a plot is shown in Figure 4.23, for AOD values at a time toward the end of Figure 4.22. The slope of the fitted line is –1.66, so the value of the Angstrom parameter a is 1.66. As shown in Table 4.2, Rayleigh scattering has a wavelength exponent of four, geometric optics zero, and Mie scattering somewhere in between. A steep slope in a plot such as Figure 4.23 indicates a large value of a, meaning that the aerosols are small compared to λ, whereas a shallow slope means a smaller a and larger particles, so the value of a gives an indication of aerosol size. The

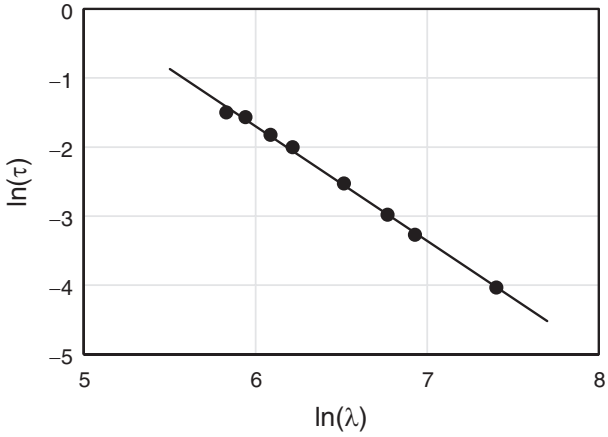


Figure 4.23 The Angstrom parameter plot. The Angstrom parameter a is the negative of the slope of a line fitted through data points, which correspond to a time near the end of the AOD series shown in Figure 4.22.

Angstrom parameter is also useful for interpolating aerosol behavior to other wavelengths. Other optical properties, such as the backscatter coefficient, are sometimes characterized in a similar way with a *wavelength exponent*, but the term Angstrom parameter strictly applies only to aerosol optical depth. The AERONET Angstrom parameter data product is shown in Figure 4.24.

The AERONET sun photometers also measure precipitable water, by means of a pair of optical filters that correspond to the "on" and "off" lines in DIAL, but for a band of absorption lines rather than an individual line. The term *precipitable water* is defined as follows: Consider a vertical column of constant area through the entire atmosphere. If all the water vapor in that column were condensed into liquid water at the bottom, the amount of precipitable water would be the thickness of the water. Precipitable water is usually expressed in cm; an example is shown in Figure 4.25.

As pointed out in Section 3.1.2, both the wavelength exponent of scattering and the phase function depend on the scattering parameter α. This fact suggests that sun photometry can provide aerosol size information, especially if the phase function can be measured. This can be accomplished with an *almucantar scan*, which is a scan in azimuth angle at the elevation of the sun (the almucantar is a circle on the celestial sphere parallel to the horizontal plane), shown in Figure 4.26. AERONET photometers perform these scans periodically to measure the radiance of sunlight scattered by aerosols at a range of scattering angles from near zero to 180 degrees. Air mass values less than 2 are generally used in such scans to minimize the chance of encountering clouds. A *principal plane scan* is also used (upward through the sun and across the zenith). Data from almucantar and principal plane scans, along with the measured AODs, are then used to develop an inversion data product, the volume size distribution $dV(r) / \ln(r)$.

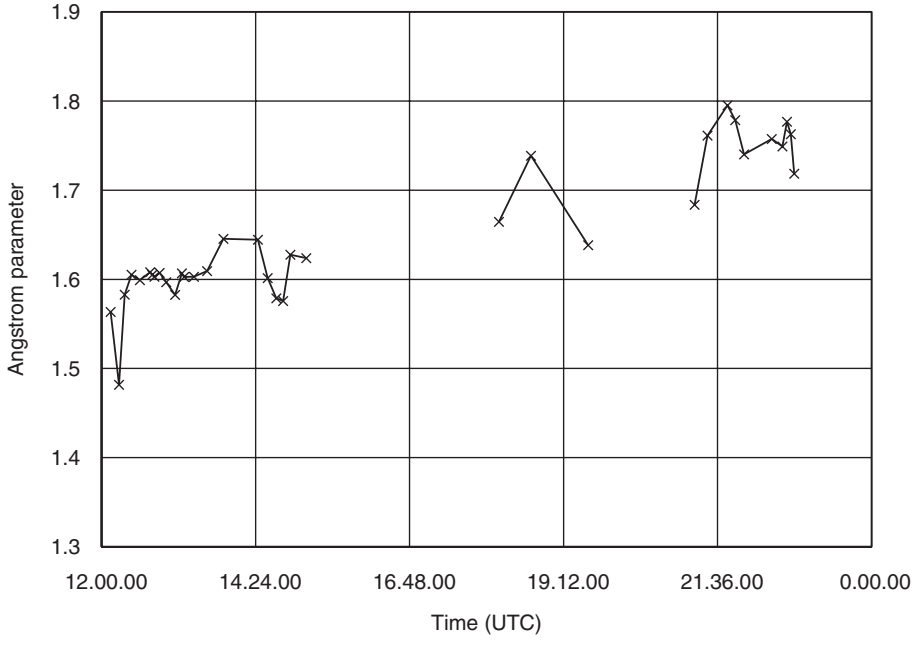


Figure 4.24 AERONET Angstrom parameter vs. time. The data were acquired at the Georgia Tech site on September 23, 2014.

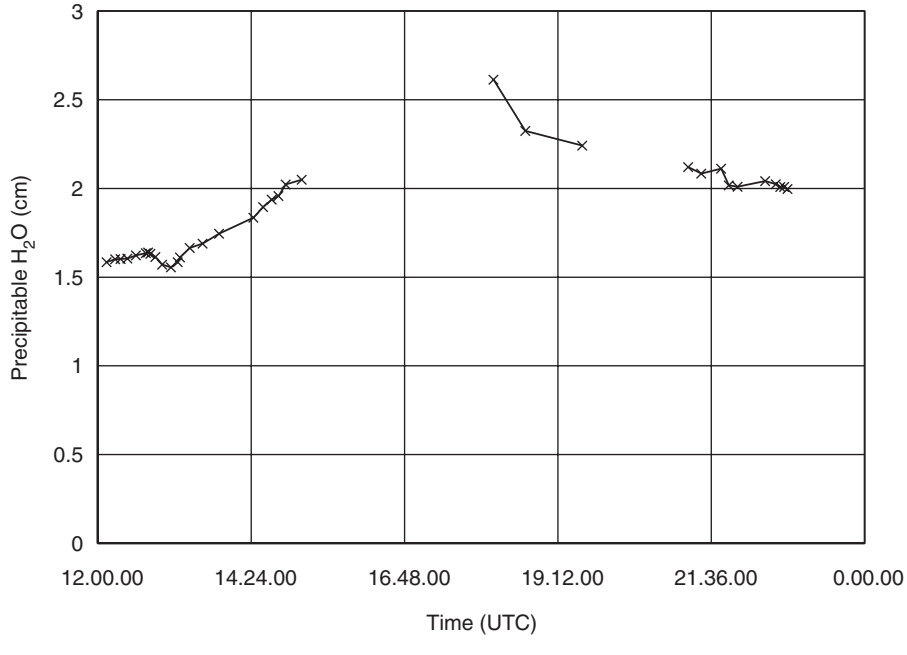


Figure 4.25 AERONET precipitable water vs. time. The data were acquired at the Georgia Tech site on September 23, 2014.

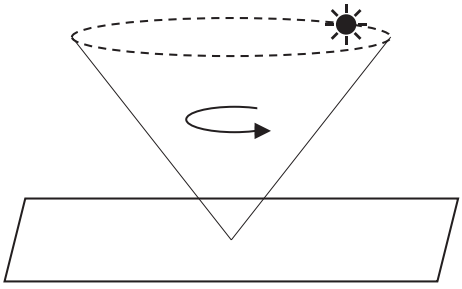


Figure 4.26 The almucantar scan.

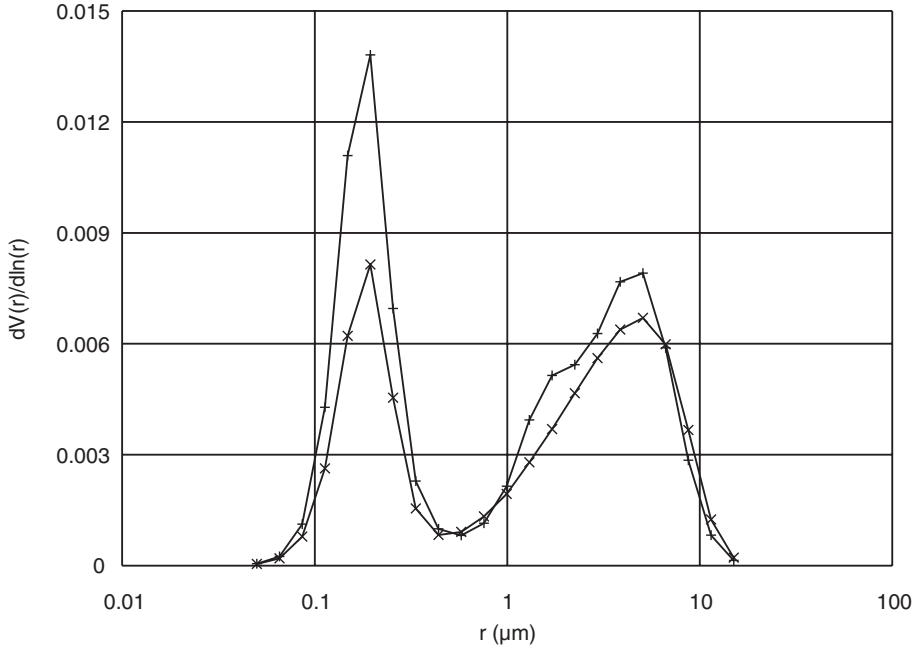


Figure 4.27 AERONET aerosol volume distribution. The data were acquired at the Georgia Tech site on September 23, 2014. Data points marked with the symbol x were recorded at 13:01 UTC; data points marked with + were at 14:27 UTC.

An example of this data product is shown in Figure 4.27 for two different times. The bimodal distribution resembles the volume distribution shown in Figure 4.7. The fine mode is more populated at the later time, which is consistent with the increase in Angstrom parameter over time shown in Figure 4.24.

High spectral resolution lidar and aerosol Raman lidar are well-established techniques for measuring σ and β in aerosol layers as mentioned in Chapter 3, but only at the specific UV-VIS wavelengths where they operate. AERONET provides a

wealth of auxiliary information for interpreting and extending aerosol lidar signals. The Angstrom parameter provided by sun photometer data can be used to extend results to other wavelengths, and sun photometer AODs can be used to constrain inversions of aerosol lidar data, as explained in Chapter 10. An extensive set of AERONET data was used to develop a table of S_a values that are used in CALIOP data analysis, and AERONET data were also used in CALIOP calibration/validation experiments.

At NIR wavelengths and longer, aerosol backscatter provides the signal for GHG DIAL lidars and other eye-safe lidars including Doppler and eddy correlation wind sounders (see Chapter 10), as shown in Figure 3.24. For those applications, a detailed knowledge of the aerosols is not required. However, remote sensing data from clouds and aerosols is hugely important for understanding global climate change as well as urban air quality, and first-principles models of their optical properties are not available except for water drops, which are spherical, and for spheroids that may approximate some aerosols. Atmospheric particles have widely varying sizes, shapes, and chemical compositions, they span all three scattering regimes, and they are highly variable in both time and space, so the values and wavelength dependencies of the lidar parameters σ, β, S_a, and δ are all variable and generally unknown. Despite these difficulties, the lidar community has developed methodologies for classifying clouds and aerosol layers with elastic backscatter lidars, such as CALIPSO, using depolarization and multiple wavelengths. Sun photometry provides auxiliary data for inverting and extending ground-based lidar data.

4.7 Further Reading

H. C. van de Hulst, *Light Scattering by Small Particles*. New York: Dover, 1981.

This classic but outdated book (first published in 1957) covers both basic scattering theory and some computations with different kinds of particles. It includes a full range of useful approximation methods because it predates the widespread availability of electronic computers.

Deirmendjian's book [3] includes models of drop size distributions in clouds that are still in use today, so it is often referenced.

C. F. Bohren and D. R. Huffman. *Absorption and Scattering of Light by Small Particles*. New York: Wiley-Interscience, 1983.

This is a more modern and rigorous book, based on electromagnetic theory. It covers absorption and scattering by arbitrary particles as well as spheres in the three scattering regimes, including their angular dependencies. It addresses many applications, and it includes a listing of a Mie scattering computer program.

A. Ansmann and D. Müller, "Lidar and atmospheric aerosol particles," in *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere*, C. Weitcamp, Ed. New York: Springer, 2005, pp. 105–141.

This chapter on aerosols is specific to lidar, and it is a good resource for lidar researchers. In addition to information on aerosol microphysical and optical properties, it also contains derivations of inversion algorithms for several types of lidar.

The book edited by Mishchenko, Hovenier, and Travis [5] presented the first systematic and unified discussion of light scattering by non-spherical particles and its practical applications in remote sensing, geophysics, astrophysics, biomedical optics, and optical engineering. It was an instant classic and the first printing quickly sold out, but it is available in pdf format on the Internet at a very reasonable cost.

Stull's Introduction to Boundary Layer Meteorology [6] is another classic text. It is comprehensive and accessible, and it has become a standard reference on boundary layer processes.

4.8 Problems

4.8.1 Equation 4.2 gives a value for the scattering cross section $\sigma_{\text{mol}}(\phi)$ of a single air molecule. Show that it is consistent with the volume angular scattering coefficient for air given in Eq. 3.9.

4.8.2 A radiation fog forms when the air temperature drops below the dew point and water vapor rapidly condenses on haze particles that were already present. The number density does not change appreciably, but the optical properties do. Consider scattering from water droplets with the radii shown in Figure 4.4 at the wavelength 0.532 μm. The refractive index of water is 1.33 at VIS wavelengths. Use an online calculator such as the Oregon Medical Laser Center, Mie Scattering Calculator at https://omlc.org/calc/mie_calc.html. to find the extinction and backscatter cross sections. Discuss their changes as functions of droplet radius.

4.8.3 The quotient $(1-\delta)/(1+\delta)$ appears occasionally in the pre-2007 linear depolarization lidar literature with no explanation of what it represents. Show that it is the polarized fraction of the received lidar signal.

4.8.4 At 22:47:38 UTC (the last data points) on September 23, 2014, the AERONET Georgia_Tech station reported the data in the following table. What value of the Angstrom parameter describes the wavelength dependence in this table? Consider the data from 380 to 1020 nm only. Is the Angstrom parameter consistent with Figure 4.24?

λ (nm)	AOD
340	0.2237
380	0.2089
440	0.1618
500	0.1352
675	0.0799
870	0.0508
1020	0.0380
1640	0.0178

References

[1] R. M. Measures, *Laser Remote Sensing: Fundamentals and Applications*. New York: Wiley-Interscience, 1984.

[2] E. J. McCartney, *Optics of the Atmosphere*. New York: Wiley, 1976.

[3] D. Deirmendjian, *Electromagnetic Scattering on Spherical Polydispersions*. New York: American Elsevier, 1969.

[4] S. Prahl, Oregon Medical Laser Center, Mie Scattering Calculator. [Online]. Available: https://omlc.org/calc/mie_calc.html. [Accessed 20 April 2021].

[5] M. I. Mishchenko, J. W. Hovenier, and L. D. Travis, Eds., *Light Scattering by Nonspherical Particles: Theory, Measurements, and Geophysical Applications*. San Diego: Academic Press, 2000.

[6] R. B. Stull, *An Introduction to Boundary Layer Meteorology*. Netherlands: Springer, 1988.

[7] Creative Commons. [Online]. Available: http://CreativeCommons.org. [Accessed 27 September 2021].

[8] C. E. Junge and J. E. Manson, "Stratospheric aerosol studies," *Journal of Geophysical Research*, vol. 66, pp. 2163–2182, 1961.

[9] H. Jäger and D. Hofmann, "Midlatitude lidar backscatter to mass, area, and extinction conversion model based on in situ aerosol measurements from 1980 to 1987," *Applied Optics*, vol. 30, pp. 127–138, 1991.

[10] H. Jäger, "Long-term record of lidar observations of the stratospheric aerosol layer at Garmisch-Partenkirchen," *Journal of Geophysical Research*, vol. 110, D08106 (9 pp.), 2005.

[11] H. N. Forrister, D. W. Roberts, A. J. Mercer, and G. G. Gimmestad, "Infrared lidar measurements of stratospheric aerosols," *Applied Optics*, vol. 53, D40–D48, 2014.

[12] I. Veselovskii, A. Kolgotin, V. Griaznov et al., "Inversion of multiwavelength Raman lidar data for retrieval of bimodal aerosol size distribution," *Applied Optics*, vol. 43, pp. 1180–1195, 2004.

[13] H. Koschmieder, "Measurements of Visibility at Danzig," *Monthly Weather Review,* November 1930, pp. 439–444, 1930.

[14] G. Roy, L. C. Bissonnette, C. Bastille, and G. Valée, "Estimation of cloud droplet size density distribution from multiple-field-of-view lidar returns," *Optical Engineering*, vol. 36, pp. 3404–3415, 1997.

[15] G. M. Hansen, M. Serwazi, and U. von Zahn, "First detection of a noctilucent cloud by lidar," *Geophysical Research Letters*, vol. 16, pp. 1445–1448, 1989.
[16] B. Schäfer, G. Baumgarten and J. Fiedler, "Small-scale structures in noctilucent clouds observed by lidar," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 208, 105384 (8 pp.), 2020.
[17] E. Hecht, *Optics*. Reading: Addison-Wesley, 1990.
[18] E. Collett, *Field Guide to Polarization*, J. E. Grievenkamp, Series Ed., Volume FG05. Bellingham, WA: SPIE Press, 2005.
[19] M. I. Mishchenko and J. W. Hovenier, "Depolarization of light backscattered by randomly oriented nonspherical particles," *Optics Letters*, vol. 20, pp. 1356–1358**,** 1995.
[20] C. J. Flynn, A. Mendoze, Y. Zheng, and S. Mathur, "Novel polarization-sensitive micropulse lidar measurement technique," *Optics Express,* vol. 15, pp. 2785–2790, 2007.
[21] G. G. Gimmestad, "Reexamination of depolarization in lidar measurements," *Applied Optics*, vol. 47, pp. 3795–3802, 2008.
[22] G. Roy, X. Cao, and R. Bernier, "On the information content of linear and circular depolarization signatures of bioaerosols," In Chemical, Biological, Radiological, Nuclear, and Explosives (CBRNE) Sensing XII, Proceedings of SPIE*,* 2011, vol. 8018, 8 pp.
[23] V. Freudenthaler, "About the effects of polarising optics on lidar signals and the Δ90-calibration," *Atmospheric Measurement Techniques*, vol. 9, pp. 4181–4255, 2016.
[24] A. Miffre, T. Mehri, M. Francis and P. Rairoux, "UV-VIS depolarization from Arizona Test Dust particles at exact backscattering angle," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 169, pp. 79–90, 2016.
[25] O. B. Toon, A. Tabazadeh, E. V. Browell, and J. Jordan, "Analysis of lidar observations of Arctic polar stratospheric clouds during January 1989," *Journal of Geophysical Research Atmospheres*, vol. 105, pp. 20589–20616, 2000.
[26] E. V. Browell, C. F. Butler, S. Ismail et al., "Airborne lidar observations in the wintertime arctic stratosphere: polar stratospheric clouds," *Geophysical Research Letters*, vol. 17, March Supplement, pp. 385–388, 1990.
[27] M. Felton and A. H. Omar, "Polar Stratospheric Cloud (PSC) classification using LIDAR measurements from the recent SAGE III Ozone Loss and Validation Experiment (SOLVE)," In Proceedings of the International Geoscience and Remote Sensing Symposium, 2002, pp. 2407–2410.
[28] S. Groß, M. Esselborn, B. Weinzierl et al., "Aerosol classification by airborne high spectral resolution lidar observations," *Atmospheric Chemistry and Physics*, vol. 13, pp. 2487–2505, 2013.
[29] M. H. Kim, A. H. Omar, J. L. Tackett et al., "The CALIPSO version 4 automated aerosol classification and lidar ratio selection algorithm," *Atmospheric Measurement Techniques*, vol. 11, pp. 6107–6135, 2018.
[30] Z. Liu, J. Kar, S. Zeng et al., "Discriminating between clouds and aerosols in the CALIOP version 4.1 data products," *Atmospheric Measurement Techniques*, vol. 12, pp. 703–734, 2019.
[31] F. M. Mims, "Solar radiometer with light-emitting diodes as spectrally-selective detectors," *Applied Optics*, vol. 39, pp. 6517–6518, 2000.
[32] A. T. Young, "Air mass and refraction," *Applied Optics*, vol. 33, pp. 1108–1110, 1994.

[33] G. E. Shaw, "Sun photometry," *Bulletin of the American Meteorological Society*, vol. 64, pp. 4–10, 1983.
[34] M. Morys, F. M. Mims, S. Hagerup et al., "Design, calibration, and performance of MICROTOPS II handheld ozone monitor and Sun photometer," *Journal of Geophysical Research*, vol. 106, pp. 14573–14582, 2001.
[35] B. N. Holben, T. F. Eck, I. Slutsker et al., "AERONET – A federated instrument network and data archive for aerosol characterization," *Remote Sensing of Environment*, vol. 66, pp. 1–16, 1998.
[36] NASA Goddard Space Flight Center, AERONET Data Synergy Tool. [Online]. Available: http://aeronet.gsfc.nasa.gov/cgi-bin/bamgomas_interactive. [Accessed: 20 April 2021].

5 Lidar Transmitters

The previous two chapters dealt with the atmospheric parameters in the lidar equation as well as depolarization, which is measurable with lidar, and this chapter and the next are mainly concerned with the instrumental parameters. The most common atmospheric lidar transmitter–receiver configurations, illustrated in Figure 5.1, are known as biaxial, coaxial, and common optics. Of these three, the most common configuration is biaxial, in which the transmitter and receiver are physically separated with some offset distance between their optical axes. In the coaxial configuration, the optical axes lie on top of each other, but the light paths are separated. In the common optics configuration, the transmitter and receiver share the same telescope. The biaxial configuration is sometimes incorrectly referred to as *bistatic*, but this term properly refers to having the transmitter and receiver widely separated, so that the received light is scattered at some angle less than 180 degrees. Bistatic lidars have been developed and operated [1], but they are not in common use and so they are not considered here. Engineering considerations for lidar transmitters are discussed in this chapter, and receivers are discussed in Chapter 6. An analysis of the transmitter–receiver combination and the crossover function $G(R)$ is also presented in Chapter 6, and that analysis applies to all three of these configurations.

5.1 Transmitter Components

The optical requirements for the lidar transmitter are that it must transmit a laser beam with the desired divergence (always less than the receiver FOV) and maintain optical alignment of that beam with the receiver. The transmitter often provides the means of aligning the lidar system. The transmitter may also be required to monitor the energy of every transmitted pulse or the average transmitted power (a good practice). The laser eye hazard is another important consideration, and the transmitter may be required to provide eye safety at the transmitting aperture, or to be eye safe beyond some known range. The main components of a lidar transmitter are shown in Figure 5.2. A "dog leg" configuration is shown, with two 90-degree turning mirrors. Having at least one turning mirror is often desirable because an adjustable turning mirror provides a convenient place to align the transmitted beam direction with the receiver FOV. The components shown in Figure 5.2 are discussed in the following paragraphs in the order in which they are encountered by the laser light as it passes through the transmitter.

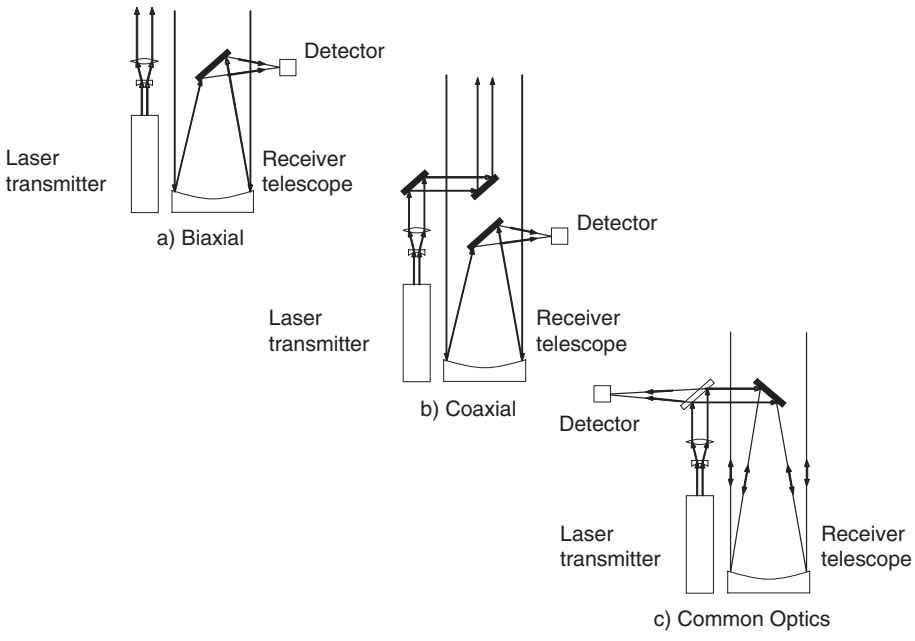


Figure 5.1 Lidar configurations. The three most common lidar transmitter–receiver configurations are shown: biaxial, coaxial, and common optics.

A first step in designing the lidar transmitter optics is to describe the "raw" laser beam, meaning the beam as it emerges from the laser. The raw beam can be described by the set of parameters listed in Table 5.1. If a monitor for transmitted pulse energy or average power is required, a partially reflecting element known as a beamsplitter or beam "pickoff" may be inserted to direct a small amount of the beam toward a power meter. A flat piece of glass with an AR coating on the second surface (optimized for the laser wavelength and a 45-degree incidence angle) makes a convenient pickoff. Note that the amount of reflected light that is picked off depends strongly on the laser beam polarization relative to the plane of incidence. For visible light, such a pickoff made of BK-7 glass will transmit 99% of the laser beam and send 1% to the power meter if the beam polarization is perpendicular to the plane of incidence. This technique was employed in EARL. A power meter was required because EARL's laser power drifted with time, and time–height plots such as Figures 4.6 and 4.16, which display received signal on a color scale or gray scale, are distorted if the power changes significantly during the data acquisition time, unless the signals are corrected for transmitted power.

Commercial laser power meters, also known as "joule meters" because they can measure energy per pulse, are widely available, although not all of them are easily interfaced to a lidar data system. The original commercial power meter in EARL had an RS-232 serial interface. It was interrogated after each laser pulse, and it

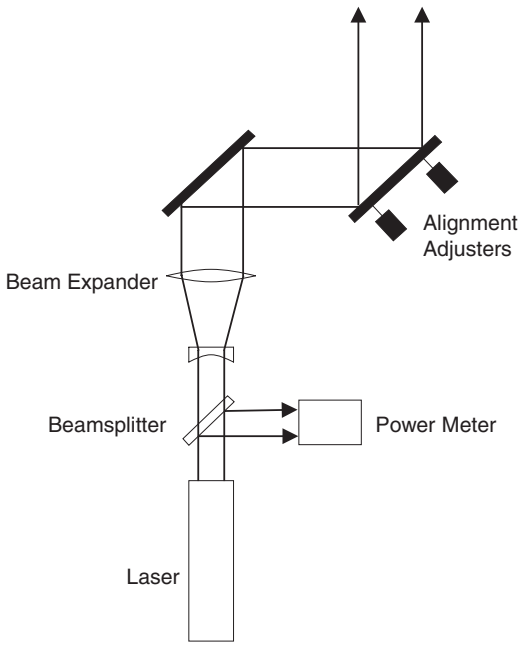


Figure 5.2 Lidar transmitter components. Typical components include the laser, a beamsplitter that sends a small fraction of the beam to a power meter, a beam expander, turning mirrors, and alignment adjusters.

occasionally failed to respond, causing data acquisition to stop. To solve this problem, GTRI developed a homebuilt power meter based on a photodetector–amplifier combination that was available as one package. EARL's power meter is described in more detail in Chapter 8.

The laser beam is often expanded to a larger diameter by a Galilean beam expander, illustrated in Figure 5.4, where the names of the two lenses, "eyepiece" and "objective," are borrowed from telescope terminology. When the two lenses are separated by the sum of their focal lengths, the expansion ratio D_o/D_i is equal to the ratio of focal lengths f_o/f_e (the focal length of the negative lens is taken as a positive number here). Many a lidar has been operated using the raw beam from a laser, but beam divergence and pointing instability both decrease inversely with the expansion ratio, so beam expansion may be useful for keeping the scattering volume within the receiver's FOV. Another reason to expand the beam is to meet eye safety criteria (see Section 5.3). Laser beam expanders can be purchased as assembled units or they can be built up from individual lenses. In either case, it is important to make the lens diameters larger than the beam diameters, as shown in Figure 5.3, to avoid truncating the edges of the laser beam and causing diffraction. The eyepiece may alternatively be a positive lens (a configuration called the Keplerian beam expander), but then the laser beam comes to a focus inside the beam expander, with a risk of causing air breakdown and

Table 5.1 Raw laser beam parameters

Parameter	Typical units
Wavelength	nm or μm
Energy per pulse	mJ or μJ
Pulse repetition frequency	Hz
Beam diameter	mm
Beam divergence	mrad
Beam quality M^2	(dimensionless)
Energy stability	% r.m.s.
Pointing stability	μrad r.m.s.
Polarization purity	%

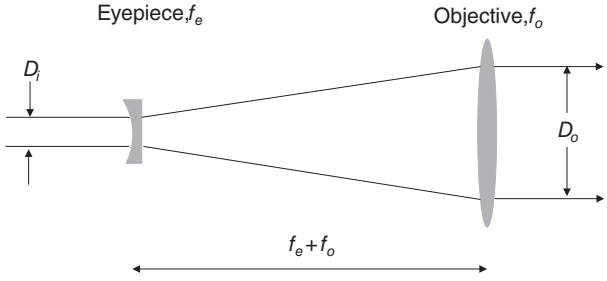


Figure 5.3 The Galilean laser beam expander. The expander consists of a negative lens and a larger positive lens separated by the sum of their focal lengths.

subsequent beam degradation. Negative lenses are usually used in laser beam expanders to avoid the air breakdown problem. When extreme beam expansion is required, such as in an eye-safe visible-light lidar, the objective lens in Figure 5.3 is often replaced with a Cassegrainian telescope. This type of objective can lead to a large loss of beam power, because laser beams usually have their highest intensity in the center, where a Cassegrainian has an obstruction (the secondary mirror). To avoid this power loss, the eyepiece can be replaced with a pair of axicons that produce an annular beam as illustrated in Figure 5.4. Axicons are cylinders of glass with one planar face and one conical face. By choosing the cone angles and the axicon separation, the emerging beam can be engineered to match the telescope, thereby transmitting the entire beam without a major power loss [2].

An off-axis paraboloid (OAP) mirror offers another way to expand a laser beam to a large diameter without a central obstruction, but large OAPs are expensive and they are sensitive to misalignment. However, beam divergences in lidar systems are small, on the order of one milliradian or less, and this fact enables the replacement of the paraboloidal mirror with a spherical mirror. This approach is inexpensive because an off-axis sphere is the same as an on-axis sphere, and the spherical figure is the simplest to fabricate. The expander, illustrated in Figure 5.5, has no central obstruction.

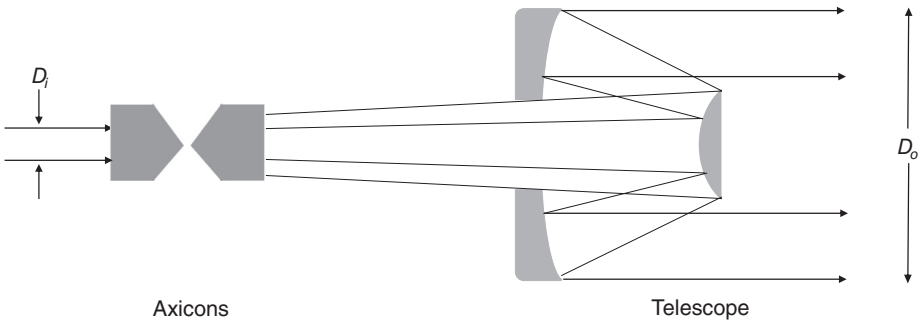


Figure 5.4 Transmitter with axicons. A pair of axicons can be used to transmit an annular laser beam through a telescope with a central obstruction, such as a Cassegrainian.

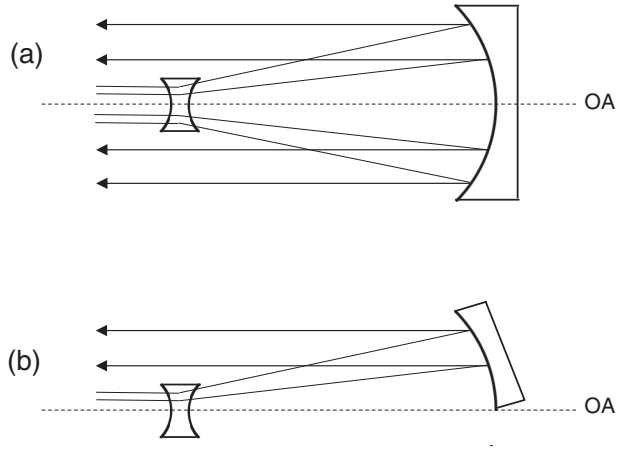


Figure 5.5 The sphere beam expander. (a) The axially symmetric expander with a central obstruction; (b) the corresponding off-axis sphere expander.

A small amount of the pulse energy is blocked in the region of the diverging lens, but this loss is typically less than 1%, because the energy density at the edge of the beam is lower than at the center. Residual aberrations can be made small compared to the lidar receiver's field of view.

Off-axis sphere beam expanders were employed in two GTRI lidars, the Astronomical Lidar for Extinction (ALE), which had an eye safe 532 nm beam with a 32 cm diameter [3], and the 1.57 μm lidar that acquired the stratospheric aerosol data mentioned in Section 4.2.2, which had a 15 cm diameter beam [4]. In both cases, the expanders were optimized using the Zemax optical design code. The ALE transmitter is pictured in Figure 5.6, sending the laser beam into the night sky through a telescope dome slit.

Figure 5.6 The ALE transmitter. The 32-cm off-axis sphere is housed in a telescope tube with ventilation so the mirror will quickly come to the temperature of the night air when the dome slit is opened. A small amount of laser light leaks through the mirror coating and it can be seen through the six holes at the back end of the tube. J. T. McGraw photo.

The next component in Figure 5.2 is a mirror with a 45-degree incidence angle, which turns the beam through 90 degrees. Laser light is usually linearly polarized, and the reflectance of this mirror may depend strongly on polarization, so it may be important to make the polarization direction perpendicular to the plane of incidence. For example, the reflectance at 355 nm of a mirror with the commonplace beryllium–aluminum coating will be a factor of 3 lower if the polarization is in the plane of incidence. A GTRI 355-nm lidar suffered from this blunder in a dog-leg configuration, resulting in a power loss of almost an order of magnitude. This comment applies to any mirror in the lidar optical system at which the beam incidence angle is off the normal.

The final transmitter component shown in Figure 5.2 is a turning mirror with alignment adjusters. Two adjusters are shown, because the alignment must be correct in two planes (in the plane of the figure and perpendicular to it). The adjusters may be manual or motorized, but they must be capable of fine adjustments and mechanically stable, because lidar FOVs are typically 1 mrad or smaller, and the laser beam direction must be maintained within a fraction of the receiver FOV. The alignment stability of the lidar is determined largely by the opto-mechanical designs of the transmitter and receiver: the designed layout of the optical elements means very little unless a mechanical system maintains their relative positions and orientations within certain tolerances.

Chapter 6 includes a simple method for discovering what some of those tolerances are, and basic opto-mechanical design principles are discussed in Chapter 7.

The optical efficiency of the transmitter will generally decrease with the number of optical surfaces, because there will be some absorption or reflectance loss at each one. For this reason, the number of optical elements in the transmitter should be kept to a minimum. For a simple transmitter, such as the one illustrated in Figure 5.2, an optical efficiency of about 0.8 is achievable for visible light by using readily available AR-coated lenses and high-quality mirrors. The transmitter's optical elements may be placed in the light path in a different order, and some types of lidar require other elements. For example, a polarization-sensitive lidar may require an element that changes or switches the polarization state of the transmitted beam. EARL employs a Pockels cell that rotates the transmitted polarization by 90 degrees for alternate bursts of 2000 laser pulses. A Pockels cell requires high voltage, on the order of a kilovolt, but nowadays the polarization state can be controlled with a liquid crystal variable phase retarder that operates at low voltage, on the order of 10 V. An eye-safe visible-light lidar may also require an IR blocking filter, because many green lasers leak a significant amount of IR light that contributes to their eye hazard.

5.2 Lidar Lasers

The modern lidar technique was first demonstrated in 1963, just three years after the invention of the laser. Beams of light from searchlights had previously been used to remotely sense atmospheric scattering, but pulse-time-of-flight measurements of range-resolved profiles were impractical until the advent of the laser, which generates short, directional pulses of monochromatic light with high energies. The term laser was originally an acronym, standing for Light Amplification by Stimulated Emission of Radiation, but it has passed into common usage as laser, in the same way that RADAR became radar. Basic laser phenomenology is described in this section, along with metrics of laser beam quality and methods of changing the wavelength.

There are three basic requirements for a laser: an active medium, a population inversion, and optical feedback. These three requirements are discussed in the paragraphs that follow.

5.2.1 Active Medium

The active laser medium (also called the gain medium or lasing medium) is the source of optical gain (amplification) within a laser. The gain results from the stimulated emission of photons as the medium transitions to a lower energy state from a higher energy state that was previously populated by an external energy source. The energy states are quantized, and the energy of the emitted photons is equal to the medium's change in energy state. For example, as discussed in Chapter 3, electrons in atoms have discrete energy levels, and transitions occur between these energy levels because of the relative instability of the levels or because of the effect of incident electromagnetic

radiation. When photons are emitted, they have discrete frequencies corresponding to the energy level changes: $h\nu = E_2 - E_1$, where E_2 is the higher energy level and E_1 is the lower. When the active medium in a laser has quantized energy levels, the emitted photons all have the same frequency. This is the reason that most laser light is *monochromatic* (meaning single color).

5.2.2 Population Inversion

In nature, systems tend to come to their lowest allowed energy state rather quickly. A medium with a large number of atoms will usually have atoms in many different states, with populations described by the Boltzmann distribution, Eq. (3.17), which implies that the ratio of the population of the upper state to the lower would be $\exp[-\Delta E/kT]$, so a higher energy state never has a larger population than a lower one when the system is in equilibrium. However, lasing requires more atoms in a higher state – a nonequilibrium situation known as a *population inversion*. Such an inversion is achieved by injecting energy into the system, most commonly by the method of *optical pumping*. The earliest, and still very common pumping technique uses flashlamps, even though they are inefficient because their output is spectrally very broad and most of it ends up heating the laser rather than pumping the population inversion. Laser diodes have a narrow spectral output, so they are much more efficient, but they are also much more expensive. Other pumping methods include electric discharges, radio frequency (RF) excitation, and chemical reactions. Whatever pumping method is used, incident energy "pumps" something in discrete energy levels, such as atomic electrons, from the ground state into a higher, unstable energy level. These pumped electrons rapidly decay to a metastable level via a non-radiative transition. They remain in the metastable level until there are more electrons there than in the ground level. The two most common pumping schemes are illustrated with energy level diagrams in Figure 5.7. The four-level scheme, which is used in the workhorse Nd:YAG laser, has the advantage that fast decay to the ground state quickly removes atoms from the lower laser transition state, which enhances the inversion.

After a population inversion is established, an atom may spontaneously decay from the metastable level into the ground or intermediate level. In doing so, the atom emits a photon with the characteristic energy $h\nu = E_2 - E_1$. When this photon strikes another atom, it may induce either an upward or a downward transition, but due to the population inversion, the second atom is more likely to be in the excited, metastable state, in which case *stimulated emission* occurs, and the second atom releases a photon of identical energy that is in phase with the original photon. Because of this process, laser light is both monochromatic and *coherent*, meaning that the electromagnetic fields of the photons are all in phase.

5.2.3 Optical Feedback

Optical feedback refers to confining photons within the active medium long enough for them to stimulate further transitions. This is most often done by placing the active medium between two mirrors, one of which is totally reflective and the other

Table 5.2 Laser light properties and causes

Monochromaticity	Emission is due to transitions between quantized energy levels.
Coherence	Emission is stimulated by incident photons.
Directionality	Only rays along the cavity axis are amplified.
Polarization	Brewster windows and Q-switches only allow one polarization.

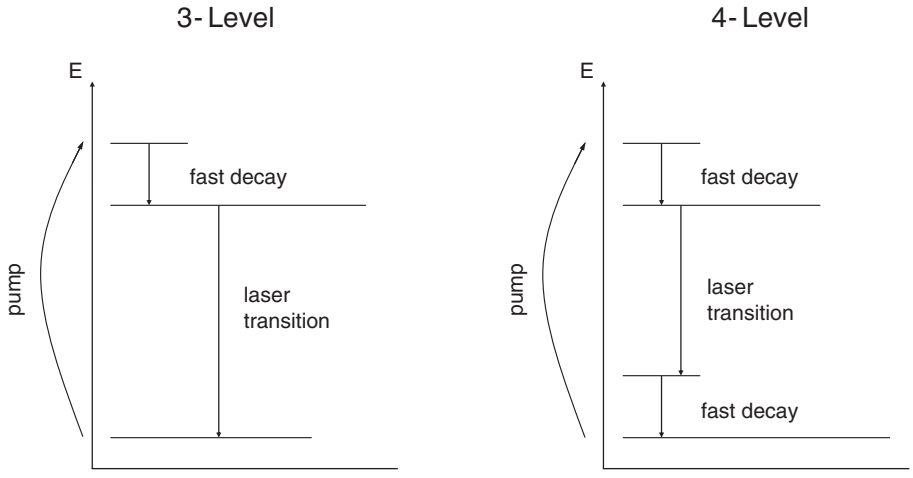


Figure 5.7 Common optical pumping schemes.

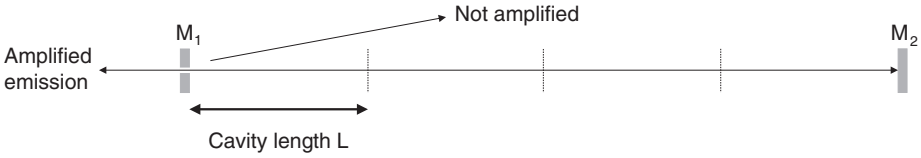


Figure 5.8 Laser directionality. M_1 and M_2 are the mirrors that form the laser cavity. The laser beam must traverse the optical resonant cavity many times to become amplified. Off-axis emissions are not amplified.

partially reflective. Usually at least one of the mirrors is concave. Light generated by spontaneous emission at the resonant frequencies becomes amplified during many passes back and forth between the mirrors. Only the rays of light travelling near the axis of the cavity can make many passes, so those rays are amplified, while others are not. This fact is the reason that laser light is usually highly directional. Laser beam directionality is illustrated in Figure 5.8. Each reflection results in a light path increase of $2L$, where L is the cavity length, and multiple reflections are shown schematically by extending the cavity for each one. Four reflections are shown for illustration in Figure 5.8, and the output coupler is shown as a hole in the mirror M_1. This type of output coupling was employed in the very first laser, in 1960.

5.2.4 Polarization

Laser light is usually linearly polarized, for two reasons. In the early days, when gas and liquid lasers were common, the laser cavity mirrors were outside the cell holding the active medium, and Brewster windows were commonly used to avoid losses due to reflection (see Section 6.2). Only light with a specific linear polarization could pass through the windows without attenuation, so that light was preferentially amplified. In modern lasers, the cavities are usually *Q-switched*. The parameter Q describes the "quality factor" of the resonant cavity. To build up a maximum population inversion, Q is held low to prevent optical feedback in the cavity while the pumping is occurring, and then it is suddenly raised to a high value, and the laser emits one large, short pulse. Q-switches usually operate by switching the linear polarization state of the laser light; hence the emitted light is linearly polarized. The properties of laser light described above, and their causes, are summarized in Table 5.2.

5.2.5 Laser Beam Quality

The properties listed in Table 5.2 make lasers an enabling technology for lidar, but they are idealizations. For example, no light source is perfectly directional, so it is necessary to characterize the beam from a specific laser to understand the degree to which it approaches the ideal limits. The *Gaussian beam*, illustrated in Figure 5.9, is a model used to characterize directionality. The curved resonator mirrors M_1 and M_2 give rise to a beam waist diameter inside the cavity of $2W_0$. The cavity type illustrated is called *confocal*, as the mirrors have a common focus in the center, and the mirror M_2 is partially transparent, so that the beam can escape from the cavity. The beam is axially symmetric, and the cross section of the beam intensity is described by a Gauss function, hence the name. The subscript zero in W_0 denotes the radius at a beam waist.

The Gaussian beam arises from the cavity mode TEM_{00} (meaning transverse electromagnetic mode 00, a terminology adapted from microwave resonators) and it has the lowest divergence of all transverse modes. The full-angle divergence is denoted by θ in Figure 5.9. The property of lowest divergence means that it can be focused on the smallest spot, which is important for laser applications such as etching materials. The beam cross section remains Gaussian as the beam propagates. Although this type of beam is not perfectly achievable in practice, the Gaussian beam model is nevertheless useful for calculations and as a baseline in characterization. Several mathematical relations hold for Gaussian beams. The full-angle beam divergence θ is given by

$$\theta = \frac{4}{\pi}\frac{\lambda}{2W_0}. \tag{5.1}$$

Note that θ is a function of only two variables, wavelength λ and waist diameter $2W_0$. As illustrated in Figure 5.9, the angle θ is measured to the *$1/e^2$ points*, that is, the radial distances where the beam irradiance has fallen to $1/e^2$ times its on-axis value. Equation (5.1) gives the lowest divergence that can be achieved by any laser, and it is

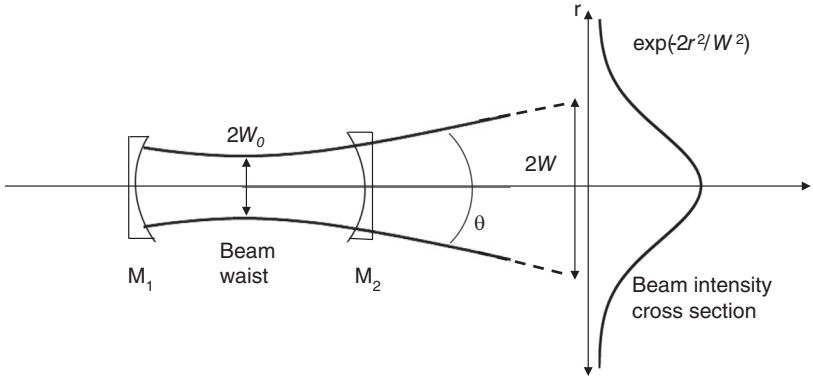


Figure 5.9 The Gaussian beam. The beam is axially symmetric with a Gauss function intensity cross section.

sometimes called the *diffraction limit*. Laser beam quality is quantified by the beam parameter product (BPP), which is the product of the beam's divergence and its waist size. The BPP for the Gaussian mode, $2\lambda/\pi$, is a function only of wavelength and it is smaller than the BPP for any other beam type. The beam irradiance (W/m^2) as a function of radial distance is given by the Gaussian function

$$E(r) = E_0 \exp[-2r^2/r_0^2], \tag{5.2}$$

where E_0 is the on-axis irradiance and r_0 is the radial distance to the $1/e^2$ points. The total beam power (W) is

$$P_{\text{total}} = \pi r_0^2 (E_0/2) \tag{5.3}$$

and the encircled power is

$$P_{\text{enc}} = P_{\text{total}}[1 - \exp[-2d^2/d_0^2], \tag{5.4}$$

where d is a diameter of interest and d_0 is the $1/e^2$ diameter. The Gaussian function approaches zero asymptotically as diameter increases, so the beam has no well-defined edge, but certain diameters are useful in lidar engineering. In particular, when $d = d_0$, the encircled power is 86.5% of the total beam power, and this relation is often used when measuring the diameter of a laser beam. When $d = 1.52d_0$, 99% of the beam power is encircled and so $1.52\,d_0$ is often used as a convenient definition of the "edges" of a laser beam when designing a lidar transmitter. Some manufacturers specify laser beam parameters in terms of the $1/e$ points rather than $1/e^2$, because it makes their products appear better. The conversion from one to the other follows from Eq. (5.2): The beam radius corresponding to the $1/e$ point is $1/\sqrt{2}$ times the $1/e^2$ radius, so a divergence quoted at the $1/e$ points must be multiplied by 1.404 to convert it to the $1/e^2$ divergence.

In real lasers, multiple cavity modes are usually excited, in both transverse and longitudinal directions. Multiple transverse modes cause non-Gaussian beams, and the modes may even change from pulse to pulse, which can cause random changes in the pointing direction. The intensity distributions of transverse modes are shown in

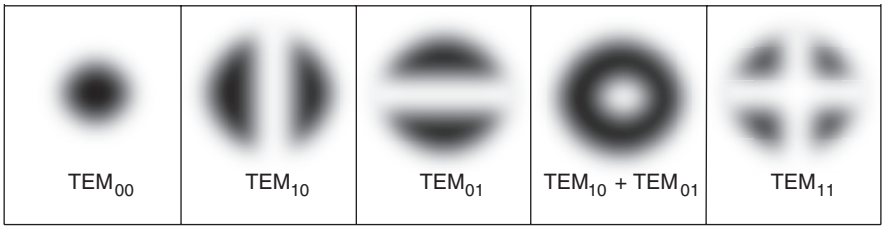


Figure 5.10 Transverse cavity modes. Intensity is shown by the gray scale, for the first four TEM cavity modes plus the donut mode.

Figure 5.10. Only the TEM_{00} mode is truly axially symmetric, although some early lasers produced TEM_{01} and TEM_{10} simultaneously, which resulted in a "donut" beam. Because of the presence of cavity modes other than TEM_{00}, the divergence of a laser beam must be characterized by means of measurements.

In direct detection lidars, the size and divergence of a laser beam are important for two reasons: First, these parameters are used in finding the crossover range, that is, the range beyond which the geometrical function $G(R)$ is unity (discussed in Chapter 6), and second, they are used in finding the range beyond which a transmitted laser beam is eye safe (Section 5.3). The simplest way to measure laser beam divergence is to measure the beam diameter d at two ranges, for example near the lidar transmitter (R_1) and at the end of a long hallway (R_2). This method was used for EARL. The full-angle divergence is then $\theta = [d(R_2) - d(R_1)]/[R_2 - R_1]$. The accuracy of this method depends on how the beam diameters are measured. If the beam is simply observed by eye using a white target board, the measurements are not very accurate because the beam has no well-defined edges and the radial distance at which it disappears will depend on background lighting and other subjective factors.

Several quantitative methods have been developed to measure laser beam parameters. A crude early method of measuring the beam size was to exploit Eq. (5.4) by passing the beam through an adjustable circular aperture to a power meter and adjusting the aperture diameter until 86.5% of the beam power passed through it. The aperture diameter was then taken to be the $1/e^2$ diameter, whether the beam approximated a Gaussian or not. Two related methods are the knife edge test, in which a shutter is slowly translated across the beam while the power is monitored, and the slit test, in which a slit is translated across the beam in the same way. Both methods can be used on orthogonal axes to check for asymmetry in the beam. Early benchtop laser beam divergence tests at GTRI employed a concept illustrated in Figure 5.11. The test fixture consists of a target in the focal plane of a lens. At the focal plane, the lens converts divergence to diameter, so $\theta = d/f$, where d is the spot size and f is the lens focal length. If the focal length is 1 m, the spot size on the target in millimeters corresponds to the beam divergence in milliradians. This method was used with a CO_2 laser operating at 10.6 μm by using an aluminum foil target. A single laser pulse blew a hole through the foil, and the hole diameter was taken as the spot diameter, but of course there was uncertainty in defining the edges of the beam with this method.

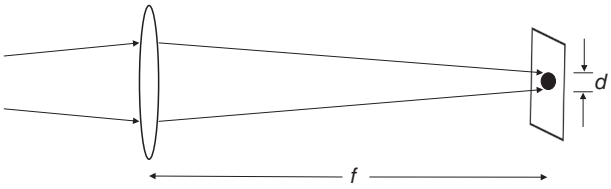


Figure 5.11 A simple beam divergence measurement. A target is placed in the focal plane of a lens, which converts beam divergence to a spot diameter.

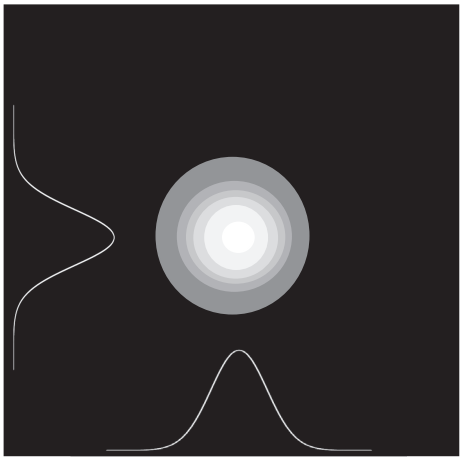


Figure 5.12 A beam profiler output. The beam irradiance is shown on a color scale or gray scale in the center, and cross sections are also plotted, on two orthogonal axes.

The problem of determining the spot size on the target has been solved by the modern technique of beam profiling, and instruments for this purpose are available from several vendors. For divergence measurements, they operate as shown in Figure 5.10 with a focal plane array (FPA) in place of the target. A great deal of attenuation is usually required in the laser beam to avoid damaging the FPA. A beam profiler can also be used with the attenuated raw beam incident on the FPA, in which case the irradiance cross section is recorded. A typical output format is illustrated in Figure 5.12. The irradiance profile is shown as concentric rings on a color scale (a gray scale is shown here, where higher values are whiter), and profiles measured as slices across the beam are shown on two orthogonal axes by the Gaussian-like curves. Profiles can also be presented as 3D plots, in which irradiance is shown by both color and vertical position.

The various beam characterization techniques that evolved over the years were used in different ways by different vendors in the laser industry, and that situation led to the development and adoption of the International Organization for Standardization (ISO) Standard 11146 for measuring laser beam widths, divergence angles, and beam propagation ratios [5]. For industrial applications such as laser etching, the

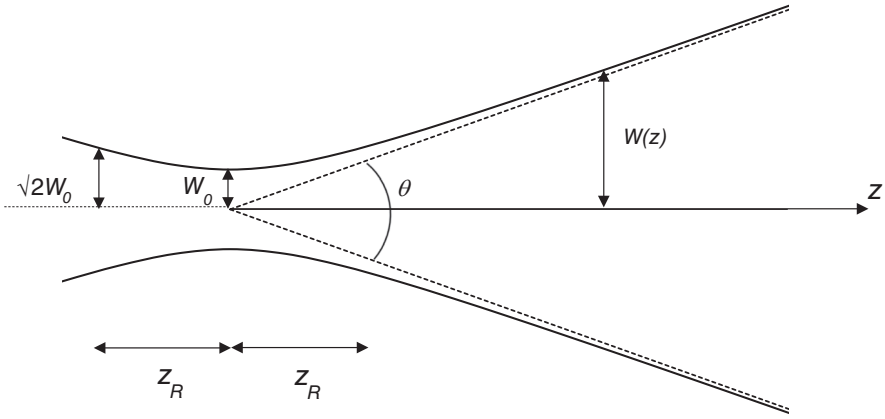


Figure 5.13 Gaussian beam propagation from a waist.

issue is how closely a laser beam approximates a Gaussian beam, and the performance of coherent detection lidars also depends strongly on beam quality. The standard is based on the propagation of a Gaussian beam over a distance along the z-axis after it is focused to form a beam waist, as illustrated in Figure 5.13. The Rayleigh range z_R is the distance along the beam axis from the beam waist to the point where the beam radius has increased by a factor of the square root of 2. The Rayleigh region extends $\pm z_R$ from the beam waist.

Radial distances are measured to the $1/e^2$ points. The minimum beam radius, at the center of the waist, is W_0 and the Rayleigh range for a Gaussian beam is given by $z_R = \pi W_0^2/\lambda$. The beam radius at any range is described by

$$W(z) = W_0\sqrt{1+\left(\frac{z}{z_R}\right)^2}. \tag{5.5}$$

As z increases, $W(z)$ asymptotically approaches the function $W_0(z/z_R)$. The beam is therefore cone shaped far from the waist and characterized by a divergence angle $\theta = 2W_0/z_R$. Substituting the definition of z_R given above, the full-angle divergence is $\theta = 2\lambda/\pi W_0$, which is the same as Eq. (5.1).

In ISO 11146, the ratio of the BPP for a real beam to the Gaussian BPP is called M^2 and it is now universally used to specify laser beam quality. The parameter M^2 is called the *beam propagation ratio*. The half-angle beam divergence is

$$\theta_{1/2} = M^2\frac{\lambda}{\pi W_0}, \tag{5.6}$$

and comparison with Eq. (5.1) shows that when M^2 is equal to unity, the beam is diffraction limited. This is the reason that M^2 is often called a factor times the diffraction limit. ISO 11146 specifies a method for measuring the waist size and divergence of a real laser beam by using a lens to create a beam waist and then measuring its radius

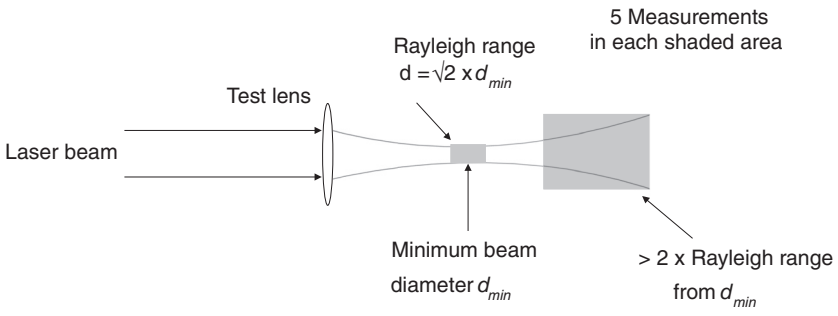


Figure 5.14 The ISO 11146 test setup.

at five points within the Rayleigh range and at five points further away, as illustrated in Figure 5.14. An analytical function is then fitted to the data points to interpolate between them, and the waist size and divergence are found from the functional fit. M^2 is then calculated as the BPP. The measurements must be done on two orthogonal axes in case the beam is not axially symmetric. The radius measurements are usually performed with a beam profiler, but the ISO standard also accommodates the legacy methods of using an adjustable aperture or a knife edge or slit test. Laser beam characterization systems with several levels of automation are available from several vendors, for both CW (continuous wave) and pulsed lasers.

5.2.6 Spectral Purity

In the previous section, laser beam quality was described in terms of the geometrical properties of the intensity profile: beam size and beam divergence. In some types of lidars, the spectral properties are also important. In Table 5.1, laser light is qualitatively described as monochromatic, but the degree of monochromaticity is quantified by the *spectral purity*. Spectral purity has several definitions. It is generally defined in terms of frequency, not wavelength, sometimes as $\Delta\nu/\nu$, where ν is the center frequency and $\Delta\nu$ is usually taken to be the full width at half maximum (FHWM). Another definition of $\Delta\nu$ is the width that encompasses 95% of the laser energy; the literature is inconsistent on this point. The laser frequency is found from $\nu = c/\lambda$, so for example, if $\lambda = 1$ μm, $\nu = (3\times10^8) / (1\times10^{-6}) = 3\times10^{14}$ Hz = 300 THz. Other useful relations are $\Delta\nu = -c\Delta\lambda/\lambda^2$ and $\Delta\nu/\nu = -\Delta\lambda/\lambda$. Another definition of spectral purity used in lidar is the percentage of the laser energy that lies within a useful passband, which is determined by the application of the lidar and the system parameters. The *spectral width* (in MHz) is a commonly used alternative to spectral purity. For pulsed lasers, the Fourier transform limit specifies the smallest possible spectral width. The line width in frequency is approximately the inverse of the pulse width in time. More precisely, there is a minimum time-bandwidth product that depends on the pulse shape. That product is 0.44 for Gaussian pulses (the term Gaussian refers to the temporal profile in this discussion, not the spatial profile).

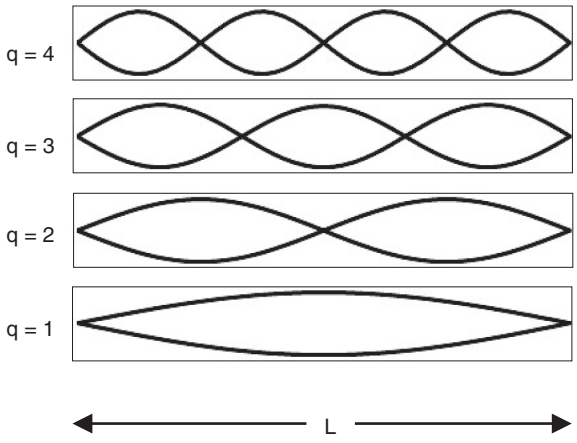


Figure 5.15 Organ pipe modes. The first four acoustic modes in a closed organ pipe are shown. The amplitude of the standing waves is plotted vertically in each pipe. Nulls must occur at both ends, so an integral number of half-waves must fit within the length L.

Example. A laser with a Gaussian pulse width of 10 ns cannot have a spectral width narrower than 44 MHz, because the time-bandwidth product is $(10 \times 10^{-9}) \times (44 \times 10^{6}) = 0.44$, and it cannot be smaller.

As mentioned earlier, the reason for non-Gaussian beams is that multiple modes are excited in the laser cavity. In addition to the transverse modes illustrated in Figure 5.10, multiple longitudinal cavity modes are also common. These are the modes referred to in the term *multi-mode laser*. A longitudinal mode is a standing wave pattern in which the modes correspond to wavelengths that are reinforced by constructive interference, so $L = q\lambda / 2$, where L is the cavity length and q is the mode order. These optical modes are analogous to the acoustic modes in an organ pipe, the first four of which are shown in Figure 5.15. Of course, the mode order is much higher in a laser. The mode number itself is not of interest in a laser, but the mode spacing is. The frequency spacing of the longitudinal modes is known as the *free spectral range*, defined by $\Delta\nu = c / 2L$ (Hz).

Example. The value of q for a 10-cm laser rod at 1-μm wavelength is $q = 2L / \lambda = 2 \times (10 \times 10^{-2})/10^{-6} = 2 \times 10^{5}$, which is a rather large number. The mode spacing in the cavity is $\Delta\nu = (3 \times 10^{8}) / (2 \times 0.1) = 1.5 \times 10^{9} = 1.5$ GHz. This is a small frequency shift from one mode to the next, and the laser may very well lase in more than one mode, in which case the laser light will have a spectral structure.

Laser output consists of all modes under the gain curve that are above a cutoff gain, as illustrated in Figure 5.16, in which five modes are active. Loop gain refers to the optical gain for one back-and-forth transit of the cavity, and the cutoff occurs where the loop gain drops below unity. Commercial Nd:YAG lasers are usually multimode

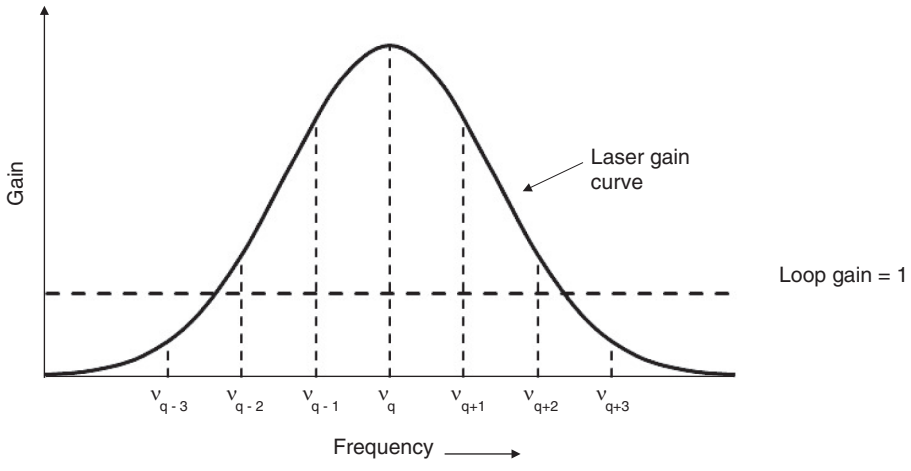


Figure 5.16 Multimode lasing. Cavity mode frequencies are shown by vertical dashed lines for the central mode q plus three other modes above and below the mode q. When the laser gain curve is above a cutoff value shown by the horizontal dashed line, the mode is active.

because the gain curve of the medium spans several modes above the cutoff for common rod lengths. Some Nd:YAG lasers have about 10 active modes, and consequently the laser fundamental line at 1064 nm appears to be about 0.1 nm wide (FWHM). For some lidar applications, the laser line must be narrowed, and the simplest method of narrowing is to insert an etalon into the laser cavity to increase the free spectral range. Etalons are thin pieces of optical material with polished plane-parallel faces. Because their thickness is typically millimeters instead of centimeters, the free spectral range of the laser cavity is increased by an order of magnitude, and hence only one mode exists under the gain curve. If high spectral purity is required in a lidar transmitter, frequency stability is also usually required, so the lasers in sophisticated lidars are often frequency stabilized. A standard way of stabilizing a laser is known as *injection seeding*, in which photons from a low power, very stable CW laser are injected into the cavity, so that those photons initiate the lasing each time the laser is fired.

5.2.7 Polarization Purity

Depolarization of laser light by atmospheric scatterers is often measured with lidars, and polarization is also exploited in other important ways in laser and lidar systems. In Chapter 4, the laser output was idealized as being completely polarized. If it is not, some error will be introduced into measurements, so it is important to define the polarization purity of the laser's output. The degree of linear polarization is most often quantified with the polarization extinction ratio (PER), defined as the ratio of optical power passing through a polarizer in the nominal polarization direction to the power with the polarizer in the orthogonal direction. Of course, the extinction ratio of the polarizer itself must be higher than that of the laser beam for this measurement to be meaningful. The polarization purity of lidar lasers is sometimes improved by passing their beams through polarizers before transmitting them.

For a simple, elastic backscatter lidar with a receiver optical bandwidth on the order of 1 nm, a commercial pulsed Nd:YAG laser will generally have adequate performance by all of the criteria discussed in Sections 5.2.5–5.2.7. More sophisticated lidars impose more demanding requirements on the laser. Also, when employing techniques to change the wavelength, described in the next sections, high laser beam quality, spectral purity, and polarization purity become important for high conversion efficiency.

5.2.8 Changing the Wavelength

As noted in Section 5.2.1, most laser light is at discrete wavelengths that arise from quantized energy levels. This fact is both a blessing and a curse. It is the reason that most laser emissions have stable wavelengths and high spectral purity, but the lasers available for lidar transmitters have only about 10 different active media, and most of them are not tunable, so the number of laser wavelengths available is not very large. For this reason, techniques for changing the wavelength of the laser light play an important role in lidar engineering.

Stimulated Raman Scattering

Stimulated Raman scattering (SRS) has become a standard frequency-shifting technique in lidar transmitters. As noted in Chapter 3, Raman frequency shifts are properties of the media doing the shifting, and the Raman cross sections are very small. SRS is a nonlinear phenomenon that overcomes the small cross sections and produces energy conversion efficiencies as high as 50%, but good spatial, spectral, and polarization quality are required in the laser beam, known as the *pump beam*, for high conversion efficiency. The most common SRS technique, illustrated in Figure 5.17, employs a pump beam focused to produce a beam waist in a high-pressure gas cell with windows. Both Stokes and anti-Stokes transitions can be excited, but Stokes is most often used. A significant fraction of the pump beam is shifted to a longer wavelength that passes through a dichroic mirror (*dichroic* means two-color), while the residual pump beam is diverted to a beam dump, where it is absorbed.

Both gaseous and solid media can be employed to shift laser wavelengths by fixed frequency increments, and the shifts in cm^{-1} for various gases are tabulated in the literature [6].

To find the final wavelength, find the laser frequency in cm^{-1}, subtract the Raman shift, and convert back to μm. Recall from Chapter 3 that a wavelength in μm is converted to cm^{-1} by dividing it into 10,000.

Example. Methane (CH_4) in a high-pressure gas cell was an early SRS medium. Methane has a Q-branch Raman shift of 2,914 cm^{-1}. Consider 1.064 μm laser light shifted by methane: dividing 1.064 μm into 10,000 yields 9,398 cm^{-1} and subtracting 2,914 cm^{-1} yields the spatial frequency of the Raman-shifted light as 6,484 cm^{-1}. Dividing this value into 10,000 yields the shifted wavelength of 1.54 μm. This is a very useful lidar wavelength because eye safety is most easily achieved in the region around 1.5 μm (see Section 5.3). It is also convenient because commonplace optical

materials such as BK7 glass can be used at both 1.064 μm and 1.54 μm, although the refractive index is slightly different. GTRI researchers demonstrated the first reported SWIR eye-safe cloud and aerosol lidar in 1989 using this technique [7].

SRS is an example of inelastic photon scattering, and in the example described above, the wavelength-shifted photons lose energy. The energy that they lose goes into the methane where it is lost by non-radiative processes, which results in heating of the gas. This heating, which occurs with each laser pulse in a small region near the beam waist, causes a problem: The hotter gas has a lower index of refraction and it is buoyant, so it causes *refractive turbulence* that can greatly degrade beam quality. Several solutions to this problem have been developed, including the use of a low PRF so that the gas can cool between laser pulses, a stirrer inside the cell to provide fresh gas for each laser pulse, and a different cell configuration with a very long path and no beam waist. This last approach was used in the scanning lidar known as REAL (Raman-shifted eye-safe aerosol lidar) [8].

Stimulated Raman scattering conversion efficiency depends so strongly on the laser beam parameters that a Raman shifter must be optimized for a particular laser by means of experiments using parameter excursions. The free parameters are the focal lengths of the lenses and the gas pressure, but the cell dimensions must also be adjusted to avoid optical damage to the windows. In addition, buffer gases such as helium are often used to improve heat transfer within the medium, so the buffer gas pressure is another free parameter. In practice, for a given laser and gas cell, a set of lens focal lengths is chosen and then gas pressures are varied over a wide range while the conversions efficiency is monitored. This process may be repeated with several sets of lens focal lengths. Once the Raman shifter is optimized, assuming that the cell windows are not being damaged by the laser beam, it can be used as a passive optical element in the lidar transmitter. Stimulated Raman scattering is most often used in lidar transmitters to achieve eye-safe wavelengths in the 1.4–1.6 μm range and to generate the wavelengths used in ozone DIAL systems shown in Figure 3.23.

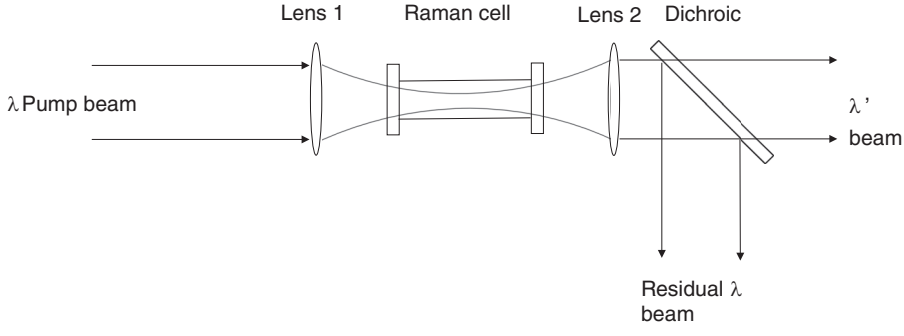


Figure 5.17 An SRS wavelength shifter. Lens 1 creates a beam waist inside the Raman cell, lens 2 re-collimates the residual beam and the wavelength-shifted beam, and the dichroic mirror separates the two beams.

Harmonic Generators

Laser harmonic generators are based on *nonlinear optics*, which is the branch of optics concerned with the behavior of light in nonlinear media, meaning that the dielectric polarization P (C/m^2) responds nonlinearly to the electric field E (V/m) of the light. The nonlinearity is typically observed only at very high light intensities such as those provided by lasers [9]. Nonlinear optics as a field of study arose in the 1960s, immediately after the invention of the laser, and it quickly produced commercial products including harmonic generators. When the polarization can be expressed as a Taylor series in electric field strength (and ignoring the vector natures of P and E for simplicity),

$$P(t) = \varepsilon_0(\chi^{(1)}E(t) + \chi^{(2)}E^2(t) + \chi^{(3)}E^3(t) + ...), \tag{5.7}$$

where ε_0 (F/m) is the dielectric constant and $\chi^{(n)}$ is the nth-order susceptibility of the medium. The relationship between the irradiance I (W/m^2) and E (V/m) is usually written $I = E^2 / 2 \times 377$, where 377 is the impedance of free space in ohms. The time-dependent electric field is proportional to $\exp[i\omega t]$ in complex notation, where ω is the angular frequency of the incident wave, so Eq. (5.7) shows that $\chi^{(2)}$ leads to twice the incident frequency, $\chi^{(3)}$ leads to thrice the incident frequency and so forth. The nonlinear interaction also leads to energy being coupled between different frequencies, which is called *wave mixing*. In particular, a frequency doubler exploits three-wave mixing in which two incident photons are annihilated and one new photon is created. A doubler is also called a *second harmonic generator* (SHG). To generate a high-intensity 2ω field, one more condition must be satisfied: The wave vectors $\vec{k}$ of the three fields must satisfy the *phase-matching condition*, which is

$$\vec{k}_3 = \vec{k}_1 + \vec{k}_2. \tag{5.8}$$

Frequency doubling in commercial lasers is accomplished by placing a nonlinear crystal in the laser beam. The most commonly used crystals are BBO (β-barium borate), KDP (potassium dihydrogen phosphate), KTP (potassium titanyl phosphate), and lithium niobate (LiNbO3). These crystals all have the necessary properties of having a specific crystal symmetry, being transparent for both the incident laser light and the frequency-doubled wavelength, and having high optical damage thresholds, which makes them resistant against the high-intensity laser light. They are also strongly birefringent, which is necessary to obtain phase matching. The term birefringence means that the crystal's index of refraction depends on the polarization and direction of the light that passes through, relative to the crystal axes. Phase matching is often achieved by choosing the polarizations of the fields and the orientation of the crystal appropriately; this technique is called angle tuning. The birefringence in lithium niobate is highly temperature dependent, so its temperature is generally controlled to achieve phase matching. The nomenclature of nonlinear optical processes is given in Table 5.3.

Harmonic generators are commonplace in commercial lasers. For example, the Nd:YAG fundamental frequency corresponds to 1,064 nm wavelength, and harmonic generators are widely available to produce 2, 3, and 4 times the fundamental frequency, corresponding to wavelengths of 1/2, 1/3, and 1/4 the fundamental wavelength (532,

Table 5.3 Nonlinear optical processes

Acronym	Meaning	Definition
SHG	Second-harmonic generation	Generation of light with a doubled frequency. Two photons are destroyed, creating a single photon at two times the frequency.
THG	Third-harmonic generation	Generation of light with a tripled frequency. Three photons are destroyed, creating a single photon at three times the frequency.
HHG	High-harmonic generation	Generation of light with frequencies much greater than the original (typically 100 to 1000 times greater).
SFG	Sum-frequency generation	Generation of light with a frequency that is the sum of two other frequencies (SHG is a special case of this).
DFG	Difference-frequency generation	Generation of light with a frequency that is the difference between two other frequencies.
OPA	Optical parametric amplification	Amplification of a signal input in the presence of a higher-frequency pump wave, at the same time generating an idler wave (can be considered as DFG).
OPO	Optical parametric oscillation	Generation of a signal and idler wave using a parametric amplifier in a resonator (with no signal input).

355, and 266 nm). Such harmonic generators have good energy conversion efficiency, which is the ratio of the pulse energy of the harmonic to the energy of the fundamental (conversion efficiency is also sometimes expressed in terms of the numbers of photons in the fundamental and the harmonic). However, the conversion efficiency depends on the beam parameters discussed in Sections 5.2.5–5.2.7. Typically, commercial Nd:YAG harmonic generators produce 1/2, 1/3, and 1/4 the pulse energy of the 1,064 nm fundamental in the second, third, and fourth harmonics.

The multi-wavelength laser configuration shown in Figure 5.18 is known as Master Oscillator Power Amplifier (MOPA). Much experience has shown that good beam quality is best obtained by engineering a high-quality low-power laser (the master oscillator) and then amplifying its output (with the power amplifier). In Figure 5.18, the master oscillator cavity lies between elements G and K. The laser rod, pumped by a flash lamp, is element H. The cavity is injection seeded by a small, stable laser (A), and when it is switched on by the Pockels cell (F), the photons from the seed laser are amplified. In this way, laser A controls and stabilizes the wavelengths of all laser outputs. The master oscillator's output is linearly polarized by element I, and the quarter-wave plates in the cavity are used in conjunction with F to enable Q-switching by rotating the plane of polarization. Elements L and M constitute a beam expander, which is used because the power amplifier rods (P) have a larger diameter than H. A larger diameter is desirable, because many of the optical elements in such a laser are susceptible to optical damage

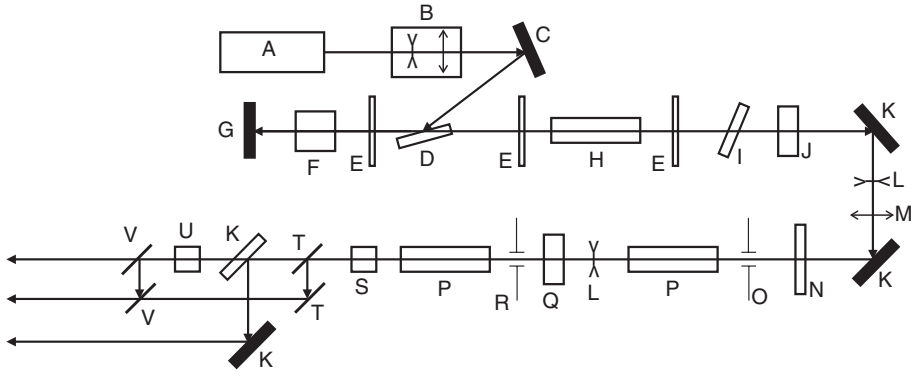


Figure 5.18 A typical MOPA configuration. A – seeder, B – seeder telescope, C – mirror – turning 30°, D – dielectric polarizer, E – l/4 wave plate, F – Pockels cell, G – mirror, H – laser rod (6 mm) with flash lamp, I – dielectric polarizer, J – output coupler, K – mirror – turning 45°, L – diverging lens, M – converging lens, N – l/2 wave plate, O – pinhole – 9.0mm, P – laser rod (9 mm) with flash lamp, Q – rotator, R – pinhole – 9.5 mm, S – SHG, T – dichroic – 532 nm, U – THG or FHG, V – Dichroics – 355 nm.

caused by the intense radiation, and spreading the beam over a larger area helps to mitigate this problem. After amplification by the two rods P, the high-power fundamental output at 1064 nm enters the second harmonic generator (SHG) S. Two wavelengths emerge from S: 1,064 nm and 532 nm. After this point, several mirrors and dichroic mirrors can be placed in the beams to produce spatially separated beams at multiple wavelengths. For example, if T and K are employed, 532 nm passes through dichroic mirror T while 1,064 is reflected at a right angle, and those two wavelengths emerge from the laser. The third harmonic 355 nm is obtained by mixing 1,064 and 532 nm in the third harmonic generator (THG) U, and separating 355 with dichroic V. The fourth harmonic at 266 is obtained by doubling 532. The harmonic generators and the dichroic mirrors are especially susceptible to optical damage, and it is not unusual to replace them periodically if the fundamental pulse energy is on the order of 1 joule. Shorter wavelengths are more likely to cause damage than longer wavelengths because the photon energy varies inversely with wavelength.

Optical Parametric Oscillators

Optical parametric oscillators are another example of three-wave mixing, in which photons in a pump beam result in photons at two other frequencies, known as the *signal* and the *idler*, such that their sum is equal to the pump frequency. The OPO is tunable, which is its great advantage. A common OPO cavity arrangement is shown in Figure 5.19. Although OPOs are tunable and efficient, they tend to have high beam divergence, and the configuration shown in Figure 5.19 creates a signal beam with a much higher divergence on one axis that on the other. Laser developers at Sandia National Laboratory solved the divergence problem by twisting half of the cavity out of plane and rotating the beam as it goes around [10]. The device, known as

Table 5.4 Lidar lasers

Type	Wavelength	Comments
Ruby	0.694 μm	The original lidar laser; still in use but rare.
CO_2	9–11 μm (many lines)	10.6 μm strongest; used in coherent lidars.
Dye	0.39–0.64 μm	Tunable.
Nd:YAG	1.064 μm	Also used as a pump laser.
Nd:YLF	1.047 μm	Also used as a pump laser.
Er:YAG	1.550 μm	Eye safe, as laser rod or fiber laser
Excimer	0.193, 0.248, 0.308, 0.353 μm	Media such as XeF only exist in excited state.
Ti:Al_2O_3	0.650–1.100 μm	Tunable. Often pumped by Nd:YAG.
Tm:YAG	1.930–2.040 μm	Tunable (narrow range).
Ho:YLF	2.08 μm	Used in coherent lidars.
Alexandrite	0.700–0.820 μm	Tunable.
QCL	MWIR – LWIR	Fabricated for chosen wavelengths.
Ce:LiSAF, Ce:LiCAF	280–316 nm	Tunable, pumped with quadrupled Nd:YAG or excimer lasers. Used in ozone DIAL.

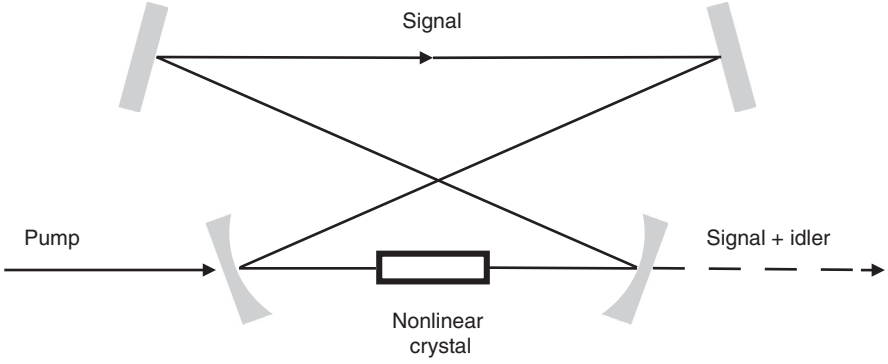


Figure 5.19 An optical parametric oscillator. The sum of the signal and idler beam frequencies is equal to the pump beam frequency.

a RISTRA (rotated image singly-resonant twisted rectangle), is now commercially available from AS-Photonics, LLC and their web site includes a wealth of technical information about it [11].

5.2.9 Lidar Lasers

Many of the types of lasers that have been used in lidar systems are listed in Table 5.4. In addition to the wavelengths shown, others can be generated by using Nd:YAG and Nd:YLF lasers to pump harmonic generators, Raman cells, and OPOs.

The basic suitability of a laser for an intended lidar measurement scenario can be assessed from the laser's wavelength, energy per pulse, PRF, and pulse width, by

using the information in Chapters 1–4, but the beam diameter and divergence must also be known to design the transmitter optics. If the system is more sophisticated than a simple elastic backscatter lidar, or if the laser will be used to pump a wavelength shifter, then other laser characteristics come into play, including beam quality, spectral purity, and polarization purity. For a lidar system to be deployed in the field, many other considerations are important, including cost; size, weight; power; expected lifetime; maintenance requirements; expendables (dye, flashlamps); shutters and warning lights; extraneous optical emissions; electromagnetic emissions; temperature range (operating and storage); cooling requirements; and wall plug efficiency. Finding the best laser for a given lidar is most often not a simple task.

5.3 Laser Safety

To this point, we have considered lasers only in terms of how appropriate they are for a given lidar instrument and measurement scenario. However, atmospheric lidar systems are deployed outdoors, so it is important to assess the hazard that they may present to personnel. In addition, lidar researchers often have access to the laser and its raw beam in the laboratory, so again, awareness of potential laser hazards is important.

5.3.1 Laser Bioeffects

Laser beams can injure biological tissue including both skin and eyes, through three processes:

(1) Thermal. An increase in temperature caused by absorption of energy.
(2) Acoustical. A mechanical shockwave caused by localized vaporization of tissue.
(3) Photochemical. A photon interacting with cells that changes cell chemistry.

The type of damage depends on the wavelength. The interaction of IR lasers with tissue is all thermal, causing burns. UV damage is photochemical, like sunburn. Visible laser light is a special case. On skin, it can cause darkening, but in the eye, it can be focused onto the retina, where it can cause permanent thermal damage. The eye has a complicated structure, illustrated in Figure 5.20, and it can also suffer damage at the cornea or the lens, depending on laser wavelength. Many laser wavelengths are stopped at the cornea, by absorption, including SWIR–LWIR (1,400 nm–1 mm) and Far-UV (180 nm–315 nm). These wavelengths can cause thermal or photochemical damage to the cornea, which is quite painful. The most serious (and permanent) damage is caused by light that is focused on the retina; this includes both visible and near IR (400 nm–1,400 nm) light. Near-UV (315 nm–390 nm) light penetrates the cornea but is stopped at the lens. Because of these effects, SWIR–LWIR laser light is the least hazardous because its effect is strictly thermal, while UV is the next least hazardous because it is stopped at the cornea or the lens. Visible and NIR light are easily the most hazardous, because they can be focused into a tiny area on the retina by the lens so that their energy is greatly concentrated.

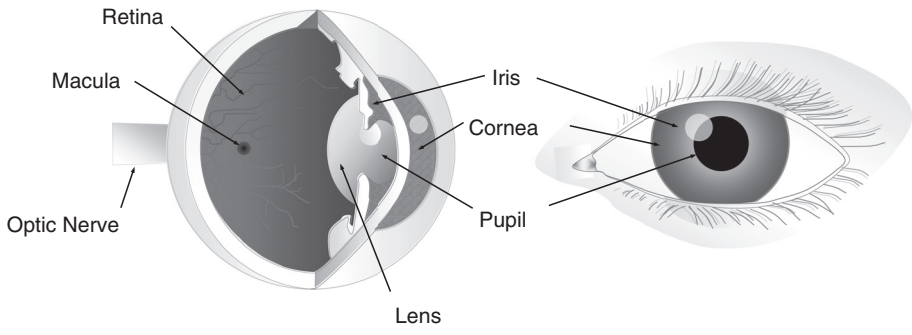


Figure 5.20 The human eye. Laser injuries may occur at the cornea, the lens, or the retina.

5.3.2 Laser Safety Regulations

There are several standards for laser safety. In the U.S., The American National Standards Institute (ANSI) publishes and updates a set of standards numbered Z136.1–Z136.9, where Z136.1 *Safe use of lasers* and Z136.6 *Safe use of lasers outdoors* are the most relevant to lidar researchers [12]. The standards for pulsed lasers are in terms of fluence (J/cm^2) at the eye, called maximum permissible exposure (MPE). A summary of the 2014 ANSI standard vs. wavelength is shown as a diagram in Figure 5.21, which should be interpreted with caution, because the ANSI standard requires three separate calculations (not just MPE) to determine laser safety, and the most restrictive result applies. Cumulative exposure is considered in the UV because of photochemical damage, whereas in the IR it is not. To put the standards on one plot, it was necessary to consider only a single pulse and a 10 ns pulse width for all wavelengths. Adjustments must be made for multiple-pulse exposures, and they differ with wavelength. The MPE in the visible light region is based on a ¼-second exposure, because the "blink reflex" protects personnel, who are expected to quickly avert their eyes when they encounter the beam. Lidar researchers sometimes claim that the term *eye safe* means that one can "stare into the beam". This statement is incorrect for visible light, because a maximum ¼-second exposure is assumed. Figure 5.21 covers the wavelength range from 0.18 μm to 5 μm. For longer wavelengths, the MPE value stays the same as at 5 μm. The range of MPE values spans almost seven orders of magnitude, which is a huge range. The MPE standard shown in Figure 5.21 matches the qualitative hazard description following Figure 5.20 with one anomaly – in the region from 1.4 μm to 1.8 μm, the MPE is 1 J/cm^2. This value is easy to achieve, so wavelengths in this range are ideal for applications such as horizontal scanning near ground level, where the absence of personnel in the beam path cannot be guaranteed. The visible region (0.4 μm–0.7 μm), where the MPEs are more than six orders of magnitude lower, is the most challenging for engineering a useful eye-safe lidar transmitter. The only solution is the micro-pulse lidar approach used in EARL: expand the beam, make the pulse energy very small, and make the PRF high to average down the noise as quickly as possible.

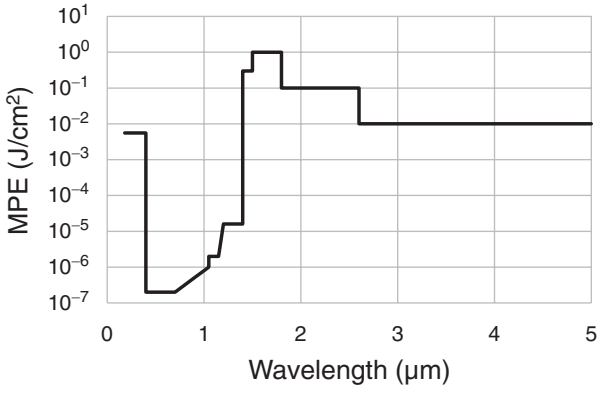


Figure 5.21 The 2014 ANSI MPEs for a single laser pulse with a 10 ns width.

The other ANSI parameter that often must be calculated by lidar researchers is the nominal ocular hazard distance (NOHD), which is defined in the ANSI Standards as "The distance along the axis of the laser beam beyond which the MPE is not exceeded …." Because laser beams diverge, at some range their fluence drops below the MPE and the other safety criteria, and they are considered eye safe. For a truly eye-safe lidar, the NOHD should be zero.

In addition to the ANSI standards, the U.S. Food and Drug Administration (FDA) Center for Devices and Radiological Health (CDRH) issued Federal Laser Standard 21 CFR 1040, which is used for classifying lasers according to their potential to cause biological damage [13]. Unfortunately, the tables in [13] are blurred copies of typewritten documents and they are illegible. However, the tables have been transcribed clearly on other web sites [14]. All use of lasers outdoors in the U.S. is governed by the Federal Aviation Administration (FAA) under FAA Order JO 7400.2M [15]. The American Council of Governmental Industrial Hygienists has also issued guidelines for radiative hazards, but they are more general, also covering sources such as flashlamps. Laser classification by 21 CFR 1040 considers the laser's output energy or power, its wavelength, the exposure duration, and the beam area at the point of interest. Lasers are classified by the Accessible Emission Limit (AEL), meaning the beam that a person could come into contact with. CW laser classes are described in Table 5.5, which is used to classify lasers by starting at the top and asking whether the laser is Class I. If not, proceed to Class II. Continue until the correct class is determined. Tables for pulsed lasers are also listed in 21 CFR 1040; see [12] or [13]. The limits are specified in tables with titles corresponding to the laser class; for example, the Class IIIa limits are listed in Table III-A.

With the exception of micro-pulse lidars, most lidars employ Class IIIb or Class IV lasers. The ANSI standards contain procedures for avoiding exposure to laser light, designation of a Laser Safety Officer, training of operators, and the posting of warning signs in laser operating areas for Class IIIb lasers. Personnel who operate Class IIIb or Class IV lasers should be familiar with these standards to ensure that the laser is used safely. The ANSI standards require that all lasers must have warning labels, and Class

Table 5.5 Laser classes (21 CFR 1040)

Class	Description	CW Power
I	Incapable of producing hazard to eyes or skin (but must be labeled)	
II	Blink response protects eyes	<1 mW
IIIa	Hazard for aided viewing (binoculars)	<5 mW
IIIb	Diffuse reflection is not a hazard	<500 mW
IV	All lasers exceeding Class III. Hazardous to eyes & skin, start fires	

II–Class IV must have an exit port label. At least one emission indicator shall be used on Class IV lasers and should be used on Class IIIb. The emission indicator must be clearly visible through protective eyewear for the laser wavelength. The laser firing switch should be clearly identified, and safety practices shall be incorporated into a standard operating procedure (SOP).

The Federal Aviation Administration (FAA) has aviation safety concerns other than bioeffects because lasers may also present indirect safety hazards, including flash-blindness, afterimage, glare, and startle. For this reason, FAA Order JO 7400.2M describes the following Aircraft Zones around airports:

(1) Sensitive Zone – visible beams may interfere with tasks but do not jeopardize safety.
(2) Critical Zone – interference would jeopardize safety.
(3) Laser-Free Zone – no lasers allowed.

Diagrams showing typical layouts and dimensions of such zones are included in ANSI Z136.6. A letter of non-objection must be obtained from the FAA before operating any laser outdoors in the U.S. FAA Order JO 4200.2M also defines roles such as Laser Safety Officer and Laser Operator.

The U.S. Department of Defense also has a concern separate from bioeffects: the possibility that orbiting satellites might be accidentally damaged by lasers used outdoors. The Office of the Under Secretary of Defense for Policy has issued DoD Instruction 3100.11, *Management of Laser Illumination of Objects in Space*, to establish policies, responsibilities, and procedures for the DoD management of risks associated with laser illuminations [16].

5.4 The EARL Transmitter

The optical configuration of the EARL transmitter is shown in Figure 5.22, and the transmitter parameters are listed in Table 5.6. As with other aspects of EARL, much of the design was constrained by the project budget. For example, the transmitter was based on a 20-cm telescope that was donated to the project by Agnes Scott College, and the non-standard 523.5-nm wavelength came from a doubled Nd:YLF laser that was the lowest cost laser available with the necessary combination of PRF and pulse energy. The telescope blocks much of the beam power, with the result that the optical efficiency k_T is only 0.30. The laser beam is expanded twice, and the final beam is

Table 5.6 EARL transmitter parameters

Parameter	Value
Wavelength	523.5 nm
Laser pulse energy	50 μJ
PRF	2.5 kHz
Optical efficiency	0.30
Pulse width	2.5 ns
Raw beam diameter	0.200 mm
Raw beam divergence	3 mrad
Expanded beam diameter	200 mm
Expanded beam divergence	~90 μrad
Polarization purity	Not measured
Spectral purity	Not measured

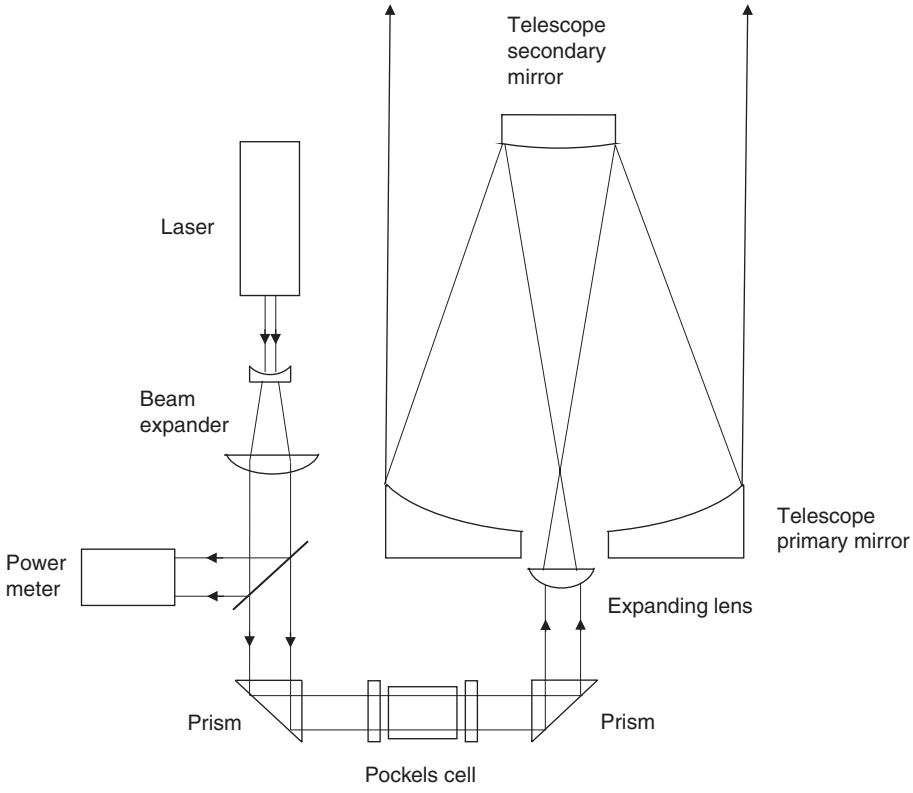


Figure 5.22 The EARL transmitter. The laser beam is expanded to 8-mm diameter before transiting the power pickoff and the Pockels cell, then expanded to 20 cm by the telescope.

annular, with an outer diameter of 20 cm. Its divergence was estimated by measuring its outer diameter before and after propagating the beam horizontally for 60 m. The expanded beam is truncated by the edges of the primary and secondary telescope mirrors, and diffraction rings are obvious at the outer edge of the beam, so it is nothing like

a Gaussian beam. However, the transmitter is eye safe, the beam divergence is less than the receiver FOV, polarization switching is implemented, and the power meter enables continuous monitoring of the output power. These features were deemed sufficient for a teaching lidar in an undergraduate environment. The beam path from the laser to the transmitting telescope is enclosed in 1-inch (2.54 cm) optical tubing for safety.

5.5 Further Reading

K. J. Kasunic, *Laser Systems Engineering*. Bellingham, Washington: SPIE Press, 2016.

This book is unusual in that it treats the fundamentals of laser-based systems in general. It includes chapters on lasers, beam propagation, optics, radiometry, and detectors, all of which are relevant to lidars.

The book by New [9] is remarkably accessible and it is recommended for newcomers to nonlinear optics.

R. W. Boyd, *Nonlinear Optics*, 3rd ed. Amsterdam: Academic Press, 2008.

This standard text on nonlinear optics is rigorous and comprehensive. It is an excellent resource for more advanced researchers.

5.6 Problems

5.6.1 Derive Eq. 5.3 for total encircled power in a Gaussian beam.

5.6.2 Is EARL eye safe? As shown in Figure 5.21, the ANSI MPE is 2×10^{-7} J/cm^2 for 2.5 ns pulses at 523.5 nm. However, there is a correction factor $C_p = n^{-0.25}$ for multiple-pulse exposures, where n is the number of pulses. Recall that an 0.25 s exposure is assumed for visible light and use the transmitter parameters listed in Table 5.6.

5.6.3 What is the 21 CFR 1040 laser class of EARL's transmitter? Table 5.5 is for CW lasers, see [13] for pulsed lasers. The parameters k_1 and k_2 are defined in a table at the end.

5.6.4 An industrial laser for etching produces a beam with wavelength 0.532 μm and BPP 2.0. If the laser beam is focused into an *f*/2 cone of light, find the spot diameter D at the beam waist.

References

[1] J. A. Reagan, D. M. Byrne, and B. M. Herman, "Bistatic Lidar: A Tool for Characterizing Atmospheric Particulates: Part 1 – The Remote Sensing Problem," *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-20, pp. 229–235, 1982.

[2] T. Shiina, K. Yoshida, M. Ito, and Y. Okamura, "In-line Type Micropulse Lidar with an Annular Beam: Experiment," *Applied Optics*, vol. 44, pp. 7407–7413, 2005.
[3] M. Dawsey, G. Gimmestad, D. Roberts, J. McGraw, P. Zimmer, and J. Fitch, "LIDAR for measuring atmospheric extinction," In Proceedings of SPIE, vol. 6270, 2006, pp. 62701F1–62701F10.
[4] H. N. Forrister, D. W. Roberts, A. J. Mercer, and G. G. Gimmestad, "Infrared Lidar Measurements of Stratospheric Aerosols," *Applied Optics*, vol. 53, pp. D40–D48, 2014.
[5] ISO 11146 has three parts, as follows: Part 1 Stigmatic and simple astigmatic beams; Part 2 General astigmatic beams; and Part 3 Intrinsic and geometrical laser beam classification, propagation and details of test methods. [Online]. Available: www.iso.org/home.html. [Accessed: 27 May 2021].
[6] R. M. Measures, *Laser Remote Sensing – Fundamentals and Applications*. New York: Wiley, 1984.
[7] E. M. Patterson, D. W. Roberts, and G. G. Gimmestad, "Initial Measurements Using a 1.54 Micron Eyesafe Raman Shifted Lidar," Letters to the Editor, *Applied Optics*, vol. 28, pp. 4978–4981, 1989.
[8] S. Mayor and S. Spuler, "Raman-Shifted Eye-Safe Aerosol Lidar," *Applied Optics*, vol. 43, pp. 3915–3924, 2004.
[9] G. New, *Introduction to Nonlinear Optics*. New York: Cambridge University Press, 2011.
[10] D. J. Armstrong and A. V. Smith, "Demonstration of Improved Beam Quality in an Image-Rotating Optical Parametric Oscillator," *Optics Letters*, vol. 27, pp. 40–42, 2002.
[11] AS-Photonics, LLC. [Online]. Available: https://as-photonics.com. [Accessed March 9, 2022].
[12] The ANSI Z136 series of laser safety standards may be purchased from the Laser Institute of America. [Online]. Available: www.lia.org/. [Accessed: May 27, 2021].
[13] Federal Laser Standard 21 CFR 1040. [Online]. Available: www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?FR=1040.10. [Accessed: May 27, 2021].
[14] A clear, transcribed version of Federal Laser Standard 21 CFR 1040. [Online]. Available: 21 CFR § 1040.10 – Laser products. | CFR | US Law | LII / Legal Information Institute (cornell.edu). [Accessed: May 31, 2021].
[15] FAA Order JO 7400.2M. [Online]. Outdoor laser regulations are covered in Part 6 Miscellaneous Procedures, Chapter 29 Outdoor Laser Operations. Available: www.faa.gov/documentLibrary/media/Order/7400.2M_Bsc_w_Chg_1_dtd_1-30-20.pdf. [Accessed: May 27, 2021].
[16] DOD INSTRUCTION 3100.11, MANAGEMENT OF LASER ILLUMINATION OF OBJECTS IN SPACE. [Online]. Available: https://fas.org/irp/doddir/dod/i3100_11.pdf. [Accessed: May 27, 2021].

6 Lidar Receivers and the Geometrical Function

This chapter continues the description of the instrumental parameters in the lidar and background equations that was started in Chapter 5. The receiver parameters and the geometrical function $G(R)$ are discussed here, starting with the basic components of a receiver. A discussion of depolarization instrumentation for implementing the theory discussed in Chapter 4 follows, with an example. Finally, $G(R)$ is discussed at length with a derivation of simple formulas for engineering the shape of the function. This chapter necessarily includes many optical diagrams and calculations, so a brief review of geometrical optics is included before the last section.

6.1 Components of Lidar Receivers

The optical requirements of a lidar receiver are that it must collect sufficient backscattered laser light for the required SNR to be achieved in the anticipated measurements and put that light on the detector; it must minimize the amount of sky background light that reaches the detector (unless the lidar is night-only); and it must maintain a stable geometrical function $G(R)$ and avoid any range-dependent distortions of the lidar signal. It may also contain a polarization analyzer, and if not, it must not have accidental polarization sensitivity. The basic components of a lidar receiver are shown in Figure 6.1. The first element is the objective mirror or lens of the telescope. Whether it is refractive or reflective, it is characterized by a diameter, a focal length, and an optical efficiency. The second element is the field stop, whose aperture diameter determines the receiver FOV. The field stop may be in the focal plane or slightly behind it. Next is a collimating lens, a narrow bandpass optical filter to block most of the sky background light, a field lens, and finally, the detector (covered in Chapter 8). The receiver forms an image of the scattering volume in or behind the focal plane, depending on range. The receiver is an imaging system because there is no other practical way to restrict the FOV while guiding the photons through all the optical elements and concentrating them onto the detector element.

6.1.1 Telescope and Field Stop

The receiver telescope may be custom made or it may be an astronomical telescope. As shown in Eqs. (2.11) and (2.12), the lidar signal is proportional to the receiver area and inversely proportional to the range squared, so telescopes for short-range

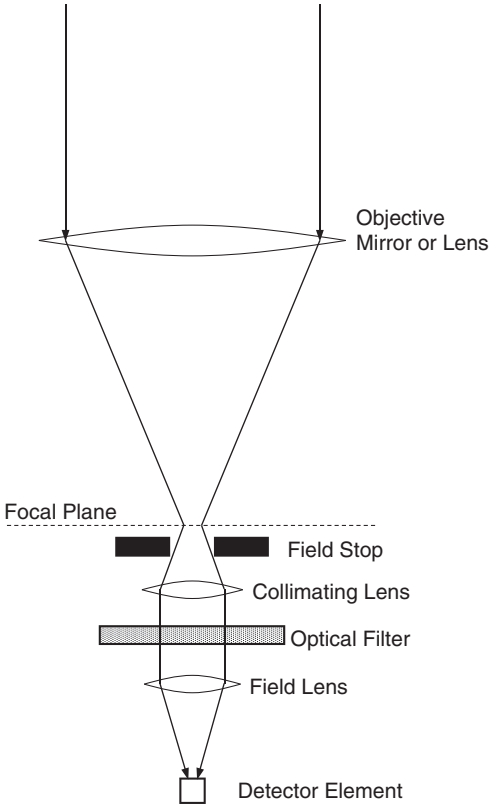


Figure 6.1 Basic components of the lidar receiver. They are the telescope, the field stop, collimating lenses, an optical filter, and a detector. A polarization analyzer may also be included.

work can be small, but as the measurement ranges get longer, they grow rapidly to diameters as large as several meters. The field stop diameter D_{FS} and the focal length f of the telescope objective determine the receiver FOV full angle as $\theta_{\mathrm{FOV}} = D_{\mathrm{FS}}/f$ (see Section 6.3.2). Lidar receiver FOVs are usually 1 mrad or smaller.

6.1.2 Lenses and Detector

The collimating lens makes the rays more nearly parallel before they pass through the optical filter. The next element after the filter is the field lens (also called a Fabry lens), which has two functions: It concentrates the rays of light so that they fit onto the detector element, and it images the objective lens or mirror onto the detector. This latter function is important because detectors often have a nonuniform sensitivity across their surfaces. If the image of the scattering volume moves on the detector element as the pulse propagates, a range-dependent distortion is introduced into the electronic signal. A main goal of good lidar engineering is to ensure that such distortions cannot

occur, and the best way to eliminate image motion is to image the objective onto the detector – the objective never moves, and it is uniformly illuminated by backscattered laser light regardless of range.

6.1.3 Bandpass Filter

The function of the optical filter is to allow the backscattered laser light to pass through while blocking most of the background light. A spectrally narrow filter is called a *bandpass* filter, and its spectral width is called its bandpass or bandwidth. Most optical filters are sensitive to the angle of the rays going through them, which is why a collimating lens is used to make the rays nearly parallel and normal to the filter surface. Lidar receivers usually employ interference filters that are fabricated from multiple layers of dielectric materials with different refractive indices, and this type of filter is sensitive to both the angle of incidence and temperature. The following discussion on the angle and temperature dependence of dielectric filters was adapted from the web site of the Andover Corporation [1]. For rays at off-normal incidence, the thickness of the filter layers effectively decreases, resulting in a shift of the pass-band to shorter wavelengths. For an all-dielectric filter, the center shift as a function of incidence angle θ is given by

$$\lambda' = \lambda_0 \left[1 - \left(\frac{1}{n_{\text{eff}}} \right)^2 \sin^2 \theta \right]^{1/2}, \tag{6.1}$$

where λ' is the shifted center wavelength, λ_0 is the normal value, and n_{eff} is the effective refractive index of the filter (which must be obtained from the manufacturer). A typical variation of center wavelength with angle of incidence is shown in Figure 6.2, where the value of n_{eff} is 1.52.

Example. A typical 532-nm lidar receiver bandwidth is 1 nm or smaller. Figure 6.2 shows that the center wavelength shifts to 531.1 nm as the incidence angle increases from 0 to 5 degrees, which means that the filter is almost opaque to 532 nm laser light. This is the reason for having a collimating lens before the filter.

The range of ray angles on the filter can be easily calculated. Referring to the simplified receiver optical diagram (without a field lens) shown in Figure 6.3, the magnification M is given by

$$M = f_o / f_c \tag{6.2}$$

and the maximum incidence angle on the filter is given by

$$\theta_{\text{filter}} = M\theta_{\text{FOV}}. \tag{6.3}$$

Equations (6.2) and (6.3) show that, even though lidar receivers have narrow fields of view, the maximum incidence angle on the detector can become sizeable because the focal length of the objective is much larger than the focal length of the

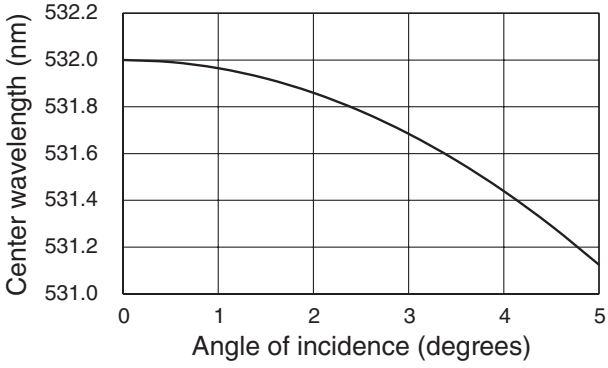


Figure 6.2 Effect of incidence angle. The curve shows the change of an optical filter's center wavelength with incidence angle.

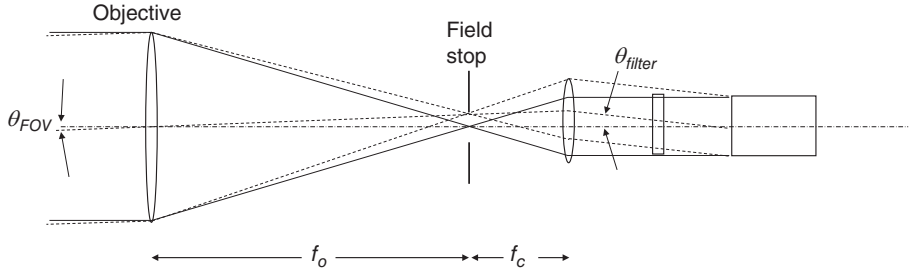


Figure 6.3 Receiver ray angles. The focal lengths shown and the FOV determine the maximum incidence angle of the received rays on the filter.

collimating lens. This fact, along with the dependence on incidence angle shown in Figure 6.2, means that controlling incidence angles is an important design consideration in lidar receivers. The collimating lens focal length can be made longer of course, but then the bundle of rays would have a larger diameter and a larger, more expensive filter would be required, so the choice of focal lengths involves engineering trade-offs.

Interference filters are also sensitive to temperature. As the filter changes temperature, the layer thicknesses and refractive indices change, and the most obvious spectral change is a shift in the center wavelength of the filter. A typical manufacturer's temperature coefficient is shown in Figure 6.4. The coefficient depends on center wavelength, and at 532 nm, it is 0.017 nm/°C in this example. For a laboratory environment where the year-round temperature is maintained at 23 ± 6°C, the center wavelength would vary by ±0.10 nm, which may be acceptable if the filter bandwidth is 1 nm. However, for a filter bandwidth of 0.1 nm it would be unacceptable, so narrow-band filters must be kept in temperature-controlled enclosures.

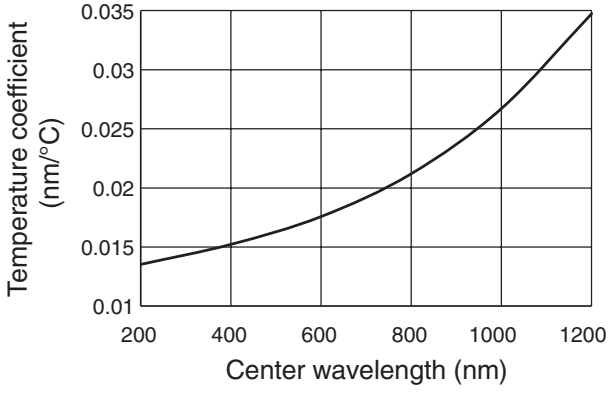


Figure 6.4 Effect of temperature. The solid curve shows the dependence of a typical filter's center wavelength on temperature.

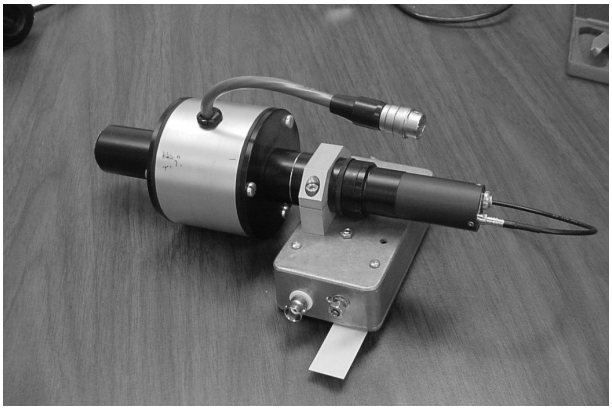


Figure 6.5 EARL's long-range receiver filter oven. The oven is the large cylinder at the left. The detector is at the right, and its amplifier is in the metal box. The collimating and Fabry lenses are inside the optical tubing.

Narrow-band receiver filters are often tuned by changing their temperature or angle. For example, the long-range receiver in EARL employs a filter with a bandwidth of about 0.14 nm, and it was fabricated with a room temperature center wavelength of about 523.2 nm. It is kept in an oven at a constant temperature of 44°C to shift its peak transmittance to the doubled Nd:YLF laser wavelength of 523.5 nm. The oven is shown in Figure 6.5, along with other components. Similarly, a receiver filter for a 355-nm water vapor Raman lidar at GTRI was fabricated with a center wavelength slightly too high, so that it could be tuned to 407.5 nm by tilting it. The calibration curve for the mechanical filter is shown in Figure 6.6, where normal incidence is at the peak. It is not necessary to know the tilt angle in absolute units if the angles are repeatable.

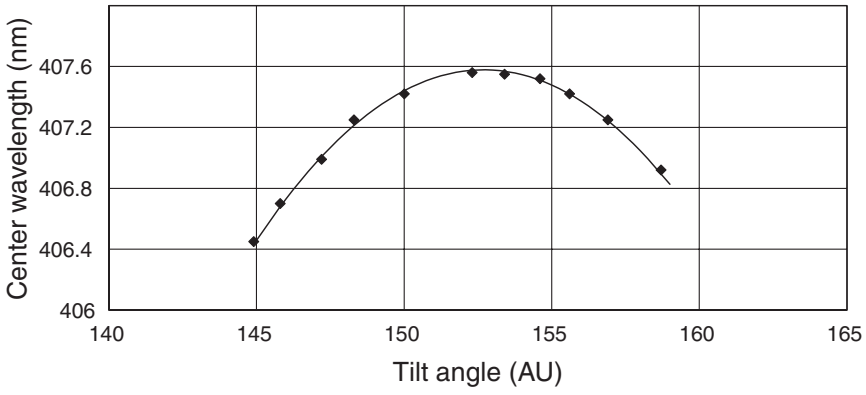


Figure 6.6 Angle tuning a filter. The diamonds are measurements of the center wavelength versus tilt angle (in arbitrary units) for a 355-nm water vapor Raman lidar receiver. The curve is a polynomial function fitted to the data.

The main characteristics of a receiver filter are that its peak transmittance should be at the desired wavelength and that it should have a narrow optical bandpass B_{opt}. Many narrowband filters exhibit one of two passband shapes, Lorentzian or Gaussian, where Lorentzian shapes are frequently used in the narrowest filters. Examples are shown in Figure 6.7. The filter functions (top row) very closely approximate the measured curves furnished by the filter vendor for the long- and short-range receiver channels in EARL, which operates at 523.5 nm. Filter widths are most often specified as FWHM, which leads to the question of what value to use for B_{opt} in the background model defined by Eqs. (2.13) and (2.14). Assuming that the sky radiance is constant across the narrow wavelength range of lidar receiver filters, the answer is found by integrating the area of each of these shape functions. A filter with a Lorentzian passband is equivalent to a rectangular passband filter with a width 1.53 times its FWHM and the same peak transmittance, and a filter with a Gaussian passband is equivalent to a rectangular passband filter with a width of 1.064 times its FWHM. The equivalent rectangular passbands for EARL are shown in the bottom row in Figure 6.7. In the background model, the peak filter transmittance is a factor in the receiver efficiency k_{R} and the equivalent rectangular width is used as B_{opt}.

There is one other important specification for lidar receiver filters: out-of-band blocking. Filters are not completely opaque, and they leak some light at wavelengths far from their centers. This can be a serious problem because detectors are usually sensitive over a wide spectral range, often hundreds of nm, and the background radiance is very broad too, especially in the UV-VIS region. For this reason, vendors specify a *blocking ratio*, generally as a power of 10, such as 10^{-4} or 10^{-5}, over a spectral band. The filter's transmittance is said to not exceed the blocking ratio anywhere in the specified band except near the filter's center wavelength. The effect of nonzero blocking on the lidar's SNR can be evaluated with a numerical convolution of the background radiance, the detector sensitivity, and the blocking ratio to compare

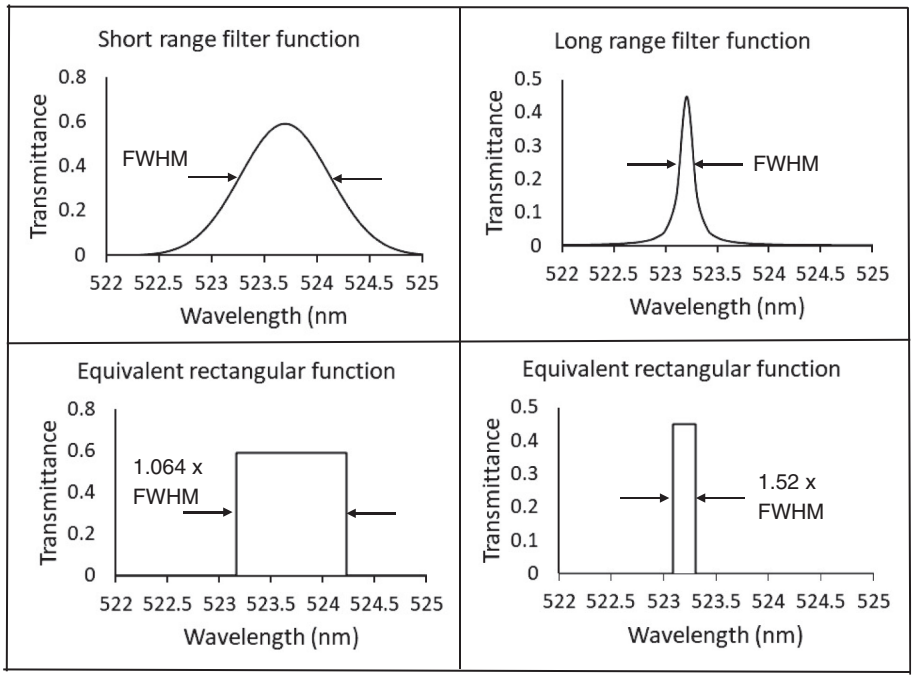


Figure 6.7 Filter bandpass shapes. The EARL receiver filter functions are shown in the top row, and their equivalent rectangular functions are shown in the bottom row.

the broadband background photon arrival rate with the laser photon rate in the filter passband. If the broadband rate is much smaller, the blocking ratio is sufficiently high.

Interference filters should be regarded as expendables because they degrade over time. One problem is fogging, in which water vapor enters the filter from the edges. EARL's long-range filter is a good example: After several years, the filter had reached the end of its useful life because the diameter of the un-fogged area in the center was the same as the received beam diameter. Filter transmission curves may also change with time. Transmission scans of EARL's short-range filter as delivered and several years later are shown as examples in Figure 6.8. The transmission at 523.5 decreased to about one-half of the original value and the FWHM increased by more than factor of 2 in six years. Interference filters should be periodically inspected for fogging and they should be re-scanned in a spectrometer.

In addition to interference filters, several other types are used in optics. Neutral density (ND) filters are used for attenuation, where "neutral" means that the attenuation is fairly constant across the visible spectrum and "density" means that the transmittance is equal to 10^{-ND}. An ND 2 filter has a transmittance of 0.01, for example. Some lidars achieve very high spectral resolution by using an atomic absorption line in a gas cell, and such a filter has the advantage that its spectral response is independent of incidence angle. A Fabry–Perot filter is a type of spectrometer that generates a series of equally spaced narrow passbands. In lidar receivers, atomic line

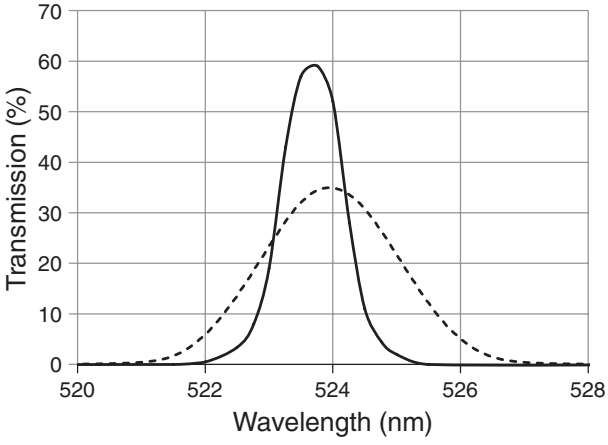


Figure 6.8 EARL's short-range filter change over time. The solid curve was recorded in April 2003 and the dashed curve in April 2009.

and Fabry–Perot filters are used in combination with other wider bandwidth filters to select a desired narrow spectral region and to block background light.

Optical efficiency considerations for the receiver are like those for the transmitter: the number of optical elements in the receiver should be kept to a minimum, mirror surfaces should be highly reflective, and other surfaces should be AR coated. A receiver's optical efficiency is often dominated by one element, the bandpass filter, which may have a peak transmittance of 0.5 or lower. For a simple receiver such as the one illustrated in Figure 6.1, the optical efficiency is typically in the range 0.25–0.5 for visible light. Example calculations for EARL's receiver are presented in Section 6.2.2.

6.2 Depolarization Lidar Receivers

Most lidars are based on linearly polarized lasers, so it is relatively easy to provide a polarized transmitted beam. To provide co-polar and cross-polar channels in the receiver as described in Chapter 4, polarization analyzers are used. The polarization analyzers used in lidar systems exploit refractive index changes at interfaces between different media. In general, when a beam of light is incident on a planar interface between media, as shown in Figure 6.9 where n_1 and n_2 are different refractive indices, part of the beam intensity is reflected and part of it is transmitted. All angles are relative to the surface normal, and their relationship is given by two rules: (1) the angle of reflection is equal to the angle of incidence and (2) the angle of the refracted ray is governed by

$$n_1 \sin\theta_1 = n_2 \sin\theta_2. \tag{6.4}$$

This relation is known as *Snell's law*. Because n_2 is assumed here to be greater than n_1, the refracted ray is bent toward the normal. The incident and reflected rays in Figure 6.9 define the *plane of incidence*. The fraction of the light that is reflected depends on the angle of incidence and the polarization state of the incident light. The state of linear

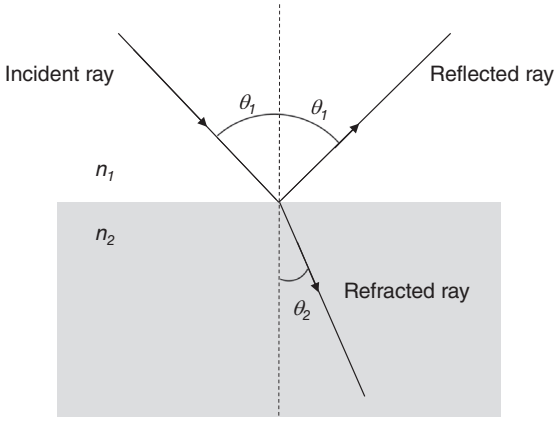


Figure 6.9 Reflection and refraction at an interface.

polarization perpendicular to the plane of reflection is called S, and polarization in the plane is called P. The reflectances R_P and R_S are given by the *Fresnel coefficients* [2] as

$$R_P = \left| \frac{n_1 \cos\theta_1 - n_2 \cos\theta_2}{n_1 \cos\theta_1 + n_2 \cos\theta_2} \right|^2 \text{ and} \tag{6.5}$$

$$R_S = \left| \frac{n_1 \cos\theta_2 - n_2 \cos\theta_1}{n_1 \cos\theta_2 + n_2 \cos\theta_1} \right|^2 . \tag{6.6}$$

Reflection is caused by a change in refractive index from one medium to the next. If n_2 is equal to n_1, Snell's law says that θ_2 is equal to θ_1, in which case Eqs. (6.5) and (6.6) predict zero reflectance. Examples of the reflectance coefficients are illustrated in Figure 6.10 for a common type of optical glass, Schott BK7, which has a mid-visible refractive index of 1.52. The S and P reflectances are significantly different except at normal and grazing incidence. In particular, the P polarization goes to zero at *Brewster's angle* θ_{Brewster}, which means that all the light is transmitted. Brewster's angle is defined as

$$\theta_{\text{Brewster}} = \arctan\left(n_2 / n_1\right) \tag{6.7}$$

and its value is 56.7 degrees for BK7. In early gas lasers, the cavity mirrors were outside the gas cells, so the cell windows were tilted at Brewster's angle to eliminate reflective losses. Crystalline rods in modern solid-state lasers are also sometimes cut at Brewster's angle.

The reflectance curves shown in Figure 6.10 have important implications for lidar receiver design: (1) both polarizations experience a loss from the first interface surface of about 4% at normal incidence and the same loss occurs for rays leaving the higher-index material. As a rule of thumb in optical design, reflectance losses on uncoated elements are 4% per surface, so a simple window will cause an 8% loss, which shows the importance of AR coatings and (2) it is easy to introduce accidental polarization sensitivity into a receiver when optical elements with off-normal incidence are used, such as the secondary mirror in a Newtonian telescope. Such sensitivity can cause an error in the measured backscatter from depolarizing particles. It is generally possible to specify and purchase mirrors, including dichroics, that are insensitive to polarization state.

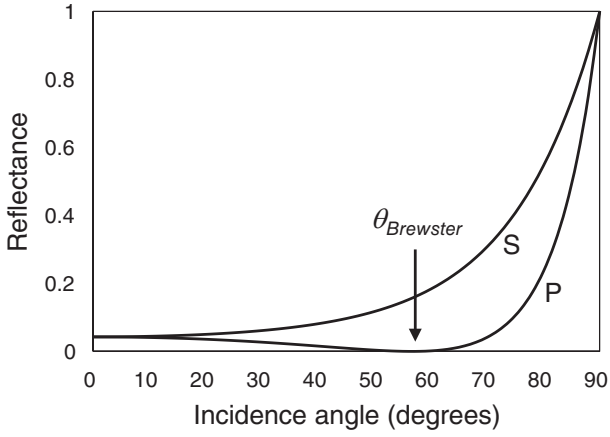


Figure 6.10 Mid-visible Fresnel coefficients for BK7 glass.

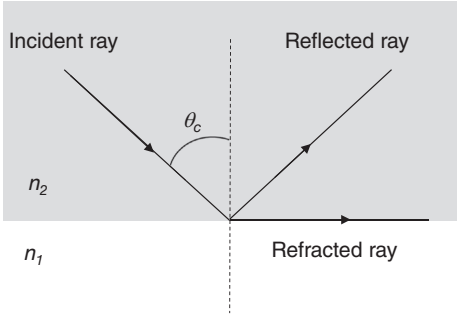


Figure 6.11 The critical angle. At the critical angle, the refracted ray is along the interface between the media. At higher angles, the incident ray is totally reflected.

For rays inside the denser medium, another useful phenomenon known as total internal reflection (TIR) occurs at the interface, as shown in Figure 6.11. The critical angle θ_c is given by

$$\theta_c = \sin^{-1}(n_2 / n_1). \tag{6.8}$$

At the critical angle, the refracted ray is in the direction of the interface boundary. At greater angles of incidence, there is no refracted ray, and the light is trapped within the denser medium. This phenomenon is often exploited in lidar systems. For example, optical fibers rely on TIR, and fiber bundles are sometimes used in high-altitude lidars to couple the light from a receiver telescope system to a remote detector, so that the detector does not have to be located near the telescope's focal plane, or to collect the light from multiple receiver telescopes onto one detector. Another example is beam steering prisms. Sometimes it is necessary to turn the raw beam from a high-power laser through 90 degrees. Metallic mirror coatings would be destroyed by the beam, so a prism is often used as shown in Figure 6.12. Because there is no mirror

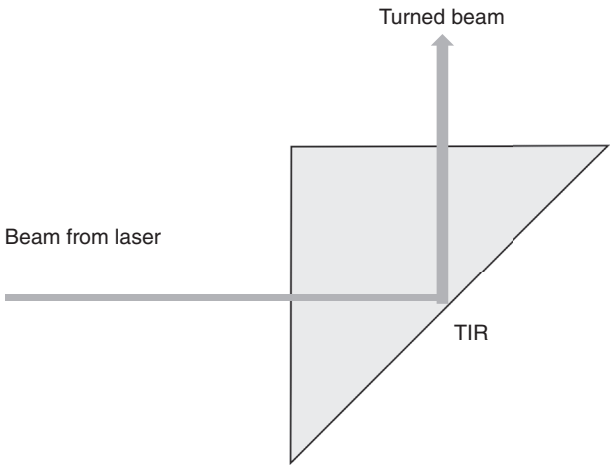


Figure 6.12 A beam steering prism. The phenomenon of TIR can be used to change the beam direction without a reflective coating.

coating involved, this type of beam steering prism is extremely resistant to optical damage. Such prisms are available for a range of wavelengths that covers the usual Nd:YAG harmonics of 1,064, 532, 355, and 266 nm. Two of them are used in the EARL transmitter, as shown in Figure 5.21.

6.2.1 Polarization Analyzers

Precision polarization analyzers make use of the optical phenomena described above plus one more: *birefringence*, which means that the material has a refractive index that depends on the polarization state and propagation direction of light. Crystals with non-cubic crystal structures, such as calcite ($CaCO_3$) are often birefringent. Many types of polarizing prism have been developed over the years, and each is named after its inventor(s). The prisms are designed to split unpolarized incident light into two polarized beams. What they have in common is that they are constructed from two prisms that are either cemented together or separated by an air gap [3]. In the Glan–Taylor prism, shown in Figure 6.13, two calcite prisms are separated by an air gap. The angles are chosen such that the S-ray experiences TIR at the calcite–air interface and is deflected, but the P-ray does not, because it experiences a smaller refractive index and therefore a larger critical angle, given by Eq. (6.8). The Glan–air prism is similar, but its crystal axis is orthogonal to the Glan–Taylor and the S and P rays are reversed. The deflected ray is not at 90 degrees to the undeflected ray (it is usually at 108 degrees) and its polarization purity is not as high, although the Glan–Taylor prism has better purity in the deflected ray than the Glan–air. Because the S and P rays from these prisms are not at right angles, a special structure must be machined to hold two photodetectors at the proper angles in a lidar receiver. Much of the lidar depolarization work over a period of decades was done using such fixtures [4]. The electronic gains in the detectors and amplifiers for the two

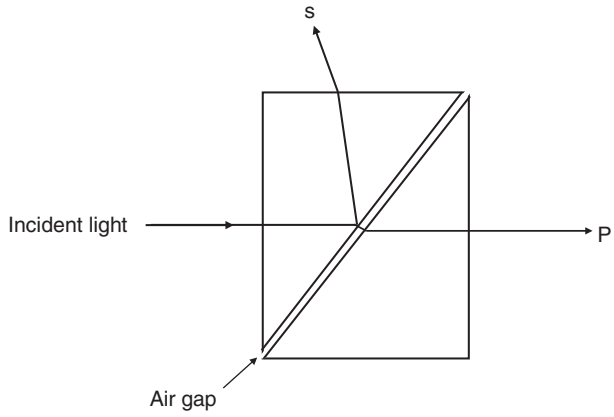


Figure 6.13 The Glan–Taylor prism. The Glan–Taylor polarization analyzer consists of two prisms with an air gap between them. Unpolarized incident light is split into S and P components, which emerge at different angles.

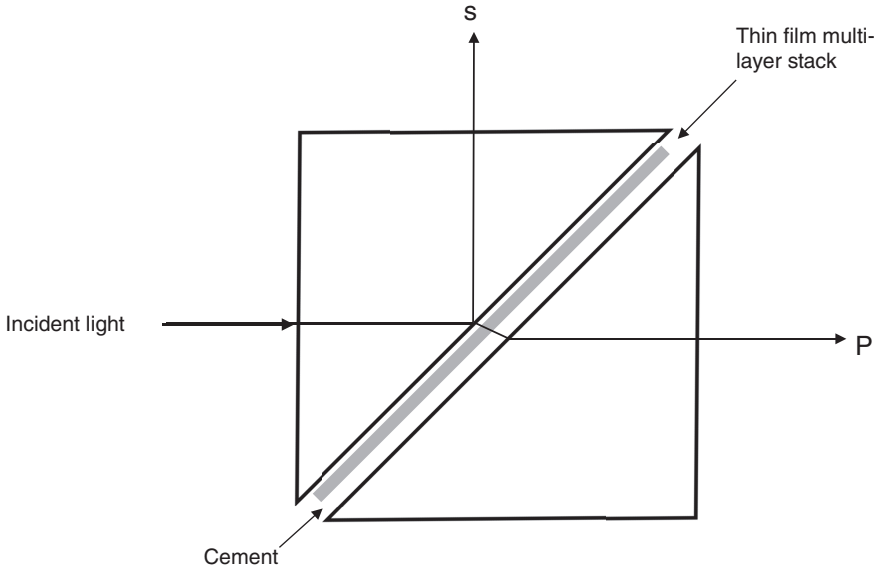


Figure 6.14 A polarizing beamsplitter cube. The two beams emerge at right angles to each other.

polarizations will not be identical, so a relative calibration must be established before calculating the quantities such as d and δ, which are defined in Chapter 4.

Nowadays, prisms that send the two beams out at right angles are commonly available; they are known as polarizing beamsplitter cubes. These devices are similar in appearance to the Glan–air prism, but they are manufactured with a multilayer dielectric coating on the hypotenuse of one of the component prisms, which are then cemented together, as shown in Figure 6.14. The disadvantage of the polarizing beamsplitter cube is that the two exiting beams do not have equal polarization purity. The S beam may have a PER of 1,000 and the P beam 100, for example. There are two ways

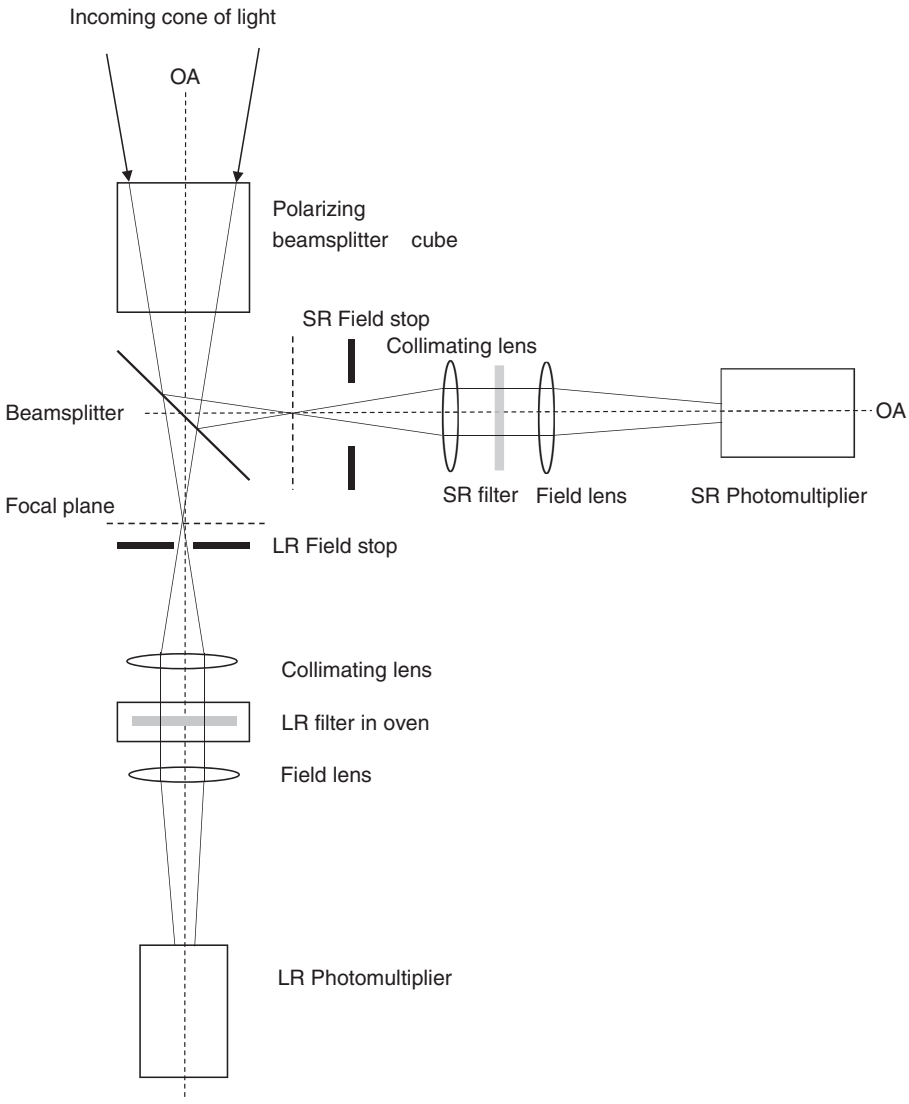


Figure 6.15 The EARL receiver. The optical axes of both channels are shown by the light dashed lines labeled OA. The receiver telescope's focal plane is shown by the heavy dashed lines. SR – short range; LR – long range.

to overcome this problem: (1) pass the deflected beam through a second cube that is rotated 90 degrees, so that the final beam has a PER of 1,000; or always use only the undeflected ray in the receiver and switch the polarization states back and forth in the transmitter. This latter solution was implemented in EARL, whose receiver components are diagrammed in Figure 6.15. This approach does not require relative calibration of two separate receiver channels, which is a big advantage. It does double the data acquisition time, but that drawback was deemed to be insignificant in the undergraduate teaching and research environment that EARL was designed for.

Table 6.1 EARL receiver geometrical parameters

Parameter	Short-range channel	Long-range channel
Telescope focal length	2.57 m	2.57 m
Fraction of received light	0.09	0.91
Field stop diameter	3 mm	1 mm
FOV	1.17 mrad	0.389 mrad
Field stop offset	3.5 mm	0.5 mm

6.2.2 The EARL Receiver

In the EARL receiver, the cone of partially depolarized light from the receiver telescope enters through a polarizing beamsplitter cube, so the polarization is always orthogonal to the plane of the diagram in Figure 6.15 from that point on. The light then encounters a beamsplitter that diverts 9% of the light to the short-range receiver channel and 91% to the long-range channel. Some of EARL's receiver parameters are listed in Table 6.1. The arrangement of optical elements in both channels is shown in Figure 6.1, and the filter functions are shown in Figure 6.7. All the components shown except for the detectors are in one assembly made from commercial optical hardware. The short-range and long-range channels were co-aligned during assembly. Alignment of the receiver with the transmitted beam is achieved with angle adjusters in two planes on the telescope's secondary mirror, which is a front-aluminized 20-cm flat as shown in Figure 2.8. The lidar system is said to be *aligned* when the cones of received light are centered on the field stops.

EARL's short-range channel generates higher signal levels than the long-range channel even though it only uses 9% of the received light, because of the $1/R^2$ factor in the lidar equation. The short-range background signal is comparable to the long-range, because the FOV plane angle is three times larger and so the solid angle is nine times larger (see Section 2.3), which compensates for the factor of 10 difference due to the beamsplitter. A low-cost filter with a FWHM of 1 nm was thought to be adequate because the SNR of the larger lidar signal is less degraded by background light, as shown by Eq. (2.8). During data analysis, the low-cost filter proved to be a mistake, as documented in Chapter 10. It should have been like the long-range filter.

It is instructive to calculate the optical efficiencies k_R of the EARL receiver channels. Best estimates of the as-built transmittance and reflectance values for the elements shown in Figure 6.15 are listed in Table 6.2, along with their products, which are the optical efficiencies. Several comments on Table 6.2 are in order: The coaxial transmitter blocks 23% of the received light, because it is larger than the secondary mirror (which would have blocked only 11%); the largest losses are at the interference filters; and adding the two optical efficiencies yields 0.211, which means that 79% of the photons that would have fallen on the primary mirror do not reach a detector! It is easy to lose photons in a lidar receiver.

The receiver components shown in Figure 6.15 are mounted in optical tubes and cubes in such a way that they can be rotated about the optical axis, so it was necessary to find the correct rotational orientation for depolarization measurements. The

Table 6.2 Optical losses in EARL receivers

Element	Long range	Short range
Blockage by transmitter	0.77	0.77
Primary mirror	0.91	0.91
Secondary mirror	0.91	0.91
Polarizing cube	0.98	0.98
Beamsplitter	0.91	0.09
Lens 1	0.92	0.92
Lens 2	0.92	0.92
Window 1	0.92	–
Window 2	0.92	–
Filter	0.45	0.59
Optical efficiency k_R	0.183	0.028

basic theory of lidar depolarization by randomly oriented small particles using Stokes vectors and Mueller matrices is in Section 4.4. After that theory was presented to the lidar community in 2008, Hayman and Thayer developed a general description of the entire measurement system that they called the Stokes Vector Lidar Equation [5]. Their method includes everything affecting the polarization state of the laser light, starting at the laser, going through all the transmitter optics to the scattering volume, back through all the receiver optical components, and finally, to the detector. Many subtleties and potential pitfalls were elucidated in the paper, which suggested that the lidar community had perhaps not been sufficiently careful about depolarization measurements and the errors in them. EARL provides examples of many potential sources of such errors. The rotational alignment between transmitter and receiver was determined in clear sky conditions, where the depolarization is very small and so the cross-polar signal is very small. The signal through two crossed polarizers is proportional to the cosine squared of the angle between them (this fact is known as *Malus' law*), so the solution was to rotate the receiver in one direction by a large angle to get a large cross-polarization signal, and then rotate it back though the minimum to the angle in the other direction where the cross-polarization signal reached the same level. The correct angle was then taken as being halfway between the two extremes. EARL could easily discriminate between liquid drops and ice crystals in clouds, which was deemed adequate for a teaching lidar, but it did not provide an estimate of the error in depolarization measurements, and there were many possible sources of less-than-ideal performance. As shown in Figure 5.21, light from the laser encounters 11 optical elements before it is transmitted into the atmosphere. Many of them are axially symmetric and perhaps not depolarizing, but three of them have surfaces at 45 degrees, so the linear polarization state of the transmitted light may not be as pure as that of the laser, which was not measured. In the receiver, shown in Figure 6.15, the light encounters just two optical elements before entering the polarizing beamsplitter cube, but it enters as a converging *f*/4.2 cone of light. The cube has a polarization extinction ratio of 1,000, but that specification is probably for normal incidence. Any error in the

transmitter-to-receiver rotational alignment will also contribute to the overall error in depolarization measurements. One simplifying feature of EARL is that the same receiver is used for the co-polar and cross-polar measurements, so there is no need for a relative calibration between channels. That feature comes with the cost of a doubled measurement time, so most research lidars do have two separate channels and hence they require a relative sensitivity calibration in addition to the rotational calibration.

Historically, most lidar stations were homebuilt, and their designers implemented depolarization sensitivity in several different ways with a variety of calibration schemes. This situation made comparison of results from different stations problematic, which was especially important in lidar networks. In 2016, to address this problem for EARLINET, Freudenthaler presented a comprehensive and general analysis of lidar depolarization for a wide-ranging set of depolarization lidar implementations [6]. His paper provided a model based on a description of the lidar optical elements in the Mueller–Stokes formalism and the polarization state of the laser light. Equations were derived for the dependence of the lidar signal on polarization parameters and the linear depolarization ratio for many different polarization setups applicable to a large variety of lidar systems. The model enabled the calculation of systematic errors from nonideal optical elements as well as errors from rotational misalignment and impure laser polarization states. The paper also advocated a calibration procedure known as Δ90 and provided recommendations for overall best practices in depolarization lidar. This paper, its Supplement, and references therein constitute a comprehensive resource that should be consulted during the design phase of a polarization-sensitive lidar system and during upgrades to such systems. It is often said that geophysical data are almost meaningless without reliable error estimates, and there is now no reason that lidar depolarization data should not have them.

6.3 The Geometrical Function

In this section, a simple geometrical optics analysis is applied to the overall lidar optical system, which includes the transmitter and the receiver. Both subsystems have important optical parameters, such as the beam divergence of the transmitted laser beam and the field of view of the receiver, and some of the properties of the geometrical function $G(R)$ depend on the transmitter and receiver parameters in ways that can be elucidated through simple, approximate derivations. Lidar optical design considerations are discussed, and a set of simple equations and a graphical presentation are derived for finding the range R at which $G(R)$ becomes equal to unity. These equations can also be used to find the lidar system's tolerance to misalignment. At the end of this section, a few words of caution are given regarding an erroneous treatment of $G(R)$ that is common in the lidar literature.

6.3.1 Description

The shape of a typical lidar signal results from the product of two terms in the lidar equation, as shown in Figure 6.16. The term $1/R^2$ is due to scattered light diverging from the scattering volume and so it cannot be changed, but the second term, $G(R)$, is

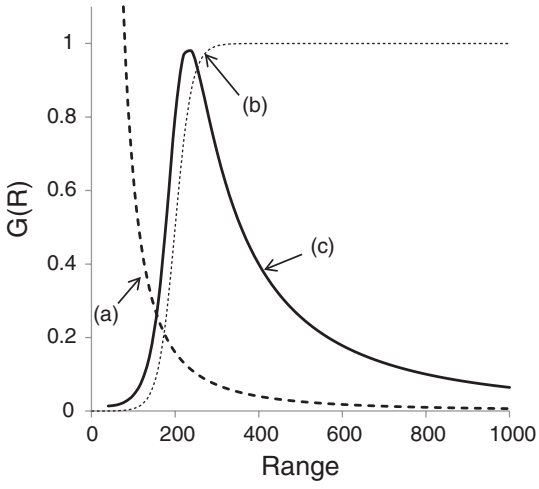


Figure 6.16 The lidar signal shape. The product of the $1/R^2$ factor (a) in the lidar equation with the geometrical function $G(R)$ (b) determines the basic shape of a typical lidar signal (c). Curves (a) and (c) have been scaled up, for clarity. The peak of the lidar signal does not occur at the range where $G(R)$ becomes unity; it occurs at a shorter range.

determined by the lidar system's optical parameters and so it can be engineered. For this reason, $G(R)$ must be understood in detail. $G(R)$ is often designated $O(R)$ in the lidar literature, and it is sometimes called the lidar "overlap" function for reasons that are discussed at the end of this chapter.

The most important reason why $G(R)$ must be carefully engineered is to avoid artifacts in the lidar data. In other words, $G(R)$ must be controlled so that the signals from the lidar system will obey the lidar equation. If they do not, an accurate analysis of the lidar data may not be possible, depending on the application. Analyzing data only from ranges where $G(R)$ has reached unity is a common practice (standard at GTRI), and in this case, the lidar's minimum range must be established. It is often desirable to make the minimum range as short as possible. At the other extreme, a minimum range as large as several kilometers is sometimes used to suppress the close-in signal and thereby reduce the dynamic range of the overall signal. The shape of the $G(R)$ curve shown in Figure 6.16 is generally very sensitive to misalignment and temperature-induced dimensional changes, and a way of calculating the lidar's tolerance to these effects is also very useful for lidar engineering. Finally, the value of the range where $G(R)$ first becomes greater than zero is somewhat sensitive to misalignment and it is measurable, so it can be a useful diagnostic when checking alignment.

Unfortunately, the function $G(R)$ is almost impossible to calculate accurately enough for detailed analysis of a lidar signal at ranges where it is not unity, but some important features of $G(R)$ can be related to optical design parameters in a straightforward way. In particular, the range R_i (initial) at which $G(R)$ departs from zero and the range R_f (final) at which $G(R)$ becomes equal to unity, illustrated in Figure 6.17, can both be calculated from the lidar's design parameters. The region between R_i and

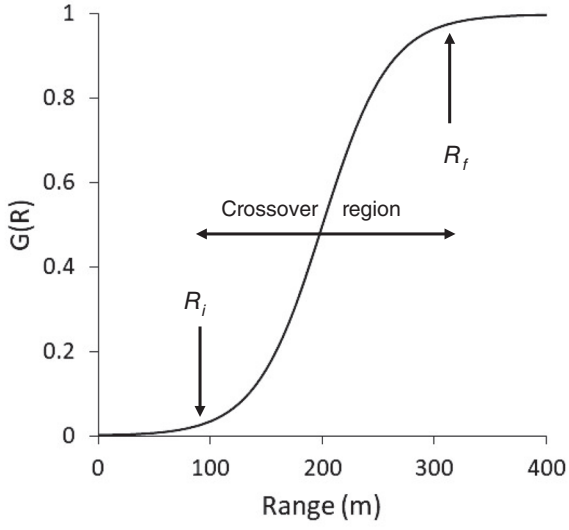


Figure 6.17 The shape of a typical lidar geometrical function. The function departs from zero at the initial range R_i and becomes unity at the final range R_f. Intermediate ranges are known as the crossover region.

R_f is known as the *crossover region*, for reasons discussed later. A simple sigmoid curve is plotted in Figure 6.17 for purposes of illustration. The curve was generated with the sigmoid function

$$G(R) = \frac{1}{1+\exp\left[\dfrac{-10(R-R_{50})}{R_{OL}}\right]}, \tag{6.9}$$

where R_{50} (200 m in this example) is the range where the function's value reaches 0.5 and R_{OL} (300 m) is the range where it becomes close to unity.

6.3.2 Geometrical Optics Review

Geometrical optics is the study of light through ray propagation without diffraction or coherent interference. The *rays* are normal everywhere to the wavefronts of propagating light waves. Some of the angles of the light rays in optics diagrams are exaggerated for clarity in this chapter, but in practice, most of the angles in a typical lidar system are quite small, on the order of 1 mrad. This fact enables use of the *paraxial approximation*: $\sin(\theta) = \tan(\theta) = \theta$, and $\cos(\theta) = 1$. With this approximation, trigonometric functions are not needed, and the analysis is based on the properties of triangles. Laser beams are treated as if they have a very sharp edge, so that the edge of a beam can be represented by a ray. The designer must relate the results of the ray optics analysis given here to his actual laser beam profile by choosing a definition for the edge of the beam, such as 1.52 times the radius at the $1/e^2$ point, as recommended

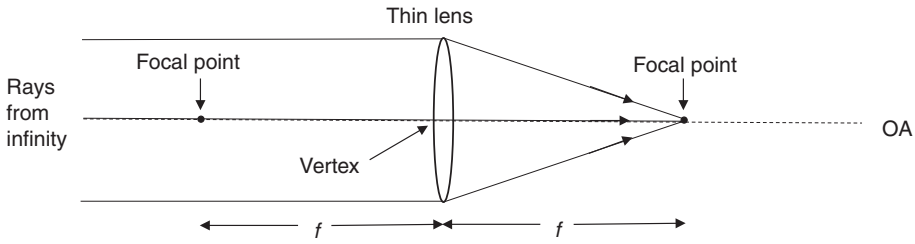


Figure 6.18 The thin lens.

in Chapter 5. Other simplifying approximations used in this chapter are noted as they occur.

For the analysis of lidar systems with geometrical optics, each optical element is assumed to have axial symmetry about its optical axis (OA), and the following simple rules and definitions are used:

(1) Rays that go through an optical system in one direction can also go in the opposite direction (this is called *reversibility of rays*).
(2) A thin lens is assumed to have zero thickness. It has a diameter and a focal length.
(3) The focal length f is the distance from the lens where rays from an object at infinity come together to form an image, as illustrated in Figure 6.18.
(4) The points on the OA at one focal length from the lens are called the *focal points*.
(5) The point where a ray of light strikes the center of the lens is called the *vertex*.
(6) A ray through the vertex is not deviated.
(7) A ray parallel to the OA goes through the lens and then through the focal point on the other side. Conversely, a ray through the focal point goes through the lens and then parallel to the OA.

Several of these definitions are illustrated in Figure 6.18, which shows rays from infinity converging at one of the focal points, which are on the OA and one focal length away from the lens. The vertex is at the center of the thin lens and on the OA. In optical diagrams, rays are conventionally drawn from left to right. The plane perpendicular to the OA at the focal point is called the *focal plane*. The ability of a lens to focus light is characterized by its focal ratio, which is the ratio of the focal length to the lens diameter. The focal ratio is called the *f-number*, which is written with an italicized f and a solidus. For example, if the focal length is twice the lens diameter, the f-number is *f*/2. Lenses with smaller f-numbers are said to be *faster* than those with larger f-numbers.

The full-angle FOV of a system is found as shown in Figure 6.19. Two rays are drawn from the edges of the field stop aperture through the vertex. Because those rays are not deviated, the angles between them on the left and right sides of the lens are *vertical angles*, so they are the same. Reversing the ray direction then shows that the system admits light from a cone of rays with the cone angle $\theta_{\mathrm{FOV}} = D_{\mathrm{FS}}/f$, as mentioned in Section 6.1.1.

As an object moves closer to the lens, the image moves farther away. Figure 6.20 illustrates the graphical method of finding the image position and size by using the

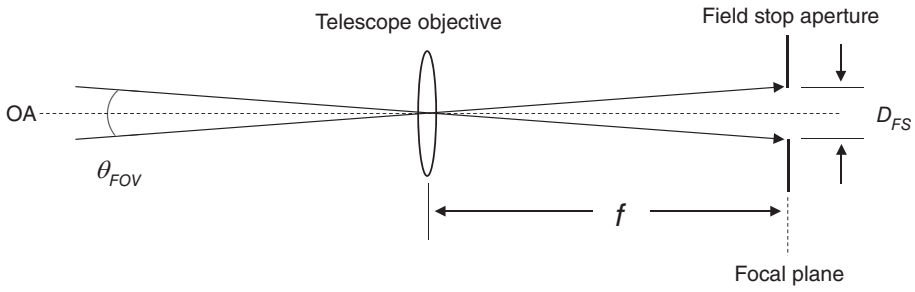


Figure 6.19 Finding the FOV.

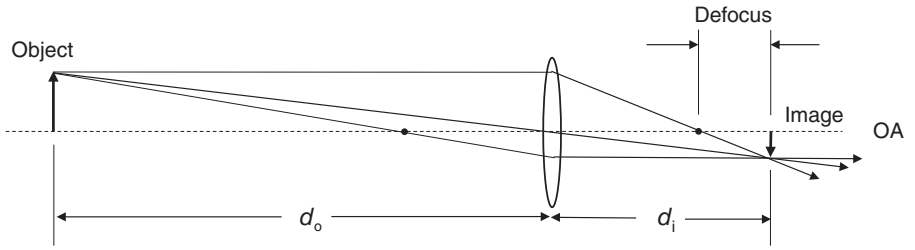


Figure 6.20 Finding image size and location by ray tracing. The parameter d_o is the object distance and d_i is the image distance.

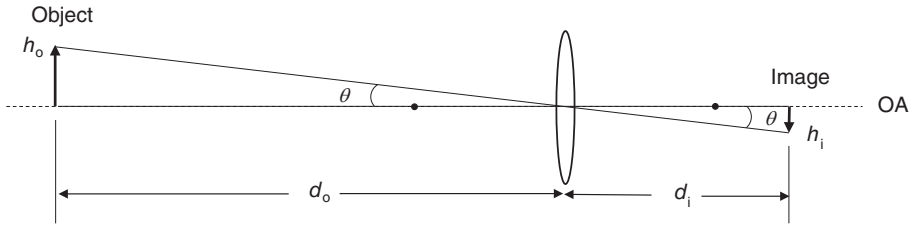


Figure 6.21 Finding image offset in lidar systems.

three rules of ray tracing through lenses. The point where the three rays converge is on the object. In Figure 6.20, the rays are drawn from the tip of the object, so they intersect at the tip of the image. In this figure, the image is closer to the focal point than the object is; it is smaller than the object, and it is on the other side of the OA.

In lidar, the object is a pulse of laser light. As the pulse travels upward through the atmosphere, its image shrinks and moves relative to the receiver optics, so the image size and its location as functions of range must be known. The image size can be calculated by using the magnification M defined by

$$M = d_i / d_o, \tag{6.10}$$

but this relation is not particularly useful in lidar because d_o is so much greater than d_i. An equivalent method based on similar triangles is illustrated in Figure 6.21, where h_o / d_o is equal to h_i / d_i, so the size of the image h_i is simply $d_i \theta$ where θ is the angular subtense θ of the object, which is h_o / d_o. The parameter h_i can also be thought of as the

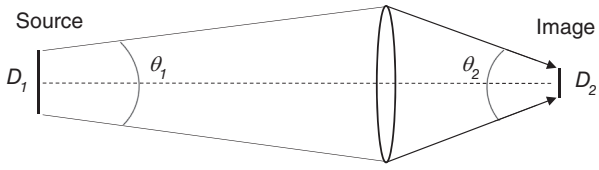


Figure 6.22 The conserved optical quantity. The product of image diameter and cone angle is conserved as light propagates through an optical system.

location of the center of the image, in which case the same relation is used to find the lateral offset of the image from the OA.

The image location along the OA can be calculated with the *thin lens equation*, which is

$$\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_i}. \tag{6.11}$$

Solving for the image distance,

$$d_i = \frac{f d_o}{(d_o - f)}. \tag{6.12}$$

The defocus distance, or simply *defocus* $(d_i - f)$ in Figure 6.20, is of special interest in understanding $G(R)$. It is given by

$$d_i - f = \frac{f d_o}{(d_o - f)} - f = \frac{f^2}{(d_o - f)} \approx \frac{f^2}{d_o}, \tag{6.13}$$

where the approximation at the end is quite accurate in lidar, because the measurement range is always much greater than the receiver telescope's focal length.

Optical systems have a conserved optical quantity known in French as the *étendue* and in English as the area-solid angle product or equivalently as the diameter-angle product (it corresponds to the BPP in laser beams). That product is conserved in an optical system. As illustrated in Figure 6.22, $D_1\theta_1 = D_2\theta_2$. This conservation law is at the heart of the Brightness Theorem in photometry and the Radiance Theorem in radiometry, which state that the brightness or radiance of a source cannot be increased by any combination of lenses and mirrors. If an optical diagram appears to violate this conservation law, it is incorrect.

6.3.3 Engineering the Function

The qualitative features of $G(R)$ that are illustrated in Figure 6.17 stem from the need to restrict the amount of background light that reaches the detector. As shown in Chapter 2, the effect of background photons in lidar signals is to decrease the SNR of the measurements, sometimes drastically. For this reason, the receiver's field of view is restricted with a field stop, to exclude as much background light as is practical. However, the field stop also prevents some of the backscattered laser light at short ranges from reaching the detector. The way that a biaxial lidar optical system causes $G(R)$ to start at zero, rises smoothly, and then stays at unity is illustrated in

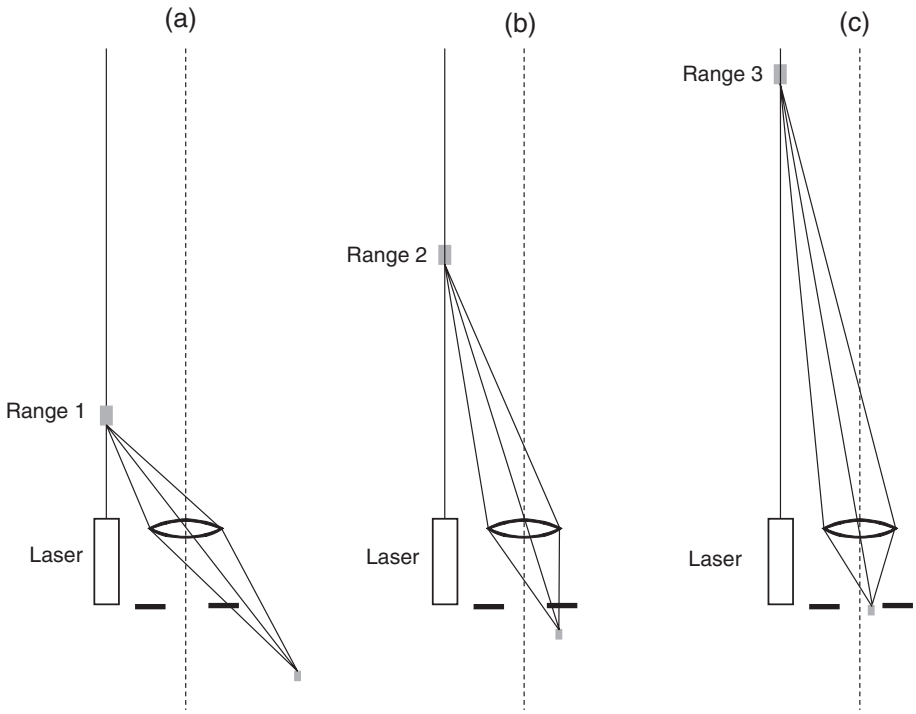


Figure 6.23 The causes of the geometrical function. The laser pulse is shown at three ranges, and the lidar receiver is shown as a lens. In (a), none of the light passes through the field stop because of defocus and lateral displacement of the image; in (b), at a larger range, some of the light passes through; and for large enough ranges such as in (c), all the light passes through.

Figure 6.23, which shows how the receiver forms images of the scattering volume at different ranges (in many of the lidar diagrams in this chapter, the light rays go from top to bottom). The laser beam is parallel to the receiver OA, so at very large ranges, the image is in the focal plane and centered on the field stop, and all the backscattered light collected by the receiver reaches the detector. However, at Range 1, none of the light passes through the field stop, because of defocus and lateral displacement of the image, so the received signal is zero. At Range 2, the image is not displaced as much, and it is closer to the focal plane, so some of the rays go through the field stop to the detector. For ranges such as Range 3 and larger, all the light passes through, meaning that $G(R)$ has reached unity. Another way of visualizing the range-dependent sensitivity is to imagine a view inside the system shown in Figure 6.23 facing the field stop, which is in the focal plane. The appearance of the backscattered laser light in the focal plane is shown in Figure 6.24. At Range 1, the light appears as a large, defocused spot that is displaced laterally from the field stop. As the scattering volume moves to the larger Range 2, the image in the focal plane shrinks and moves toward the field stop aperture. Eventually Range 3 is reached, beyond which all the light goes through the stop. Because the image crosses over into the field stop aperture, the term *crossover*

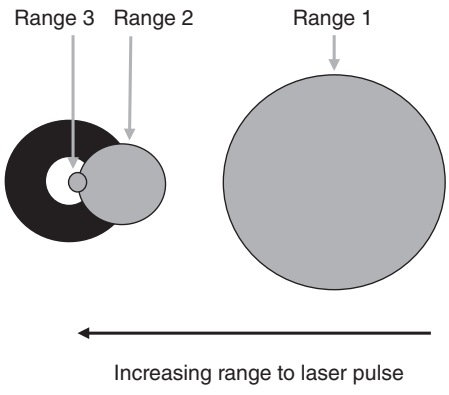


Figure 6.24 The field stop view. The appearance of the received light in the plane the field stop of a lidar receiver, for the three ranges illustrated in Figure 6.23. The field stop is the black annulus, and the gray circles are the received laser light.

is used here, to describe the phenomena that occur at the field stop and determine the shape of $G(R)$. This qualitative discussion illustrates that the general shape of the geometrical function $G(R)$ is caused at the field stop aperture by defocus and lateral image displacement due to the transmitter–receiver offset. In the following sections, a simple set of equations is derived for properly engineering $G(R)$ using all the optical parameters that are available. In optical design, a ray tracing computer application should always be used as the final step, but the equations derived here are adequate for understanding $G(R)$ and for performing many of the important engineering tradeoffs in lidar optical design.

The discussion above also sheds light on a common problem in lidar – the phenomena of defocus and lateral image displacement will always cause $G(R)$ to look much like Figure 6.17, even if the lidar is seriously out of alignment, or the field stop is improperly located relative to the focal plane, or there are other optical problems. Consequently, with the $1/R^2$ range dependence factored in, the lidar signal will always have the general shape shown in Figure 6.16: It will rise rapidly to a peak and then decrease more gradually with increasing range. The signal lacks features that the operator can use to determine that the lidar is operating properly. Therefore, the lidar's optical system must be designed and constructed carefully, and a reliable alignment procedure must be developed, so that the operator can have confidence in the data that the lidar produces.

Lidar Optical Design Parameters

In this section, all the lidar system's optical parameters that influence $G(R)$ are elucidated, and engineering tradeoffs involving these parameters are discussed. In the remainder of this chapter, a general way of finding R_i and R_f is developed for all the lidar system configurations shown in Figure 5.1, based on the receiver and

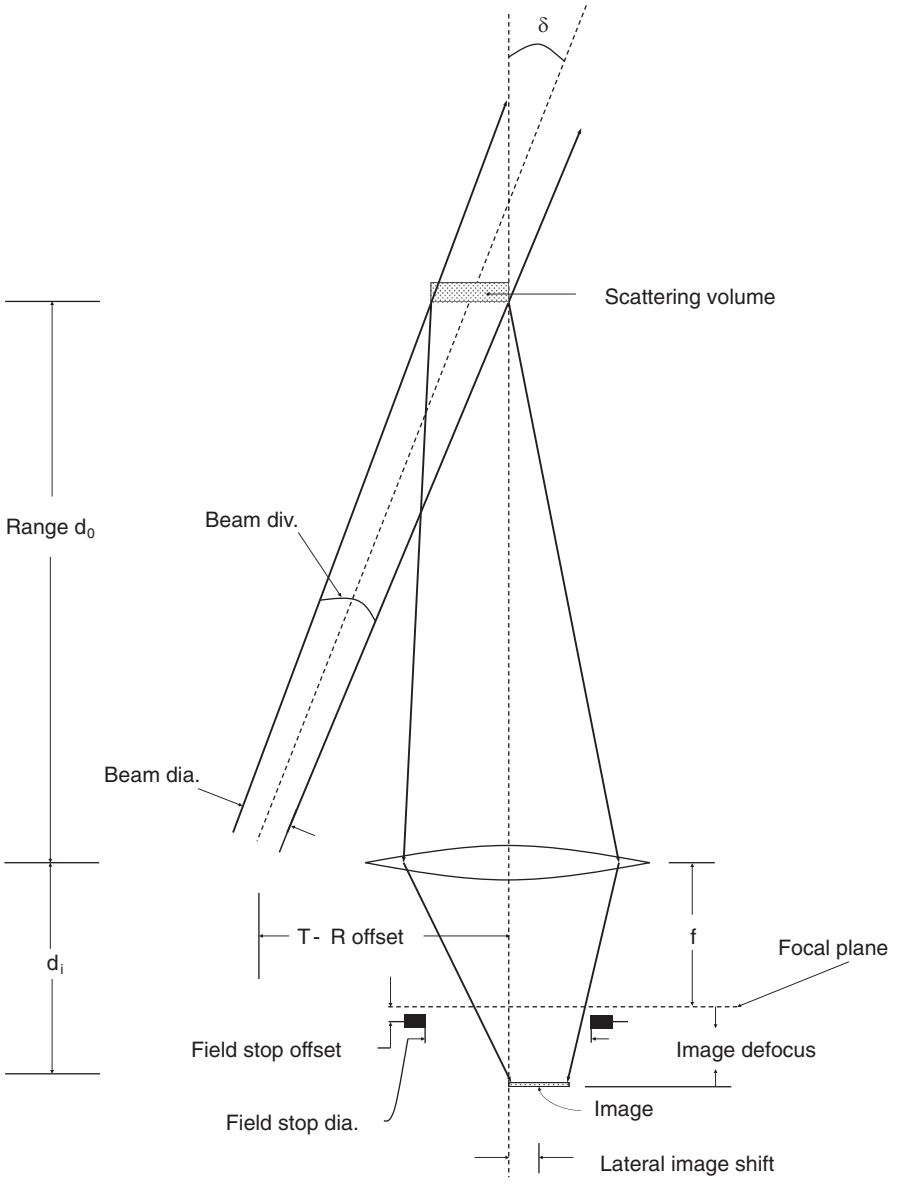


Figure 6.25 The lidar optical system. Seven design parameters are available for engineering the geometrical function $G(R)$ in a biaxial lidar.

transmitter parameters, their relative alignment, and the lateral offset of their optical axes. The transmitter contributes two design parameters, the laser beam diameter and its divergence. The receiver contributes its focal length, the FOV, and the field stop offset distance. In addition, the transmitter–receiver alignment angle is an adjustable parameter. The transmitter–receiver combination for the biaxial configuration is shown in Figure 6.25, in which one more design parameter is included: the

Table 6.3 Lidar optical system parameters for engineering $G(R)$

Parameter	Symbol	Dimensions
Transmitted beam diameter	D_T	length
Transmitted beam divergence	θ_T	angle
Receiver diameter	D_R	length
Receiver FOV	θ_{FOV}	angle
Field stop offset	d_{FSO}	length
Transmitter–receiver offset	d_{TR}	length
Transmitter–receiver alignment angle	δ	angle

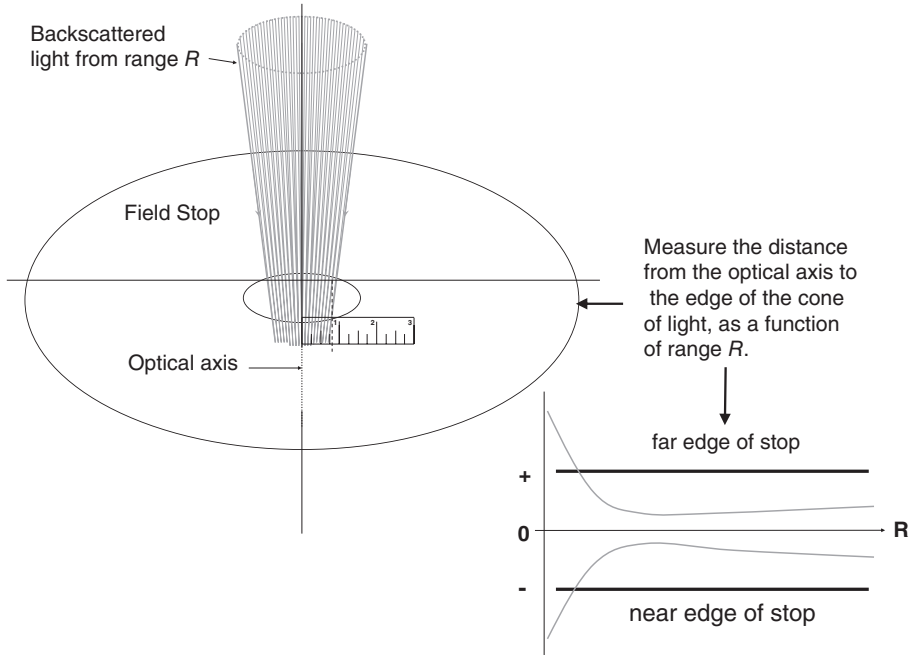


Figure 6.26 The crossover plot. Conceptually, the radial distances of the edges of the received cone of light from the OA are measured and plotted versus range, as the laser pulse propagates.

transmitter-to-receiver offset distance (the value of this parameter is zero for coaxial and common optics systems). A maximum of seven design parameters is therefore available for engineering $G(R)$: two for the transmitter, three for the receiver, the transmitter–receiver alignment angle, and the transmitter–receiver offset distance. These parameters are listed in Table 6.2. Note that a capital D in the diagrams denotes a diameter, whereas a lower case d always refers to a distance.

The Crossover Model

The approach taken here is to first consider the simplest possible system, which is a coaxial lidar with no misalignment and with the field stop in the focal plane. The challenge, illustrated conceptually as a set of measurements in Figure 6.26, is to find

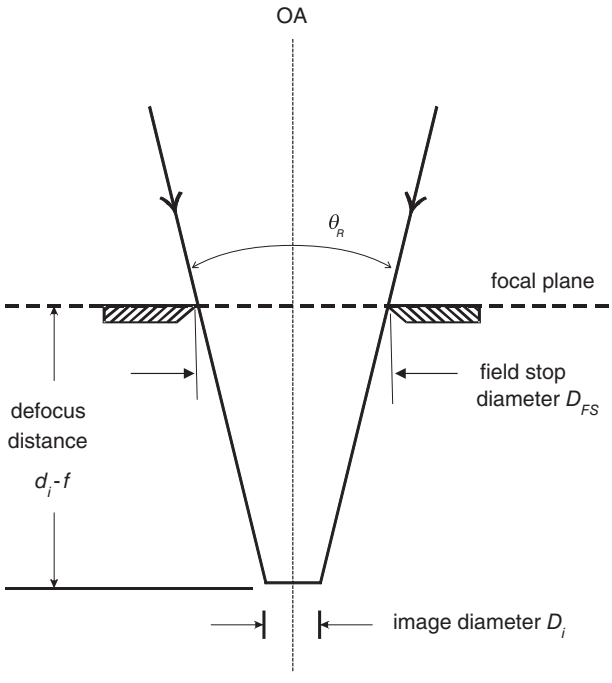


Figure 6.27 Finding the final range. The ray diagram corresponds to the range R_f at which $G(R)$ reaches unity for a perfectly aligned coaxial lidar.

the radial distance from the OA to the edge of the receiver's cone of light in the plane of the field stop as a function of range. Of course, the pulse of light cannot really be stopped at various ranges, so the procedure is rather to calculate the image position and size for any range using geometrical optics, find the edges of the cone of light in the plane of the field stop, and then plot the edges versus range along with the field stop edges, as shown in the lower right-hand corner of Figure 6.26. At GTRI, this plot is called a *crossover plot* (it has not previously been presented in the lidar literature). The geometry of the problem is shown in Figure 6.27 for the range R_f, at which the received light, with cone angle θ_R, just fits through the field stop aperture. The image is located behind the focal plane because of defocus. Recall from Section 6.3.2 that the defocus distance is

$$(d_i - f) = \frac{f^2}{d_o - f} \approx \frac{f^2}{R}, \tag{6.14}$$

where we have approximated the object distance minus the focal length as the range R.

To determine the image size, first find the object size, which is the scattering volume diameter at range R, then use the magnification to find the image size. The scattering volume diameter is the initial beam diameter plus the beam divergence times the range,

$$D_o = D_T + R\theta_T. \tag{6.15}$$

The image size is then

$$D_{\mathrm{i}} = \frac{d_{\mathrm{i}}}{d_{\mathrm{o}}} D_{\mathrm{o}} \approx \frac{f}{R} D_{\mathrm{o}} = f\left(\frac{D_{\mathrm{T}}}{R} + \theta_{\mathrm{T}}\right). \tag{6.16}$$

Now find the edges of the cone of light at the field stop by using the cone angle, which is just the receiver diameter D_{R} divided by the image distance,

$$\theta_{\mathrm{R}} \approx D_{\mathrm{R}} / f, \tag{6.17}$$

assuming that the receiver diameter is much larger than the image diameter and approximating the image distance as the focal length f. The edges of the cone of light $S(R)_{\mathrm{light}}$ at the field stop relative to the optical axis are then one-half of the image size plus one-half of the defocus distance times the receiver cone angle, so

$$S(R)_{\mathrm{light}} = \pm\frac{1}{2}\left[f\left(\frac{D_{\mathrm{T}}}{R} + \theta_{\mathrm{T}}\right) + \frac{f^2}{R}\left(\frac{D_{\mathrm{R}}}{f}\right)\right], \text{ which reduces to} \tag{6.18}$$

$$S(R)_{\mathrm{light}} = \pm\frac{1}{2} f\left[\frac{D_{\mathrm{T}} + D_{\mathrm{R}}}{R} + \theta_{\mathrm{T}}\right]. \tag{6.19}$$

At short ranges, the first term in the square brackets in Eq. (6.19) dominates, so the edge locations start large and decrease as $1/R$. As the range R becomes large, the first term becomes very small, and the edges of the cone of light asymptotically approach the constant value $\pm f\theta_{\mathrm{T}} / 2$. Both extremes are shown in the plot of the edge locations versus range in the lower right-hand corner of Figure 6.26. In a computer application such as a spreadsheet, the designer can choose a value of f and vary the parameters D_{T}, D_{R}, and θ_{T} while observing the behavior of the received cone of light and monitoring the range R_{f}, in order to optimize $G(R)$ for the intended lidar measurements.

The plot in Figure 6.26 is also useful for finding the lidar system's tolerance to misalignment, which is given by the distance from the edge of the field stop to the cone of light, at long ranges. Misalignment is usually thought of in terms of angles, so for this purpose, the plot is more useful if the edge locations are converted to angles rather than distances. This conversion is accomplished by simply dividing the distances in the focal plane by the focal length f. In angle units (radians), the edges of the cone of light are given by

$$\theta(R)_{\mathrm{light}} = \pm\frac{1}{2}\left[\frac{D_{\mathrm{T}} + D_{\mathrm{R}}}{R} + \theta_{\mathrm{T}}\right] \tag{6.20}$$

and the edges of the field stop in radians are given by one-half the field of view angle $(1/2)\theta_{\mathrm{FOV}}$, which is defined by

$$\theta_{\mathrm{FOV}} = D_{\mathrm{FS}} / f. \tag{6.21}$$

The diagram in Figure 6.27 corresponds to the range R_{f} at which $G(R)$ becomes unity, that is, when the cone of light just fits through the field stop. To find the range at which this occurs, set the edge distance given by Eq. (6.19) equal to the field stop radius $D_{\mathrm{FS}} / 2$:

$$\frac{D_{\mathrm{FS}}}{2} = \frac{f}{2}\left[\frac{D_{\mathrm{T}} + D_{\mathrm{R}}}{R} + \theta_{\mathrm{T}}\right]. \tag{6.22}$$

Solving for R,

$$R_f = \frac{D_R + D_T}{\theta_{FOV} - \theta_T}, \tag{6.23}$$

where we have used the definition of θ_{FOV} in Eq. (6.21). Equation (6.23) reveals three important features of a lidar's optical system:

(1) The range R_f at which $G(R)$ becomes unity is independent of the focal length of the receiver telescope.
(2) The range R_f increases linearly with both D_R and D_T.
(3) The range R_f depends inversely on the difference $\theta_{FOV} - \theta_T$.

The first feature explains why astronomical telescopes (which have long focal lengths) are often used as lidar receivers – focal length does not matter. Simply choose the desired FOV by adjusting the field stop diameter. This approach works if the FOV is set with an aperture. In lidars with small detectors such as the APDs discussed in Chapter 8, the field stop is usually the detector element itself. In that case, a telescope with a long focal length may result in an unacceptably small FOV. The second feature, that the range R_f increases linearly with both D_R and D_T, shows that larger lidar systems tend to have longer crossover ranges. This is one reason that lidars sometimes employ two receiver telescopes – a large telescope for the weaker signals from long ranges and a smaller telescope with a shorter crossover range, to extend the measurements to close-in ranges. The third feature, that the range R_f depends inversely on the difference $\theta_{FOV} - \theta_T$, suggests that one should make the laser beam divergence small to achieve a short crossover range, or conversely, if a long crossover range is desired, make the FOV and the laser beam divergence comparable. Both extremes are employed in practical lidar systems. Note that θ_{FOV} must be larger than θ_T for the lidar equation to be valid.

Example. The commercial eye-safe lidar known as MPL (for micro-pulse lidar) employs a 178-mm Cassegrainian telescope in a common-optics configuration. The transmitted beam divergence is about 40 μrad and the receiver FOV is 83 μrad. The range R_f given by Eq. (6.23) is therefore (178 mm + 178 mm) / [(83 − 40) × 10^{-6}] = 8.3 km. This is an example of a lidar system in which the function $G(R)$ is deliberately engineered so that the crossover region extends to a long range. The reason for doing this is to compress the dynamic range of the lidar signal. The MPL is a photon-counting lidar, and the signal at short ranges must be suppressed to avoid exceeding the maximum count rate of the data system. The long crossover range is achieved by using common optics and a laser beam divergence that is approximately one-half of the FOV. Such a system cannot tolerate misalignment angles greater than tens of μrad.

The linear functional relationships in Eq. (6.23) are required for proper scaling. If a drawing of a lidar optical system such as Figure 6.25 is scaled to a larger size, all the distances will be increased by a common scale factor and all the angles will be unchanged. Therefore, the value of R_f (which is a distance) will be increased by the scale factor. This scaling consistency is the reason that the numerator in Eq. (6.23)

contains only distances, and the denominator contains only angles. Equation (6.23) can also be viewed as a manifestation of the conserved optical quantity defined in Section 6.3.2 – it is consistent with the fact that the diameter of the scattering volume times the cone angle of its rays to the receiver must be equal to the size of the image times the cone angle of its rays from the receiver. Proving consistency with the conserved optical quantity is left as a problem for the student.

Central Obstructions

The crossover plot shown in Figure 6.26 does not address the initial range R_i at which $G(R)$ begins to increase. This range is zero if there is no central obstruction in the receiver telescope. However, most telescopes have a secondary mirror that obstructs the center of the telescope, and a shadow of that mirror extends all the way to the image, as shown in Figure 6.28. Note that the receiver telescope in this figure is represented by an equivalent lens, and that the central obstruction has been approximated by a disk placed in the plane of that lens. The shadow determines the initial range R_i when there is a central obstruction, so it should be included in the model. The analysis is illustrated in Figure 6.29, which corresponds to the range R_i at which the shadow cone just fills the field stop and the cone of light rays is about to enter the stop as the range to the scattering volume increases.

The shadow edges at the field stop are given by

$$S(R)_{\text{shadow}} = \pm(1/2)\theta_S \frac{f^2}{R}, \tag{6.24}$$

where θ_S is the shadow cone angle defined by

$$\theta_S = D_{CO} / f, \tag{6.25}$$

where D_{CO} is the diameter of the central obstruction, and we have again used the approximation that the image distance is the focal length f. Substituting Eq. (6.25) into Eq. (6.24), we have

$$S(R)_{\text{shadow}} = \pm(1/2)D_{CO} \frac{f}{R}, \tag{6.26}$$

which is expressed in angle as

$$\theta(R)_{\text{shadow}} = \pm(1/2)\frac{D_{CO}}{R}. \tag{6.27}$$

Adding the shadow of the secondary obstruction results in the more complete crossover plot shown in Figure 6.30, which enables the designer to find both R_i and R_f.

The range of complete crossover R_f can be decreased by moving the field stop back from the focal plane slightly, as shown in Figure 6.25. This change focuses the lidar receiver at a fixed distance rather than at infinity. The trade-off is that the lidar will have a lower tolerance to misalignment. To derive the equation for the edges of the cone of light, note that in Eq. (6.18) the defocus distance f^2/R represents the distance from the image to the field stop. If the stop location is offset a distance d_{FSO} from the focal plane, the image-to-field stop distance becomes $(f^2/R) - d_{FSO}$. Simply

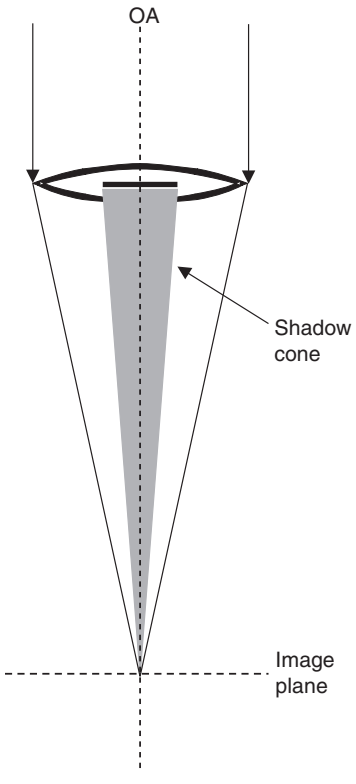


Figure 6.28 The central obstruction model. A central obstruction in a receiver telescope casts a conical shadow that extends to the image plane.

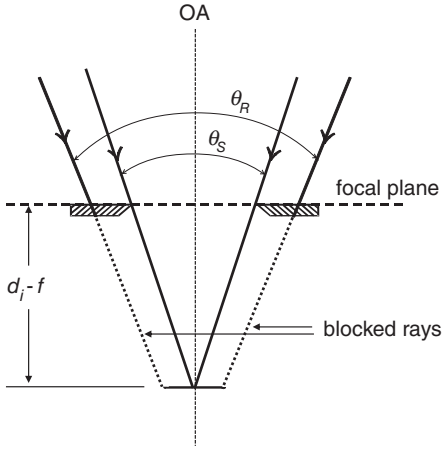


Figure 6.29 The shadow cone. Illustration of light rays when the shadow cone just fills the field stop.

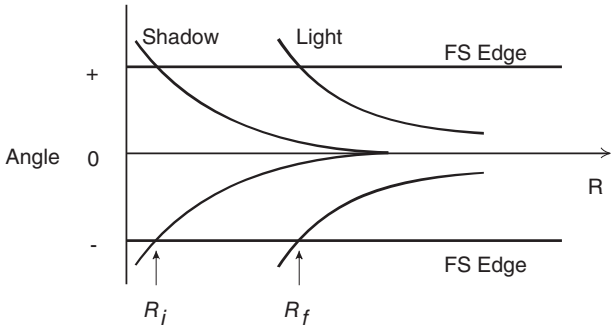


Figure 6.30 Crossover with a shadow. The complete crossover plot includes the shadow of a central obstruction.

making this substitution would lead to the correct equations, except for one detail: as the range R increases, the image in such a receiver moves through the field stop on its way to the focal plane, and the quantity $(f^2/R) - d_{\mathrm{FSO}}$ becomes negative, yielding an incorrect answer. The geometry is illustrated in Figure 6.31. In our simple model, the rays of light converge to the image and then diverge at the same cone angle (in reality, there is a waist, like the waists shown in Chapter 5 for laser beams). The width of the cone of light is therefore the same whether the image is above or below the field stop, and we can obtain the correct answer by using the absolute value of the distance $(f^2/R) - d_{\mathrm{FSO}}$. With a derivation like that following Eq. (6.18), we arrive at the result

$$\theta(R)_{\mathrm{light}} = \pm\frac{1}{2R}\left[D_{\mathrm{T}} + R\theta_{\mathrm{T}} + \left|D_{\mathrm{R}}\left(1 - \frac{Rd_{\mathrm{FSO}}}{f^2}\right)\right|\right]. \tag{6.28}$$

When d_{FSO} is zero, Eq. (6.28) reduces to Eq. (6.20), as it must, and when R becomes large, the edges of the cone of light are determined not only by the transmitter divergence θ_{T} but also by another term, which is the receiver cone angle defined in Eq. (6.17) times the field stop offset distance. This additional term is the cause of the trade-off mentioned above, that is, the range R_{f} can be decreased by moving the field stop, but then the lidar will have a lower misalignment tolerance.

The equation for the edge of the shadow cone in the field stop is derived from Eq. (6.26) by making the same substitution, replacing the defocus distance f^2/R with the quantity $(f^2/R) - d_{\mathrm{FSO}}$. The result is

$$\theta(R)_{\mathrm{shadow}} = \pm(1/2)\frac{D_{\mathrm{CO}}}{R}\left[1 - \frac{Rd_{\mathrm{FSO}}}{f^2}\right]. \tag{6.29}$$

In this case, the use of an absolute value is not required, but the value of $\theta(R)_{\mathrm{shadow}}$ must be set equal to zero for all ranges beyond $R = f^2/d_{\mathrm{FSO}}$, because the shadow cone is no longer in the field stop for those ranges. Offsetting the field stop in a coaxial or common optics lidar makes the shadow cone smaller in the field stop, decreasing

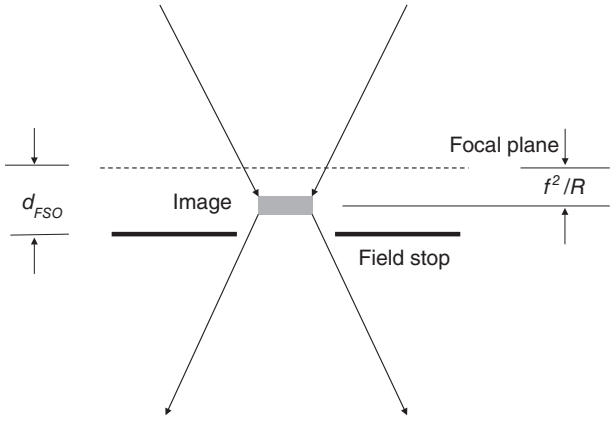


Figure 6.31 The diverging cone model. When the image is between the focal plane and the field stop, the diverging rays of light are modeled as a cone with the same angle as the converging cone.

the initial range at which light begins to reach the detector, so both R_i and R_f are decreased by the offset.

EARL Crossover Plots

The EARL receiver is described in Section 6.2.2. The long-range field stop is offset 0.5 mm behind the focal plane, and crossover is predicted to be complete at about 2350 m, as shown in Figure 6.32(a). Note that Eq. (6.23), without the field stop offset, predicts a crossover range of 2710 m, so an offset of only 0.5 mm reduces the crossover range by 360 m. Crossover is very sensitive to the field stop offset, so it is essential to locate the focal plane very accurately. The goal was to locate EARL's focal plane within 1/5 of the planned offset, or 0.1 mm, which is about the thickness of a sheet of notebook paper. This accuracy is difficult to achieve for a 3-m-tall instrument with a focal length of 2.57 m, so the focal plane was located on a clear night with an FPA TV camera, by observing images of stars transiting across the FOV and adjusting for best focus. The short-range field stop is offset 3.5 mm behind the focal plane, and crossover is complete at about 600 m, as shown in Figure 6.32(b). Equation (6.23) yields 752 m without the offset. The angular distance between the received cone of light and the edge of the field stop aperture is the misalignment tolerance. For the short-range receiver it is 387 μrad, and the misalignment tolerance of the long-range receiver is 137 μrad. The performance simulations in Figure 2.9 start at ranges of about 600 and 2350 m because of the designed crossover behavior shown in Figure 6.32.

EARL was designed with two receivers partly to provide enough dynamic range for the electronics during measurements from well within the mixed layer up into the stratosphere. Having two co-aligned receivers with different FOVs also creates another valuable feature: At altitudes above complete short-range crossover, the ratio of the long-range to short-range signals is proportional to the long-range $G(R)$. As such, the channel ratio should start at zero and increase smoothly to a constant value. A measured

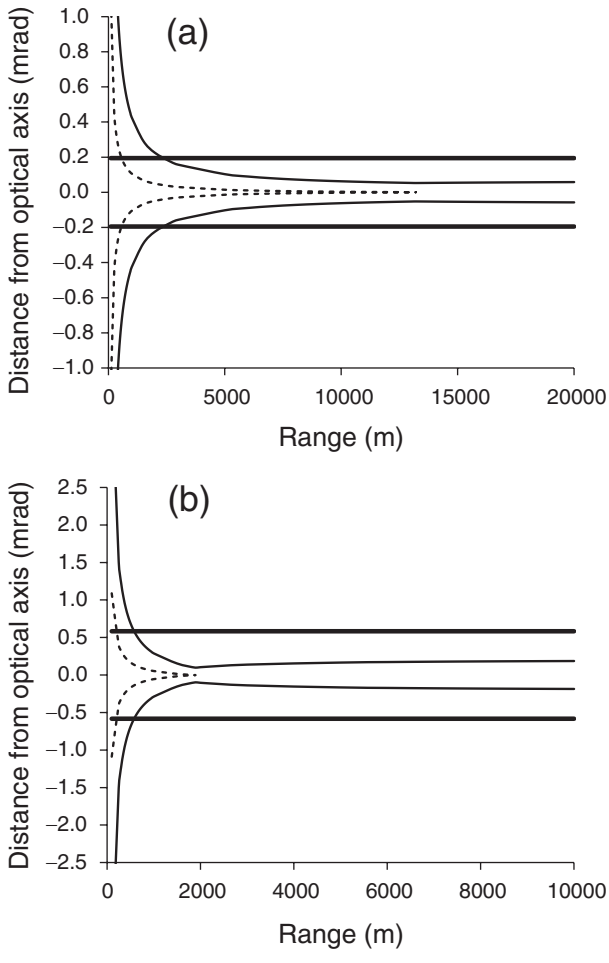


Figure 6.32 EARL Crossover plots. Crossover plots are shown for (a) the long-range receiver channel and (b) the short-range channel. The solid curves are the edges of the cones of light and the dashed curves are the edges of the shadow cones. The heavy black lines are the edges of the field stop apertures.

example, normalized to unity at long ranges, is shown in Figure 6.33. The background levels must be subtracted from both channels before they are ratioed; that process is described in Chapter 10. Plots such as Figure 6.33 are routinely made from archived data sets to verify that EARL was properly aligned during the data acquisition, and they can also be made in real time. A software utility was developed for EARL that produced such plots for the 0–3,000 m range at one-half second intervals to assist with alignment. At night, the plots resemble Figure 6.33. During daytime, the plots become quite noisy above about 1,500 m, but it is still possible to align the system by noting the nighttime value of the curve at a lower altitude; for example, in the plot shown, the channel ratio value is 0.5 at 1,000 m altitude. When that value is reached, the system is aligned. In this way, daytime alignment of EARL using atmospheric signals is possible. EARL is

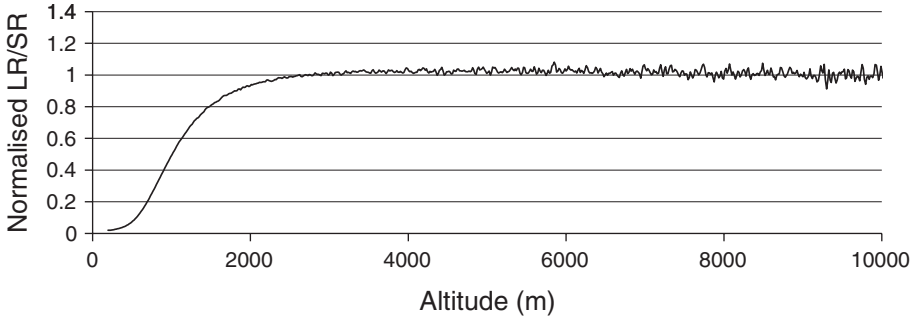


Figure 6.33 Measured long-range crossover function for EARL. The curve is the normalized ratio of long-range to short-range signals that were averaged for one hour at night. The data were acquired and analyzed in 2014 as part of the study referenced as [4] in Chapter 5.

probably the only micro-pulse lidar to achieve such daytime alignment. This alignment technique is not restricted to coaxial lidars; it can be done with biaxial systems; the only requirements are that the receiver channels must be co-aligned and they must have different FOVs. As shown in Figure 6.30, complete long-range crossover was expected at 2,350 m, but Figure 6.31 shows that it was closer to 3,000 m. This difference may be due to the approximations in the simple model described here, as well as the facts that diffraction is ignored and the transmitted laser beam is not well characterized. The model is easy to implement in a spreadsheet and it is very useful for doing parameter excursions and tradeoffs, but it should be followed up with a proper ray tracing analysis.

Biaxial Lidars

To this point, only coaxial and common optics lidars have been considered, with no misalignment between the transmitter and receiver optical axes. Equation (6.20), which determines the edges of the cone of light shown in the crossover plots, can be generalized to include all seven of the lidar optical system parameters that can be used to engineer the function $G(R)$. That is, a similar analysis can be performed for a biaxial lidar with an intentional misalignment. This analysis is left as a problem for the student. The result is

$$\theta(R)_{\text{light}} = \pm\frac{1}{2R}\left[D_{\text{T}} + R\theta_{\text{T}} + \left|D_{\text{R}}\left(1 - \frac{Rd_{\text{FSO}}}{f^2}\right)\right| + d_{\text{TR}} - R\delta\right]. \tag{6.30}$$

A similar generalization can be done for the shadow equation, but it is not particularly interesting because the shadow of a central obstruction does not influence either R_{i} or R_{f} in the case of a bistatic lidar. Solving Eq. (6.30) for the range R_{f}, we have the relation

$$R_{\text{f}} = \frac{D_{\text{R}} + D_{\text{T}} + 2d_{\text{TR}}}{\theta_{\text{FOV}} - \theta_{\text{T}} + 2\delta + (D_{\text{R}}d_{\text{FSO}}/f^2)}. \tag{6.31}$$

Note that the transmitter-to-receiver offset distance d_{TR} appears in Eq. (6.31) in exactly the same way as D_{R} and D_{T}, and that δ and $(D_{\text{R}}d_{\text{FSO}}/f^2)$ appear in the

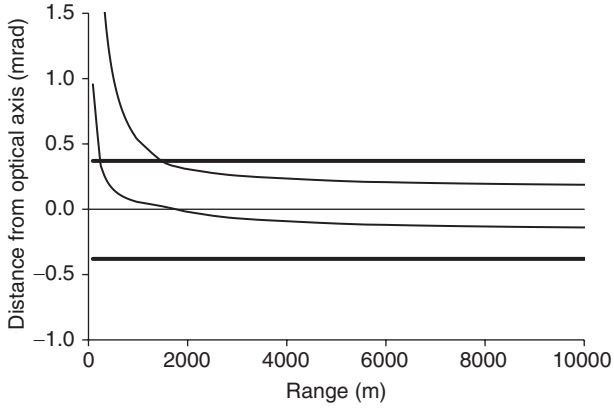


Figure 6.34 Crossover plot for a biaxial lidar. The cone of light has its minimum diameter in the field stop at a range of 1600 m, where the receiver is focused.

denominator with θ_R and θ_T, as required for proper scaling. The ratio $(D_R d_{FSO}/f^2)$ will be unchanged by an enlargement of a diagram, like the angles, which explains why d_{FSO} enters the equation in a dimensionless ratio rather than by itself. The range of complete crossover R_f can be made arbitrarily large by simply increasing d_{TR}, and this fact is sometimes exploited in lidar systems for stratospheric research, in which the signal from lower altitudes must be suppressed. Equation (6.31) also shows that increasing either δ or d_{FSO} has the effect of decreasing R_f.

In addition to determining the crossover range, there is another important reason for engineering $G(R)$: to control the dynamic range of the signals that the lidar's electronics must accommodate. This problem was mentioned earlier in connection with the MPL. As is shown in Figure 6.16, the peak signal occurs before the crossover range. If only data after crossover are to be analyzed, the pre-crossover peak makes the overall dynamic range larger than the useful data range, sometimes by a factor of 2 or more. Exceeding the dynamic range of detectors and electronics must be avoided because it can cause artifacts in the recorded lidar signal, so dynamic range is an important consideration in lidar engineering. In this sense, the biaxial configuration is better than the others because it has two more design parameters.

Equation (6.30) can be implemented in a computer application such as a spreadsheet, enabling the lidar designer to investigate tradeoffs between various parameters. The result of such an investigation is shown in Figure 6.34 for a lidar proposed as an updated version of EARL. The biaxial configuration employs 20-cm *f*/4 telescopes for both the transmitter and the receiver. The optical axes are parallel and offset from each other by 30 cm, and the transmitted beam divergence is 0.2 mrad. The 0.6 mm field stop has been offset from the focal plane by 0.4 mm to decrease R_f. This offset amounts to focusing the receiver at 1,600 m, which is why the cone of light has its minimum diameter at that range.

In this section, the simple, approximate treatment of the crossover phenomenon in lidar systems elucidated the roles of all the optical design parameters that control $G(R)$.

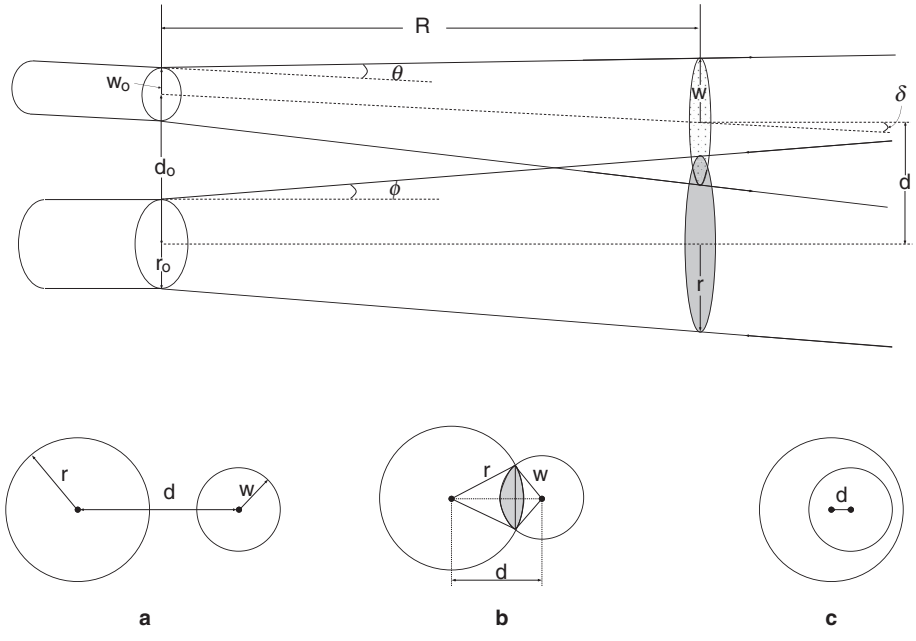


Figure 6.35 The incorrect picture. Inset (b) is a head-on view at the range indicated in the figure above by the shaded circles. The shading in inset (b) is intended to show a region of overlap between two circles representing the transmitted beam and the receiver FOV. Adapted with permission from [7].

A graphical method, the crossover plot, provides a quick visual overview of the crossover behavior and the lidar system's misalignment tolerance. These tools can be used to discover how accurately the lidar engineer must know parameters such as the field stop offset distance d_{FSO} and the misalignment angle δ. The misalignment tolerance sets some requirements for the system's optomechanics, for example, the maximum allowable amount of mechanical expansion and flexing of its structure throughout the temperature range that the lidar is designed for. Several approximations were used in the derivations, starting with the paraxial approximation and geometrical optics. The analysis of the effects of defocus was based on cones and triangles, and the receiver's central obstruction was assumed to be in the plane of a thin lens that represented the receiver telescope. The treatment presented here is adequate to give the designer a good conceptual grasp of the design parameters that determine $G(R)$ and to investigate tradeoffs through simple calculations, but it should be followed by an optical analysis using a ray tracing computer application.

A Few Words of Caution

The dependence of the geometrical function on the lidar system's optical parameters was well understood by some of the earliest lidar researchers, who realized that $G(R)$ depends on what happens at the field stop. However, an alternate (and incorrect) picture, illustrated in Figure 6.35, has become prevalent in the lidar literature.

The insets a, b, and c in Figure 6.35 are meant to correspond to the three ranges illustrated in Figure 6.21. In this picture, the value of $G(R)$ is related to the degree of "overlap" of two cones representing the transmitted laser beam and the receiver FOV, and it becomes unity at the range where the laser beam has "crossed over" into the FOV cone. This picture is the origin of the terms *overlap* and *crossover*, which have become standard. In the lidar literature, $G(R)$ is most often called $O(R)$, where the letter O stands for overlap. It is essential to realize that Figure 6.35 is not an optical diagram; for optical calculations, the FOV cannot be represented by a cone emerging from the receiver telescope. When the optical axes are parallel, the picture in Figure 6.35 results in the equation

$$R_f = \frac{2d_{TR} - D_R + D_T}{\theta_{FOV} - \theta_T} \tag{6.32}$$

which is different from the correct result given by Eq. (6.31) and is not consistent with the conserved optical quantity. The derivation of Eq. (6.32) is left as a problem for the student.

6.3.4 Comments on Alignment

After a lidar optical system has been designed and constructed, the transmitter–receiver alignment must be adjusted. Like everything else, the alignment technique should be thought out in advance during the design process, and the alignment capability should be implemented in hardware and software. Several alignment techniques are discussed in the lidar literature, including the use of special fixtures and the development of auto-alignment systems [8–10], but the discussion here is limited to accomplishing and verifying correct alignment using only the lidar signals. The goal of alignment is to put the image of the laser pulse at long ranges in a known position in the field stop aperture, usually the center. The alignment adjusters should have position indicators whether they are manual or motorized and they must have no backlash, so that the adjustments are repeatable.

In the early days of lidar, there was no choice but to use the raw signal observed with an oscilloscope. This method is difficult because the signal lacks features for assessing alignment, and the $1/R^2$ factor in the lidar equation causes the signal from long ranges to be very small and hence noisy and difficult to use. One solution is to use the signal from targets of opportunity such as high clouds. The drawbacks to this method are that clouds are not always present, and that they usually have some structure that makes their backscatter vary with time. This method has been used successfully; for example, the GTRI 1.57-μm lidar mentioned in Section 4.2.2 was always aligned by observing the signal from the stratospheric aerosol layer. However, alignment using the raw signal from targets of opportunity is, in general, inconvenient at best.

Better approaches rely on software utilities with adjustable multi-pulse averaging. The 1.57-μm lidar alignment was performed using 10-pulse (0.5 s) signal averaging of the raw signal, and experience has shown that an update time of 1 second or less is desirable, especially for manual adjustments. Such a short averaging time means that alignment is usually done with noisy signals, especially during daytime. The channel

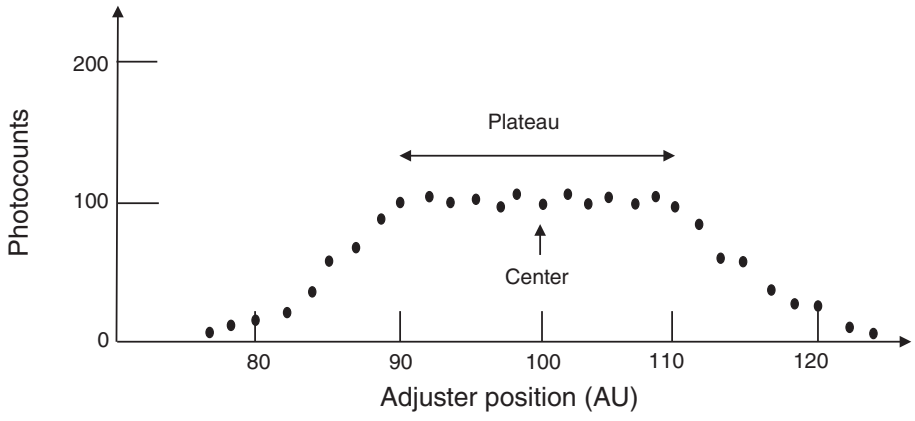


Figure 6.36 The plateau method. Such scans of photocounts versus adjuster position are made on two orthogonal axes to locate the plateau centers.

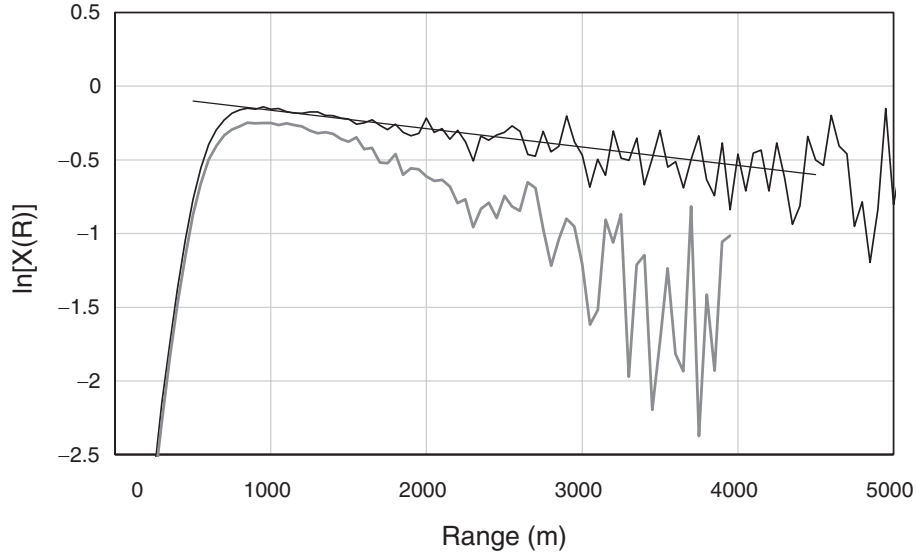


Figure 6.37 The logarithm of the range-corrected signal versus range. Solid black curve – aligned; gray curve – misaligned; black line – exponential density decrease. The simulation is based on a density scale height of 7 km and a crossover range of 1 km. Gaussian noise was added to the raw signals.

ratio display described in Section 6.3.3 and illustrated by Figure 6.33 is an excellent and recommended alignment tool, but it requires two co-aligned receiver channels with different FOVs, so it cannot be used in all lidars. Photon-counting lidars for high-altitude measurements often use the plateau method, which is schematically illustrated in Figure 6.36. The software utility counts the number of photons received from a set of range bins at a certain altitude. Both the number of bins and the altitude can be user

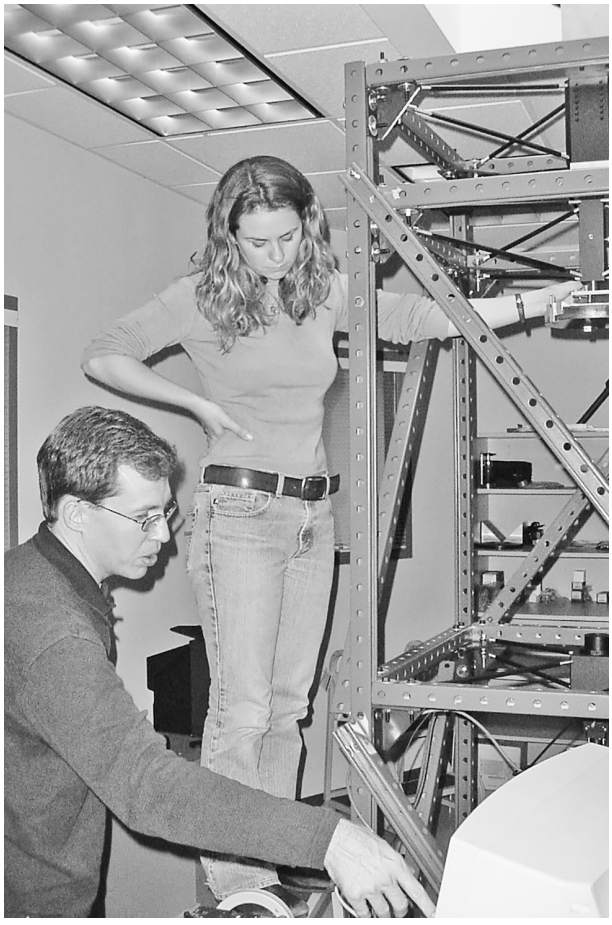


Figure 6.38 EARL alignment. David Roberts and Agnes Scott College student Martha Dawsey aligning EARL in 2005.

selectable. The alignment adjusters are scanned back and forth on two axes, yielding plots of plateaus, which occur when the long-range laser pulse image is fully within the field stop aperture. The adjusters can be returned to any positions, so the long-range laser pulse image can be placed anywhere within the field stop, although it is usually centered. The center of the plateau in Figure 6.36 is at adjuster position 100, so that reading corresponds to centering the image in the field stop aperture on one axis. Most of the methods described here rely on two-axis scans to find center positions, which is why position indicators are helpful and backlash must be avoided.

For UV-VIS lidars, another simple and effective utility provides a real-time plot of the logarithm of the range-corrected signal $\ln[X(R)]$. A simulation of this plot is shown in Figure 6.37 for the cases when the long-range laser pulse image is fully within the field stop aperture and when it is not. In clear air, the nearly exponential decrease of air density with altitude causes the linear slope in the plot indicated by the straight black

line. As with other methods, the adjusters are scanned back and forth one at a time on two orthogonal axes, and the positions where $\ln[X(R)]$ begins to depart from the linear slope are noted, and then the adjusters are returned to their mid-points. This alignment method is simple to implement, and it is easy to use because the shape of the plot changes dramatically when the laser pulse image starts to cross the edge of the field stop aperture. Alignment using the $\ln[X(R)]$ plot has been employed successfully on several GTRI lidars, including EARL before the channel ratio alignment software had been developed. EARL was aligned by turning micrometer handle adjusters behind the secondary mirror while watching a near-real-time plot of $\ln[X(R)]$ on a monitor, as shown in Figure 6.38.

6.4 Further Reading

For an excellent introductory text on optics, see the book by Hecht [2].

6.5 Problems

6.5.1 The paraxial approximation.

(a) How many milliradians are in one degree?

(b) Fill in the table below, to eight digits beyond the decimal point. How good is the paraxial approximation for these angles?

Angle θ	θ (radians)	$\sin(\theta)$	$\tan(\theta)$
10 degrees			
1 degree			
10 mrad			
1 mrad			
100 μrad			

6.5.2 Lidar receiver parameters.

(a) EARL has an *f*/4.21 receiver based on a parabolic mirror with a diameter of 0.61 m. What is the defocus distance for backscattered light from a range of 500 m?

(b) What is the receiver cone angle θ_R ?

(c) If the field stop diameter is 3.00 mm, what is the FOV?

6.5.3 Show that Eq. (6.23) is consistent with conservation of the product of image size and cone angle, as discussed in Section 6.3.2, within the limits of the approximations used to derive the equation.

6.5.4 Derive Equation (6.30) for biaxial crossover plots.

6.5.5 Derive Equation (6.32) for the incorrect crossover range.

References

[1] Andover Corporation. [Online]. Available: www.andovercorp.com. [Accessed: June 22, 2021].
[2] E. Hecht, *Optics*, 5th ed. London: Pearson (2017).
[3] J. M. Bennett, "Polarization," in *Handbook of Optics*, M. Bass, Ed. New York: McGraw-Hill, 1995.
[4] K. Sassen, "Lidar Backscatter Depolarization Technique for Cloud and Aerosol Research," in *Light Scattering by Nonspherical Particles: Theory, Measurements, and Geophysical Applications*, M. I. Mishchenko, J. W. Hovenier, and L. D. Travis, Eds. San Diego: Academic Press, 2000.
[5] M. Hayman and J. P. Thayer, "General Description of Polarization in Lidar Using Stokes Vectors and Polar Decomposition of Mueller Matrices," *Journal of the Optical Society of America*, vol. A 29, pp. 400–409, 2012.
[6] V. Freudenthaler, "About the Effects of Polarising Optics on Lidar Signals and the Δ90-Calibration," *Atmospheric Measurement Techniques*, vol. 9, pp. 4181–4255, 2016.
[7] R. M. Measures, *Laser Remote Sensing – Fundamentals and Applications*. New York: Wiley, 1984.
[8] S. A. Young, "Lidar System Optical Alignment and Its Verification," *Applied Optics*, vol. 26, pp. 1612–1616, 1987.
[9] L. Fiorani, M. Armenante, R. Capobianco, N. Spinelli, and X. Wang, "Self-Aligning Lidar for the Continuous Monitoring of the Atmosphere," *Applied Optics*, vol. 37, pp. 4758–4764, 1998.
[10] B. Liu, F. Yi, and C. M. Yu, "Methods for Optical Adjustment in Lidar Systems," *Applied Optics*, vol. 44, pp. 1480–1484, 2005.

7 Optomechanics

The procedures in the preceding chapters are used in developing a set of design parameters for the optical elements of a lidar system, such as the diameters and focal lengths of lenses and mirrors, but the performance of an optical system depends on the relative positions of its elements and the quality of their surfaces. Reflection and refraction happen on those surfaces, so it is essential to avoid distorting them. Optomechanics is the design process for tying the optics together with a mechanical package, with the goal of designing in its required durability, reliability, and maintainability from the start, considering its environmental, operational, and performance requirements [1]. Designing a simple elastic backscatter lidar built around an astronomical telescope using commercial components on an optical table may not need to be a formal process, because the telescope, table, and components already embody sound optomechanical engineering. On the other hand, many lidars are used in the field, including very sophisticated types, and they must be rugged enough to withstand the shock and vibration associated with transporting them and they must operate in the wide temperature range of the outdoor environment. Airborne lidars have even more stringent requirements, and of course, spaceborne lidars have the most extreme sets of requirements. The more sophisticated lidar types do require a formal design process, and any lidar will probably benefit from adherence to a few basic principles described in this chapter.

Optomechanics requires expertise in both optics and mechanics. The need for it arises because light wavelengths are very small compared to the dimensions of the elements in an optical system. For this reason, *tolerances*, which are the acceptable errors in dimensions, locations or deflections, may be 2.5 to 0.025 μm (10^{-4} to 10^{-6} inches) in optics whereas the acceptable error in a dimension or deflection may be 0.25 to 0.025 mm (0.01 to 0.001 inches) in mechanics. Deflections that are insignificant in ordinary mechanical engineering can be critical in optomechanics. One reason for the small tolerances in optics is to avoid wavefront distortions, as illustrated in Figure 7.1. The commonly used Rayleigh criterion states that, for the best image, a distorted wavefront must not depart more than $\lambda/4$ from a sphere, although this restriction can be loosened for systems that do not require extremely sharp images, such as a lidar receiver with a reasonably large field of view (FOV). A main goal of optomechanical design is to preserve the wavefront quality that the optical elements can provide. Mirrors are particularly sensitive to deformation and misalignment because they double tilt angle errors and they double pathlength errors. The doubling of tilt angle errors

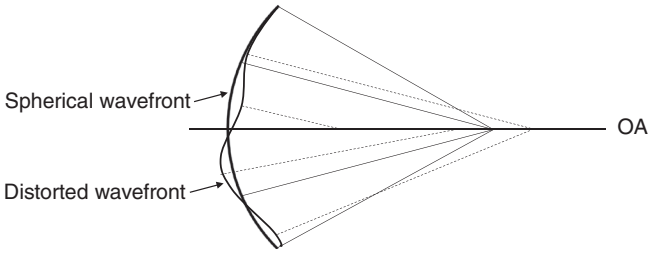


Figure 7.1 Effect of wavefront distortions. If the wavefront is spherical (heavy black curve), the rays converge to a point on the optical axis (OA). If it is distorted (lighter black curve), they cross the OA at various points.

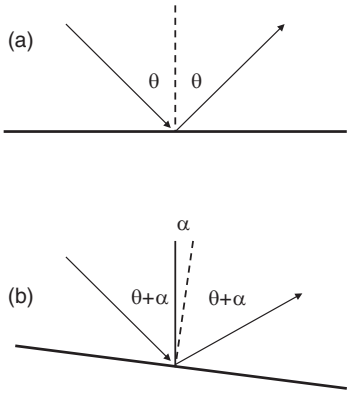


Figure 7.2 Effect of mirror tilt errors. (a) The angles of incidence and reflection are equal; (b) if the surface normal is tilted by an angle α, the reflected ray angle changes by 2α. All angles shown are relative to the surface normal.

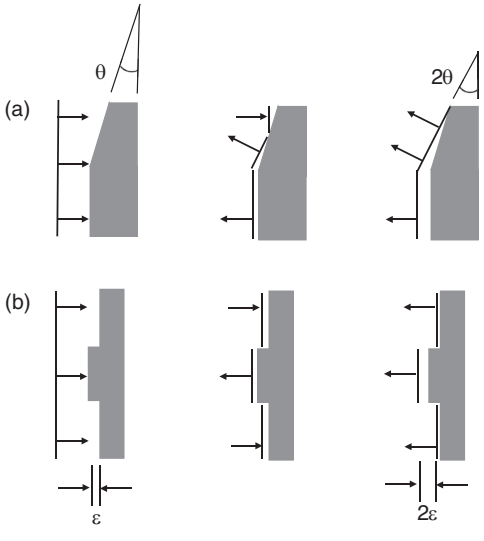


Figure 7.3 Effects of mirror deformations. A plane wave is incident from the left. (a) An angular error in the mirror surface is doubled in the wavefront angle; (b) a bump on the surface is doubled in the wavefront position.

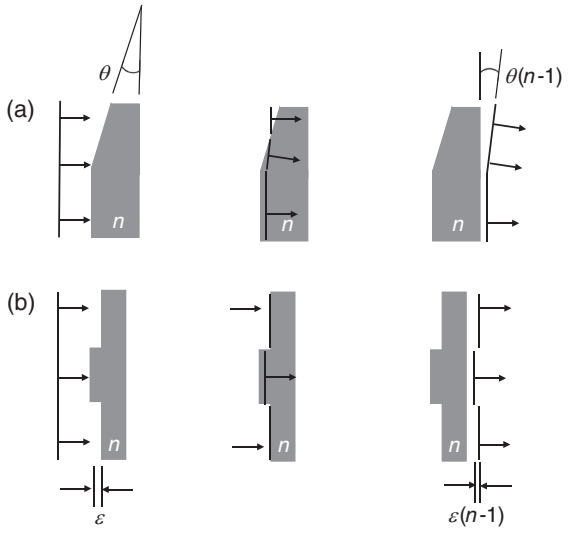


Figure 7.4 Effects of lens deformations. The effects on the wavefronts are reduced by the refractive index.

stems from the law of reflection, as illustrated in Figure 7.2. If the surface normal is tilted by an angle α, the reflected ray angle changes by 2α. The doubling of pathlength errors is shown in Figure 7.3.

Lenses are less sensitive to deformations than mirrors. As shown in Figure 7.4, pathlength and slope errors are reduced by the refractive index difference. For $n = 1.5$, wavefront errors are ¼ as large as similar errors produced by mirrors (there is a reduction only for materials with $n < 2$). Material properties are discussed in the next section, which is followed by techniques for mounting lenses and mirrors, and mechanical engineering considerations for the overall instrument structure are described in the last section of this chapter.

7.1 Optical Instrument Materials

Two types of material are typically used in optical systems: glass for the optical elements and metal for the structural elements. The two major considerations for optomechanics are how much the materials deform under stress (often due to the force of gravity) and how much their dimensions change with temperature. The deformation of optical materials, such as the sagging of mirrors, due to *stress* σ, illustrated in Figure 7.5 for the stretching of a cylindrical object, is governed by the *elastic modulus* E. The word *elastic* means that there are no permanent deformations; the material will return to its original dimension or position when the stress is removed. Stress is the applied force per unit area, N/m^2, which has the name pascals (Pa), defined by

$$\sigma = F/A. \tag{7.1}$$

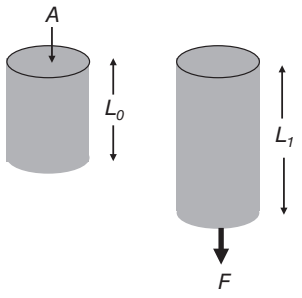


Figure 7.5 The elastic modulus.

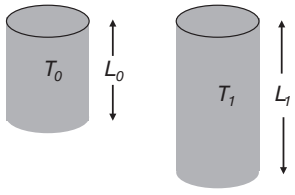


Figure 7.6 Thermal expansion. T_1 is greater than T_0.

The *strain* ε is the change in length normalized by the original length,

$$\varepsilon = (L_1 - L_0) / L_0 \tag{7.2}$$

so it is dimensionless. The elastic modulus is defined as

$$E = \sigma / \varepsilon. \tag{7.3}$$

In addition to the deformations of mirrors due to their self-weight, the elastic modulus is used in calculations of bending in structural members. The word *stiffness* means the ability of something to resist deformation due to forces on it. A higher elastic modulus means a stiffer material or structure.

The change in dimensions of materials due to a temperature change, illustrated in Figure 7.6, is governed by the coefficient of thermal expansion α, (often called the CTE) which is the change in length per degree Celsius normalized by the original length:

$$\alpha = \frac{1}{T_1 - T_0}\left[\frac{L_1 - L_0}{L_0}\right]. \tag{7.4}$$

The units of α are 1/°C. The CTE is used in Eq. (7.5) for calculating focus shifts, thermal tilting, and mount clearances.

$$\Delta L / L_0 = \alpha \Delta T. \tag{7.5}$$

The values of the elastic modulus and the coefficient of thermal expansion for common optical system materials are listed in Table 7.1, along with the material densities. ULE is a tradename of Corning Inc. that stands for ultra-low expansion. The CTEs are often quoted as ppm/°C.

Table 7.1 Properties of common optical system materials

Material	Density	Elastic modulus	Coefficient of thermal expansion
Aluminum (6061)	2710 kg/m^3	69.0 GPa	23.0×10^{-6}/°C
Steel (1025)	7858 kg/m^3	200 GPa	12.1×10^{-6}/°C
Stainless steel (310)	8030 kg/m^3	193 GPa	16.6×10^{-6}/°C
Glass (BK7)	2530 kg/m^3	82 GPa	7.1×10^{-6}/°C
Corning ULE	2200 kg/m^3	67.7 GPa	0.03×10^{-6}/°C

Table 7.1 shows that the stiffness and density of glass are nearly the same as aluminum. For this reason, aluminum properties are sometimes substituted for glass properties in the early stages of a design, until the types of glass have been determined. For calculations of deflections and deformations of materials, one more property of materials is needed, which is *Poisson's ratio* ν. This ratio is a measure of the Poisson effect, which describes the expansion or contraction of a material in directions perpendicular to the direction of loading; for example, the radius of the rod in Figure 7.5 will become smaller as it is stretched. The materials in Table 7.1 have ν values of 0.2–0.3. Mechanical engineers have developed formulas for deformation in many special geometries, including some that apply to mirror sag. If a horizontal disk of material is supported by a ring at its edge, as shown in Figure 7.7, the sag in the middle ΔY is calculated as

$$\Delta Y = \frac{-3W(m-1)(5m+1)D^2}{64m^2\pi Et^3}, \tag{7.6}$$

where W (N) is the weight, m is $1/\nu$, D (m) is the diameter, and t (m) is the thickness. The deflection increases linearly with the mirror's weight and as the square of the diameter but inversely with the cube of the thickness, which is the reason that optical mirrors are generally quite thick. As a rule of thumb, mirrors should have a diameter-to-thickness ratio of 6:1 or less; otherwise, they are considered thin mirrors. A thin mirror with a diameter like EARL's receiver primary mirror would have a center sag of several wavelengths if supported only at the edge (the calculation is left as a problem for the student).

Example. A mirror will also change dimension with temperature changes. For example, if a 25-mm thick BK7 mirror is heated by 10°C, its thickness change as given by Eq. (7.5) is $\Delta L = (7.1 \times 10^{-6}) \times (10) \times (25\text{ mm}) = 1.78 \times 10^{-3}$ mm = 1.78 μm. This is about three wavelengths of visible light, and the pathlength change is doubled by the mirror. This change can cause focus shifts and aberration changes. For instruments that must operate in the open air outdoors, such as astronomical telescopes, temperature changes can be a serious problem, which is the reason for using special glasses such as ULE.

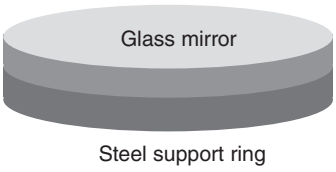


Figure 7.7 Ring support.

The lidar's mechanical structure will also change dimensions with temperature, and lidar structures must not be exposed to direct sunlight because of differential expansion: The sunlit side will expand more than the side in shade. If this happens to a telescope tube, for example, the secondary mirror will tilt, potentially causing misalignment. EARL was installed in a laboratory under a roof hatch, and during a few days in mid-summer, direct sunlight came through the hatch and heated one side of the structure and caused a complete misalignment, causing the long-range signal to drop to zero. A calculation related to this issue is provided as a problem for the student.

7.2 Mounting

The most common approach to building optical instruments is to attach all the optical components to one stiff structure known as the *optical bench*. For one-of-a-kind instruments, the bench is almost always an optical table or an optical breadboard, both of which are engineered to be very stiff. They have tapped holes at regular intervals in two dimensions. In the U.S., tables and breadboards have ¼-20 tapped holes at 1- or 2-inch spacings. If the instrument is to be manufactured in multiple identical copies or if it must be compact, engineering a custom bench is usually justified, but using commercial mounts fastened to a commercial bench is the most cost-effective way of getting the required stability in a prototype or a one-of-a-kind lidar system. An optical system built on a breadboard is inherently two-dimensional and hence not compact, but that geometry gives the researcher easy access to all optical components, and this approach has been highly successful in lidar.

7.2.1 Optics Mounts

The means of attachment of the optical elements to the bench is through *optical mounts*. Lenses are often mounted in a lens *cell*, which centers the lens and holds it in the correct position. Lenses are most often held in cells by a retaining ring, as shown in Figure 7.8. Such cells, which are usually made of aluminum alloy, can provide both centration and positioning without putting stresses on lenses that would distort them, but thermal expansion must be considered because the CTE is much higher for aluminum than for glass. Lenses are sometimes held in cells by setscrews that press on the lens radially, but that is a bad practice because the radial force can deform the

Figure 7.8 Lens cells with retaining rings. Permission to use granted by Newport Corporation. All rights reserved.

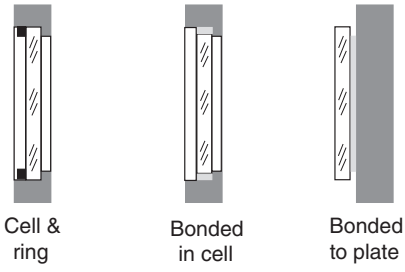


Figure 7.9 Mounting small mirrors. Three common methods are shown in cross section. The bonding material is shown in light gray; darker gray and black areas are aluminum.

lens and degrade its performance. Best practices for designing the cells and retaining rings were explained by Yoder [2]. Lenses are mounted in a similar way in optical tubing. Such tubes are threaded on the inside, and retaining rings are simply screwed in against lenses and other elements from both sides, which makes positioning them easy. EARL's receiver was assembled using optical tubes in this manner.

Lidar mirrors are of the front surface type, in which the surface has a reflective coating of aluminum or often beryllium–aluminum and a protective overcoat such as SiO to prevent oxidation. Both the coating and the overcoat must be chosen to optimize reflectance at the operating wavelength(s) of the lidar. Small mirrors (less than about 25-mm diameter) can be treated like lenses and mounted in cells, or they may be bonded to metal plates, as shown in Figure 7.9. Thin mirrors up to 15 cm in diameter are often bonded to metal plates with room temperature vulcanizing silicone rubber elastomer.

Lens and mirror mounts such as those shown in Figures 7.8 and 7.9 must of course be attached to the optical bench, and several styles of hardware are commercially available for this purpose. The post and fork mount shown in Figure 7.10 is a very stable option. The heavy post is secured to the optical table or breadboard by the fork,

Figure 7.10 The post and fork mount. The post is secured to the optical bench by the fork, which is slotted so that the closest available hole can be used. Permission to use granted by Newport Corporation. All rights reserved.

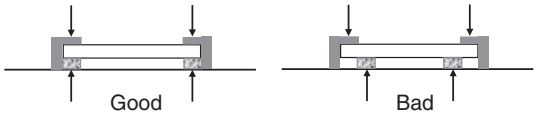


Figure 7.11 Mounting larger mirrors. The clamps must be opposite the support pads to avoid bending the mirror.

which is slotted so that the closest available threaded hole can be used. The lens or mirror is usually in a mount that provides angle adjustments on two orthogonal axes.

Larger mirrors are usually not in cells or bonded to plates; they are clamped to their mounts, but care is required: Bending of the mirror must be avoided by clamping opposite the support pads, as shown in Figure 7.11. The 20-cm secondary mirror in EARL, shown in Figure 7.12, is clamped at three points in the correct manner. Mounting even larger mirrors is a different issue because large mirrors can deform themselves through self-weight deflection. The purpose of all the glass in a mirror is only to support the reflective coating, so when the glass deforms, the image quality is degraded.

This problem is exacerbated because the wavefront error is double the surface error, as shown in Figure 7.3. For these reasons, special techniques have been developed to support large mirrors. These techniques all make use of the fact that three points determine a plane, and the simplest mount is in fact a three-point support on pads that are attached to some sort of base structure, as shown in Figure 7.13, where R_1 is the radius of the mirror and R_2 is the radius of the circle on which the three points

Figure 7.12 EARL's secondary mirror. The mirror clamps are opposite the support pads.

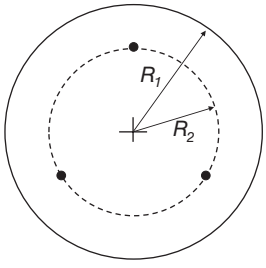


Figure 7.13 The three-point mirror mount. The mirror radius is R_1 and the three pads (shown by black dots) are on a circle of radius R_2 that results in the least self-deflection due to gravity.

are equally spaced. The mirror is in a horizontal plane, and the design principle is that stress is minimized when the mirror weight inside the support circle is equal to the weight outside. That principle is satisfied when $R_2 = R_1/\sqrt{2}$ (assuming uniform mirror thickness). Each pad supports 1/3 of the mirror weight, by symmetry. As a rule of thumb, this type of support is adequate for mirrors 25 cm or less in diameter that have a diameter-to-thickness ratio of at least 6:1. Larger mirrors and thin mirrors are supported by *whiffletrees*, which incorporate multiple three-point contacts with the mirror. The three-point supports may sit on their own pads or in pairs on rocker arms with pivots, so that the whole support system is compliant, meaning that it accommodates the small imperfections in the back surface of the mirror and guarantees that the load is equally distributed among the pads. The nine-point whiffletree mount shown in Figure 7.14 was first described in 1945 by J. H. Hindle, in a book on amateur telescope

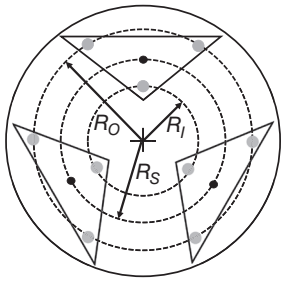


Figure 7.14 The nine-point whiffletree mount. Three triangular plates each carry three mirror support pads (graydots). The plates are supported on pivots (black dots). All the supports lie on three circles with radii labeled as shown.

making [3]. A key concept in Hindle's design was that each support pad should carry an equal amount of the mirror's weight. In the nine-point mount, the pivot points for the triangles are on a circle of radius R_S and the support pads contact the mirror at two different radii R_I and R_O. The EARL primary mirror is supported on a nine-point whiffletree mount. The three radii are given by the expressions in Eqs. (7.7)–(7.9) for a mirror of diameter D. The next level of sophistication, which is useful for very thin mirrors, is the 18-point mount, which employs six three-point mounts on rocker arms. The approach of using multiple three-point supports in pairs can be extended; whiffletree supports have been constructed with at least 54 support pads.

$$R_I = \frac{\sqrt{6}}{12} D \tag{7.7}$$

$$R_O = \frac{\sqrt{6}}{6} D \tag{7.8}$$

$$R_S = 0.30374 D. \tag{7.9}$$

7.2.2 Kinematic Mounts

Optical system components must be mounted so they are not distorted, and they should be mounted in a repeatable manner, for convenience. Rigid objects possess six degrees of freedom, three rotational and three translational, and a mount is said to be *kinematic* if all degrees of freedom are constrained. In addition to providing stability, a kinematic mount enables the accurate and repeatable location of one part relative to another, making it easy to remove and replace components. The relationship of mounting points to degrees of freedom is illustrated in Figure 7.15. The number of degrees of freedom of a mounted object is six minus the number of constraints. An important principle in kinematic mount design is that three points determine a plane, so contact at more points on the plane may distort the object. For this reason, three support points are shown under the objects in Figure 7.15; they are constraints against motion along the z-axis. The support points prevent downward

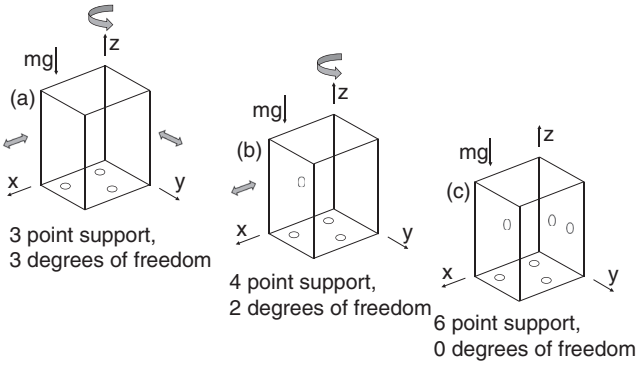


Figure 7.15 Support points and degrees of freedom. Gray arrows show unconstrained motions.

motion and upward motion is prevented by the force of gravity, mg. In the absence of gravity, the support points are replaced by attachment points. Kinematic design assumes perfectly rigid bodies with constraints touching only at points, but these are idealizations. Nothing is perfectly rigid, and point contacts would produce enormous pressures. In practice, semi-kinematic mounting is usually the best that can be achieved, because everything flexes to some degree, and the constraints usually have area contacts, such as pads.

The adjustable mirror mount shown in Figure 7.10 deserves further description. Typically, the movable frame that holds the mirror pivots on a ball bearing that is set into a hole in the fixed frame. Ideally this hole should be trihedral (pyramid-shaped), but a conical hole is often used due to easier manufacture. The angles of the movable frame are adjusted by means of 2 micrometers or fine-thread screws tipped with steel ball bearings. One of these balls rests in a V-groove, and the other rests on a flat surface. The first ball (ideally) contacts the fixed frame at exactly three points, the second ball at two, and the third ball at just one. Springs in tension pull the two frames together [4]. The six points of contact exactly constrain the six degrees of freedom for motion of the movable frame. This type of kinematic design can be built without precision manufacturing, and in fact the mount for the EARL secondary mirror shown in Figure 7.12 was custom made in the GTRI machine shop in exactly this way.

7.2.3 Fasteners

The optics mounts must be fastened to the bench or to the lidar system's structural members in some manner, generally with machine screws. The U.S., Canada, and the U.K. adopted the Unified Thread Standard (UTS) in 1949 for inch-based screw threads. The standards specify both UNC (coarse) and UNF (fine) threads. However, the Society of Automotive Engineers (SAE) standard pre-dated the UTS by more than 30 years, and in the U.S., UTS screws and the wrenches that fit them are still referred to commercially as SAE. Worldwide, the ISO metric screw thread is now the standard, and it is slowly displacing all former standards, including UTS.

UTS screws are characterized by their thread diameter, length, material, and head style. As an example, a screw might be specified as a ¼-20 × 1-inch steel socket head cap screw, which means that the threads have ¼-inch diameter with 20 threads per inch, the screw is 1 inch long, and the screw material is steel. The head of such a screw has a hexagonal hole for a 3⁄16-inch hex key. In the U.S., the most common threads in optomechanical hardware are ¼-20 and 8–32. In the latter, the prefix 8 is a designator for a certain diameter, and the 32 means 32 threads per inch. For convenience, manufactured parts of a lidar should use these same threads so that both the screws and the tools that fit them will be compatible with the commercial components. Metric threads are specified by their thread diameter and the thread width, both in mm. The optomechanical metric threads closest to the UTS standards are M6 × 1 and M4 × 0.7. These numbers mean 6-mm diameter and 1 mm per thread and 4-mm diameter and 0.7 mm per thread, respectively. Optical breadboards, tables, and mounts are manufactured with both UTS and metric threads.

7.2.4 Precision Motion

When motion is needed, the tight alignment tolerances in optics usually require precision motion. Optical elements are moved by actuators, and the simplest actuator is the screw. One rotation of a screw moves the screw one thread pitch. In the U.S., machine shop micrometers are based on screws with 40 threads per inch, so one complete rotation corresponds to 0.025 inch, and the handles on micrometers have 25 gradations, each marking 0.001 inch. When EARL was first constructed, the alignment actuators on top of the secondary mirror were micrometer handles in a homebuilt kinematic mount, as shown in Figure 7.12. They controlled the mirror angles by pushing on the mounting plate at 4 inches (10 cm) from the ball bearing, so one gradation on a micrometer handle meant an angle change of 0.001/4 or 0.25 mrad. The law of reflection doubled this angle to 0.5 mrad, or 500 μrad. As listed in Chapter 6, EARL's misalignment tolerances are 387 μrad for short range and 137 μrad for long range, so this result means that the entire alignment adjustment was within one gradation. For this reason, the micrometer handles were replaced with precision actuators of the type shown in Figure 7.16, which moves 5 μm per major division. This corresponds to an alignment angle change of 100 μrad, which is much more appropriate.

In addition to precision angle adjustments, precision translation is sometimes required in optical instruments. For this purpose, rails that embody dovetail slides are available from several manufacturers. Rails permit translation of optical elements along a single direction, as shown in Figure 7.17, and the dovetail design is semikinematic because it constrains five degrees of freedom with area contacts, as opposed to point contacts.

7.2.5 Athermalization

The term *athermalization* refers to the design process of achieving thermal stability in optomechanical systems, which means minimizing variations in optical performance

Figure 7.16 A precision actuator.

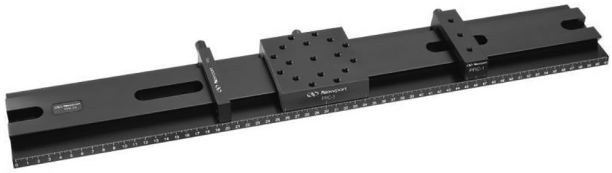


Figure 7.17 An optical rail with dovetail slides. Permission to use granted by Newport Corporation. All rights reserved.

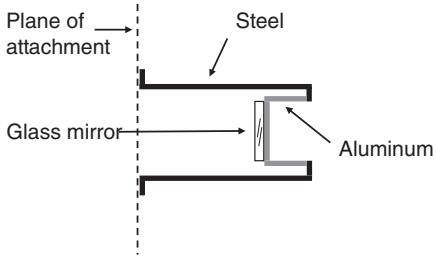


Figure 7.18 A simple athermalized mount. If the mirror thickness and the lengths of the steel outer tube and the aluminum inner tube are chosen properly, the distance from the plane of attachment to the mirror surface will be independent of temperature.

over a range of temperatures. Athermalization is a subdiscipline of optomechanics that includes several different aspects, from engineering the overall structure to clever ways of mounting and constraining mirrors and beamsplitters. Large astronomical telescopes are often athermalized because they must operate outdoors in a wide range of temperatures. Spaceborne instruments in polar orbits experience rapid temperature changes as they transit from full solar illumination to being in Earth's shadow and back, typically every 45 minutes. A simple example of an athermal mirror mount cross section is illustrated in Figure 7.18, where the expansions of the glass mirror and the aluminum tube exactly compensate for the expansion of the steel tube as the temperature changes.

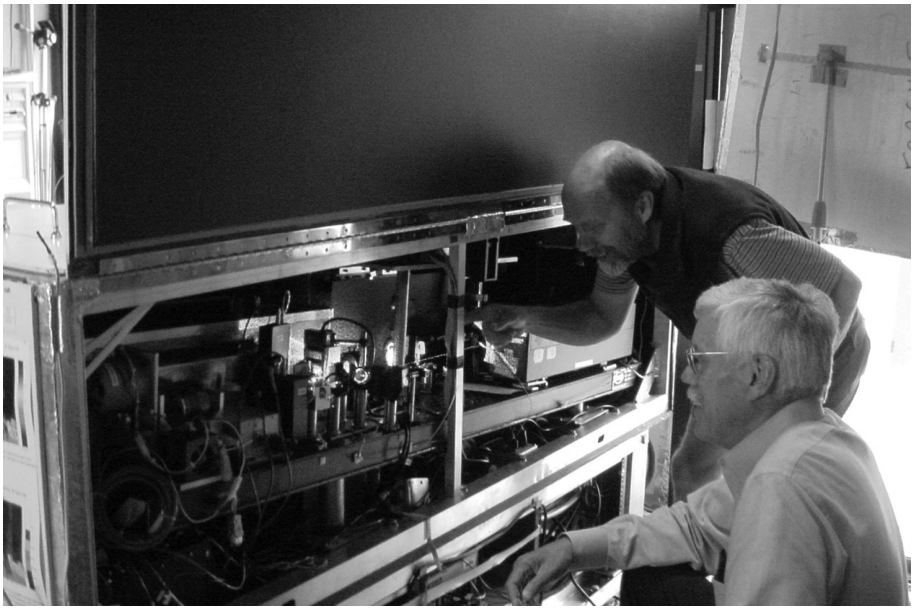


Figure 7.19 The Arctic HSRL. Ed Eloranta (rear) explains the Arctic HSRL to Gary Gimmestad.

The simplest way to avoid thermal problems in a lidar is to enclose the entire optical bench in a temperature-controlled housing. A good example, shown in Figure 7.19, is the University of Wisconsin Arctic HSRL. This common-optics lidar had the extremely small receiver FOV of 40 μrad and it was successfully controlled remotely using the Internet. It operated in Barrow, Alaska for three months in 2004 and in Eureka, Canadian territory of Nunavut, from July 2005 to June 2010. All UW HSRL operations are unattended, except for periodic checks on coolant level and window cleaning. Their best record for unattended operation was three years between maintenance visits, for a later HSRL version that operated in Barrow. The Arctic HSRL also illustrates the design approach of obtaining the necessary stiffness with a commercial breadboard and commercial optics mounts. As shown in Figure 7.19, the bench is tilted several degrees to move the lidar's FOV off the zenith, to avoid the problem with oriented ice crystals mentioned in Chapter 4.

7.3 Lidar Structures

The purpose of a lidar structure is to provide stable positioning and alignment for the optical elements or subsystems, including the transmitter-to-receiver alignment. The structure must be sufficiently strong and rigid for this purpose, and it must be either insensitive to temperature changes or else kept in a temperature-controlled environment. In mechanical engineering, the term *strong* means that the structure will not break, either under normal load or under some user error, and the term *rigid*,

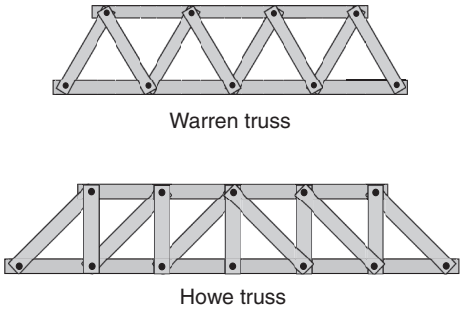


Figure 7.20 Common types of trusses. Both types derive their rigidity from triangles.

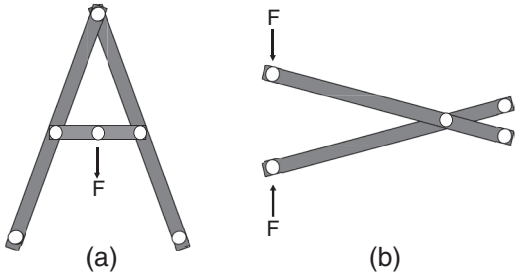


Figure 7.21 Mechanical frames. (a) A frame to support a load; (b) a frame to provide mechanical advantage.

like stiff, means that it will not flex excessively due to external forces (also called loads) on the structure. The loads on lidar structures are generally gravitational, and for zenith-pointing ground-based lidars, they are straight down. However, some lidars are scanned in zenith angle by tilting the entire lidar structure, in which case they must be engineered for changing gravitational loads.

7.3.1 Types of Structures

A *truss* is a structure composed entirely of members that are pinned or welded together. The two long beams are connected by braces, and each brace is connected at two points only. The rigidity of trusses is based on triangles. All parts of a truss are either in simple tension or compression; there are no bending forces on the members. Two common types of trusses are illustrated in Figure 7.20. They are simple to build, and they have a high strength-to-weight ratio.

A *frame* is a structure that consists of members pinned together, as shown in Figure 7.21, but an individual member can have more than two connections. Members can carry bending loads. Frames provide more design flexibility, and mechanical advantage can be obtained through frames (for example, a pair of pliers). If members are connected at joints that have some stiffness, then they constitute a *general structure*. Almost all real structures fall into this category, but most real structures can be modeled well as trusses or frames. A lidar structure must be strong enough for its

intended use, and it must also be stable, meaning that there is no clearance or backlash, commonly called "slop." The structure should be based on triangles, and load paths should be carefully designed. For example, all gravitational loads in EARL are carried straight down through four vertical frame members.

For a stable structure, conventional bolted connections are to be avoided. The term *bolting* means drilling clearance holes through two members and fastening them together with a machine screw and a nut. Washers are often used under the screw head and the nut. The term *clearance* means that the holes are slightly larger than the machine screw diameter, for ease of assembly. Bolting does not precisely locate two members with respect to each other; instead, it relies on friction to keep them from moving. For these reasons, welded or *pinned* joints are much preferred in optical structures. Pins fit tightly, with no clearance. The *shoulder bolt* is a variation on the pin that has a constant, tight-fitting diameter within the joint and a smaller diameter on the threaded end.

7.3.2 EARL Mechanical Structure

The need for a tall structure for EARL arose from limited funding. The transmitter telescope, which is a 20-cm Schmidt-Cassegrain, was donated to the project by Agnes Scott College, so the transmitter beam diameter D_T was fixed at 20 cm. Based on that diameter, calculations (using the pre-2014 ANSI MPE) showed that the maximum eye-safe energy per pulse would be 31.3 μJ. A set of measurement SNR goals was developed for four scenarios with a range of altitudes from the mixed layer to the stratospheric aerosol layer, and performance modeling results such as those shown in Figure 2.9 showed the need for a large receiver area. EARL's 61-cm receiver primary mirror was within the project budget, but a full astronomical telescope of that diameter was not. Mirrors with shorter focal lengths were available, but at much higher cost, so the long cone of light from the *f*/4.21 mirror was folded down with a flat secondary mirror in order to reduce the system's height. However, a coaxial configuration was chosen for lowest crossover altitude while avoiding the complexity and cost of common optics, and that decision put the transmitter above the secondary mirror spider and led to the overall system height of 2.97 m (about 10 ft). For these reasons, a tall structure was required with a mount for the primary mirror at the bottom and the receiver assembly, the secondary mirror, and the transmitter mounted above it, as shown in Figure 7.22. Supporting structures for secondary mirrors and other coaxial items are known as spiders, because they employ slender radial members to avoid blocking the light path.

Such a tall structure should be built with proper engineering and welded or pinned joints, but the project funding level dictated the use of steel U-channel with large holes at 2-inch (50.8 mm) intervals. The sides resemble Howe trusses, with horizontal members and angled braces. All the joints in the U-channel structure were bolted, which is contrary to the principle of welding, pinning, or using shoulder bolts. The holes in the U-channel were so large that oversized washers, known as fender washers, were required and the joints relied on friction for stability. In that way, the EARL

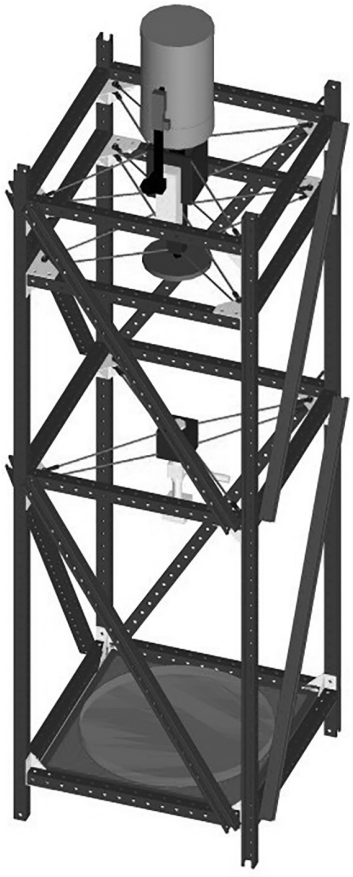


Figure 7.22 The EARL structure. All gravitational loads are carried by the four vertical frame members. Three spiders are used to support the receiver, the secondary mirror, and the transmitter. The primary mirror rests on a nine-point whiffletree mount.

structure is an example of wrong ways to do things. Such a structure is not very rigid, but an open framework has a great advantage in teaching: The students can easily envision the ray paths through it (shown in Figure 2.8) and explain them to their colleagues. It also had the advantage that the students themselves were able to assemble EARL in their lidar laboratory using simple hand tools.

The spider construction shown in Figure 7.23 also incorporated the concept of triangles. The spiders were built with rods that worked like turnbuckles, so that they put radially inward forces on the four corner posts of the structure. The rods were made from hexagonal steel, and they were threaded on both ends. The inner ends are threaded into a steel block at the center of each spider. The outer ends had a left-hand thread. They were screwed into spherical rod ends (also known as Heim joints) which accommodate their angles. The rod ends are bolted through gussets. There are four horizontal frame members at the level of each spider. As EARL was assembled,

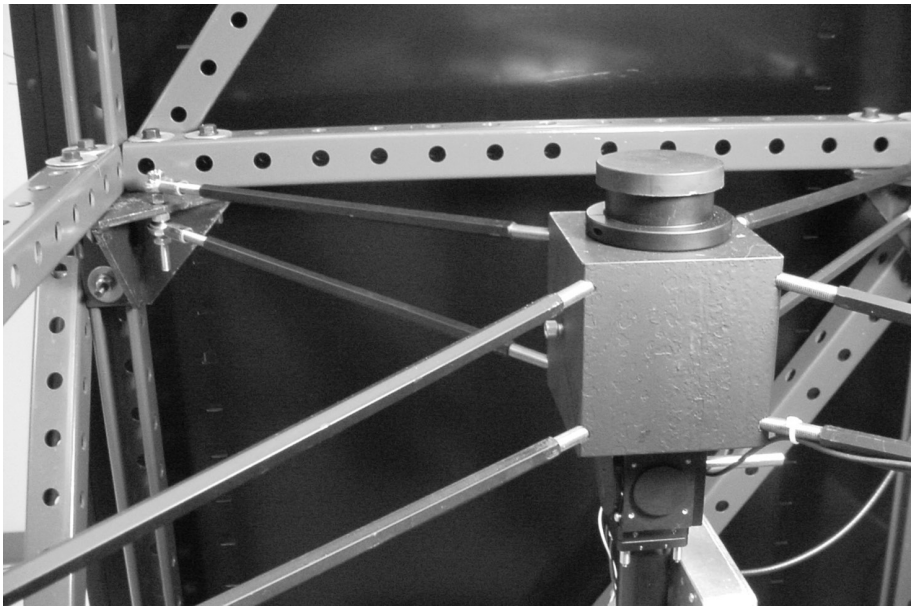


Figure 7.23 The receiver spider assembly.

the spider rods were rotated with a wrench to center and level their respective center blocks, while also putting large radially inward forces on the gussets. Those forces eliminated any clearance in the rod end bolts and helped to stiffen the overall structure by putting the horizontal side members in compression.

The EARL structure embodies some of the engineering concepts discussed in this section and it violates others. Because it is a fixed installation aimed at the zenith, it is rigid enough to maintain alignment for several weeks (if it is undisturbed), which is adequate for a teaching environment, but it is very sensitive to differential temperatures due to its great height, and it has an unforeseen torsional mode of oscillation in which the transmitter rotates back and forth about the OA. The torsional mode has a frequency of several Hz and it persists for a long time once it is excited, indicating that it is lightly damped. The low frequency and the persistence of this mode are due to the massive transmitter mounted on the top and the fact that nothing in the structure works to prevent the oscillation or damp it. The student operators were cautioned to avoid touching the structure during measurements because of the torsional mode of oscillation and its unknown effect on alignment. The transmitter and receiver are not on optical benches; they are three-dimensional because they are in the light path and must be compact. The receiver is built from 1- and 2-inch optical tubing and cubes, some of which can be seen in Figure 7.23 below the block of steel at the center of the spider.

Optomechanics is challenging because the mechanical parts must hold the optical parts securely without deforming them, while maintaining their relative positions and angles precisely. The best optical design in the world is worthless without adequate

optomechanical design, and the performance of an optical instrument is strongly determined by the mechanics. For prototypes and one-of-a-kind lidars, the use of commercial breadboards and mounts may very well suffice, especially if the environmental temperature is controlled. More sophisticated optical instruments require a formal optomechanical design process, such as that described by Yoder [1]. For working with any lidar or constructing a new lidar, a researcher should at minimum be aware of the principles described in this chapter, and the examples of how *not* to do things.

7.4 Further Reading

D. Vukobratovich and P. Yoder, *Fundamentals of Optomechanics*. Boca Raton: CRC Press (2018).

This classic textbook is known for its many tutorial examples and exercises, which are based on the vast experience of the authors. It is appropriate for self-study as well as classroom instruction. The topics covered include materials, optical components such as windows, prisms, and lenses, and kinematic design.

P. Yoder and D. Vukobratovich, *Opto-Mechanical Systems Design*, 4th ed. Boca Raton: CRC Press (2015).

This two-volume set by the same authors also has several chapters and sections of chapters contributed by other experts. It covers materials, opto-mechanical design, analysis of optical instruments, structures, kinematics, and applications of flexures. It is notable for having a total of 110 worked-out design examples to show how the theory, equations, and analytical methods can be applied by the reader.

J. E. Shigley and C. R. Mischke, *Mechanical Engineering Design*, 5th ed. Boston: McGraw-Hill (1989).

This basic text has been a standard in machine design for decades, covering its elements while emphasizing good design and problem-solving skills. It includes case studies and examples of real engineering situations.

7.5 Problems

7.5.1 Consider a thin BK7 mirror with a diameter of 0.61 m and a thickness of 25 mm. If it is supported only on its edge, as in Figure 7.7, how much does the mirror sag in the middle, and how many wavelengths of light at 532 nm is the sag? Assume Poisson's ratio ν is 0.2.

7.5.2 If a mirror has a focal length of 2.00 m, how much will the focal length change with a 10°C change in temperature, if (a) the mirror is made of BK7 and (b) the mirror is ULE?

7.5.3 A lidar engineer has a pendulum clock that keeps accurate time during winter but loses about one minute per month during summer. He suggests that the reason is that his home is warmer in the summer, so the pendulum is longer due to thermal expansion. Is this reasonable? The pendulum is steel, and the indoor temperatures are 20°C in winter and 24°C in summer. The period T of a pendulum is given by the relation $T = 2\pi\sqrt{l/g}$, where l is the length and g is the acceleration due to gravity.

7.5.4 What temperature difference between two opposite sides of the EARL structure illustrated in Figure 7.22 would cause misalignment of the long-range receiver channel by changing the transmitted beam direction? Assume that the steel framework is 2.5-m tall and 0.7-m wide. As stated in Chapter 6, EARL's misalignment tolerance is 137 μrad.

7.5.5 If a 1-cm thick mirror is athermalized by the mounting method shown in Figure 7.18 and the steel tube is 2-cm long, what is the length of the aluminum tube?

References

[1] P. R. Yoder, "Designing the durable optical instrument," in Proceedings of SPIE, 1988, vol. 0959.

[2] P. R. Yoder, *Mounting Optics in Optical Instruments*, 2nd ed. Bellingham, WA: SPIE Press (2008).

[3] J. H. Hindle, "Mechanical Flotation of Telescope Mirrors," in *Amateur Telescope Making, Book One*, A. G. Ingalls, Ed. (originally published in 1945). Richmond, VA: Willmann-Bell (1996).

[4] Optical Mirror Mounts Infographic. [Online]. Available: www.newport.com. [Accessed: July 24, 2021].

8 Optical Detection

Up to this point, we have considered the lidar signal only in terms of photons per range bin that reach the detector as a function of range, or equivalently, power on the detector as a function of range, and the only type of noise mentioned has been the unavoidable statistical fluctuation known as shot noise or Poisson noise. To store, average, and analyze lidar data in a computer, a digital electronic signal is required, and creating, processing, and storing such signals is squarely in the province of electrical engineering. All the components in the electronic signal chain can add electrical noise, so ways of including other noise sources in SNR calculations are required. Also, the specifications of commercial components are listed in the terminology of electrical engineering, so it is necessary for the lidar practitioner to be familiar with it. For these reasons, some of the language, terminology, and concepts of electrical engineering are introduced in this chapter. The SNR equations appear to be quite different, but they are consistent with previous chapters and parallels are drawn as appropriate.

From an electrical engineering point of view, acquisition of a lidar signal proceeds as a sequence of modifications to the initial light pulse from the laser, as illustrated in Figure 8.1. In electrical engineering, each modification is formally described by a *transfer function*, which is the mathematical representation of the relation between the input and output of a system. The description here is more heuristic, in keeping with previous chapters. The signal starts out as an energetic pulse of light from the laser that is only a few nanoseconds long, and what happens to that pulse is described by the stages shown by gray boxes in Figure 8.1, where the logical flow is from the top left corner through a backwards S pattern, ending in the lower right corner. The transmit optics control the laser beam parameters and cause some loss of pulse energy. The pulse then enters the atmosphere, which returns a tiny fraction of the pulse energy to the receive optics as a transient burst of light, known as a *waveform*, which is described by the lidar equation. The time duration of the transient is usually hundreds of microseconds (which is several orders of magnitude longer than the pulse width) and the atmosphere also adds noise in the form of the background light that is described by the background equation. The receive optics then transfer the optical waveform (with some loss) onto the detector element. What happens after the detector depends on the type of detection. Three types of optical detection are used in lidar systems: analog, photon counting, and coherent. The first two are called *direct detection*.

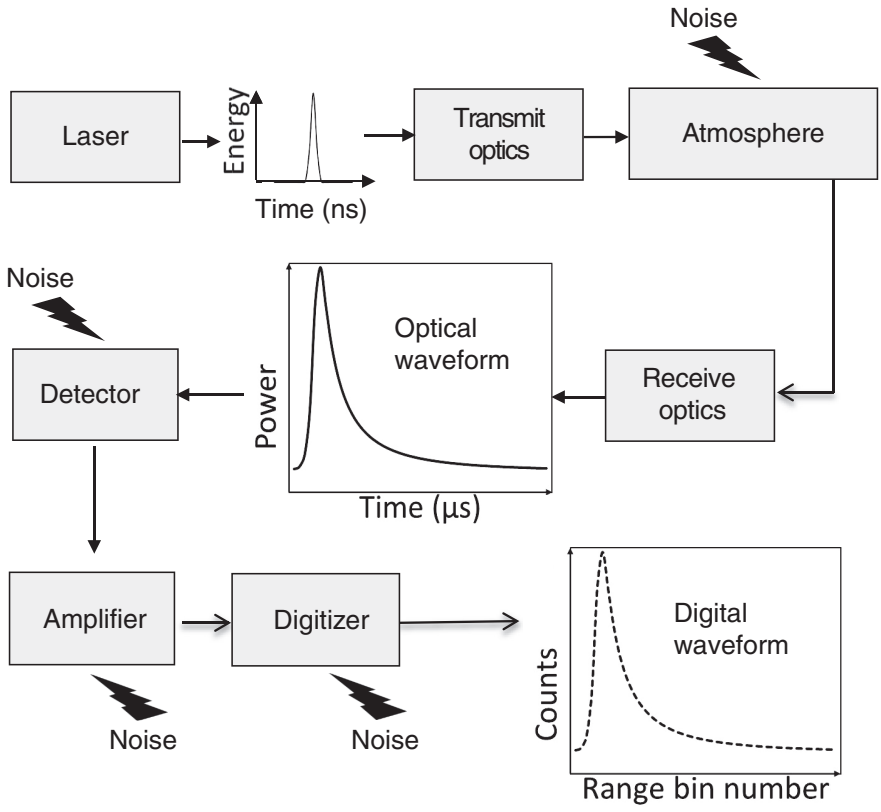


Figure 8.1 The stages in analog detection lidar data acquisition. The signal begins as a very short laser pulse that is greatly expanded, modified, and added to by the atmosphere, creating an optical waveform that is then detected, amplified, and digitized. Noise is added in several of the stages.

(1) In *analog detection*, the received optical power causes a current, which is converted to a voltage, amplified, and digitized. The detector usually incorporates a gain mechanism, meaning that its output current has been amplified internally. Analog detection is shown in Figure 8.1 and discussed in Sections 8.2, 8.3, and 8.4.

(2) In *photon counting*, the individual photons cause current pulses that are converted to voltage pulses and electronically counted. This technique eliminates several problems that analog detection suffers from, but it is not always applicable because the maximum rate at which the pulses can be counted is limited, so it is used when the optical signal is small. Photon counting is discussed in Section 8.5.

(3) *Coherent detection* is achieved by mixing the received light with coherent light from another source, thereby generating sum and difference frequencies. It provides advantages in both SNR and elimination of background light, and it is almost universally used in tropospheric wind-sounding lidars, which measure Doppler shifts due to the motion of the aerosols that cause the signal. However, the receiver aperture diameter is limited to about 10 cm by optical turbulence in

the atmosphere, so coherent systems are generally restricted to fairly short ranges. The basic principles of coherent detection are discussed in Section 8.6, but the practical aspects of coherent detection are sufficiently different from the first two techniques that they are outside the scope of this book.

Four of the stages in Figure 8.1 are susceptible to added noise of several kinds, shown by black lightning bolts, and distortions of the waveform can occur in the last three stages, so potential sources of noise and distortion and ways of avoiding them are discussed in the sections that follow. Some basic electronics concepts and terms are reviewed in Section 8.1, which is followed by three sections on optical detection techniques as well as detectors.

8.1 Basic Electronics

Relationships between three electrical quantities are illustrated in Figure 8.2, which is an electrical schematic diagram, and they are defined in Eqs. (8.1) and (8.2). If a voltage V in volts is applied across a resistance R in ohms (Ω), it causes a current i in amperes (A). Voltage is an electrical potential difference, always measured between two points in a circuit. The small circles in the diagram indicate connections or test points. Electrons flow from negative to positive, but current is defined as a flow of electrical charge from positive to negative, for historical reasons. The ampere is 1 C/s, and the coulomb (C) is the charge carried by 6.24×10^{18} electrons. The relationship between the three quantities defined in Eq. (8.1) is known as *Ohm's law*. The electrical power dissipated in the resistor is stated three ways in Eq. (8.2).

$$V = iR \tag{8.1}$$

$$P = iV = i^2R = V^2 / R. \tag{8.2}$$

A schematic diagram for a basic voltage amplifier circuit is shown in Figure 8.3. The triangle in the center is the symbol for an operational amplifier, also known as an *op-amp*. An op-amp is a DC-coupled high-gain electronic voltage amplifier with a differential input and (usually) a single-ended output. The term was first coined in connection with analog computers, referring to an amplifier for performing mathematical operations. The input and output voltages are relative to ground, and V+ and V– are the power supply voltages. The circuit has a characteristic input impedance that is usually 50 Ω. This impedance is important because other components of the electronic circuitry, including their interconnecting cables, have characteristic impedances and it is important to avoid any impedance mismatches. A mismatch will cause a reflection, in the same way, a refractive index mismatch causes the optical Fresnel reflections shown by Eqs. (6.5) and (6.6). Such reflections can cause a distortion of the waveform, which must be avoided at all costs in lidar. The gain of the amplifier can be expressed as the linear multiplicative factor R_f/R_{in} as it is in Figure 8.3, but more often in electronics it is in decibels (dB). A power gain is defined as $10\log_{10}(P_{out}/P_{in})$, which is equal to $20\log_{10}(V_{out}/V_{in})$, because power is proportional to voltage squared.

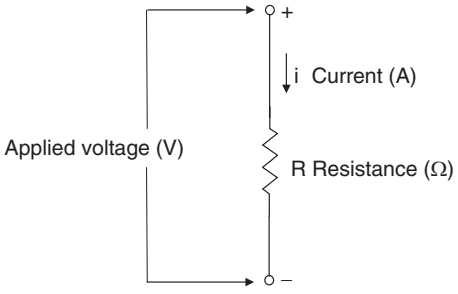


Figure 8.2 Electrical circuit relationships.

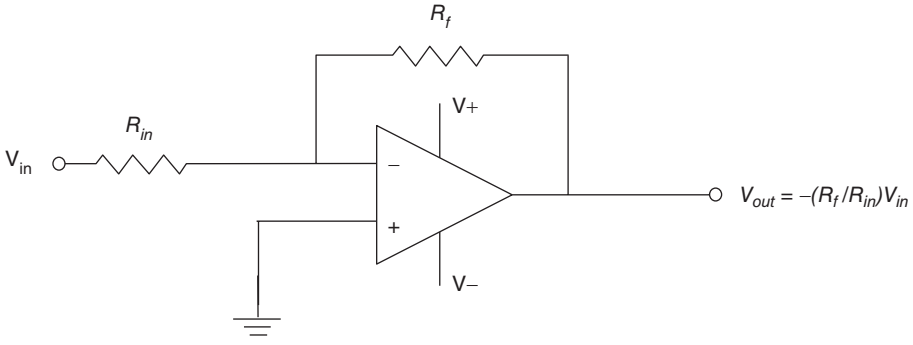


Figure 8.3 A voltage amplifier circuit. The triangle is the symbol for an amplifier, and the voltage gain is R_f/R_{in}.

The frequency response of circuitry is a major consideration in electrical engineering, and attention to it is required to avoid electronic distortions of the lidar signal, so the *bandwidth* B of an amplifier is an important characteristic for lidar. Electronic circuits cannot respond to arbitrarily high frequencies, but rather their response, in terms of power, has a roll-off toward zero at some point, as shown in Figure 8.4. This type of response is called a *low-pass filter*, and ideally the response is constant from zero frequency out to the point where it begins decreasing, and the frequency where the response has decreased by a factor of 2 (–3 dB) is called the *critical frequency* f_c. The *equivalent bandwidth* is defined by the width of a rectangular box that has the same area as the area under the response curve.

The power version of the lidar equation models the lidar signal in the *time domain* as a waveform, but any waveform can be disassembled into its frequency components by Fourier transformation. A few examples are shown in Figure 8.5, which illustrates several important points. First, there is only one function that transforms into itself, and that is the Gaussian. Time and frequency are inversely related, so a small time interval implies a large frequency interval; the frequency range Δf corresponding to a time interval Δt is roughly $1/\Delta t$. This is the basis for the constraint on the time-bandwidth product in Section 5.2.6 (the factor of 0.44 rather than unity comes from measuring the pulse width as FWHM). Second, the sine wave (shown

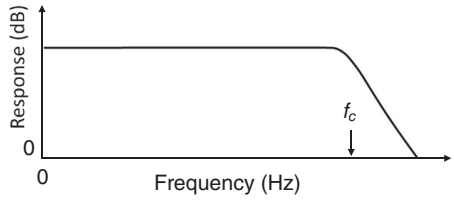


Figure 8.4 Electronic bandwidth. Filter response curves are normally logarithmic on both axes.

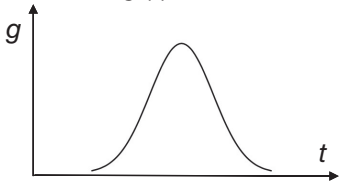
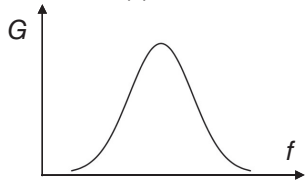
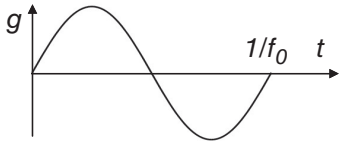
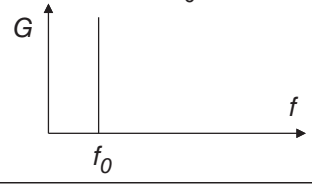
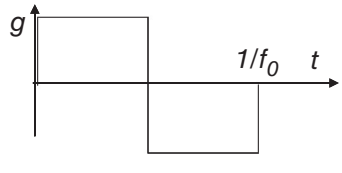
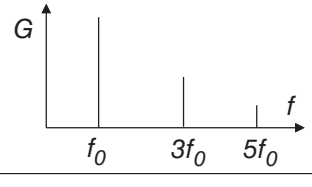
Function	Time domain	Frequency domain
Gaussian	$g(t) = e^{-t^2}$ 	$G(f) = e^{-f^2}$ 
Sine wave	$g(t) = \sin(2\pi f_0 t)$ 	$G(f) = \delta(f_0)$ 
Square wave	$g(t) = \text{sgn}[\sin(2\pi f_0 t)]$ 	$G(f) = (4/\pi)[\sin(2\pi f_0 t) + 1/3 \sin(6\pi f_0 t) + ...]$ 

Figure 8.5 Fourier transform examples. In each case, the time domain is shown with its corresponding frequency domain. A function $g(t)$ is in the time domain and $G(f)$ is in the frequency domain. Only the first three frequency components are shown for the square wave transform.

as one cycle but assumed to have an infinite extent in the time domain) appears at a single point in the frequency domain. Third, the square wave, with its very fast rise, corresponds to an infinite series of odd-harmonic sine waves, including frequencies that increase without limit. The square wave is not a continuous function; the sign function (sgn) is positive when the sine function is positive, negative when it is negative, and zero at the discontinuities. This last example illustrates a general fact:

The faster a time domain signal rises, the higher the frequency components it contains. The risetime of a lidar signal is determined by the geometrical function $G(R)$, and care must be taken when designing it to avoid exceeding the bandpass of amplifiers in the data system.

Example. To estimate the highest frequency in the optical waveform at the center of Figure 8.1, we note that the spatial scale of information is $c\tau_{\text{pulse}} / 2$. Assuming a typical transmitted laser pulse width of 6.0 ns, that scale is $[(3 \times 10^8) \times (6.0 \times 10^{-9})] / 2 = 0.90$ m. Spatial structures this small have been observed in the atmosphere's residual aerosol layer. The corresponding maximum frequencies in the lidar waveform would be on the order of $1 / \tau_{pulse}$, which $1 / (6.0 \times 10^{-9}) = 1.7 \times 10^8$ Hz or 170 MHz, and most lidars have much coarser spatial resolution due to the bandwidth of their detector–amplifier combination. EARL has 15-m range bins and a 5 MHz bandpass. Because of lidar electronic bandwidth limitations, the final digital waveform in Figure 8.1 almost never contains all the spatial information that is in the optical waveform. This is not usually a problem, because spatial resolutions on the order of tens to hundreds of meters are sufficient for most atmospheric science applications.

8.2 The Direct Detection Process

There are several different types of optical detector, but for lidar receivers, only *quantum detectors* are of interest. Quantum detectors respond to individual photons, as opposed to a physical effect of many photons, such as heating. There are two basic types of quantum detector, as illustrated in Figure 8.6. The photoelectric devices include *vacuum photodiodes* and *photomultiplier tubes* (PMTs), and the semiconductor devices are known as *photodiodes*.

The photoconductive detection process is illustrated in Figure 8.7. Photons incident on the photodiode raise electrons to a conduction band, and the resulting stream of electrons constitutes a current. The current is converted to a voltage signal by a *transimpedance amplifier* (TIA). The photodiode is reverse biased, so the arrow in the diode symbol points toward a positive bias voltage. The output voltage is proportional to the photodiode current and the gain, in V/A, is simply the value of the feedback resistor, R_f. The gain is sometimes specified in ohms, because the ohm is voltage/amperage, or V/A.

The homebuilt EARL power meter mentioned in Section 5.1 was based on the circuit shown in Figure 8.7. The detector was a large-area silicon photodiode with an op-amp in the same package. The gain and the time constant of the op-amp were determined by an external resistor and a capacitor. EARL acquired data in bursts of 2000 pulses that occurred in 0.8 s. The design time constant was 0.27 s, which meant that the power meter's response to a change would reach 95% of its full value in 0.8 s (explained below). The power meter's analog output voltage was

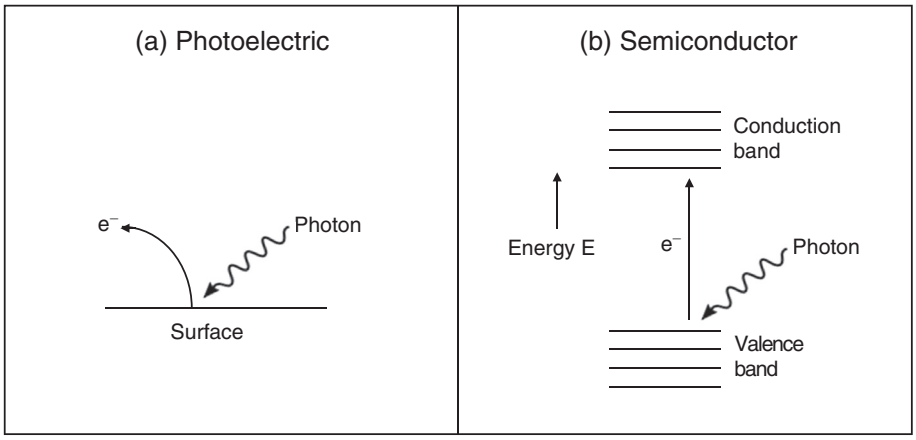


Figure 8.6 The two types of quantum detector. The symbol e^- denotes an electron. In (a), an incident photon ejects an electron from a surface in a vacuum tube device; in (b), an incident photon raises a valence electron to a conduction band in a solid-state device.

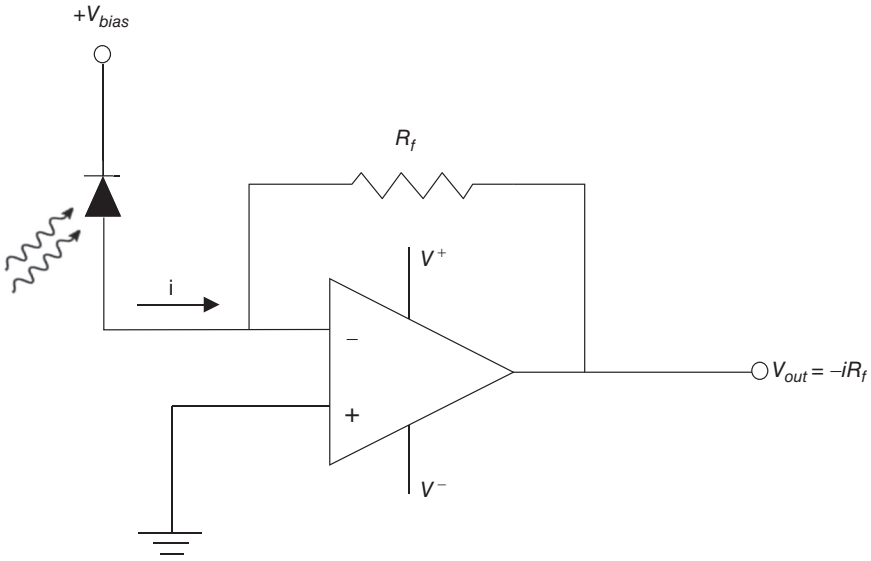


Figure 8.7 The photodetector-amplifier circuit. The current from a photodiode is converted to a voltage by a transimpedance amplifier with gain R_f (V/A).

digitized by a LabJack ™ model U12, which is a multichannel input/output device with a USB interface. The power meter was tested to verify linearity and long-term stability, and it was calibrated against a commercial power meter by using an off-axis paraboloidal mirror to catch the entire 20-cm transmitted beam and focus it onto the commercial power meter. Photographs of the homebuilt power meter are shown in Figure 8.8.

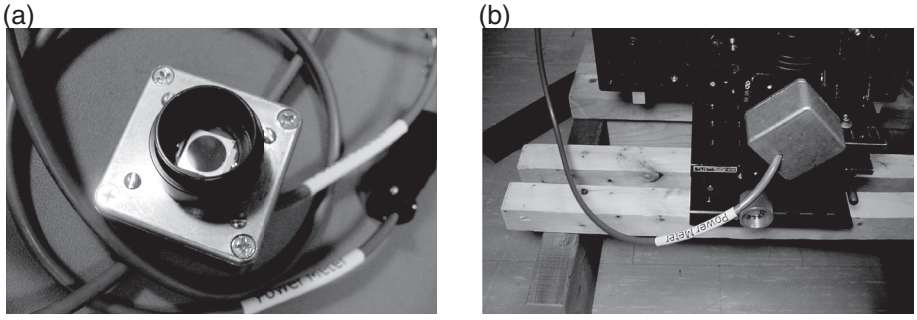


Figure 8.8 The homebuilt EARL power meter. (a) view of the 1-cm diameter detector; (b) the power meter installed on the EARL transmitter.

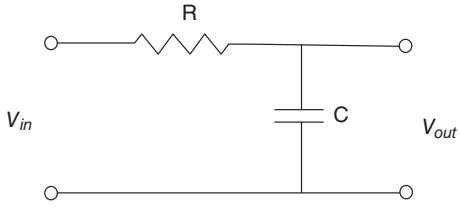


Figure 8.9 The RC filter. The time constant in seconds is the product RC.

One other important circuit is the RC filter, shown in Figure 8.9. The RC product is in units of seconds, because the units of resistance are V/A, which is V/C/s, and the units of capacitance are C/V (farads). The RC product is known as the *time constant* of the circuit shown in Figure 8.9. This circuit is the most widely used low-pass filter in electronics because of its simplicity, but for limiting the bandwidth in lidar electronics its performance is poor. In the time domain, the circuit in Figure 8.9 smooths out sudden changes due to high frequencies, which is a valuable feature for filtering out noise, but it has a slow response to changes. If $V_{\text{in}}(t)$ is a step function that increases from zero to V_{in} at time $t = 0$, $V_{\text{out}}(t)$ is given by

$$V_{\text{out}}(t) = V_{\text{in}}[1 - \exp(-t / RC)], \tag{8.3}$$

which is shown graphically in Figure 8.10 for $V_{\text{in}} = 1$. The time axis is in units of RC. A common rule of thumb is that it takes at least three times RC to get to the final value of the change, which is an unnecessarily long time: If an experimenter wants to average a signal over a time RC, he must wait for a time of at least 3 *RC* to be assured of getting within 95% of the correct value. In the frequency domain, the *RC* filter response expressed as a gain is

$$G(f) = \frac{1}{1 + (2\pi RCf)^2}, \tag{8.4}$$

where the range of G is 0–1. A plot of the frequency response of the filter in Figure 8.9 is shown in Figure 8.11, for R = 50 Ω and C = 1 nf. Electronic gains are usually

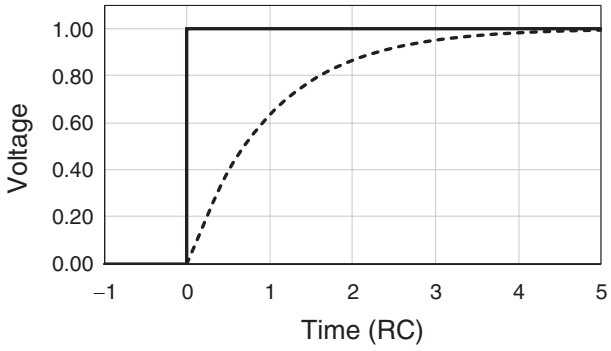


Figure 8.10 The RC filter time response. The solid line is the step function input; the dashed line is the output voltage $V_{\text{out}}(t)$.

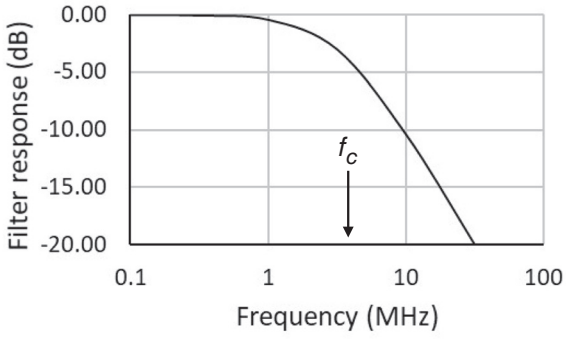


Figure 8.11 The RC filter frequency response. The curve is the filter response for $R = 50\ \Omega$ and $C = 1$ nf.

expressed in the logarithmic units of decibels (dB), defined by $G = 20\log_{10}(V_{\text{out}} / V_{\text{in}})$. Attenuations are gains less than unity, so their values in dB are negative. The half-power point (–3 dB) occurs at the *critical frequency* $f_c = \frac{1}{2}\pi RC$, which is 3.18 MHz for the example shown. More sophisticated bandpass filters are generally used in high-quality electronics, but accidental capacitance, known as *parasitic capacitance*, is commonplace in circuitry and cabling, and RC time constants due to parasitic capacitance often determine the upper limit to bandwidth. This problem in TIAs is addressed in Section 8.4.

The equivalent bandwidth is the width of a rectangular response with the same area as the actual response. Integrating Eq. (8.4) from zero to infinity yields an equivalent bandwidth $B = \frac{1}{4}RC$, or 5 MHz in the example discussed here. Equation 8.4 is plotted on linear scales in Figure 8.12, along with the rectangular function with the same area. The filter response below about 1 MHz appears to be constant on a log-log plot, but Figure 8.12 shows that it is actually not constant. Logarithmic plots accommodate many orders of magnitude, but they smooth out detail.

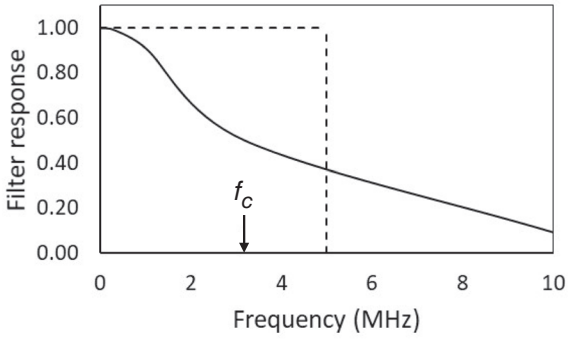


Figure 8.12 Equivalent bandwidth. The solid curve is the RC filter response defined by Eq. (8.4) with the same parameters as in Figure 8.11. The dashed lines are the rectangular response with the same area, which defines the equivalent filter bandwidth. The critical frequency is 3.18 MHz.

8.3 Analog Detection and SNR

As illustrated in Figure 8.1, the photodetector bridges the gap between the receiver optics and the signal electronics. Its job is to produce a time-varying current that is proportional to the rate at which photons are incident on its surface. The frequency response of a signal is sometimes constrained by the detector, and the electronic signal will be distorted if the detector output is nonlinear with respect to the input. The detector and amplifier together produce a voltage waveform that is intended to be a faithful (though filtered) reproduction of the optical waveform, shown in the center of Figure 8.1, that is provided by the lidar optics. When consecutive photons are close enough together in time so that their pulses from the detector–amplifier stages overlap, a digitizer (also known as an analog-to-digital converter or ADC, see Chapter 9) converts the electronic signal to discrete values while adding its own type of noise. This approach is called *analog signal processing*. If the voltage pulses are far enough apart in time, the pulses are shaped and counted in a process called *photon counting*, which is covered in Section 8.5. In the interest of simplicity, only detectors without internal gain are discussed at first, even though they are the exception in lidar receivers. Detector gain is covered later in Section 8.7. A great variety of photodetectors is available, so the question of which one is best for a given application must be addressed. This question is partially answered by the figures of merit listed in Table 8.1 and discussed in the paragraphs that follow. Much of the SNR analysis in this chapter was adapted from [1].

The starting point for the definitions in Table 8.1 and for SNR calculations is to note that when an optical power P is incident on such a quantum detector, electrons are generated at an average rate r_{avg} (electrons per second) given by

$$r_{\text{avg}} = \eta \frac{P}{h\nu} = \frac{\bar{i}}{q}, \tag{8.5}$$

Table 8.1 Detector figures of merit

Term	Definition	Units	Comments
Quantum efficiency (η)	$\frac{electrons\ out}{photons\ in}$	Unitless	Before detector gain is applied
Responsivity (S)	$\frac{output}{input}$	A/W	Often as a function of wavelength
Noise equivalent power (NEP)	$\frac{Noise\ (V)}{Reponsivity\ (V/W)}$	W	Power to create SNR = 1
Spectral NEP	$\frac{Noise\ current\ (A)}{Responsivity\ (A/W)\sqrt{Bandwidth\ (Hz)}}$	W/Hz$^{1/2}$	Power to create SNR = 1 in a given bandwidth
Detector integration time (τ_{int})	$\tau_{int} = \frac{1}{2B}$	s	Measure of noise bandwidth

where $\bar{i}$ is the detector's mean output current in amperes and q is the charge on the electron, 1.602×10^{-19} coulombs. Because electric charge is in discrete units of q, electron generation obeys the Poisson distribution described in Eq. (2.5) and illustrated in Figure 2.1. The number of electrons generated in a time interval τ is $r_{avg}\tau$, so the probability of generating n electrons in an interval τ is given by

$$p(n,\tau) = \frac{(r_{avg}\tau)^n \exp[-r_{avg}\tau]}{n!}, \tag{8.6}$$

and the variance σ^2 is equal to the mean number $r_{avg}\tau$.

Equation 8.5 shows that the output current i is proportional to the optical power P. However, the power in an electronic circuit is proportional to i^2, so the electronic power is proportional to the square of the optical power. This may seem counter-intuitive at first, but the electronic power does not come from the incident photons; it comes from a power supply. The incident photon merely liberates an electron that is moved through the detector by an applied voltage. The square of the current corresponding to an optical power P is

$$i^2 = \left(\eta \frac{q}{h\nu}\right)^2 P^2. \tag{8.7}$$

The quantum efficiency η, also called QE, is the ratio of electrons out to photons in. The optical SNR defined in Eq. (2.8) is proportional to the square root of η, so a large value is highly desired. The QE can be derived from Eq. (8.5) as

$$\eta(\lambda) = \frac{h\nu\bar{i}}{qP}. \tag{8.8}$$

The QE calculated from photocurrent and incident power according to Eq. (8.8) is cited by manufacturers, but it is often specified only at the wavelength of maximum responsivity.

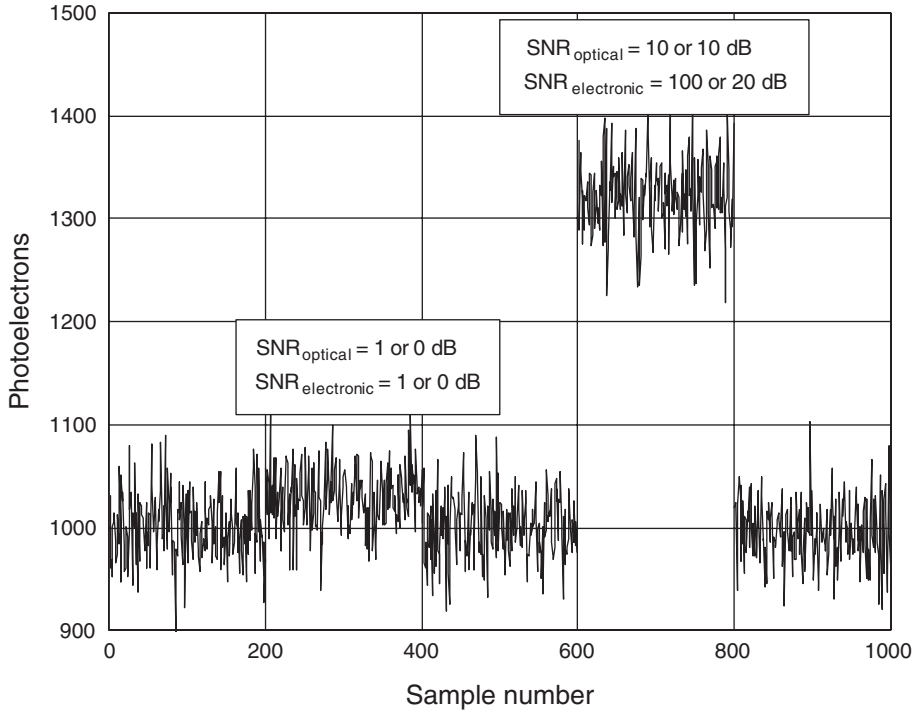


Figure 8.13 Optical and electronic SNRs.

The *responsivity* S is the output divided by the input, i / P, in amperes per watt. The output signal current depends on the incident optical power and the detector quantum efficiency η (and the gain G if the detector has gain). The responsivity increases with wavelength, but this is merely an artifact of the units used to define S: More photons are needed to make a watt as the wavelength increases. Equation (8.5) can be rearranged to express S in terms of other parameters as

$$S = \frac{i}{P} = \frac{\eta q}{h\nu}. \tag{8.9}$$

To understand the next parameter in Table 8.1, NEP, it is helpful to use the electronic definition of SNR, which is the mean squared divided by the variance, or the square of optical SNR, because the electronic power is the square of the optical power. Electronic SNR is usually expressed in dB, so $\text{SNR} = 10\log_{10}(P_{\text{signal}}/P_{\text{noise}})$ where the powers are electronic, not optical. Optical and electronic SNRs, linear and logarithmic, are compared in Figure 8.13, which is the same as Figure 2.4 but with additional notations.

In Chapter 2, SNR was discussed in terms of photoelectrons per range bin, but the output of the detector in Figure 8.1 is a continuous current waveform, so range bins do not apply at that stage of analog detection. However, the detector does have a bandwidth that limits its frequency response. The bandwidth also limits the noise, because

the noise spectrum is assumed to be white (independent of frequency), so the total noise is proportional to the bandwidth. As shown in Eq. (8.2), electronic power is proportional to current squared. For this reason, electronic SNR is defined as the ratio of signal power to noise power, which is equal to a ratio of squared currents, defined by

$$\mathrm{SNR}_{\mathrm{power}} = \frac{i_{\mathrm{S}}^2}{\overline{i_{\mathrm{N}}^2}}, \tag{8.10}$$

where i_{S} is the signal current and $\overline{i_{\mathrm{N}}^2}$ is the mean-square noise current. In general, for j independent noise sources,

$$\mathrm{SNR}_{\mathrm{power}} = \frac{i_{\mathrm{S}}^2}{\sum_j \overline{i_{\mathrm{N}_j}^2}}, \tag{8.11}$$

because statistically independent variances add together. The statistical limit to SNR is shot noise, which is governed by the Poisson probability distribution discussed in Chapter 2 and shown in Eq. (8.6) for photoelectrons. The familiar relationship between the mean and the variance is that, with an average generation rate of photoelectrons r_{avg}, the variance σ^2 in the number in a time interval τ is the mean number $r_{\mathrm{avg}}\tau$, as mentioned after Eq. (8.6). There is shot noise associated with a current because the charge carriers are discrete particles, so sources of electronic noise are modeled as *noise currents*. The question then is how to find the mean-square noise current. Recalling that r_{avg} electrons are generated in a time τ, the mean current is $\bar{i} = r_{\mathrm{avg}}(q/\tau)$. If many measurements are made over equal time intervals τ, the mean squared current fluctuation will be $\overline{i_{\mathrm{N}}^2} = \overline{r_{\mathrm{avg}}^2}(q/\tau)^2$, but the mean squared value of r_{avg} is just r_{avg} because of Poisson statistics, so $\overline{i_{\mathrm{N}}^2} = r_{\mathrm{avg}}(q/\tau)^2$, which is $\bar{i}(q/\tau)$. From Table 8.1, the bandwidth of the system is given by $B = 1/2\tau$, so the mean-square noise current is

$$\overline{i_{\mathrm{N}}^2} = 2q\bar{i}B, \tag{8.12}$$

where q is the electron charge, $\bar{i}$ is the average current, and B is the equivalent bandwidth as defined before Figure 8.4 [1]. In electronics, all noise sources are modeled as currents, so the mean current $\bar{i}$, which is usually written as simply i, can be signal, background, dark current, or current from other sources. Amplifier noise is modeled as a noise current. The fact that the mean-square noise current is proportional to bandwidth is a key feature of electronic noise analysis. If the detector integrates during a time τ_{int}, its effective bandwidth is $B = 1/2\tau$, as shown in Table 8.1. If the frequency response is determined by a simple RC filter, the effective bandwidth is $B = 1/4RC$.

If a series of N range bins is averaged together, the bandwidth is B/N. This fact represents a trade-off: With averaging, the bandwidth decreases and therefore the noise current variance in Eq. (8.10) decreases, but decreasing bandwidth means that the frequency content of the lidar signal decreases, which means that the spatial resolution of the lidar increases. Lidar waveforms are routinely smoothed during analysis with various levels of averaging (described in Chapter 10) and that averaging has the effect of lowering the bandwidth.

In addition to shot noise, *Johnson noise*, or *Johnson-Nyquist* noise, arises from the thermal agitation of charge carriers in a conductor. It behaves exactly like shot noise, but the mean current $\bar{i}$ is zero [2]. Johnson noise is modeled as a current source in parallel with a resistor, with the mean-square noise current given by

$$\overline{i_{\mathrm{N}}^2} = \frac{4kTB}{R}, \tag{8.13}$$

where k is the Boltzmann constant (1.38×10^{-23} J/K) and R is the resistance in ohms.

Example. For a 10 kΩ resistor at 300 K with a 10 MHz bandwidth, the mean squared noise current is $4 \times (1.38 \times 10^{-23}) \times 300 \times 10^7 / 10^4 = 1.66 \times 10^{-17}$ A^2.

The phenomenon of detector dark counts was mentioned in Chapter 2. In electronics, those counts constitute a current. Dark current has two main sources: (1) thermal charge carrier generation in both photodiodes and PMTs; and (2) leakage current in reverse biased diodes. Dark current produces a signal offset because its average value adds directly to the signal current, and it exhibits shot noise just like any other current, so it degrades the SNR. The mean-square noise current is found as shown in Eq. (8.12) as

$$(\overline{i_{\mathrm{N}}^2})_{\mathrm{dark}} = 2q\bar{i}_{\mathrm{dark}}B. \tag{8.14}$$

If the dark current and other noise currents are negligible, the limiting electronic SNR results from shot noise in the photodetector signal and background currents. Those currents, which are found from the optical powers on the detector using Eq. (8.7), are

$$i_S = \eta \frac{q}{h\nu} P_S \text{ and} \tag{8.15}$$

$$i_B = \eta \frac{q}{h\nu} P_B. \tag{8.16}$$

The mean-square shot noise current calculated using Eq. (8.12) is then

$$\overline{i_N^2} = 2q(i_S + i_B)B = 2q^2\eta \frac{P_S + P_B}{h\nu} B. \tag{8.17}$$

The electronic SNR is the ratio of the signal power to the noise power, so

$$\mathrm{SNR}_{\mathrm{power}} = \frac{i_{\mathrm{S}}^2}{\overline{i_{\mathrm{N}}^2}} = \frac{\eta P_{\mathrm{S}}^2}{2h\nu B(P_{\mathrm{S}} + P_{\mathrm{B}})}. \tag{8.18}$$

Equation (8.18) may look unfamiliar, but it is just the square of the shot noise-limited optical SNR derived in Chapter 2 (the proof is left as a problem for the student). A main reason for switching to the language of electrical engineering is to accommodate noise sources other than shot noise, such as Johnson noise and amplifier noise. Those noise sources are added in Section 8.4.

The third figure of merit in Table 8.1 is the noise equivalent power (NEP), which is defined as the optical signal power required to make the SNR equal to unity. Smaller

NEP values are better because they correspond to more sensitive detection. The NEP in the statistical limit is found by setting the SNR in Eq. (8.18) equal to 1 and solving for the signal power, yielding

$$\mathrm{NEP} = \frac{h\nu B}{\eta} + \sqrt{\left(\frac{h\nu B}{\eta}\right)^2 + \frac{2h\nu B P_\mathrm{B}}{\eta}}. \tag{8.19}$$

The signal-limited and background-limited values of the NEP are then

$$\mathrm{NEP}_\mathrm{SL} = \frac{2h\nu B}{\eta} \quad P_\mathrm{B} = 0, \text{ and} \tag{8.20}$$

$$\mathrm{NEP}_\mathrm{BL} = \sqrt{\frac{2h\nu B P_\mathrm{B}}{\eta}} \quad P_\mathrm{B} \gg P_\mathrm{S}. \tag{8.21}$$

Equations (8.20) and (8.21) may also look unfamiliar, but like Eq. (8.18) they are consistent with the equations derived in Chapter 2. Equation (8.20) illustrates the importance of detector QE for signal-limited detection. Because the NEP has an inverse dependence on η, as the QE becomes small, the NEP rises rapidly. By definition, NEP values only correspond to situations where SNR is unity, but NEP is a quite common figure of merit for detectors, especially in the infrared regions. At wavelengths longer than about 3 μm, the NEP values quoted by vendors are almost always for background-limited detection, because such detectors are used in thermal imaging systems that have wide FOVs and large optical bandwidths. Such systems operate in spectral regions where the background radiance is due to thermal radiation, which is always present.

If a detector creates a dark current, the current can be thought of as being produced by an additional source of photons illuminating a detector without dark current. The dark current is proportional to the average optical power P_D of the additional photons times the responsivity S defined in Eq. (8.9):

$$i_\mathrm{D} = \eta \frac{q}{h\nu} P_\mathrm{D}. \tag{8.22}$$

To be compatible with Eq. (8.17), we solve Eq. (8.22) for P_D and follow the same procedure as for Eqs. (8.15)–(8.18). Equation (8.17) gets an additional term as

$$\overline{i_\mathrm{N}^2} = 2q(i_\mathrm{S} + i_\mathrm{B} + i_\mathrm{D})B = 2q^2\eta \frac{P_\mathrm{S} + P_\mathrm{B} + \dfrac{h\nu}{\eta q} i_\mathrm{D}}{h\nu} B, \tag{8.23}$$

and the SNR becomes

$$\mathrm{SNR}_\mathrm{power} = \frac{i_\mathrm{S}^2}{\overline{i_\mathrm{N}^2}} = \frac{\eta P_\mathrm{S}^2}{2h\nu B\left(P_\mathrm{S} + P_\mathrm{B} + \dfrac{h\nu}{\eta q} i_\mathrm{D}\right)}. \tag{8.24}$$

Equations (8.18) and (8.24) are only valid for detectors without internal current gain. Photomultipliers and avalanche photodiodes have substantial gain, so methods to include detectors with gain in SNR calculations are covered in Section 8.6. The

mean-square noise currents in the denominator of Eq. (8.24) are summed because variances add for independent Gauss-distributed random processes. When a circuit has multiple noise sources, as in the detector–amplifier shown in Figure 8.7, additional terms appear in the denominator and SNR can be conveniently expressed as

$$\mathrm{SNR}_{\mathrm{power}} = \frac{P_S^2}{\mathrm{NEP}_{\mathrm{SL}} P_S + \mathrm{NEP}_{\mathrm{BL}}^2 + \mathrm{NEP}_{\mathrm{D}}^2 + \mathrm{NEP}_{\mathrm{AL}}^2}, \tag{8.25}$$

where signal fluctuations are described by $\mathrm{NEP}_{\mathrm{SL}}$, background fluctuations by $\mathrm{NEP}_{\mathrm{BL}}$, detector dark current by $\mathrm{NEP}_{\mathrm{D}}$, and the amplifier noise by $\mathrm{NEP}_{\mathrm{AL}}$. The first two terms in the denominator are statistical and unavoidable, but the last two are technological and they can therefore be minimized by good engineering and proper selection of the best available components.

The first lidar designed and constructed at GTRI was called the Cloud Climatology Field Unit (CCFU) [3]. It was advanced for its time (1986) in that it operated unattended in a transportable shelter, acquiring cloud data every 15 minutes. The CCFU used a wavelength of 10.6 μm for eye safety, where detectors with gain are unavailable and SNR is limited by amplifier noise. The measurement requirement was to detect all clouds including high, thin cirrus, and the CCFU failed to achieve that goal. The designer had extensive experience with lidar systems, but he had never worked in the LWIR and did not realize that the NEP values of commercially available detector–amplifier combinations were too high (on the order of 10^{-9} W) in conjunction with the other system parameters, to enable cirrus detection.

The MWIR is also eye safe, as shown in Chapter 5, and it is a region of great interest to lidar because many molecules of interest including GHGs have absorption spectra there. However, MWIR detectors with gain are not available, and care must be taken to achieve sufficiently low NEPs. The detectors are in cryostats, often cooled with liquid nitrogen to 77 K, so the background power can be lowered with cold baffles in the cryostats to restrict the FOV. The zenith sky spectral radiance in the LWIR is about one order of magnitude smaller than the daytime VIS value, and in the MWIR it is at least two orders of magnitude lower [4], but thermal sky background is present during both day and night. A cold bandpass filter can also be used to minimize the background power reaching the detector. Johnson noise in the TIA's feedback resistor can be decreased by placing it in the cryostat. Such techniques increase the complexity and cost of the receiver, but they are sometimes necessary for obtaining useful SNR. Kruse et al. [5] showed an example in which a cold baffle with a cold lowpass filter improved the NEP by an order of magnitude, for a 3–5 μm lead selenide detector operating at 77 K.

8.4 Analog Detection Circuitry

After the detector, the next block in Figure 8.1 is the amplifier. The TIA shown in Figure 8.7 turns the detector current into a measurable voltage. The amplifier will add noise of its own, and it may limit the bandwidth. Amplifier noise (separate from

Johnson noise) is specified by vendors as an equivalent *noise current density* at the input, usually in $\text{pAHz}^{-1/2}$ (called picoamps per root hertz). Manufacturer's specifications for amplifiers typically state the input noise current density at a specific frequency, but it varies with frequency, and the specified frequency will often correspond to a region where the noise is lowest. Designating the input noise as a current by $(i_N)_{\sqrt{f}}$, the total amplifier noise is then

$$(\overline{i_N^2})_{amp} = \frac{4kTB}{R_f} + (i_N)^2_{\sqrt{f}} B. \tag{8.26}$$

If detection is amplifier noise-limited, the SNR is

$$\text{SNR} = \frac{\left(\frac{\eta q P_S}{h\nu}\right)^2}{\frac{4kTB}{R_f} + (i_N)^2_{\sqrt{f}} B}, \tag{8.27}$$

and the NEP, which is found by setting the SNR equal to unity, is

$$\text{NEP}_{AL} = \frac{1}{S}\sqrt{\frac{4kTB}{R_f} + (i_N)^2_{\sqrt{f}} B} \tag{8.28}$$

using the definition of responsivity S in Eq. (8.9).

Example. Consider a detector with a responsivity S of 10 A/W at a temperature of 300 K, and an amplifier with a 10 MHz bandwidth, a 5.0 kΩ feedback resistor, and an input noise density of 2.0 $\text{pAHz}^{-1/2}$. The Johnson noise term is $4 \times (1.38 \times 10^{-23}) \times (300) \times (1.0 \times 10^7) / (5.0 \times 10^3) = 3.3 \times 10^{-17}$ A^2 and the amplifier noise term is $(2.0 \times 10^{-12})^2 \times (1.0 \times 10^7) = 4.0 \times 10^{-17}$ A^2, so the two noise terms have similar magnitudes. The NEP is $(1/10) \times (73 \times 10^{-18})^{1/2} = 8.6 \times 10^{-10}$ W. The great advantage of detectors with gain is that their signal and background currents are so large that they overwhelm the amplifier noise currents, making them inconsequential. The SNR can therefore approach the statistical limit set by the signal and background shot noises.

The equivalent circuit shown in Figure 8.14 is often used for calculating the signal and noise voltages at the output of the TIA. The double-circle symbol indicates a source, and the sources include the signal, background, and dark current; the photocurrent associated with them; the Johnson noise; and the amplifier noise. All noises are proportional to the bandwidth. The total current with its fluctuations flows through a 1-Ω resistor so that the amplifier input voltage (V) is equal to the current (A). The amplifier voltage gain is then the value of the feedback resistor.

The TIA modeled in Figure 8.7 does not have an arbitrarily high bandwidth; the bandwidth is often limited by parasitic capacitance. A capacitor is just two parallel conductors separated by air or a dielectric material, and electronic circuit boards are rife with such configurations, so parasitic capacitance is impossible to avoid. Typical values in TIAs range from tens of picofarads to hundreds of femtofarads.

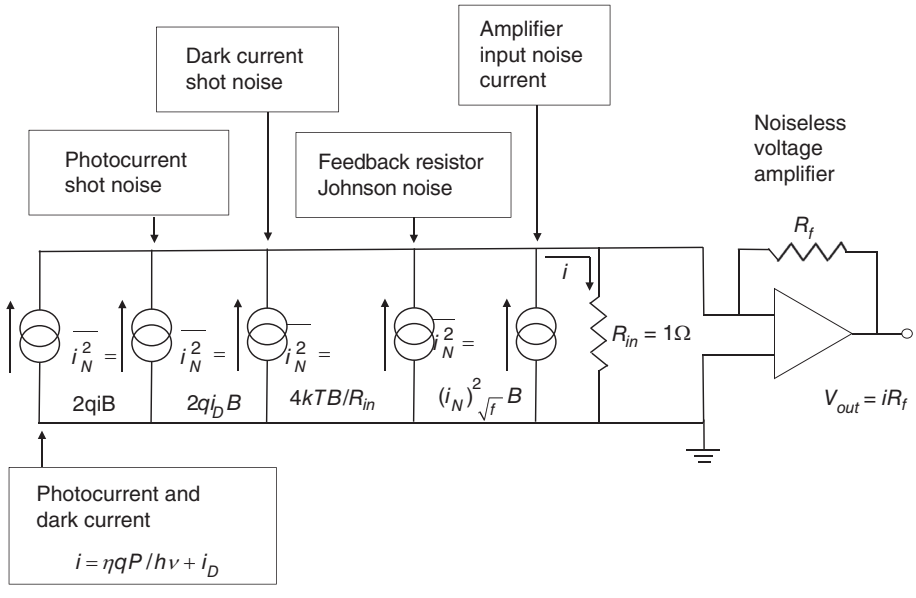


Figure 8.14 An equivalent circuit. This circuit is used to calculate the signal and noise voltages including shot noise and amplifier noise. The optical power P includes both signal and background.

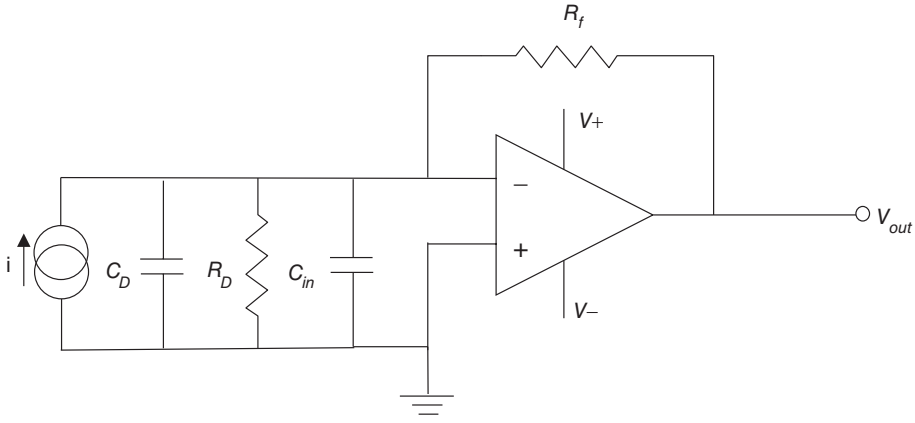


Figure 8.15 Detector–TIA equivalent circuit with parasitic capacitance.

In addition, the detector will have its own capacitance and resistance. An equivalent circuit incorporating those capacitances and the detector resistance is shown in Figure 8.15.

The bandwidth can be increased by decreasing R_f, but that change will reduce the gain. In general, there is a trade-off between gain and bandwidth characterized by the *gain-bandwidth product*, which is about 10^{12} Ω Hz, as shown in Figure 8.16, which covers about ten orders of magnitude on both axes.

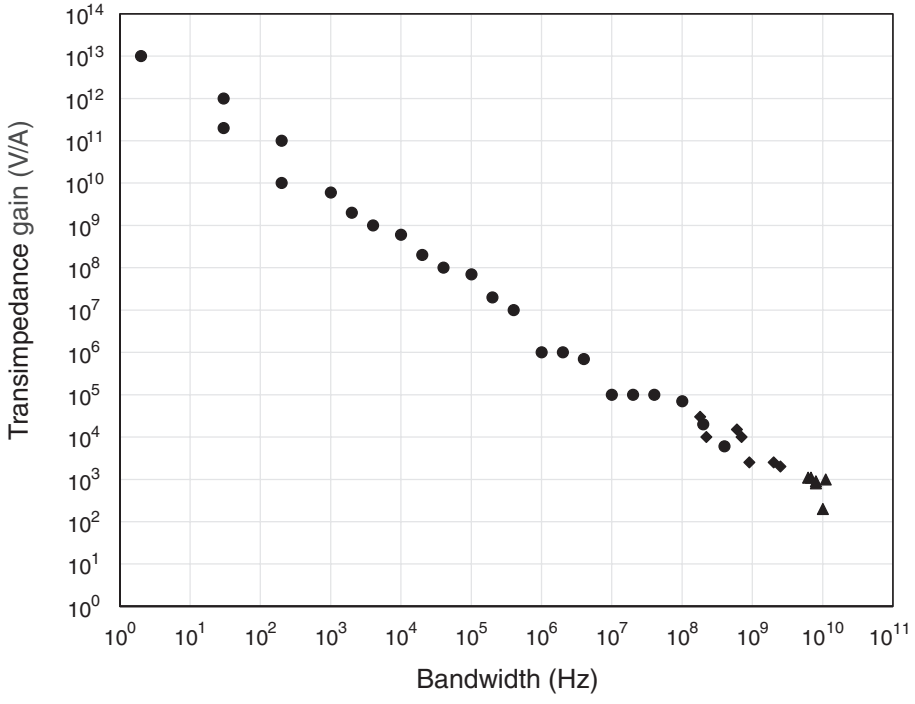


Figure 8.16 TIA gains and bandwidths. The graph summarizes the properties of commercially available TIAs as well as reported research-grade TIAs. Circles – commercial TIAs; diamonds – commercial chips; triangles – research grade TIAs (data compiled by C. R. Valenta).

The TIA must be operated in its linear region to avoid distorting the lidar waveform. All amplifiers become nonlinear at some point because an amplifier's output voltage cannot exceed its power supply voltage. Another consideration for lidar TIAs is that they tend to ring if the input frequencies are too high. Ringing, by analogy with acoustic vibrations in bells, is a damped oscillation, and it can seriously distort the lidar waveform. An exaggerated picture of the effect is shown in Figure 8.17, in which the range-corrected signal goes negative. This sort of distortion is generally fatal because the signal cannot physically be negative and analysis algorithms cannot deal with negative values. A waveform with such a ring was observed at GTRI during integration of the lidar known as ALE because the crossover was too fast, which caused frequencies in the waveform that were too high for the TIA. Fortunately, the cause was that the alignment angle δ (defined in Figure 6.25) was too high, so a realignment of the optical system eliminated the ring. This is a good example of the interplay between optical, mechanical, and electronic engineering in lidar systems. The bandwidth requirement can be estimated from the risetime of the lidar signal by using a common rule of thumb in electronics: $B = 0.35 / \tau_R$, where the risetime τ_R is the time required for the signal to rise from 10% of its peak value to 90%.

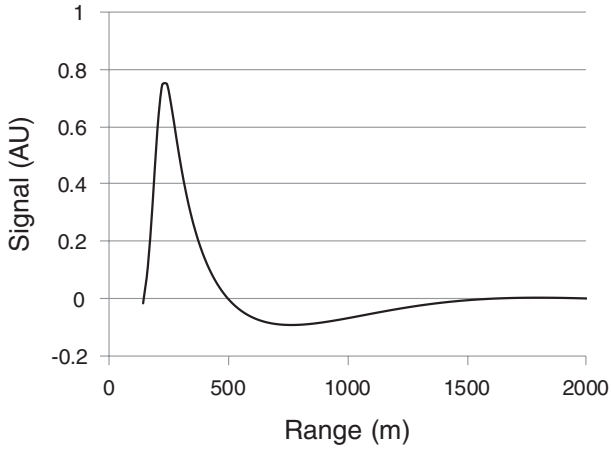


Figure 8.17 Electronic ringing. The range-corrected signal may drop below zero if the initial rise in the signal is too fast for the TIA.

One other effect can add noise or distortion to lidar signals: electromagnetic interference (EMI). There are two types of EMI, electric and magnetic. Electric EMI can be caused by electric fields when the wiring in the lidar system acts as an antenna. The High Frequency (HF) and Very High Frequency (VHF) telecommunication bands cover 3 to 300 MHz, which include frequencies important to lidar. HF-VHF uses include shortwave radio broadcasts, aviation and marine communications, television, microwave ovens, remote control transmitters, and amateur radio. In the U.S., Citizens Band (CB) radio is in the HF, around 27 MHz. Proper shielding, such as the use of coaxial cable with the outer shield grounded through high-quality connectors, should eliminate this type of EMI. A more troublesome problem is caused by unwanted frequencies that appear on the DC power supplied to the amplifiers in the lidar system, because such frequencies tend to appear in their outputs. Sources include electric motors and especially switch mode power supplies, which have become commonplace. Switch mode power supplies regulate their output voltage by rapidly switching on and off a current that charges a capacitor. They are compact, efficient, and low cost, but the switching causes transient voltage spikes that can migrate into all areas of the circuits that they power if the spikes are not properly filtered. Additionally, the spikes can cause RF interference that can affect other items of electronic equipment. Some laser power supplies are notorious for their high-frequency noise. The problems with switch mode power supplies can be eliminated by using linear power supplies.

Magnetic EMI is different: The induced voltage in a loop of wire is governed by Faraday's law of induction,

$$\text{emf} = -d\phi / dt, \tag{8.29}$$

where emf (electromotive force) is the induced voltage and $d\phi / dt$ is the time rate of change of the magnetic flux ϕ through the loop. The minus sign means that, if a current is induced in the loop, it will create a magnetic field that opposes the change

(this fact is known as Lenz's law). If the loop has N turns, the emf is N times larger. An ozone DIAL lidar developed at GTRI was initially plagued with magnetic EMI that resulted in a distorted waveform like that shown in Figure 8.17, except that the distortion was on a shorter time scale. The laser was flashlamp pumped, so a large current pulse travelled through the umbilical between the laser power supply and the laser head for each laser pulse. The umbilical was examined for pulsed magnetic fields in a very simple way: A spool of hookup wire (with many turns) was attached to an oscilloscope, and the spool was used to probe the umbilical's magnetic field. Bipolar voltage pulses appeared on the oscilloscope because the field rapidly increased and then rapidly collapsed, changing the sign of $d\phi / dt$. The bipolar pulses looked very much like the distortion in the lidar signal. The problem then was to find the loops of wire in which the emf was induced. All wiring was suspect because the shielding on coaxial cable does not block magnetic fields. The loops turned out to be in the wires that supplied DC power to the detectors and their TIAs, and the solution was to twist all those sets of wires. Twisting eliminates the emf because the induced voltages in adjacent half-turns cancel each other. The ground conductors in power cords can also form loops, known as ground loops. A classic laboratory example is two grounded pieces of equipment connected with a coaxial cable. The coaxial shield and the ground wires form a loop, which is susceptible to transient magnetic fields. A related problem, the loose ground, occurred with EARL, when the lidar signal's background level was seen to be drifting around independently of the sky condition. The problem was traced to a commercial multiple-outlet strip that was made by installing four duplex outlet receptacles in a metal box, with AC power jumpers between them. The ground connections were only made through the center attachment screw on each receptacle, and the screw had fallen out of one of them, so the equipment plugged into it was ungrounded. Voltage levels in electronic equipment are usually referenced to the chassis ground, so a loose ground can cause a drifting voltage level. As best practices, only high-quality electric power distribution components should be used, and a single-point ground should be established in a lidar system with all grounding cables and straps connected to it.

8.5 Photon Counting

Some photodetectors, such as PMTs, are capable of single-photon detection, which means that they generate a measurable current pulse for each absorbed photon. Using electronic devices to count those pulses is known as photon counting. This technique can only be used when the pulses are separated in time, which means that the signal must be small, but when the lidar signal is small enough, counting photons has a significant SNR advantage over measuring a current. In the region from ~10^7 counts per second (cps) down to ~10^5 cps, either of the two detection methods can be used. Analog and photon counting regimes are illustrated graphically in Figure 8.18, for 532 nm wavelength. Photon counting instrumentation is described in Chapter 9, along with count rate limitations.

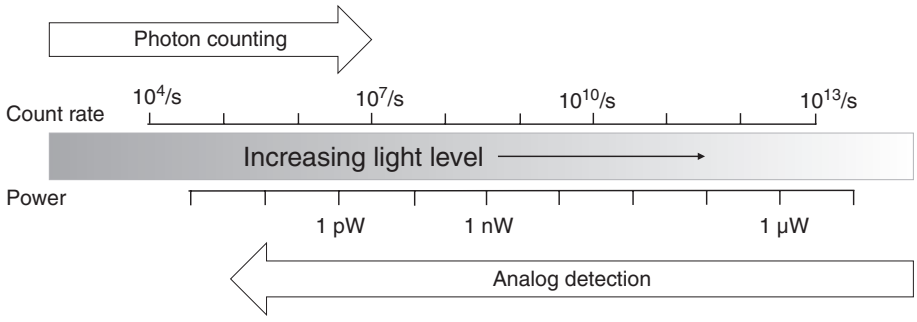


Figure 8.18 Analog detection and photon counting regimes. Photon counting is used for small signals and analog detection is used for large signals. Either can be used in a region from roughly 10^5 to 10^7 cps.

8.6 Coherent Detection

Direct detection measures the incident power on the photodetector. If the detector has a linear response over some range of power values, then the output current is proportional to the average optical power. Coherent detection has its roots in radio technology, and it is very different from direct detection: The received signal is mixed on a photodetector with a *local oscillator* (LO) beam, the two beams interfere, the output current is proportional to the square of the combined electric fields, and the beat frequency and amplitude contain information about the signal. Coherent detection allows the amplitude and phase of the incident light to be measured, which means that the Doppler shift of laser light backscattered by aerosols moving with the wind can be measured. The two detection methods are illustrated in Figure 8.19. The angled brackets denote the average values of squared electric fields.

In the coherent technique, mixing the LO and signal electric fields on the detector produces terms at the sum and difference of the optical frequencies of the two fields. Optical detectors cannot respond to the optical frequencies nor their sums because they are too high (~10^{14} Hz), so those terms may be ignored. The remaining terms consist of DC terms proportional to the LO and signal powers and a term at the difference, or *intermediate frequency* ω_{if}, which is usually designed to be tens of MHz to be within the detector–TIA bandwidth. The optical power density (W/m^2) is proportional to E^2 where E is the wave amplitude (V/m). The optical power on the detector is therefore proportional to the square of the electric field, which is expressed as

$$E^2(t) = [E_S \cos(\omega_S t + \phi) + E_{LO} \cos(\omega_{LO} t)]^2, \tag{8.30}$$

where E_S and E_{LO} are the magnitudes of the cosine waveforms, ω_S and ω_{LO} are the angular frequencies of the two waves ($\omega = 2\pi f$), and ϕ is a phase angle. Writing out the squared term in Eq. (8.30) and using trigonometric identities, we have

$$\begin{aligned} E^2(t) = {} & \frac{E_S^2}{2}[1 + \cos(2\omega_S t + 2\phi)] + \frac{E_{LO}^2}{2}[1 + \cos(2\omega_{LO} t)] \\ & + E_S E_{LO} \left\{\cos[(\omega_S - \omega_{LO})t + \phi] + \cos[(\omega_S + \omega_{LO})t + \phi]\right\}. \end{aligned} \tag{8.31}$$

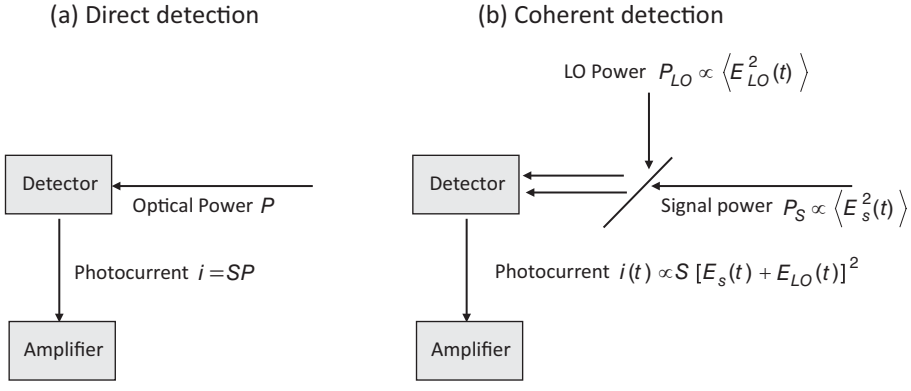


Figure 8.19 Direct and coherent detection. (a) In direct analog detection, the signal current is proportional to the received optical power; (b) in coherent detection, the signal current is proportional to the squared sum of the electric fields of the signal power and the LO power.

Dropping the optical frequency terms then yields

$$E^2(t) = \frac{E_S^2}{2} + \frac{E_{LO}^2}{2} + E_S E_{LO} \cos(\omega_{if} t + \phi), \tag{8.32}$$

where $\omega_{if} = \omega_S - \omega_{LO}$. The first two terms are DC because they are not modulated, but the last is the product of the signal and LO amplitudes modulated by a cosine oscillation at the frequency ω_{if}. The photodetector current is proportional to the incident power as before, but now there is an additional AC current at the angular frequency ω_{if} that is proportional to the signal electric field, as opposed to the square of electric field. This fact makes the electronic power proportional to the optical signal power in coherent detection. The photodetector current is a replica of the signal field in amplitude and phase, and if the signal field changes frequency, ω_{if} changes too. Equation (8.27) implies that there are three components of the coherent photodetector current shown in Figure 8.17:

$$i(t) = \mathrm{SP}(t) = i_S + i_{LO} + i_{if} \cos(\omega_{if} t + \phi). \tag{8.33}$$

SNR is calculated according to Eq. (8.10) as

$$\mathrm{SNR}_{power} = \frac{\overline{i_S^2}}{(\overline{i_N^2})_S + (\overline{i_N^2})_B + (\overline{i_N^2})_{LO} + (\overline{i_N^2})_{elec}}, \tag{8.34}$$

where the noise currents in the denominator are due to the signal power, the background power, the LO power, and electronic noise sources. In coherent detection, the LO power is typically adjusted such that $i_S \ll i_{if} \ll i_{LO}$ so that the LO shot noise dominates all other noise sources, making their contributions negligible. This is one great advantage of coherent detection lidar, and it is accomplished by simply increasing the LO optical power. There is a limit, of course, because too much LO power will saturate the detector, making the system unusable. Adjusting the LO power appropriately yields a coherent detection NEP that is one-half of the direct detection NEP. When LO shot noise dominates, Eq. (8.34) reduces to

$$\mathrm{SNR}_{\mathrm{power}} = \frac{\overline{i_{\mathrm{S}}^2}}{(\overline{i_{\mathrm{N}}^2})_{\mathrm{LO}}}, \tag{8.35}$$

but the signal current we care about is i_{if} because it carries the lidar waveform and phase information. Noting that $i_{\mathrm{S}} \propto E_{\mathrm{S}}^2 / 2$ and $i_{\mathrm{LO}} \propto E_{\mathrm{LO}}^2 / 2$,

$$\overline{i_{\mathrm{if}}^2} \propto \left(2\sqrt{i_{\mathrm{S}} i_{\mathrm{LO}}}\right)^2 \overline{\left(\cos^2(\omega_{\mathrm{if}} t + \phi)\right)} = 2 i_{\mathrm{S}} i_{\mathrm{LO}}, \tag{8.36}$$

because the average value of $\cos^2(\omega)$ is ½. Remembering that the noise current is given by Eq. (8.12) as $(\overline{i_{\mathrm{N}}^2})_{\mathrm{LO}} = 2qi_{\mathrm{LO}}B$, we have

$$\mathrm{SNR}_{\mathrm{power}} = \frac{2 i_{\mathrm{S}} i_{\mathrm{LO}}}{2qi_{\mathrm{LO}}B} = \frac{i_{\mathrm{S}}}{qB} = \frac{\eta P_{\mathrm{S}}}{h\nu B}. \tag{8.37}$$

Setting the SNR equal to unity and solving for P_{S} yields the ideal coherent detection NEP as

$$\mathrm{NEP}_{\mathrm{coherent}} = \frac{h\nu B}{\eta}, \tag{8.38}$$

which is one-half the value for signal-limited direct detection given in Eq. (8.20). This result may at first appear to violate the arguments in Chapter 2, because it predicts an SNR twice as high as the Poisson statistical limit based on signal photons. Perhaps the simplest explanation is that, on average, two photons are required to generate the signal at ω_{if}, one from the received signal and one from the LO, and doubling the number of photons doubles the power SNR and hence halves the NEP.

A second advantage of coherent detection is that it discriminates against background power, having an effect like a very narrow optical filter. Coherent lidars generally limit the range of f_{if} with a bandpass filter, and the spectral width of the signal has the same value, that is, $\Delta f_{\mathrm{if}} = \Delta\nu_{\mathrm{S}}$, where $\Delta\nu_{\mathrm{S}}$ is the corresponding range of signal frequencies. To find the range in wavelength units, recall that

$$\Delta\lambda = -\Delta\nu \frac{\lambda^2}{c}. \tag{8.39}$$

Example. If f_{if} is 10 MHz and λ is 2.1 μm, $\Delta\lambda = (10^7) \times (2.1 \times 10^{-6})^2 / (3 \times 10^8) = 1.5 \times 10^{-4}$ nm (where the absolute value is used in calculating a range of frequencies), which is an extremely narrow optical bandpass. An optical filter is still required in coherent lidar receivers to avoid saturating the detector with broadband background light, but that filter can have a much wider bandpass.

A third advantage of coherent detection is that it enables a lidar to measure wind speed. The signal in infrared lidars is due to backscattering from aerosols that are advected with winds, so that signal is Doppler shifted. In wavelength units, the shifts are very small, but as changes in the electronic frequency f_{if} they are conveniently measured. The relation is

$$\Delta\nu = \Delta f_{\mathrm{if}} = -2\frac{V_{\mathrm{wind}}}{\lambda}, \tag{8.40}$$

where Δf_{if} is the shift in f_{if} and V_{wind} is the component of the wind vector along the lidar line of sight. If the wind is moving toward the receiver, $V_{wind} < 0$. The factor of 2 in Eq. (8.40) arises from the fact that the lidar light goes out to the aerosols and back. Equation (8.40) shows that the Doppler shift Δf_{if} varies inversely with wavelength. The earliest coherent lidars all operated at the CO_2 laser wavelength of 10.6 μm, but over the years they moved to the 2-μm region and later to the 1.5-μm region as appropriate lasers became available, partly to exploit the higher Doppler shifts at shorter wavelengths. Coherent wind-sounding lidars are shown in the LWIR and SWIR regions in Figure 3.24.

With all the advantages of coherent detection described above, it may seem surprising that it is not exploited by all atmospheric lidars. There is one basic reason why it is not, with two manifestations: *mixing efficiency*. The mixing process illustrated in Figure 8.19 requires controlling and matching two wavefronts on the detector – one from the receiver and one from the LO. The foregoing analysis assumed flat and parallel LO and signal wavefronts that have a constant phase relationship across the surface of the detector. In practice, wavefronts become harder to control as the wavelength becomes shorter, because light wavelengths are so tiny compared to macroscopic objects such as detectors. Coherent lidar is limited to SWIR wavelengths and longer, where molecular backscatter is negligible and lidar signals are due to aerosol backscatter only. A more fundamental limitation arises from the phenomenon of atmospheric refractive turbulence, which is caused by random temperature fluctuations in the air. For optical wavelengths, the refractive index of air is given by

$$n = 1 + 79 \times 10^{-6} \left(\frac{P}{T} \right), \tag{8.41}$$

where P is the air pressure in mbar and T is the air temperature in kelvins. At typical surface-level values of 1000 mbar and 300 K, n is 1.000263 (1 + 263 ppm). The fact that n varies inversely with T means that temperature fluctuations cause refractive index fluctuations. The fluctuations occur when different parcels of air, known as eddies, have different temperatures. The r.m.s values of random temperature fluctuations near the surface typically range from 0.01 K to 1.0 K, depending on time of day and weather parameters. If the temperature is 299 K, Eq. (8.36) shows that the refractive index is 1.000264, so at typical surface values of P and T, a 1 K temperature change caused a 1 ppm refractive index change. This may seem like a small change, but it has a major effect on wavefronts. If two rays of light at 1-μm wavelength propagate through two eddies at 300 K and 299 K they will be completely out of phase after propagating only ½ m (the calculation is left as a problem for the student). Considering that lidar ranges are usually expressed in km, it is easy to see that wavefronts can be seriously distorted by this phenomenon, causing a serious loss of mixing efficiency. The diameter of a receiver over which a wavefront can be considered to have constant phase (within one radian) is given by the *Fried parameter* r_0, which depends on path length and the variance of the refractive index fluctuations, but is typically on the order of a few cm. For this reason, ground-based coherent lidars usually have receiver diameters limited to about 10 cm. Such a small

receiver is only adequate for short ranges, so ground-based coherent wind lidars are mainly employed at airfields and wind farms. Airborne and space-based coherent lidars may have larger receiver apertures.

8.7 Photodetectors

As mentioned in Section 8.2, only quantum detectors are applicable to lidar receivers. As shown in Figure 8.6, such detectors absorb photons and either eject electrons from a surface, such as the cathode in a photomultiplier, or raise valence electrons to a conduction band in solid-state photodiodes. In either case, a bias voltage is applied to move the electrons, resulting in a current.

8.7.1 Vacuum Tubes

Vacuum tube optical detectors are based on the photoelectric effect, in which a free electron may be emitted when a photon is absorbed on a surface. An electric field is applied to cause the electron to move from the surface (the cathode) to a collector (the anode). An early version of the vacuum tube detector known as the *vacuum photodiode* is illustrated in Figure 8.20. It was packaged using the technology of the vacuum tubes used in radios, and its main significance is probably that it was used to convert the optical soundtracks of motion pictures to electronic signals. When used in mechanisms such as automatic door openers, it was called the "electric eye." The photocathode was a section of a cylinder, and the anode was a wire placed at its center of curvature so that the electric field lines pointed radially inward to the anode, and the anode did not shadow the cathode appreciably. The photocathode was often coated with Ag-O-Cs, which was designated S-1. This material has a low quantum efficiency, but it has a broad spectral response covering the VIS spectrum. Modern versions of the vacuum photodiode are still available for specialized instruments, for one reason: They are fast. Commercial devices have pulse FWHM values as low as 60 ps [6]. They also have high dynamic range. The vacuum photodiode is the predecessor of PMTs, and its S-1 photocathode material is still in use.

In the photoelectric effect, when photons strike certain materials, electrons absorb their energy and are ejected from the surface. The kinetic energy of the ejected electron is $\frac{1}{2}mv^2$ and the *work function* Φ is the amount of energy required to eject an electron from the surface with zero kinetic energy, so the two must add up to the photon's energy, which implies that

$$\frac{mv^2}{2} = h\nu - \Phi. \tag{8.42}$$

If the photon energy is less than Φ, an electron is not ejected, so there is a cutoff wavelength for any material above which photoemission does not occur. In discussions of the photoelectric effect, energies are conventionally expressed in electron volts (eV). A convenient relation is that the photon energy in eV is equal to $1.2398/\lambda$ (μm).

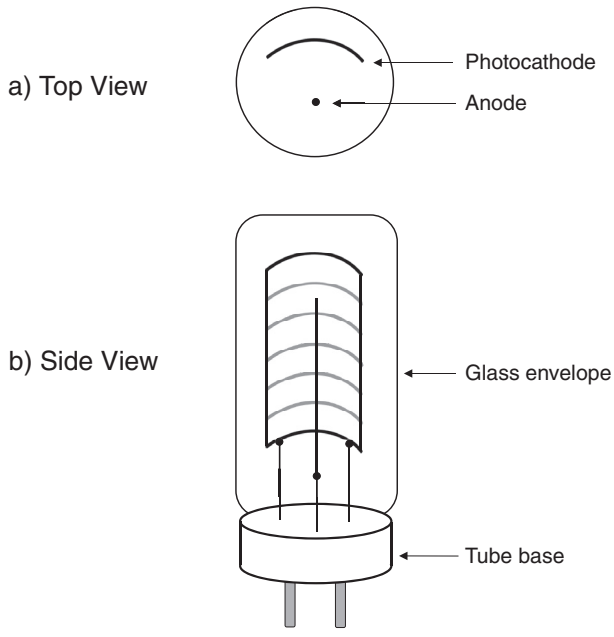


Figure 8.20 An early vacuum photodiode. As shown in a), the photocathode is a section of a cylinder, and the anode is a wire placed at its center of curvature. This configuration was incorporated into standard radio vacuum tube technology, as shown in b), the front view.

Historically, the work functions of materials used as PMT photocathodes were 1 eV or greater, so the cutoff wavelengths were around 1.2 μm and PMTs were restricted to the UV-VIS-NIR region. A material's work function is also the phenomenon that resists *thermionic emission* of electrons, which is the cause of PMT dark current, so photocathodes with low work functions have large dark currents and hence require cooling.

The first multistage PMT was reported in 1936 by Zworykin, Morton, and Malter of RCA [7]. The classic PMT configuration, known as head-on and linear, is illustrated in Figure 8.21. Whereas the incident photon has an energy of about 1 eV, the ejected electron is accelerated by a potential difference on the order of 100 V, so it is much more energetic, and it causes 5–6 electrons to be ejected from the first dynode. This process is repeated through a chain of 5–10 dynodes, producing a current gain G at the anode of 10^5–10^8. Such huge gains make single photon detection easy, and they make the TIA gain requirement minimal. Johnson noise and amplifier noise are generally not factors in SNR when using PMTs, because they are insignificant compared to the shot noise associated with the large signal, background, and dark currents. The configuration shown in Figure 8.21 is the simplest to explain and it is commonly used in lidar receivers, but there are several other configurations [8]. The accelerating voltages between the dynodes are determined by a voltage divider network in which all the resistors often have equal values. In Figure 8.21 for example, if the power supply voltage is 700 V, each stage will have a 100 V accelerating voltage. Operating in pulsed mode, the currents in the last two stages can become so high that their voltages

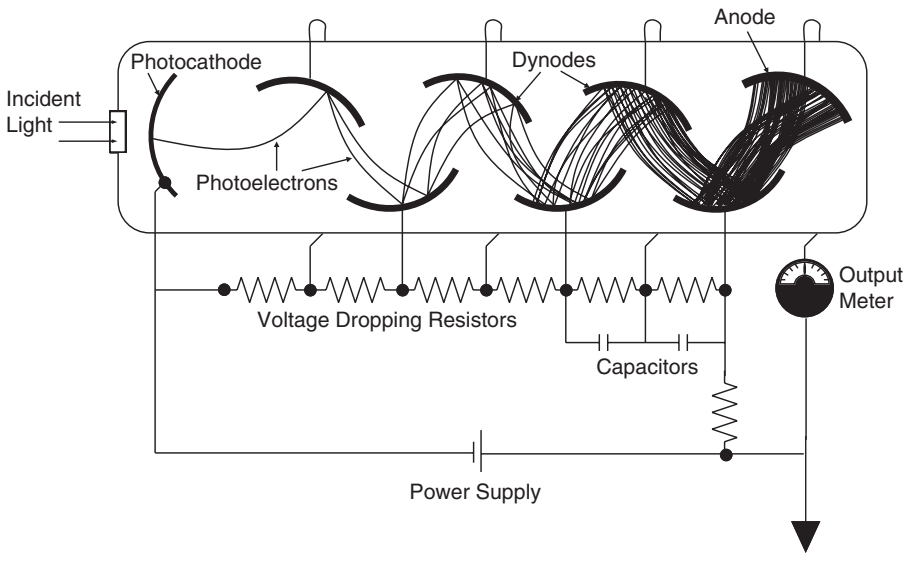


Figure 8.21 The photomultiplier tube. The rectangle with rounded corners denotes the glass envelope of the tube, which hides the connections to the top row of dynodes and the anode.

are not maintained. For this reason, capacitors are often added to the last two stages. The capacitors charge up to the full voltage between pulses so they can supply extra current if needed. The PMT amplification process is fairly fast – a typical transit time is 1 ns, and pulses resulting from single photons have FWHMs of 1.3–5 ns. At low light levels, the process is also linear, meaning that the gain does not depend on the light intensity. At higher light levels, the gain will eventually be decreased by space charge near the last dynode.

The quantum efficiency of a PMT is largely determined by its photocathode material, and it is always a function of wavelength. There are more than 10 different materials in commercial photocathodes, but only a few are relevant to lidar:

(1) Ag-O-Cs. The S-1 photocathode is sensitive from 400 to 1200 nm. It was used in 1064-nm lidars but its QE was low, and it was supplanted by solid-state avalanche photodiodes when they became available.
(2) Bialkali. These are made from antimony (Sb) plus two alkali metals from column I in the Periodic Table of the Elements. Sb-Rb-Cs and Sb-K-Cs are common examples. Peak sensitivities are in the UV-VIS region. Bialkali was designated as S-20.
(3) Multialkali. These are made from antimony (Sb) plus three or more alkali metals. They are also broadband, extending from the UV to 850 nm, extendable to 900 nm.
(4) InP/InGaAsP(Cs) and InP/InGaAs(Cs). These are semiconductor crystals activated with Cs. They are NIR-SWIR photocathodes with sensitivities extending to 1700 nm. They are operated at –60°C to –80°C, and they are easily damaged by high light intensities.

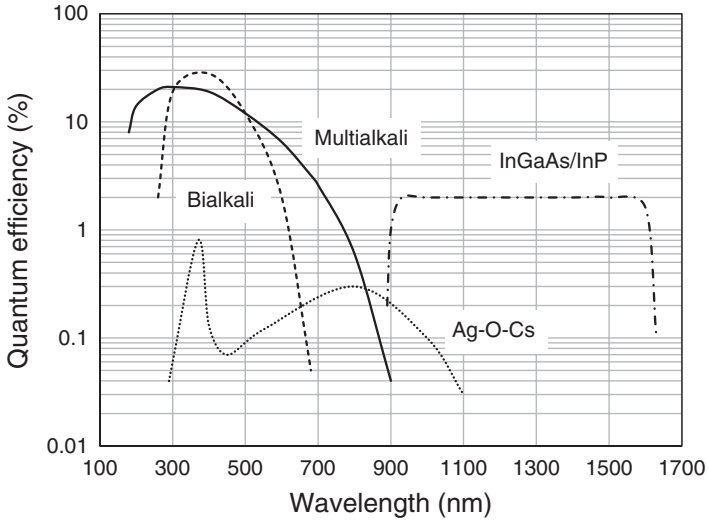


Figure 8.22 Typical quantum efficiencies of lidar PMTs with four photocathode materials (adapted from product data sheets).

Typical quantum efficiencies for PMTs with these photocathode materials are shown in Figure 8.22. Their long wavelength cutoffs are determined by their work functions, and their short wavelength cutoffs are determined by the PMT window materials, for the visible-light PMTs. Historically, lidar engineers relied mostly on S-20 photocathodes with a QE of ~10% for visible light and S-1 photocathodes with a QE of less than 1% for NIR light. Great progress has been made over the years however, and PMTs now span the range from 200 to 1700 nm with peak QEs much higher than what was previously available. For example, whereas a former bialkali photocathode had a peak QE in the UV of 25%, the Hamamatsu superbialkali PMT introduced in 2007 has a peak QE of 35% and the ultrabialkali PMT has a peak QE of almost 45%. A photocathode with higher QE may also have a higher dark count rate, but detailed specifications are provided by vendors so that a lidar engineer can select the optimum PMT and accurately model a lidar receiver's performance in terms of SNR.

The PMT current gain G depends on the secondary emission ratio δ (average number of electrons emitted by a dynode when a single electron strikes it); the voltage E between dynodes, which is a fraction of the total bias voltage V_{bias} on the PMT; and the number of dynode stages N in the PMT. Note that the secondary emission ratio δ depends on V_{bias} and it is typically 5 to 6 when the PMT is operating at optimum V_{bias}. The secondary emission ratio is modeled as

$$\delta = aE^{\alpha}, \tag{8.43}$$

where the parameter a ranges from 0.1 to 0.3 and the parameter α ranges from 0.7 to 0.9 [7]. The voltage between dynodes is calculated as $E = V_{\text{bias}} / (N+1)$, where V_{bias} is taken to be a positive number. The PMT gain is then modeled as

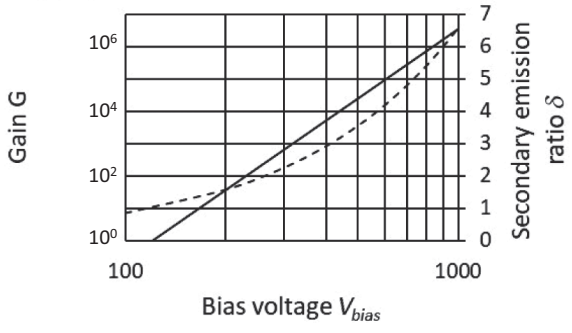


Figure 8.23 R7400U gain and secondary emission ratio. The solid line is the gain; the dashed line is the secondary emission ratio.

$$G = \delta^N = a^N \left(\frac{V_{\text{bias}}}{N+1} \right)^{\alpha N}. \tag{8.44}$$

Example. The Hamamatsu R7400U, which was the original PMT in EARL, has the following parameters: $a = 0.1$, $N = 8$, and $\alpha = 0.888$. The secondary emission ratio and gain predicted by Eqs. (8.38) and (8.39) are plotted in Figure 8.23. At a bias voltage of 800 V, the gain is 7×10^5 and the secondary emission ratio is 5.4.

Equation 8.44 and Figure 8.23 show that PMT gain is sensitive to bias voltage. Because almost all lidars operate in an uncalibrated mode, changes in gain may appear unimportant. However, a power supply ripple (a small oscillation) will appear in the data at the ripple frequency because of the gain change it causes, so care must be taken to avoid ripple. For example, an R7400U operated at 800 V with a 1% ripple in the bias voltage would have a gain ripple of 7% (calculating this value is left as a problem for the student). This problem occurred in an ozone DIAL lidar built at GTRI, when the boundary layer ozone concentration appeared to be layered, whereas the aerosols appeared to be well mixed. The problem was traced to PMTs with internal power supplies based on oscillators at 333 kHz. The power supply ripple caused a gain ripple that appeared in the lidar signals, and the DIAL analysis algorithm (see Chapter 11) was very sensitive to it, with the result that the ozone appeared to be layered. The signal ripple was not visible in plots, but a Fourier transform of the signal revealed a spike at 333 kHz plus several harmonics. The solution was to switch to PMTs with external linear power supplies.

As mentioned earlier, dark current is mainly produced by the thermionic emission of electrons from the photocathode and dynodes (although there are several other sources). The emission rate for metals as a function of temperature and work function is described by the Richardson–Dushman equation

$$J = 120T^2 \exp(-\Phi / kT), \tag{8.45}$$

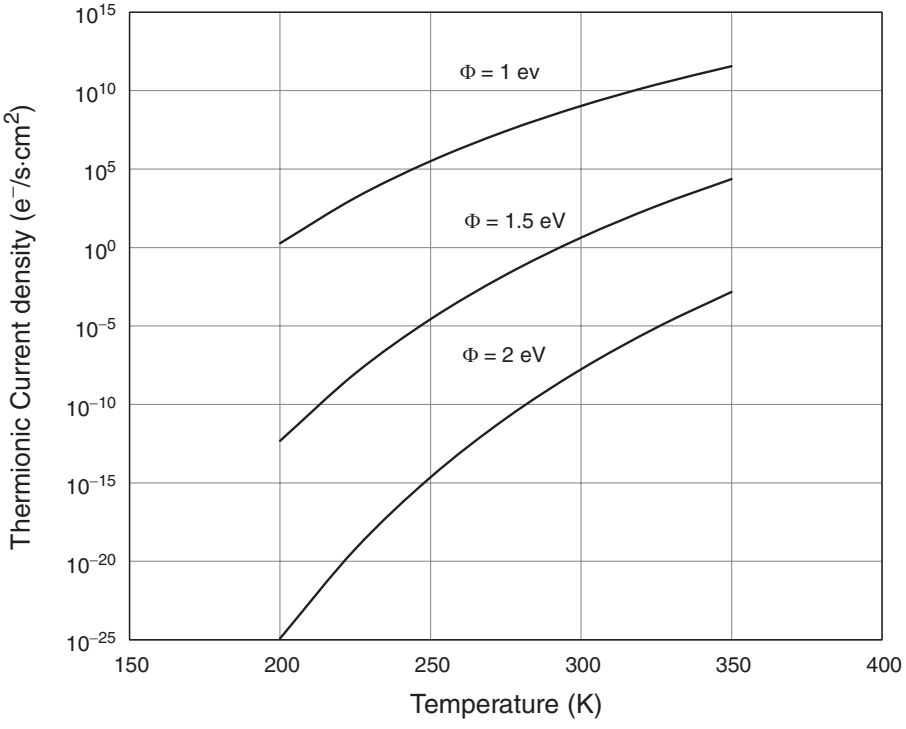


Figure 8.24 Thermionic current density. The current density predicted by the Richardson–Dushman equation is plotted versus temperature for three values of the work function.

where J is the emission rate in electrons per second per square centimeter. This equation was developed by British physicist Owen W. Richardson, who received a Nobel prize in physics in 1928 for his work on thermionic phenomena. In practice, the factor of 120 must be modified by correction factors for different metals. Equation (8.45) is not accurate for the materials in photocathodes; it is used here only to elucidate the dependence of the dark count rate on the work function and temperature. Plots of J are shown versus temperature in Figure 8.24 for three different values of the work function. For perspective, recall that a dark count of 10^7 cps corresponds to one count per 15-m range bin.

Figure 8.24 shows that the Richardson–Dushman model describes an extremely strong dependence on work function and temperature, as the vertical scale spans 40 orders of magnitude. The horizontal gridlines are spaced five orders of magnitude apart, and that much reduction in emission rate can be achieved by lowering the temperature by about 50 degrees, which can be accomplished with a thermoelectric cooler. On the other hand, changing the work function from 1.00 to 1.50 decreases the emission rate by about 10 orders of magnitude at a temperature of 250 K. In a PMT, the dark current count rate is proportional to the photocathode area, so PMT diameters of 1 cm or less are generally preferred in lidar. GTRI's Megalidar used a

PMT with 2-inch (50.8-mm) diameter S-20 photocathode which had a significant dark count rate, but the tube's housing included a thermoelectric cooler. Cooling the tube to –25°C caused a dramatic reduction in dark counts [9]. Hamamatsu uses a somewhat different model for PMT dark current that predicts four orders of magnitude increase with a 50°C temperature increase, and dark current values at 20°C span an order of magnitude for a selection of bialkali and multialkali photocathodes [8]. More modern lidar PMTs such as the Hamamatsu R7400 used in EARL have a much smaller photocathode (9-mm diameter) and a dark count rate of about 50 cps. Considering that a complete EARL profile of 2048 range bins (0–30.7 km) requires only 205 μs, on average only 10 dark counts would be expected when averaging profiles from about 1000 laser shots, so cooling was not necessary. At the other extreme, GTRI developed a photon-counting lidar at 1.57 μm using the Hamamatsu H10330A-75 IR PMT, which is in a sealed cryostat cooled to –60°C, and the photocathode diameter is only 1.6 mm. Despite the small size and low temperature, the measured dark count rate was 2.7×10^5 cps. The reason for the high dark count rate is that the cutoff wavelength for this PMT is 1700 nm, which implies a work function of 0.73 eV or lower. The high dark count rate is consistent with the dependence on work function shown in Figure 8.24.

Photomultiplier tubes contribute a type of statistical noise to the signal that is accounted for by the excess noise factor Γ. This noise arises mainly at the first dynode, because the secondary electron emission process is described by Poisson statistics. The average secondary emission ratio is called δ, and when $\delta = 4$, the number of emitted electrons with significant probability ranges from zero to roughly ten, as shown in Figure 2.1. The emission variations result in photon-to-photon variations in the size of the current pulse, with the result that the mean-square noise current is increased by a factor Γ, which is determined by the secondary emission ratio δ according to the relation

$$\Gamma = \frac{\delta}{\delta - 1}. \tag{8.46}$$

The ratio δ increases with V_{bias}, and Γ is lowest for large δ as shown in Figure 8.25, so keeping V_{bias} high helps reduce this noise. Fortunately, Γ is usually around 1.2 for typical lidar PMTs, meaning it only decreases the power SNR by about 20%. Also, the excess noise can be eliminated by photon counting, as described in Section 8.5.

In a detector with unity gain, the variance associated with a noise current (shot noise) is given by Eq. (8.10) as $\overline{i_{\text{N}}^2} = 2q\bar{i}B$. For a PMT, this becomes

$$\overline{I_{\text{N}}^2} = 2q\overline{I_{\text{T}}}B\Gamma G, \tag{8.47}$$

where the capitalization of currents indicates that they are measured at the anode (after the gain), and the total current is given by $I_{\text{T}} = I_{\text{S}} + I_{\text{B}} + I_{\text{D}}$. The first two currents are simply the lower-case currents (at the photocathode) multiplied by the PMT gain G, but the dark current is generally specified at the anode by manufacturers because it arises from dynodes as well as the photocathode, so it remains I_{D}. The SNR for analog direct detection with a PMT can be calculated as it was in Eqs. (8.22)–(8.24) from the signal and noise currents. The signal current is

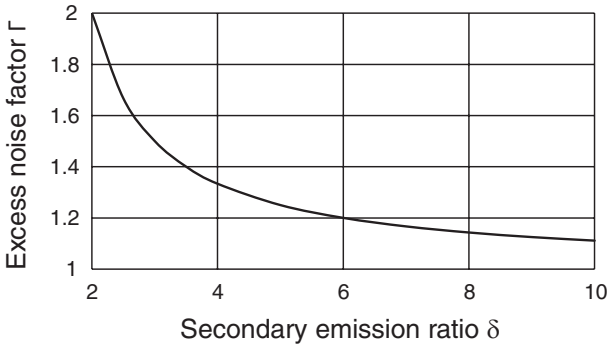


Figure 8.25 The PMT Excess noise factor versus secondary emission ratio.

$$I_S = Gi_S = \frac{G\eta q P_S}{h\nu}, \tag{8.48}$$

and the noise current is

$$\overline{I_N^2} = 2qI_T B\Gamma G = \frac{2\eta q^2 B\Gamma G^2}{h\nu}\left(P_S + P_B + \frac{h\nu I_D}{\eta q G}\right), \tag{8.49}$$

where an equivalent power at the cathode is found from I_D. The power SNR, including Johnson noise and amplifier noise, is then

$$\text{SNR} = \frac{I_S^2}{\overline{I_N^2} + \dfrac{4kTB}{R_f} + (i_N)^2_{\sqrt{f}} B}, \tag{8.50}$$

which can be manipulated, after substituting the right side of Eq. (8.49) for $\overline{I_N^2}$, to yield the expression

$$\text{SNR} = \frac{\eta P_S^2}{2\Gamma h\nu B\left(P_S + P_B + \dfrac{h\nu I_D}{\eta q G}\right)\left[1 + \dfrac{1}{\Gamma G^2}\dfrac{h\nu}{2\eta q^2\left(P_S + P_B + \dfrac{h\nu I_D}{\eta q G}\right)}\left(\dfrac{4kT}{R_f} + (i_N)^2_{\sqrt{f}}\right)\right]}. \tag{8.51}$$

The reason for writing SNR in this fashion is to show explicitly how a large gain makes the last two noise terms in the denominator irrelevant: they are multiplied by $1/\Gamma G^2$. For this reason, Eq. (8.51) reduces to

$$\text{SNR}_{\text{power}} = \frac{\eta P_S^2}{2h\nu B\Gamma\left(P_S + P_B + \dfrac{h\nu I_D}{\eta q G}\right)}, \tag{8.52}$$

which is Eq. (8.24) with an added excess noise factor Γ in the denominator plus a different expression for the dark current. In an analog detection system, a PMT adds excess noise which lowers the SNR before the TIA, but a PMT is still

preferable. A detector without gain would produce a very weak signal that requires a large amplification before it reaches the digitizer, and the amplification stages would add their own noise and possibly limit the bandwidth. In addition, the high gain of a PMT enables single photon detection and photon counting, so PMTs are very common in lidar receivers.

PMTs do have a few other drawbacks in addition to the excess noise, one of which is after-pulsing, in which a PMT emits a secondary pulse after a signal-induced pulse. This problem has been traced to the presence of residual gases, mainly helium, in the tube that are ionized by the electron cascade due to the signal pulse. The ions are attracted to the cathode, where they produce photoelectrons that result in after-pulses. The amplitude of an after-pulse depends on where the ion was generated, and the time delay ranges from hundreds of nanoseconds to several microseconds. Natural helium atoms slowly penetrate glass-envelope tubes over time, so this problem gets worse as tubes age. Another issue is that PMTs have a limitation on the maximum photon rate they can accept because of their high gains. PMTs can be damaged by intense pulses of light, so manufacturers specify a maximum safe anode current. As a rule of thumb, the anode current should not exceed 100 μA (the manufacturer's specifications should be consulted for more accurate values). The maximum optical power can be found from

$$P_{\max} = I_{\max} \frac{hc}{\lambda \eta q G}. \tag{8.53}$$

Example. The Hamamatsu R9880u-20 PMT is rated at 100 μA maximum average output current over 30 seconds. It has a QE of 18% at 532 nm. If it is operated with a gain of 10^5, the maximum optical power is $(100 \times 10^{-6}) \times (2 \times 10^{-25}) / (0.532 \times 10^{-6}) \times (.18) \times (1.60 \times 10^{-19}) \times (1 \times 10^5) = 1.31 \times 10^{-8}$ W.

PMTs generate a major artifact if they are saturated by too much signal power, as illustrated in Figure 8.26. Rather than recovering quickly to follow the optical signal, their output has a long, slow exponential decrease with time. Saturation sometimes arises because of the extreme dynamic range of lidar signals, especially for ground-based lidar measurements at high altitudes, in which the desired signal is very weak but the nearby tropospheric signal is many orders of magnitude stronger. Saturating the PMT causes a fatal artifact, so it must be avoided by some means. For example, an early GTRI lidar was based on a 2.5-m diameter receiver telescope that collected a huge amount of nearby backscatter, so it employed a mechanical shutter that blocked all light from altitudes below 40 km [9]. A later incarnation of that lidar simply had a $G(R)$ with a very long crossover range, which was accomplished with a transmitter-to-receiver distance on the order of 10 m. Other approaches have been tried, including electronically gating the PMT bias voltage on and off, or replacing the voltage divider resistor chain with active circuitry so that the PMT gain vs. time can be engineered. Modifying the voltage divider network to approximate a logarithmic response to the optical signal has also been achieved. The reason that GTRI only used a shutter or a long crossover range to avoid saturation was to ensure that no distortions of the lidar waveform were introduced.

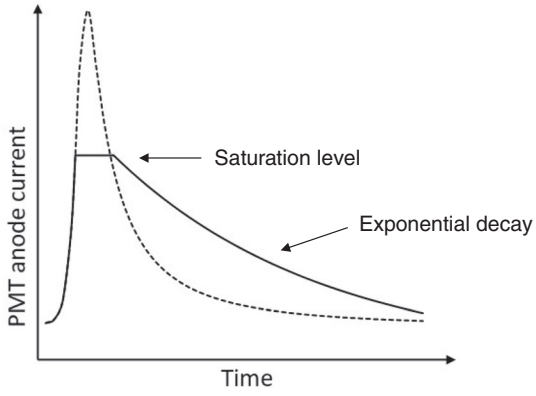


Figure 8.26 PMT saturation. An undistorted response to the optical signal is shown by the dotted line. If PMT saturation occurs, the signal will have a flat top followed by a slow exponential decrease, as shown by the solid line.

Manufacturers recommend keeping PMTs in darkness when they are not in use, because exposure to light causes a higher dark count rate, and this effect persists for hours. For this reason, the EARL receiver was manually covered with a cap at the end of operations, but GTRI eventually adopted a practice of including electrically operated shutters in all lidar receivers. The shutters are normally closed, and they open in response to an applied voltage during lidar operations.

8.7.2 Solid-State Detectors

Much of this section was adapted from [10] and [11]. Photoconductive solid-state detectors operate as illustrated in Figure 8.6(b): Photons are absorbed within the material, their energy raises electrons in a *valence band* up to a *conduction band*, and the electric field due to an applied bias voltage causes the electrons to move, establishing a current. The unfilled electron states left in the valence band are known as *holes*, which have positive polarity, and they move in the opposite direction to the electrons. Electrons and holes are not necessarily created at the same rate, and in general they do not have the same mobility, which is the speed at which they move in response to an applied field. Photoconductive detection is illustrated in Figure 8.27 for the simplest case, a bulk semiconductor. This kind of detector is called *intrinsic*. Figure 8.27(a) is an electrical schematic in which plus signs indicate holes and minus signs indicate electrons. The dark areas on the ends of the semiconductor are electrical connections, known in the trade as *ohmic* contacts. Figure 8.27(b) is an energy band diagram, where E_c is the energy at the bottom of the conduction band, E_v is the energy at the top of the valence band, and the band gap is defined as $E_g = E_c - E_v$. The detector in the 10.6-μm CCFU [3] was of this type, made of HgCdTe. This material can be used in the SWIR, MWIR, and LWIR regions because its spectral response can be tuned by adjusting the ratio of Hg to Cd. The occupancy of states $F(E)$ in the conduction band is determined by Fermi–Dirac statistics as

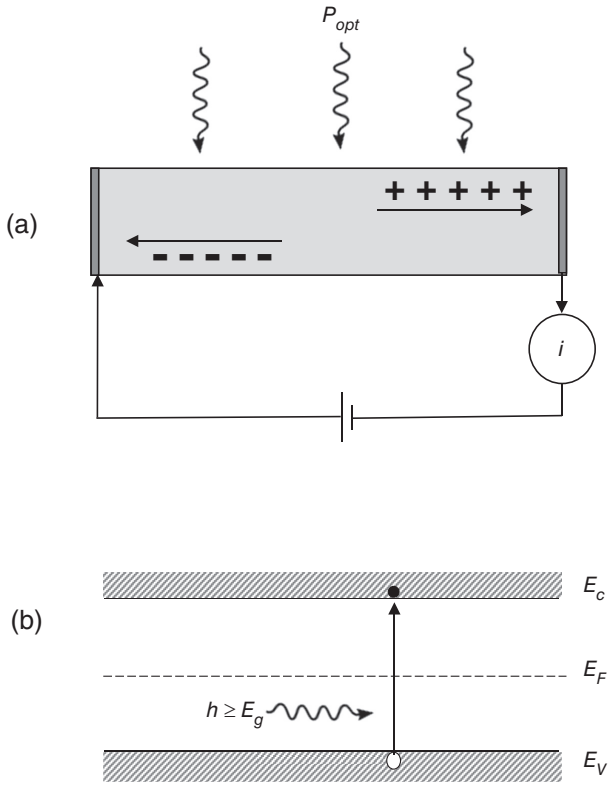


Figure 8.27 The intrinsic solid-state detector. Photons absorbed by a semiconductor raise electrons from the valence band to the conduction band, and an applied voltage causes electrons and holes to move in opposite directions. The movement of the charge carriers constitutes a current. Adapted with permission from Figures 1 and 2 in Chapter 13 of [10].

$$F(E) = \frac{1}{1 + \exp\left[\left((E - E_F)/kT\right)\right]}, \tag{8.54}$$

where E_F is the Fermi energy, which lies near the middle of the band gap for intrinsic semiconductors.

More sophisticated detectors are achieved by doping, which is the process of introducing two types of impurities to semiconductors: donor and acceptor. An n-type semiconductor is obtained by adding an impurity to increase the number of free charge carriers (in this case, negative). When a donor doping material is added, it gives away (donates) weakly bound outer electrons to the semiconductor atoms. A p-type semiconductor is obtained by adding a different type of atoms to the semiconductor to increase the number of free charge carriers (in this case, positive). When the acceptor doping material is added, it takes away (accepts) weakly bound outer electrons from the semiconductor atoms. The Fermi energy lies closer to the valence band in p-type semiconductors and closer to the conduction band in n-type semiconductors. A p–n junction is a boundary or interface between the two types of semiconductor materials,

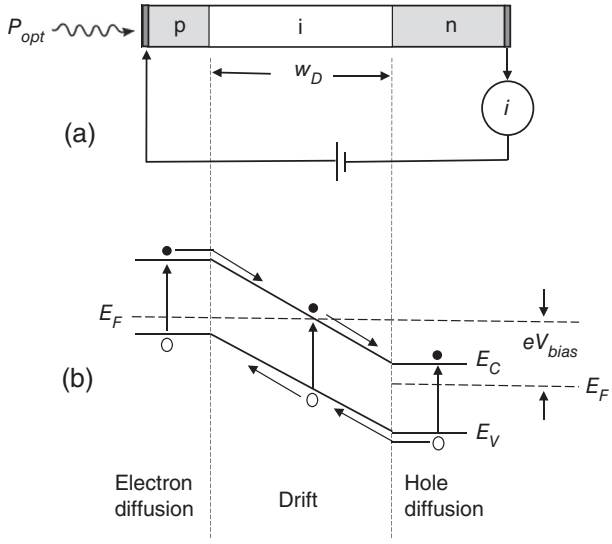


Figure 8.28 The PIN diode. A schematic diagram and cross-sectional view are shown in (a); the corresponding reverse bias energy band diagram is shown in (b). Adapted with permission from Figure 10 in Chapter 13 of [10].

p-type and n-type, inside a single crystal of semiconductor. The "p" (positive) side contains an excess of holes, while the "n" (negative) side contains an excess of electrons in the outer shells of the electrically neutral atoms. A p–n junction can be used as a photodetector, but detectors are more often p–i–n, or PIN junctions, where the i stands for intrinsic. The PIN configuration is illustrated in Figure 8.28, where the optical power is incident from the left. The intrinsic region is known as the *depletion* region, and its width is w_D. The figure is exaggerated for clarity; the doped regions are actually kept thin so that most of the electron–hole pairs are generated in the depletion region. The pairs become separated in the depletion region, which creates a current in the external circuit. Large silicon PIN diodes are commercially available. The EARL power meter detector shown in Figure 8.8 is a 1-cm diameter PIN diode.

The parameter w_D affects the quantum efficiency and the bandwidth. To achieve a high QE, it is necessary to absorb most of the incident photons according to Beer's law, with an absorption coefficient α. The condition for full absorption is $w_D\alpha \gg 1$. This means that the QE of solid-state detectors can be engineered, and QE values of 80% and greater are common, with some AR-coated silicon detectors achieving almost 100%. However, increasing w_D increases the time required for carriers to traverse the depletion region, which decreases the bandwidth. This trade-off occurs for any type of junction photodiode, and it is dependent on wavelength because α generally varies with wavelength.

As mentioned earlier, detectors without internal gain are not used in direct detection lidar receivers unless there is no alternative, because gain can render thermal noise and amplifier noise inconsequential and greatly improve SNR, as shown by

Eq. (8.51). Photomultiplier tubes are used almost exclusively in the UV-VIS region and they can also be used in the NIR–SWIR, but at the longer wavelengths they tend to have high dark counts due to their low work functions, along with low QE values. In the early days of lidar there were no alternatives to PMTs, but by the mid-1980s, avalanche photodiodes (APDs) had become available and most 1064-nm lidars used silicon APDs. The APD is a solid-state analog of the PMT, achieving its gain through the phenomenon of *avalanche multiplication*, which occurs when the carriers gain enough energy from the applied field that they produce additional electron–hole pairs, which in turn produce further carriers. This process is called *impact ionization*. Lidars operating at 1064 nm use silicon APDs with gains of about 100, and SWIR lidars often use InGaAs APDs with gains of about 10. Impact ionization occurs at a field strength of ~10^5 V/cm, which may seem high, but the devices are so thin that the required bias voltages are only on the order of volts or tens of volts. If the bias voltage is kept lower than the *breakdown voltage* (at which the current becomes unlimited), an APD operates as a linear detector.

The single-photon avalanche diode (SPAD) is a related device that operates in the Geiger mode. It is an APD biased above the reverse bias breakdown voltage, so that a single photon will cause a self-sustaining avalanche. The current must be quickly interrupted before the device destroys itself, and this process is called *quenching*. The SPAD cannot detect another photon while being quenched, which means there is a *dead time* after each detection. There are two types of quenching circuit, active and passive, and the active version enables higher count rates by having a shorter dead time. SPADs are sometimes used in photon-counting lidars because they have higher QE than PMTs.

Linear APDs are modeled in a manner analogous to PMTs, with a gain M and a noise factor F. The power SNR is given by

$$\mathrm{SNR} = \frac{\eta P_\mathrm{S}^2}{2Fh\nu B\left(P_\mathrm{S} + P_\mathrm{B} + \dfrac{h\nu I_\mathrm{D}}{\eta q M}\right)\left[1 + \dfrac{1}{FM^2}\dfrac{h\nu}{2\eta q^2\left(P_\mathrm{S} + P_\mathrm{B} + \dfrac{h\nu I_\mathrm{D}}{\eta q M}\right)}\left(\dfrac{4kT}{R_\mathrm{f}} + (i_\mathrm{N})^2_{\sqrt{f}}\right)\right]} \tag{8.55}$$

which is Eq. (8.51) with the alternate variable names. The big difference is that the noise factor depends on both the gain and the *hole-to-electron ionization coefficient ratio* k_ion. The ionization coefficients describe the probability of creating an electron or a hole. Assuming that k_ion is independent of the electric field and that electron injection occurs on the p-side of the junction, the generation-recombination Poisson noise factor can be shown to be

$$F = k_\mathrm{ion} M + \left(2 - \frac{1}{M}\right)\left(1 - k_\mathrm{ion}\right). \tag{8.56}$$

The behavior of the function defined in Eq. (8.56) is not obvious, so it is plotted in Figure 8.29 for six values of k_ion. The noise factor is much larger for APDs than for PMTs, and in fact the noise factor is equal to the gain when k_ion is unity. Typical APD

Table 8.2 APD parameters

Detector	Ionization ratio k_{ion}	Typical gain M_{typ}	Noise factor F at M_{typ}
Silicon reach-through	0.02	150	4.9
Silicon epitaxial	0.06	100	7.9
InGaAs	0.45	10	5.5
Germanium	0.9	10	9.2

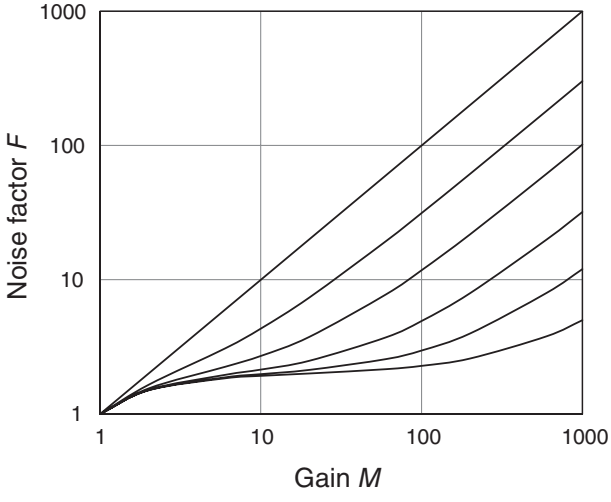


Figure 8.29 APD noise factor vs. gain. The values of k_{ion}, starting with the lowest curve, are 0.003, 0.001, 0.03, 0.01, 0.3, and 1.0. Adapted with permission from Figure 5.12 in [11].

gains of 10–100 are much smaller than the PMT gains of 10^5–10^7, but detector gain still helps to reduce the SNR-degrading effects of preamp noise. However, the second term in the square brackets of the denominator in Eq. (8.55) is much more significant for APDs than for PMTs, and as the signal power P_S decreases with range, the second term grows with range, and the SNR decreases faster than it does in the PMT case where that term is insignificant.

Another difference between PMTs and APDs is a lack of data for use in calculating the gain and the noise current as a function of the bias voltage. For this reason, the APD user must rely on gain and noise values that the manufacturer provides, or else measure them. Some APD parameters gleaned from sales literature are listed in Table 8.2.

In practical lidar engineering, the SNR formalism in this chapter results in simple equations for detectors with gain. For a lidar with a PMT detector and analog signal processing, such as EARL, Eq. (2.8) for the expected photon SNR becomes

$$\mathrm{SNR} = \sqrt{\frac{\eta}{\Gamma}} \frac{N_S}{\sqrt{N_S + N_B + N_D}}, \tag{8.57}$$

where N_S and N_B are calculated with Eqs. (2.11) and (2.14) and N_D is the equivalent dark count photon number, n_D/η. Dark count rates and QEs are generally provided by vendors, along with the secondary emission ratios that are used in finding the excess noise factor. For APDs, the equation is similar except that their lower gains may not make amplifier noise negligible. For that reason, vendors often provide a spectral NEP value in $W/Hz^{1/2}$ that includes all noise sources in the detector–amplifier package, in which case the expected SNR is

$$\text{SNR} = \sqrt{\frac{\eta}{F}} \frac{N_S}{\sqrt{N_S + N_B + N_{det}}}, \tag{8.58}$$

where

$$N_{det} = \frac{\eta\tau}{2F}\left(\frac{\lambda}{hc} NEP_{det}\right)^2 \tag{8.59}$$

and the parameter NEP_{det} is the spectral NEP value [12]. The derivation of Eq. (8.59) is left as a problem for the student.

To summarize this section, lidar optical signals are usually very small, and detectors with gain are required for approaching the statistical SNR limit because of amplifier noise. PMTs have several advantages: They can be used with either analog or photon counting signal processing; they are available in convenient sizes for lidar receivers; they are easy to model; they have gains of 10^5–10^7; their excess noise factor is small (~1.2) when operated at high gain; and their dark current is low in the UV-VIS region, where QE values are as high as 45%. Their disadvantages include the need for a high voltage power supply, and they have low QE and high dark current in the NIR–SWIR region. APDs offer high QE and gains of 10–100, and they operate with low-voltage power supplies. They tend to be small because many models were developed primarily for fiber-optic telecommunications. Their excess noise factors are much higher than those of PMTs, ranging from about 3 to 9. But even with their relatively low gains and high noise factors, their performance is sufficient to mitigate some amplifier noise, and they are usually the best choice for NIR–SWIR lidars. Intrinsic photoconductors and PIN diodes are seldom used in lidar receivers, with one exception: coherent detection systems. As pointed out in Section 8.6, coherent detection can in principle produce an NEP lower than signal-limited direct detection while also effectively limiting background light to a very small spectral region.

8.8 Further Reading

R. H. Kingston's first book [1], published in 1978, is the classic on optical and infrared detection. His derivations stem from basic physics, so the applications are very general. The book is remarkably terse, definitely at the graduate level. SNR is in terms of power (mean squared over variance). There are occasional errors in the equations.

R. H. Kingston's second book [11], published in 1995, is an update and extension of his first. Again, the book is at the graduate level, but it is more explanatory than the

first. The measure of merit throughout is SNR, but this time it is in terms of voltage (mean over standard deviation).

The Hamamatsu PMT Handbook [8] contains a wealth of technical information in an accessible format. It is well worth consulting for anyone using PMTs.

The book on semiconductor devices by Sze and Ng [10] is another classic; it has become the standard text on this topic, and it is widely referenced. It was written for advanced undergraduates, so it has much explanatory detail.

8.9 Problems

8.9.1 What is the maximum value of $S(\lambda)$ for a detector with unity gain? Derive an equation for $S(\lambda)$ for a detector with $\eta = G = 1$ as a function of wavelength in nm.

8.9.2 Using Eqs. (8.8) and (8.10), show that the signal-limited power SNR is just the mean number of photoelectrons, as expected from Poisson statistics.

8.9.3 Show that the SNR in Eq. (8.16) is the square of Eq. (2.8) when dark current can be neglected.

8.9.4 A certain lidar has a wavelength of 532 nm, a bandwidth of 10 MHz, a detector QE of 50%, and a daytime background power on the detector of 2.6×10^{-10} W. Considering only statistical noise, is the NEP signal-limited or background-limited?

8.9.5 Derive Eq. (8.25).

8.9.6 Derive coherent detection Eq. (8.32) from Eq. (8.30).

8.9.7 Show that two rays of light at 1-μm wavelength travelling through two different eddies at 300 K and 299 K and a pressure of 1000 mbar will be completely out of phase after propagating only ½ m.

8.9.8 Show that a 1% change in the 800 V bias voltage on a Hamamatsu R7400U PMT will cause a gain change of about 7%. The R7400U has eight dynodes and model parameters are $a = 0.1$ and $\alpha = 0.888$.

8.9.9 Derive Eq. (8.59) for APDs.

References

[1] R. H. Kingston, *Detection of Optical and Infrared Radiation*. Berlin: Springer-Verlag, 1978.
[2] J. Johnson, "Thermal Agitation of Electricity in Conductors," *Physical Review*, vol. 32, p. 97, 1928.

[3] G. G. Gimmestad and S. C. Bauman "Autonomous Lidar System for Cloud Climatology," in Optical Remote Sensing of the Atmosphere: summaries of papers presented at the Optical Remote Sensing of the Atmosphere Topical Meeting, Incline Village, Nevada, 1990, pp. 560–561.

[4] R. M. Measures, *Laser Remote Sensing – Fundamentals and Applications*. New York: John Wiley & Sons, 1984.

[5] P. W. Kruse, L. D. McGlauchlin, and R. B. McQuistan, *Elements of Infrared Technology: Generation, Transmission, and Detection*. New York: John Wiley & Sons, 1963.

[6] Photek photodiode data sheet. [Online]. Available: DS041 Photodiode Datasheet.pdf (photek.com). [Accessed: October 31, 2021].

[7] V. K. Zworykin, G. A. Morton, and L. Malter, "The Secondary Emission Multiplier-A New Electronic Device." *Proceedings of the IRE*, vol. 24, p. 351, 1936.

[8] Hamamatsu PMT handbook, 4th ed. [Online]. Available: www.hamamatsu.com/resources/pdf/etd/PMT_handbook_v4E.pdf. [Accessed: October 31, 2021].

[9] D. W. Roberts, G. G. Gimmestad, A. K. Garrison et al., "Design and Performance of a 100-Inch Lidar Facility," *Optical Engineering*, vol. 30, pp. 79–87, 1991.

[10] S. M. Sze and K. N. Ng, *Physics of Semiconductor Devices*, 3rd ed., Hoboken: John Wiley & Sons, 2007.

[11] R. H. Kingston, *Optical Sources, Detectors, and Systems*. San Diego: Academic Press, 1995.

[12] S. Ismail and E. V. Browell, "Airborne and Spaceborne Lidar Measurements of Water Vapor Profiles: A Sensitivity Analysis," *Applied Optics*, vol. 28, 3603–3616, 1989.

9 Data Systems

The last step in analog signal processing is digitization, which introduces its own unavoidable type of noise into the lidar signal and can also introduce phantom frequencies unless a certain precaution is followed. The first part of this chapter is an explanation of these noise and distortion phenomena. The next part, on figures of merit and testing methods, introduces the specialized jargon and terminology of digitizers so that lidar researchers can understand the specifications of commercial digitizers. Finally, some considerations for photon counting and hybrid systems are included.

9.1 Analog Data Systems

As shown in Figure 8.1, analog detection concludes with a digitizer, which converts an analog voltage waveform into digital signal levels at discrete time intervals, known as range bins, and transfers the data to a computer. Digitizers are also known as transient recorders, analog-to-digital converters, ADCs, and A-D, A to D, or A/D Converters. In the early days of lidar, such devices did not exist and the only electronic instrument fast enough to display a lidar waveform was the oscilloscope. The displayed waveforms were photographed; the film was developed, paper prints were made in a darkroom; and finally, the researcher analyzed the prints with a ruler to make a table of signal levels vs. range. Electronic digitizers date back to the 1950s and 1960s, but they were mostly for military radars and were not affordable for early lidar researchers. In addition, data processing was done on mainframe computers, so an additional step of recording data on punch cards or magnetic tape was required. By the 1980s, general purpose electronic digitizers were commercialized. For example, the Biomation Model 6500, from a medical instruments company, could digitize 1024 sequential range bins at 6-bit resolution (64 voltage levels) and it could be interfaced to a computer, and digitizers quickly became standard lidar equipment. Laboratory computers became commonplace with the advent of the IBM Personal Computer (PC) in 1981. Digital oscilloscopes also became available, some with signal averaging capability, and they provided another digitizing option. During the next three decades, digitizers and PCs matured together and became more and more capable and more affordable. A comprehensive history of electronic data converters is presented in Chapter 1 of *The Data Conversion Handbook* [1], and most of the material that follows in Section 9.1 is covered in other chapters of [1], in various forms.

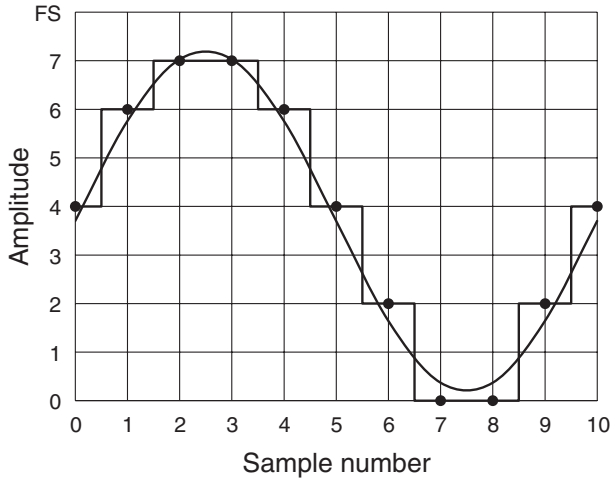


Figure 9.1 Sampling an analog waveform. The sine wave (solid curve) is continuous, but the samples, shown by the solid circles, are discrete digital values at discrete time intervals.

Digitizers convert a continuous, analog waveform in time and voltage into a *digital signal*, which is defined as a signal in which both time and voltage are discrete. They must minimize additional noise and nonlinearity, and they should be selected in tandem with the detector and amplifier to match the required dynamic range, bandwidth, and input impedance. Digitizers also need sufficient data transfer speed for moving the data to the computer. The process of converting a continuous analog signal to a discrete set of digital numbers is known as *sampling*. Quantization within a digitizer is performed by sampling the amplitude of an analog signal at discrete time intervals and assigning that amplitude an integer value, as illustrated in Figure 9.1 for a 3-bit digitizer. Such a digitizer has only $2^3 = 8$ possible output levels (the numbers 0–7) and so its applications are limited, but it is used as an illustration in this chapter for simplicity and clarity. The dashed stairstep line through the solid circles in Figure 9.1 illustrates that the output of such a digitizer represents the input waveform rather coarsely. The resolution on both axes is inherently limited by digitization, so it is important to understand digitizers in detail to accommodate the dynamic range and SNR of the analog lidar signal while avoiding distortions.

The time resolution of a digitizer (shown by the vertical gridlines in Figure 9.1) is determined by its *sample rate,* which is the speed at which sampling occurs in samples per second. In practice, sample rates are given in Hz, for example, a rate of 10 MHz means ten million samples per second. There is a crucial relationship between sampling rate and the frequency content of the signal, given by the *Nyquist Theorem*: the sampling rate must be at least twice as high as the highest frequency in the signal [2]. The reason for this constraint is to avoid *aliasing*, which is the phenomenon of higher frequencies appearing in the digital data as lower frequencies. That phenomenon is illustrated in Figure 9.2, which shows that a sine wave with a frequency slightly lower than the sampling frequency is aliased into the digital data as a much lower frequency.

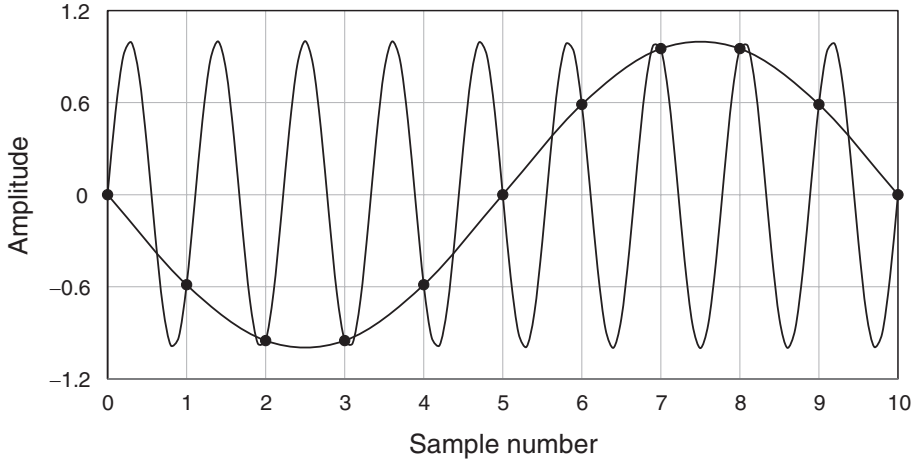


Figure 9.2 Aliasing. The solid black line is a sine wave with a frequency that is 90% of the digitizer's sampling frequency, the black circles are the sampled amplitude values, and the line connecting the black circles is the resulting aliased sine wave.

Once a too-high frequency has been aliased into digital data, it cannot be removed by any sort of tricks such as digital filtering, so aliasing must be avoided by one of two methods: (1) use a higher sampling rate. Using a sampling rate more than twice the highest frequency will ensure that aliasing does not occur; or (2) use a low-pass filter (called an anti-aliasing filter) to ensure that all frequencies entering the digitizer are lower than one-half the sampling rate. For example, the digitizer in EARL had a fixed 10 MHz sampling rate and a built-in 5 MHz low-pass filter so that aliasing could not occur. However, many digitizers have variable sampling rates under software control, in which case the user must ensure that the Nyquist criterion is not violated by the frequencies that are present in the input waveform.

A digitizer's full-scale range (FSR) specifies the range of voltages that a digitizer can properly digitize, and they usually have several ranges to select from. Some digitizers have zero volts as the center of the FSR, so that FSR = 1 V often means the max/min voltages are ±500 mV. The FSR cannot be arbitrarily low, which is the reason that sufficient amplification must be provided in the circuitry before the digitizer. Because a digitizer's resolution is the limiting factor in the fidelity of a digital lidar signal, the FSR should be fully utilized so that most of the output levels are used. Ideally, a voltage range should be used that will allow the input waveform to occupy 90–95% of the FSR. Also, digitizers often have a settable offset that can be used to take advantage of the maximum bits of resolution over a signal's range. An offset is often useful in lidar because the signal is all positive, whereas a digitizer's range may be centered on zero volts.

Discretization on the voltage axis is a larger consideration than on the time axis because it adds an uncertainty called *digitizer noise*. The input waveform can only be represented by a finite number of discrete levels, specified in binary bits (ones and

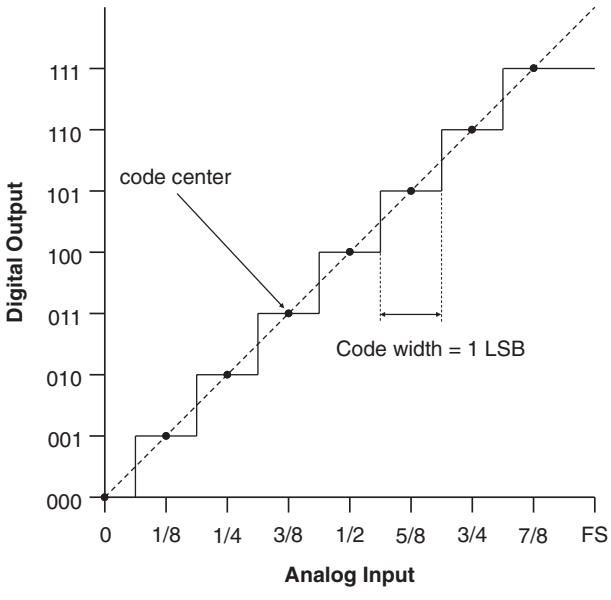


Figure 9.3 The ideal characteristic of a 3-bit digitizer. Adapted from Figure 2.5 in [1], used by permission. Owned by Analog Devices, Inc., © 2022. All rights reserved. This figure is reproduced under license and with permission by ADI. No unauthorized reproduction, distribution, or usage is permitted without ADI's prior written consent.

zeros). As examples, 8 digitizer bits provide $2^8 = 256$ discrete voltage levels, 10 bits $2^{10} = 1024$ levels, and 12 bits $2^{12} = 4096$ levels. Digitizers are designed to operate as illustrated in Figure 9.3, which is known as the *ideal characteristic* for a 3-bit digitizer. The output values (on the vertical axis) are binary numbers known as *codes*, and each unique code corresponds to a small range of analog input voltages, which is one *least significant bit* (LSB) wide. Input voltages are said to *resolve* to the code of the nearest code center. Formally, the input voltage V_{in} is related to the bits of a digitizer's binary output codes by the relation

$$V_{\text{in}} = V_Q \sum_{k=0}^{N-1} b_k 2^k + \varepsilon, \tag{9.1}$$

where the quantum voltage level $V_{\text{Q}} = V_{\text{FS}} / 2^N = 1\,\text{LSB}$. V_{FS} is the full-scale voltage, N is the number of bits, the b_k are the individual output bits, and ε is the quantization error. The index k runs from zero to $N-1$. By convention, the analog value represented by the all-ones code is not full-scale (abbreviated FS), but rather FS – 1 LSB. For example, in a 3-bit digitizer the code 111 corresponds to 7/8 of the full-scale voltage. Some quantization error is always present, because the voltage uncertainty at each code is ±½ LSB. The error vs. input voltage for a 3-bit digitizer is plotted in Figure 9.4. The quantization error limits the voltage resolution of a digitizer, and that resolution is characterized by the digitizer's SNR [3]. The electronic SNR is defined

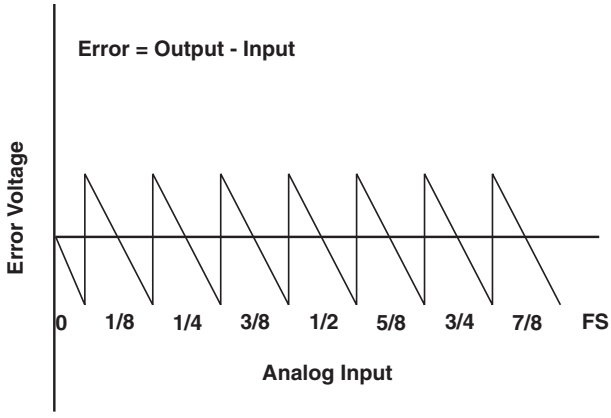


Figure 9.4 Quantization error for a 3-bit digitizer. Adapted from Figure 2.37 in [1], used by permission. Owned by Analog Devices, Inc., © 2022. All rights reserved. This figure is reproduced under license and with permission by ADI. No unauthorized reproduction, distribution, or usage is permitted without ADI's prior written consent.

as 20 times the common logarithm of the mean voltage divided by the standard deviation of the noise, which in this case is the r.m.s. voltage divided by the standard deviation of the error shown in Figure 9.4. In the analysis of digitizers, the input waveform is usually assumed to be a sine wave. The r.m.s. value of a sine wave with amplitude A is given by $V_{\text{rms}} = A/\sqrt{2}$, so the r.m.s. value of a full-scale sine wave with peak-to-peak voltage V_{FS} is

$$V_{\text{rms}} = \frac{V_{\text{FS}}}{2\sqrt{2}} = \frac{2^N V_Q}{2\sqrt{2}}, \tag{9.2}$$

The range of the voltage error ε is $\pm V_Q/2$. To find the r.m.s value of ε, we must find the expectation value of the squared error voltage and then find the square root. The mean-square error ε^2 is

$$\overline{\langle \varepsilon^2 \rangle} = \frac{1}{V_Q} \int_{-V_Q/2}^{+V_Q/2} \varepsilon^2 d\varepsilon = \frac{1}{V_Q} \left| \frac{\varepsilon^3}{3} \right|_{-V_Q/2}^{+V_Q/2} = \frac{V_Q^2}{12} \tag{9.3}$$

and the signal-to-noise ratio due only to quantization is therefore

$$\text{SNR} = 20\log\left(\frac{V_{\text{rms}}}{\sqrt{\overline{\langle \varepsilon^2 \rangle}}} \right) = 20\log\left(2^N \sqrt{1.5}\right) = 6.02N + 1.76\,\text{dB}. \tag{9.4}$$

The digitizer resolution in bits, digitizer levels, and SNR is shown in Table 9.1 for several values of N. It should be stressed that these SNRs are *ideal* values, meaning the best that can be achieved with a given number of bits. Reasons that a real digitizer may not achieve these values are described in Sections 9.1.1–9.1.3. In digital

Table 9.1 Ideal digitizer SNR values

Resolution (bits)	Quantizing levels	SNR (dB)
6	64	37.9
8	256	49.9
10	1,024	62.0
12	4,096	74.0
14	16,384	86.0
16	65,536	98.1

electronics, the notation K means 2^{10} or 1,024, rather than 1,000, so a 12-bit digitizer has 4 K levels, for example.

Table 9.1 suggests a question: How could early lidars operate with 6-bit or 8-bit digitizers, with so few quantizing levels? The answer is that atmospheric lidar systems almost always use multi-pulse averaging, which increases the digitizer SNR according to the relation

$$\mathrm{SNR}_{\mathrm{ideal}}(N,n) = 20\log(2^N\sqrt{1.5n}), \tag{9.5}$$

where n is the number of pulses averaged. There are therefore two reasons why multi-pulse averaging is important in analog detection lidars: It increases the SNR of the signal, and it increases the digitizer's SNR so that it can accommodate the signal SNR. The increase in SNR with averaging is predicated on having enough electronic gain before the digitizer so that the noise excursions span several digitizer levels, because averaging will have no effect if the noise is all within one level.

In practice, a real digitizer may deviate from the ideal characteristic in several ways. For this reason, several different figures of merit have been developed for digitizers. They are described in the next three sections, and testing methods are described in Section 9.1.4.

9.1.1 Figures of Merit: Static

Offset and gain errors are shown in Figure 9.5. Offset error is the deviation at zero, where the first transition should be at ½ LSB, as it is in Figure 9.3. Gain error is the deviation of the slope of the line through the code centers from $2^N/V_{\mathrm{FS}}$ codes per volt. These errors are not usually a problem in lidar, because most lidars are uncalibrated so gain is not important, and because removing any offset in signals is usually the first step in data analysis.

Nonlinearities, illustrated in Figure 9.6, are a potential problem for lidars because they distort the signal. Differential nonlinearity (DNL) is the deviation of the code transition widths from the ideal of 1 LSB (caused by the missing code 100 in Figure 9.6), whereas integral nonlinearity (INL) is the distance of the code centers from the ideal line. A best linear fit is shown in Figure 9.6, but the INL can be nonlinear, which would distort the signal; see Chapter 3 of [1].

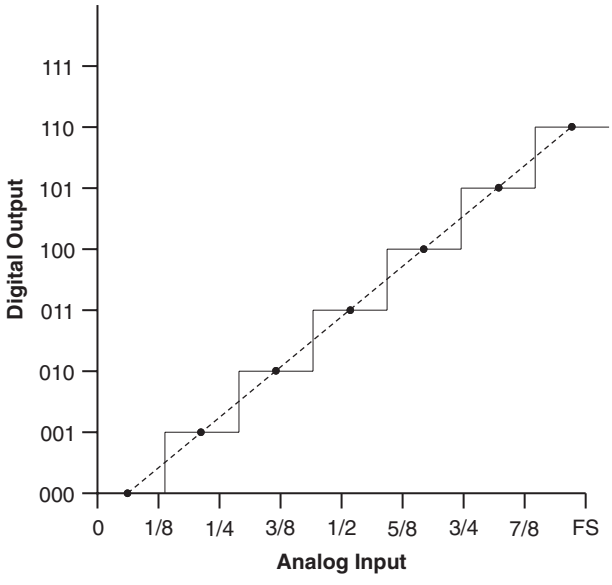


Figure 9.5 Digitizer offset and gain errors. Adapted from Figure 2.17 in [1], used by permission. Owned by Analog Devices, Inc., © 2022. All rights reserved. This figure is reproduced under license and with permission by ADI. No unauthorized reproduction, distribution, or usage is permitted without ADI's prior written consent.

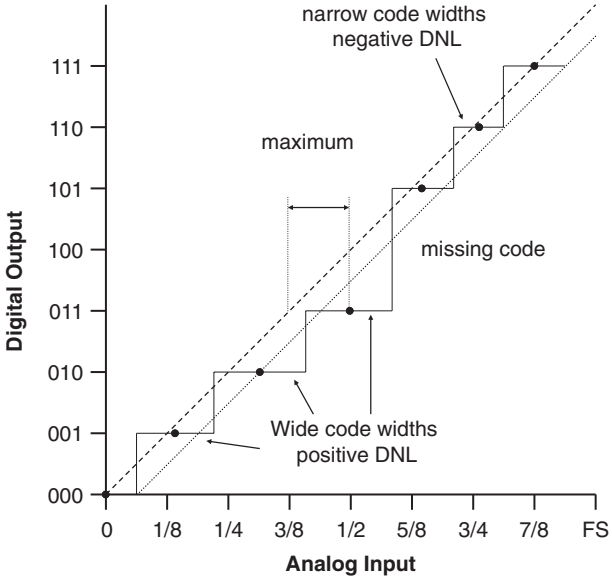


Figure 9.6 Nonlinearities. The solid line is the ideal characteristic, and the dashed line is the best linear fit through the code centers. Adapted from Figure 5.27 in [1], used by permission. Owned by Analog Devices, Inc., © 2022. All rights reserved. This figure is reproduced under license and with permission by ADI. No unauthorized reproduction, distribution, or usage is permitted without ADI's prior written consent.

9.1.2 Figures of Merit: Dynamic, Frequency Domain

In addition to its number of bits, a digitizer has several other figures of merit that pertain to dynamic performance: signal-to-noise and distortion ratio; total harmonic distortion; effective number of bits (ENOB); spurious-free dynamic range (SFDR), and intermodulation distortion (IMD). All these figures of merit are concisely defined in [3]. Their values are generally provided in the specifications of commercial digitizers, so it is important to understand what they mean when selecting a digitizer for a lidar data system. All the dynamic figures of merit pertain to pure sine wave inputs, and they tend to vary with the amplitude and frequency of the input signal. In the frequency domain, they are generally measured by examining a fast Fourier transform (FFT) of a digitizer's output, as illustrated in Figure 9.7, which is a synthetic FFT of a digitizer's output when the input is a 1 kHz sine wave. The presence of harmonics of the input frequency in the digitized output is a common problem, and some constant level of noise, known as the *noise floor*, is always present because it is due at least in part to the quantization error. The noise floor shown is at –120 dB.

The actual SNR is the digitized signal power divided by the noise. The signal-to-noise-and-distortion ratio (SNDR), also called S/N + D or SINAD, is the input signal amplitude divided by the r.m.s. sum of all other spectral components (harmonics plus noise). It is a measure of the real SNR of the digitizer. For an M-point FFT of a sine wave test, with the fundamental in bin m with amplitude A_m,

$$\mathrm{SNDR} = 10\log\left[\frac{A_m^2}{\sum_{k=1}^{m-1} A_k^2 + \sum_{k=m+1}^{M/2} A_k^2}\right]. \tag{9.6}$$

SNDR degrades as frequency increases, so test results are often plotted versus frequency, and the input signal frequency where the SNDR of the digitizer has fallen by 3 dB from its low-frequency value is called the effective resolution bandwidth.

Total harmonic distortion (THD) is also explained by reference to Figure 9.7. It is a measurement of the distortion due to harmonics of the fundamental signal, and it is computed as the inverse ratio of the power of the fundamental signal and the sum of the powers of all harmonics, or according to some authors, the first five harmonics, in which case it is

$$\mathrm{THD} = 20\log\left(\frac{V_2^2 + V_3^2 + V_4^2 + V_5^2 + V_6^2}{V_1^2}\right), \tag{9.7}$$

where V_1 is the amplitude of the fundamental, and V_n is the amplitude of the nth harmonic.

At low frequencies, the resolution of a digitizer should be close to that specified in Table 9.1. At higher input signal frequencies, the effective resolution of a digitizer will generally be reduced. This phenomenon is characterized by the digitizer's ENOB, which is a measure of the actual voltage resolution of a digitizer at various frequencies. Effective number of bits must be specified with relation to amplitude and

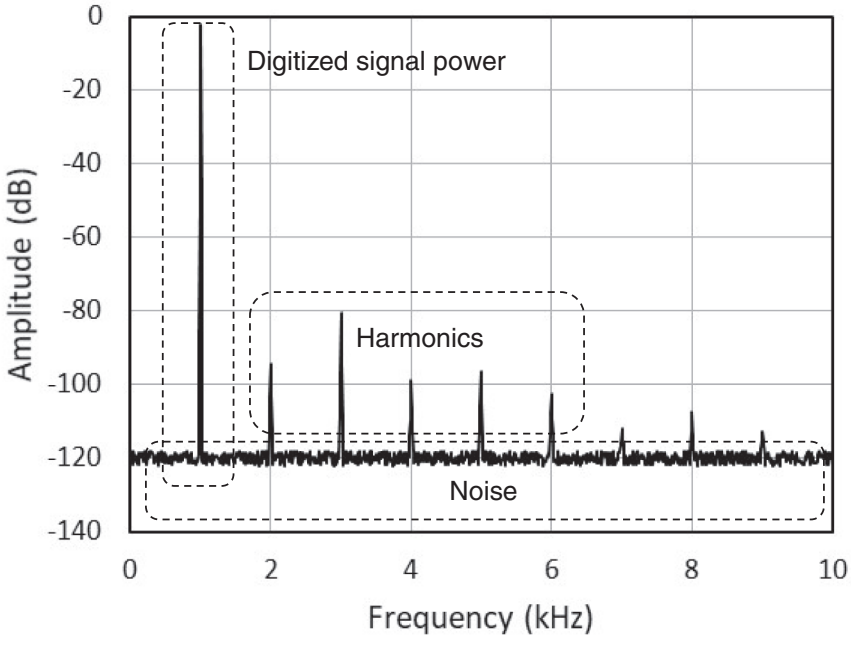


Figure 9.7 Synthetic FFT of digitizer output. The values within the labeled dashed boxes are used to calculate several figures of merit. Adapted from Figure 1 in www.planetanalog.com, "Signal Chain Basics #80: Optimizing Power vs. Performance for a SAR-ADC Drive Amplifier," used by permission.

frequency. If we interpret SNDR as the real SNR of the digitizer, then substituting SNDR for SNR in Eq. (9.5) and solving for N yields the relation

$$\text{ENOB} = \frac{\text{SNDR} - 1.76}{6.02}. \tag{9.8}$$

Another way of assessing ENOB is explained in Section 9.1.4.

The harmonics of the input frequency are caused by nonlinearities in the digitizer, and they are known as *spurs*. The amplitude of the spurs does not decrease with averaging. Spurious-free dynamic range is the input signal divided by the peak harmonic component. In Figure 9.7 for example, that component is at 3 kHz. SFDR can be larger than SNDR, and like many other figures of merit, it depends on the frequency and amplitude of the input signal.

If the input to a digitizer is two different frequencies, sum and difference frequencies may appear in the FFT in addition to harmonics. This phenomenon is known as IMD, which is defined as

$$\text{IMD} = 10\log\left(\frac{V_+^2 + V_-^2}{V_1^2 + V_2^2}\right), \tag{9.9}$$

where V_1 and V_2 are the r.m.s. amplitudes of the input signals, and V_+ and V_- are the r.m.s. amplitudes of the sum and difference intermodulation products.

9.1.3 Figures of Merit: Dynamic, Time Domain

Dynamic figures of merit in the time domain include aperture error, transient response, and overvoltage recovery. *Aperture error* is named by analogy with a snapshot camera, which freezes an image at a specific instant in time that ends when the shutter closes. High-speed digitizers use a *sample and hold* technology, where the signal is measured (sampled) somewhat like the way a camera aperture is closed. While the digitizer aperture is open, the input signal charges a capacitor. When the aperture closes, the voltage of the capacitor is read and held. Any variation in the time when the aperture closes introduces a voltage error known as aperture error, and the variation in time can be caused by clock jitter. The higher the input frequency and amplitude of a signal, the more the aperture error will introduce uncertainty in a signal, as illustrated in Figure 9.8.

Aperture error can cause a surprisingly large decrease in ENOB. An example is shown in the next section, where the measured high-frequency ENOB values for a state-of-the-art 12-bit digitizer proved to be in the range of 8.2 to 8.8. This loss of resolution may not be an issue for atmospheric lidars with sufficient multi-pulse averaging, but it is serious for waveform lidars that rely on single-pulse signals, such as mapping bathymetric lidars and biomass lidars that document tree density in forests. If we assume that the input waveform is a full-scale sine wave, $V_{\text{in}} = V_{\text{FS}} \sin \omega t$, the maximum slope of the waveform is ωV_{FS}. If the r.m.s. aperture error is t_{a}, the r.m.s. voltage error is

$$V_{\text{rms}} = \omega V_{\text{FS}} t_{\text{a}} = 2\pi f V_{\text{FS}} t_{\text{a}}. \tag{9.10}$$

The signal-to-noise ratio caused by aperture error for a full-scale signal is then

$$\text{SNR} = 20 \log\left(\frac{V_{\text{FS}}}{V_{\text{rms}}}\right) = 20 \log\left(\frac{1}{2\pi f t_{\text{a}}}\right). \tag{9.11}$$

Example. For a 5.00 MHz sine wave and a 250 ps aperture error, the SNR would be $20 \log[1/(6.28) \times (5 \times 10^6) \times (250 \times 10^{-12})] = 20\log[127] = 42$ dB.

Transient response is the settling time for the digitizer to achieve full accuracy (within ± ½ LSB) after a step in the input voltage from zero to full scale. Overvoltage recovery is the settling time for the digitizer to achieve full accuracy after a step in the input voltage from outside the full scale. Neither of these figures of merit should be important for atmospheric lidar because atmospheric signals generally do not have such steps, and the signal must never exceed the full-scale voltage V_{FS}.

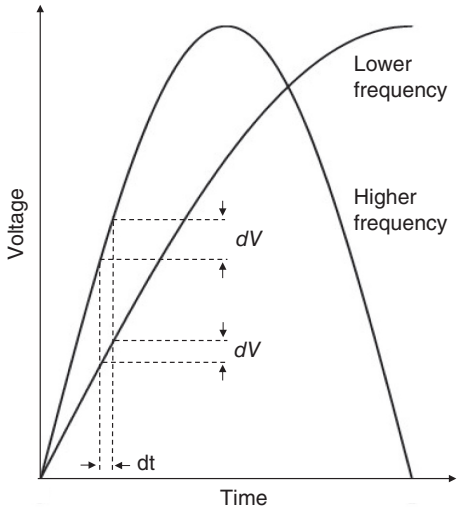


Figure 9.8 Aperture error. The voltage uncertainty (*dV*) caused by aperture error (*dt*) increases with the amplitude and frequency of the input waveform. Adapted from Figure 1 in [4], used by permission. Owned by Analog Devices, Inc., © 2022. All rights reserved. This figure is reproduced under license and with permission by ADI. No unauthorized reproduction, distribution, or usage is permitted without ADI's prior written consent.

9.1.4 Dynamic Digitizer Testing Methods

The first step in testing should be to conduct some common sense tests. Many problems can be detected simply by observing the noise of a digitizer using manufacturer-supplied software. Test should be performed with all the required options exercised, for example all input ranges, all sample rates, with and without averaging, with and without an internal filter (if applicable) and recording pre-trigger and post-trigger data (if desired). After these tests, dynamic testing methods include the use of FFTs, the histogram linearity test, and the sine wave curve fit for ENOB.

With a pure sine wave input, an FFT should result in a plot like Figure 9.7 that can be used to calculate SNR, SINAD, SFDR, and THD. However, there are several subtleties to this method that cause serious distortions in the results if they are ignored. The sampled data must contain an integral number of periods of the input waveform, but the number of periods must not be a non-prime submultiple of the record length. For example, a 4096-point FFT of 128 periods of a sine wave will result in seriously misleading results. A non-divisor prime number of cycles is needed; for example, 127 periods of a sine wave in a 4096-point FFT will give good results.

The *code density*, or *histogram*, test is a dynamic method that will reveal digitizer problems related to nonlinearities, missing codes, gain error, and offset error. The concept of this test is simple: Apply a signal to the digitizer (usually a sine wave) that has a known probability density function (PDF). Sample the

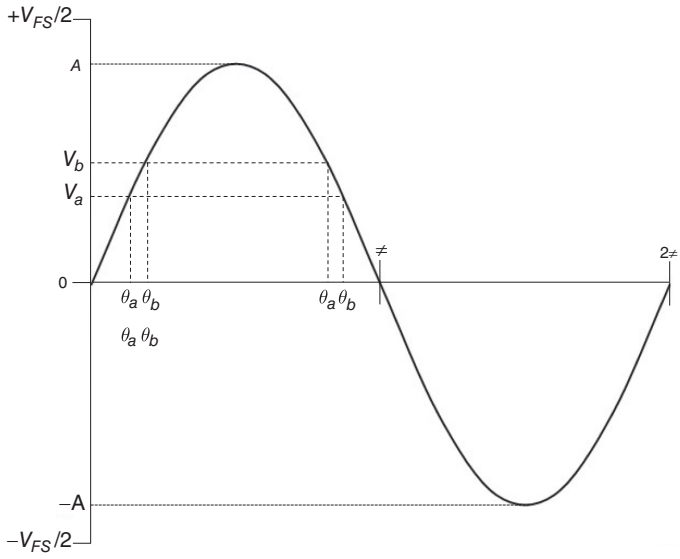


Figure 9.9 Code probabilities. The probability that a code lies within a given voltage range is the corresponding angle range divided by 2π.

waveform repeatedly while forming a histogram showing the probability of occurrence of each code. Errors can be calculated by comparing the sampled PDF with the expected PDF. As illustrated in Figure 9.9 for continuous variables, for a sine wave $V = A\sin\omega t$ the probability of a sample being in the interval $[V_a, V_b]$ is just the angular range corresponding to those two voltages divided by the full wave angle 2π, so the probability is

$$p(V_a, V_b) = \frac{1}{\pi}\left(\sin^{-1}\frac{V_b}{A} - \sin^{-1}\frac{V_a}{A}\right). \tag{9.12}$$

Note that the voltage pairs appear twice in each half-wave. When Eq. (9.12) is translated to discrete variables for an N-bit digitizer, it becomes

$$p(n) = \frac{1}{\pi}\left[\sin^{-1}\left(\frac{V_{\mathrm{FS}}(n-2^{N-1})}{A\cdot 2^N}\right) - \sin^{-1}\left(\frac{V_{\mathrm{FS}}(n-1-2^{N-1})}{A\cdot 2^N}\right)\right], \tag{9.13}$$

where n is the code index with range $1-N$; see Chapter 5 of [1]. An example of this PDF is shown in Figure 9.10 for a 6-bit digitizer, which has codes from 0 to 63 (decimal). The maxima are at the extreme high and low values because they are the most likely for a sine wave, whereas the lowest probability is for the zero crossing, in the center.

The histogram method has subtleties, like the FFT method. One is that the frequency of the input sine wave must not be a subharmonic of the sampling frequency, because the method is predicated on random sampling. Another is that a large number

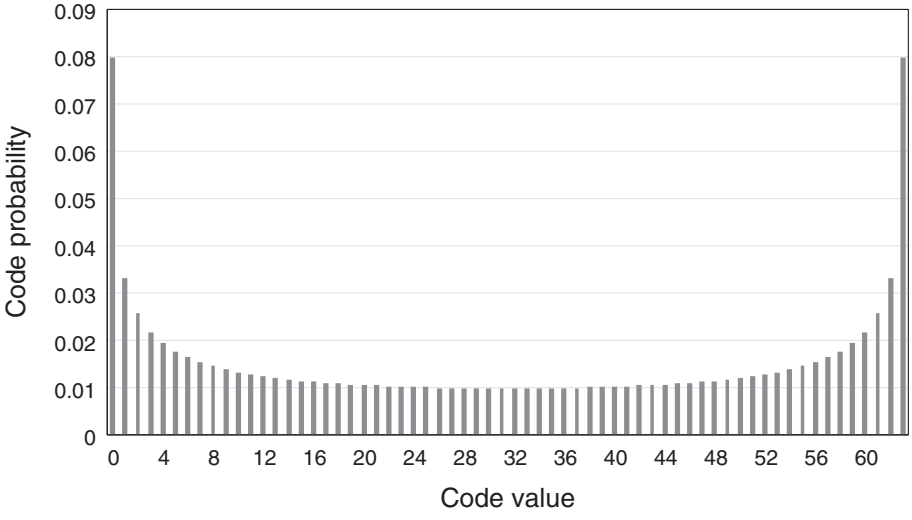


Figure 9.10 Code probability histogram for an ideal 6-bit digitizer.

of waveforms must be digitized to get a good statistical measure of the PDF. The total number of samples needed to compute the histogram can be calculated from the user's accuracy requirement. For a Gaussian error distribution, the number of samples M required for a high-confidence measurement is

$$M = \frac{\pi 2^{N-1}(z^*)^2}{\beta^2}, \tag{9.14}$$

where N is the number of bits, z^* is the confidence level expressed as the number of standard deviations from the mean, and β is the desired accuracy [5]. For a 6-bit digitizer, a confidence level of 99% (2.58 standard deviations), and a DNL measurement resolution of ±0.1 LSB,

$$M = \frac{\pi 2^5 (2.58)^2}{(0.1)^2} \approx 67{,}000. \tag{9.15}$$

The number of samples required doubles with each added bit and it may seem unwieldy, but the measurements are automated. The Georgia Tech Research Institute performed this test on a 12-bit digitizer for an ozone lidar in the early 2000s with good results. In practice, the histogram test does not use all the codes illustrated in Figure 9.10, because of the difficulty of adjusting the amplitude A to exactly match $V_{FS}/2$. Instead, the digitizer is slightly overdriven and the two endpoint codes (all zeros and all ones) are not used in the analysis, so the total number of codes used becomes $2^N - 2$. Any voltage offset in the input sine wave is assessed by inspecting the total statistics above and below zero volts. To assess DNL, the number of "hits" (occurrences) for code n is divided by the number

of samples M to find $h(n)$, and the theoretical probability for code n is given by Eq. (9.13). The DNL is then

$$\mathrm{DNL}_n = 1\,\mathrm{LSB}\left(\frac{h(n)}{p(n)} - 1\right), \tag{9.16}$$

and the INL is

$$\mathrm{INL}_n = \sum_{i=0}^{n} \mathrm{DNL}_i. \tag{9.17}$$

The effective number of bits can be calculated from SINAD using Eq. (9.8), but another method is to compare an ideally digitized sine wave to an actual one. As illustrated in Figures 9.1 and 9.4, the digital output codes do not represent the input perfectly because of quantization error, and this phenomenon leads to the SNR shown by Eq. (9.4) for an ideal digitizer. This concept can be expanded to create another way to measure ENOB: Digitize a sine wave and calculate the r.m.s. difference between the input and the digital output. ENOB is then calculated from

$$\mathrm{ENOB} = N - \log_2\left(\frac{E_{\mathrm{rms}}}{V_Q / \sqrt{12}}\right), \tag{9.18}$$

where E_{rms} is the r.m.s. difference between the input and the digital output and the denominator is the r.m.s. quantization error given in Eq. (9.3). In practice, several cycles of a sine wave are digitized, a function of the form $y = A\sin(\omega t + \phi) + C$ is fitted to the data by means of a curve-fitting algorithm, and then E_{rms} is calculated. This method is of course insensitive to errors in amplitude, phase, and offset because they are all calculated from the output. As with all statistical measures, large numbers of samples are required for reasonable accuracy.

In 2015, GTRI evaluated a state-of-the-art high-speed digitizer that was a candidate for an airborne bathymetric lidar. In this type of lidar, an analyzed backscatter profile must be developed from each laser pulse, so multi-pulse averaging is not possible, and ENOB is therefore a key figure of merit. The digitizer had four input channels, 12 bits of resolution, and a maximum sample rate of 3.2 GHz. The curve fitting method with Eq. (9.18) was used to measure ENOB over a range of frequencies at two different digitization rates. Some of the resulting data are shown in Figure 9.11, for a digitization rate of 1.6 GHz and input frequencies of 550, 600, and 650 MHz. Measured ENOB values ranged from 8.2 to 8.8 bits (294 to 445 useful digitization levels, as opposed to 4096). This is a good example of the fact that ENOB may be significantly less than 2^N at high input frequencies.

Digitizer figures of merit and testing methods are summarized in Table 9.2, which was adapted from [6]. An entry of "Yes" in this table indicates that a test is sensitive to an error type, not that it is a specific test for it.

The foregoing descriptions of digitizer performance metrics and test procedures may suggest that a digitizer with a constant input voltage will always produce the

Table 9.2 Digitizer testing summary

Digitizer error	ENOB test	FFT test	Histogram test
DNL	Yes	Yes	Yes
INL	Yes	Yes	Yes
Missing codes	Yes	Yes	Yes
Aperture error	Yes	Yes	No
Noise	Yes	Yes	No
Gain error	No	No	Yes
Offset error	No	No	Yes

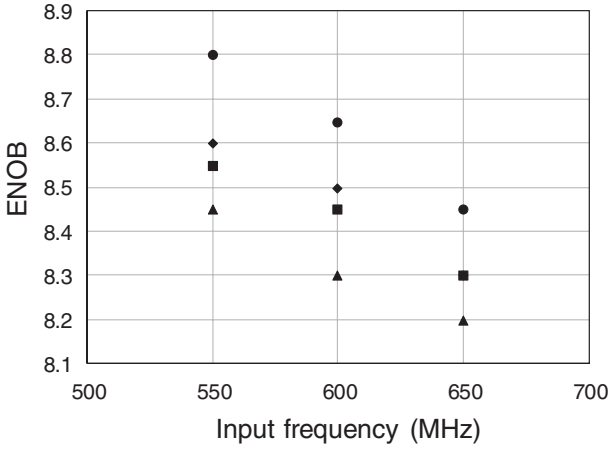


Figure 9.11 Measured ENOB values of a 12-bit digitizer. The digitization rate was 1.6 GHz, and the four symbols refer to the four input channels.

same output code. In a fast digitizer, this is generally not true. For example, EARL was tested with 50-Ω terminations on the digitizer inputs, so that the input voltages were zero. The data acquisition software was used to acquire a single sweep of the 2048 range bins, and then the standard deviation of the data values was calculated from a sample of about 300 bins. The standard deviation was 0.17 digitizer levels, which presumably means that zero voltage resolved to the code for zero most of the time, but it occasionally resolved to one code above or below zero, which shows that a source of noise has been omitted in the discussions above. That source is Johnson noise [7]. The noise power is given by $P = kTB$, where the symbols have their usual meaning. Including Johnson noise yields a noise floor consistent with Figure 9.7 and consistent with the measured EARL standard deviation (the calculations are left as a problem for the student). The EARL tests were repeated with averages, increasing the number of sweeps by a factor of 4 each time, to see whether the standard deviation

decreased by a factor of 2. Initially it did, but after several quadruplings it approached a constant value. This test revealed a software blunder: Digitizer codes are integers of course, but averages must be saved as real numbers and the software was only recording two decimal digits, which was insufficient for long-term averages. The same tests were performed with the PMTs and TIAs powered up and connected, but with a cap over the receiver. The single-sweep standard deviation was 0.59, or about 3.5 times higher, due to TIA noise. Both Johnson noise and TIA noise were very small compared to the signal, which was on the order of 100 digitizer levels for a single photon, but simple tests such as these are well worth doing once a lidar receiver and data system are assembled, to insure that these parts of the lidar (and the software) are operating as expected.

Lidar signals tend to have very large dynamic ranges, so accommodating them is another issue for data systems. Estimates start with the lidar and background equations along with atmospheric and operating scenario parameters. After ensuring that the detector and TIA can accommodate the signal's full range (by splitting the optical signal into short-range/long-range receiver channels for example, as in EARL), the last step is to ensure that the digitizer has adequate dynamic range. If not, the TIA output can be split, and two digitizers can be used, for weaker and stronger signals. This technique was employed on the spaceborne CALIOP lidar that had to accommodate strong backscatter from Earth's surface as well as weak Rayleigh backscatter from high in the atmosphere, which was used for calibration [8]. CALIOP employed two overlapping 14-bit digitizers on each receiver channel, yielding an effective 22-bit resolution.

9.2 Photon Counting Systems

Following the TIA, a photon-counting lidar has different components from those shown in Figure 8.1. Photon counting components are illustrated in Figure 9.12, where the inset graphs illustrate the voltage waveforms that are passed from each component to the next. The sequence of operations is that, after the photodetector produces narrow current pulses in response to incident photons and dark current pulses, the current pulses are converted to voltage by the TIA (and often inverted); a discriminator only passes pulses above a set voltage level; a pulse shaper converts the analog pulses from the discriminator into identical digital pulses; and a counter counts the pulses in equal time intervals (range bins). The result is again a digital waveform, plotted as photocounts vs. range bin number. The time bases for the three inset waveforms shown in Figure 9.12 are identical: Of the five pulses coming from the amplifier, only the three largest ones are passed on by the discriminator. Those three pulses are then converted to identical logic pulses, which are counted and ultimately transferred to a computer for averaging, storage, and analysis. In practice, some or all the functions shown in the gray boxes may be incorporated into one piece of electronic equipment.

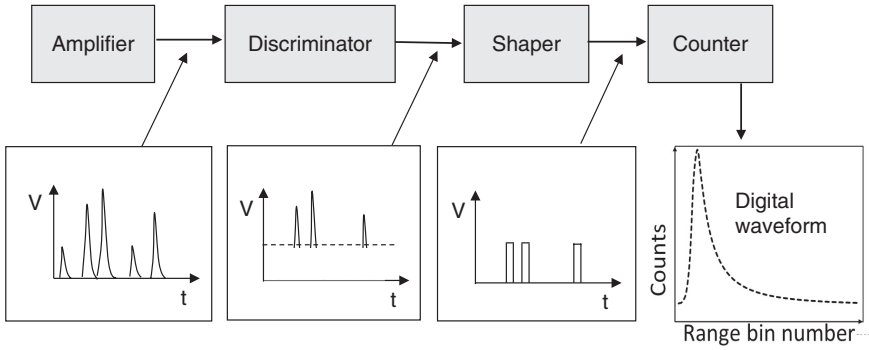


Figure 9.12 Photon counting system components. The discriminator, shaper, and counter replace the digitizer shown in Figure 8.1.

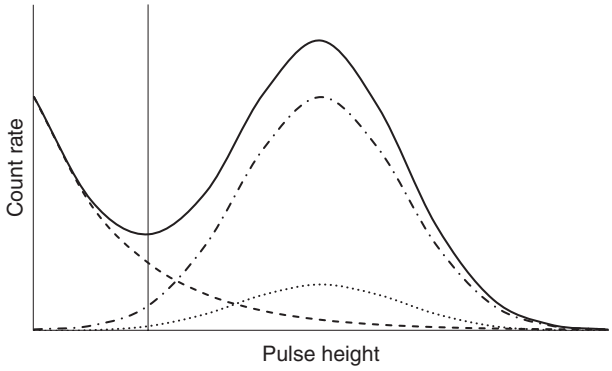


Figure 9.13 PMT pulse height distribution. Dashed curve – dark counts from dynodes; dotted curve – dark counts from photocathode; chain-dot curve – photocounts; solid curve – total counts; vertical line – optimum discriminator setting.

The reason for using a discriminator is to remove noise (dark counts), if the noise pulses have a smaller amplitude than the pulses due to photons. The PMT's pulse height distribution (PHD) is a histogram that shows the frequency of occurrence of pulse heights. The PHD is the result of several contributions: pulses with large amplitudes produced by electrons from the photocathode (photoelectrons and thermally generated electrons) and pulses with smaller amplitudes due to thermally generated electrons from the dynodes. The amplification of electrons from the dynodes is lower because the full dynode chain is usually not involved. If the PHD looks like Figure 9.13, the discriminator level is set at the "valley" between the thermal counts and the photoelectron peak. Although illustrations like Figure 9.13 are common in the PMT literature, in practice the PHD for a lidar PMT may not have a minimum.

The measured PHD of the Hamamatsu H10330A-75 IR PMT mentioned in Chapter 8 decreased approximately exponentially with pulse height, so the discriminator setting was not important. With 1 hour of integration (72,000 laser shots) at

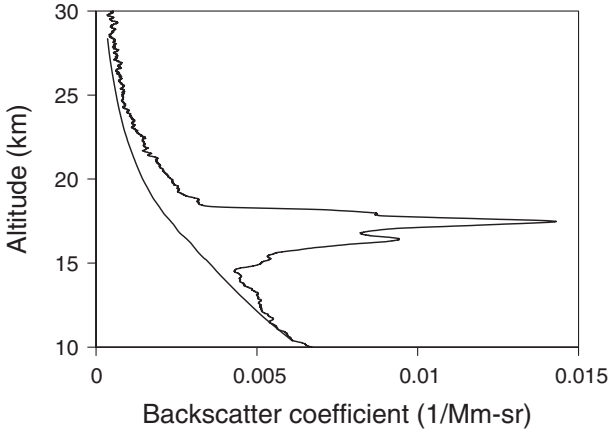


Figure 9.14 Stratospheric backscatter at 1574 nm. Solid line – total backscatter; dashed line – molecular backscatter calculated from upper air data. The data were recorded in 3.75-m range bins and smoothed to 150-m resolution.

night, that lidar accurately measured the stratospheric aerosol layer backscatter, as shown in Figure 9.14 [9]. It is commonly supposed that SWIR lidars cannot see stratospheric molecular backscatter, but it is obviously a part of the measured profile. This example illustrates the ability of photon-counting lidars to approach the statistical limit of SNR, and it is also an example of the conclusion in Figure 2.3 that a useful lidar SNR can be obtained even in the presence of a high dark or background count rate if sufficient averaging is used (the dark count rate was 2.7×10^5 cps).

In photon counting TIAs, the electronic bandwidth is significantly higher than in analog detection to produce narrow pulses that overlap as little as possible. However, due to the random nature of photon arrival times, it is still possible for detector output pulses to overlap in time. When this happens, the pulses are not counted correctly and the observed count rate is less than the true count rate, and for high count rates, the count rate is no longer linearly proportional to the irradiance on the detector. There is a time interval after a photon arrival called the dead time τ_d, when the counting system cannot respond to the arrival of the next photon. This phenomenon is illustrated in Figure 9.15. The counter (the last gray box in Figure 9.12) can count at high rates, typically hundreds of Mcps, but the dead time problem occurs before it, in the shaper.

Count rates are high even at very low levels of detector irradiance. In typical lidars, the received power level from low altitudes can easily exceed 10^{-9} W, and nonlinearities produced by dead time effects can be substantial. Count rates for an irradiance of 10^{-11} W and a QE of 20% are illustrated in Figure 9.16 for the first three harmonics of the Nd:YAG laser. A count rate of 10^7 s^{-1} corresponds to one count per range bin for 15-m bins.

There are two types of photon counting systems: in a *non-paralyzable* system, the arrival of a second photon during the dead time of the previous photon is ignored; and in a *paralyzable* system, the arrival of a second photon during the dead time of the previous photon is ignored but it also extends the dead time. Theoretical models have been developed for both types [10]. The observed count rate in a paralyzable system is modeled as

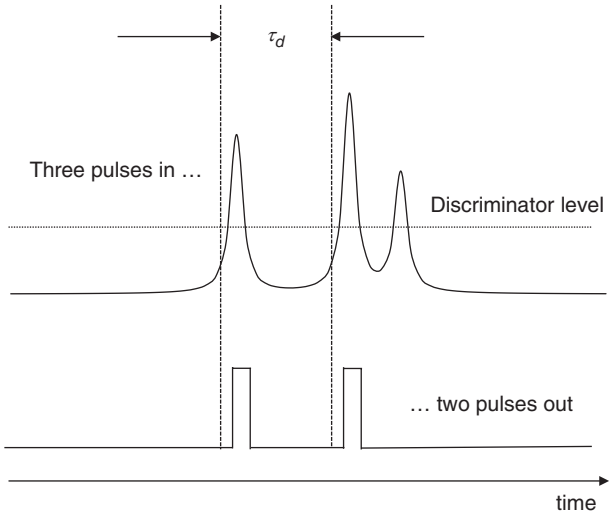


Figure 9.15 Dead time. If one pulse is followed by another in less than the dead time, the shaper does not respond to the second pulse.

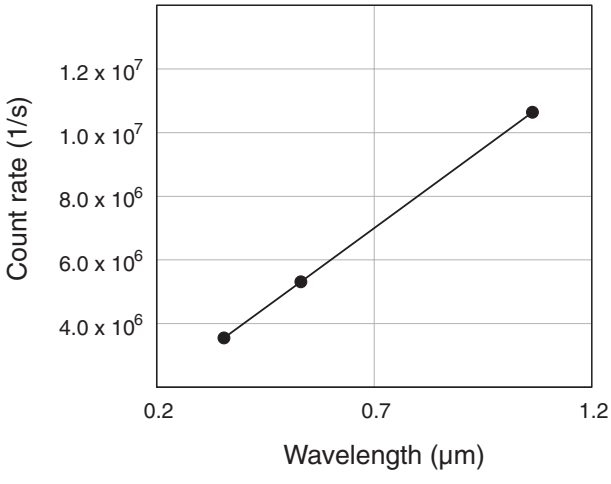


Figure 9.16 Count rates. The solid circles show the count rates for the first three harmonics of the Nd:YAG laser for a received power of 10^{-11} W and a 20% QE.

$$R_{\text{obs}} = R_{\text{true}} \exp(-R_{\text{true}}\tau_{\text{d}}), \tag{9.19}$$

and the observed count rate in a non-paralyzable system is modeled as

$$R_{\text{obs}} = \frac{R_{\text{true}}}{1 + R_{\text{true}}\tau_{\text{d}}}, \tag{9.20}$$

where R_{obs} is the observed count rate and R_{true} is the true count rate. In both cases, the observed count rate approaches the true count rate when the product of the true

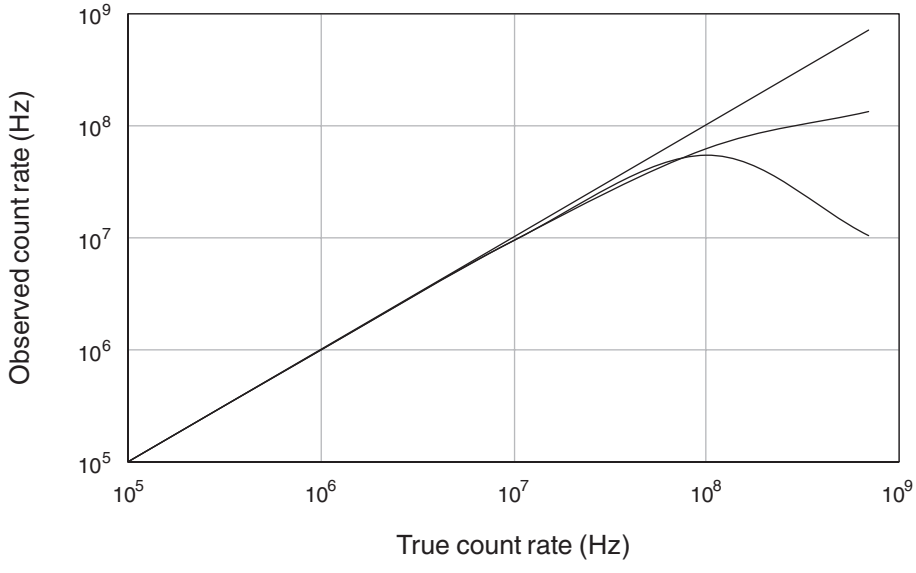


Figure 9.17 Models of photon counters. Observed count rates versus true count rate for a dead time of 6 ns. Top line – true count rate; lowest line – paralyzable; middle line – non-paralyzable.

count rate R_{true} and the dead time τ_{d} is much less than one. The predictions of the two models are illustrated in Figure 9.17 for a dead time of 6 ns. The observed count rate begins to depart from the true count rate at about 10 Mcps in this case, and in practice, photon-counting lidars are often limited to count rates of a few Mcps. The count rate models are useful for estimating the linear range of a photon counting system, and they may also be used to develop dead time corrections. The model given in Eq. (9.20) may be inverted to yield

$$R_{\text{true}} = \frac{R_{\text{obs}}}{1 - R_{\text{obs}}\tau_{\text{d}}}, \tag{9.21}$$

with the restriction that $R_{\text{obs}} < 1/\tau_{\text{d}}$ so that the denominator does not become zero or a negative value. The accuracy of such a correction depends on the fidelity of the model and on knowledge of τ_{d}, so such corrections must be used with caution, if at all.

Example. EARL simulations like those shown in Figure 2.9 but for a single pulse and a QE of 12% show that the maximum count rate for the specified conditions (clear air at 4-km altitude) would be about 10^7/s. A photon counter considered for EARL had a dead time of 22 ns. Using Eq. (9.19) to find the ratio of observed to true count rate, the argument of the exponential function would be $(-10^7) \times (22 \times 10^{-9}) = 0.22$ and the ratio would be 0.80. Even a micro-pulse lidar may experience tropospheric signals that are too high for photon counting.

There are several engineering approaches to lowering the photon count rate at low altitudes: the crossover range can be increased by adjusting the transmitter–receiver alignment or separation, or by reducing the field of view, and the overall count rate can be reduced by using a smaller laser and/or a smaller receiver. The sky background can only be reduced by using a narrower bandpass filter and/or a narrower field of view.

9.3 Hybrid Systems

One other valuable lidar data acquisition technique is to use a hybrid data system that splits the signal into an analog channel and a photon counting channel, and records both at all times. The choice of the range for switching from analog to photon counting detection is then made post hoc, based on SNR. The data shown in Figure 9.14 were recorded by such a system, the Licel Transient Recorder. The main purpose of the measurements was to record data in the 0.5- to 10-km altitude range using the analog channel, and the photon counting channel provided the stratospheric results as an unplanned bonus.

9.4 Further Reading

The IEEE Standard for Terminology and Test Methods for Analog-to-Digital Converters, IEEE Std 1241–2010.

This standard is an authoritative resource on digitizers.

Analog-Digital Conversion, Analog Devices, Inc. 2004, ISBN 0-916550-27-3.

This invaluable reference is available as a free downloadable nine-chapter book called The Data Conversion Handbook; see [1]. It is a comprehensive resource for information on digitizers.

9.5 Problems

9.5.1 The DIAL technique requires very high SNR. An ozone DIAL developer requires a voltage SNR of 10,000 and plans to use a 10-bit digitizer. If the digitizer is ideal, will it provide sufficient SNR for a single laser shot? If not, how many pulses must be averaged to provide it?

9.5.2 Derive the code probabilities for the code density test, as given by Eq. (9.13).

9.5.3 If the histogram test is employed on a 12-bit digitizer, how many samples are required for an accuracy of 0.1 LSB with a confidence of 99%?

9.5.4 Is Johnson noise a consistent explanation for EARL's digitizer noise measurements?

(a) Calculate EARL's noise floor in dB due to Johnson noise. The bandwidth is 5 MHz. Assume that room temperature is 290 K, and compare the noise power to the maximum signal power, assuming a sine wave input voltage range of ±1 V. Recall that power is V^2 / R; in this case, V is the r.m.s. signal voltage and R is the input impedance of 50 Ω.

(b) What noise floor (in dB) is represented by EARL's measured standard deviation of 0.17 digitizer levels? The digitizer has 16 bits.

9.5.5 A lidar data system starts acquiring data when a laser pulse is transmitted, and it must stop data acquisition before the next pulse is transmitted. The lidar's maximum range is therefore limited by its PRF.

(a) Derive an expression for the maximum lidar range.

(b) The GTRI lidar at the Starfire Optical Range had a PRF of 5 kHz. What was its maximum range? The digitization rate was 10 MHz. What was the maximum number of range bins?

References

[1] The Data Conversion Handbook. [Online]. Available: www.analog.com/en/education/education-library/data-conversion-handbook.html. [Accessed September 15, 2021].

[2] H. Nyquist, "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, vol. 3, April 1924, pp. 324–346. [Online]. Available: https://archive.org/details/bstj3-2-324. [Accessed September 15, 2021].

[3] Tutorials 728: Defining and testing dynamic parameters in high-speed ADCs, Part 1. [Online]. Available: www.maximintegrated.com/en/design/technical-documents/tutorials/7/728.html. [Accessed September 15, 2021].

[4] Linear Technology Design Notes: Understanding the effect of clock jitter on high speed ADCs. Design Note 1013. [Online]. Available: www.analog.com/media/en/reference-design-documentation/design-notes/dn1013f.pdf. [Accessed September 15, 2010].

[5] B. R. Frieden, *Probability, Statistical Optics, and Data Testing*. New York: Springer, 2001.

[6] Effective Bits Testing Evaluates Dynamic Performance of Digitizing Instruments, Tektronix, Inc. 2020. [Online]. Available: www.tek.com/document/application-note/effective-bits-testing-evaluates-dynamic-performance-digitizing-instrument. [Accessed September 15, 2010].

[7] Tutorials 1197: How quantization and thermal noise determine an ADC's effective noise figure. [Online]. Available: www.maximintegrated.com/en/design/technical-documents/tutorials/1/1197.html. [Accessed September 17, 2021].

[8] CALIPSO Spacecraft and Instruments. [Online]. Available: www.nasa.gov/mission_pages/calipso/spacecraft/index.html. [Accessed September 15, 2021].

[9] H. N. Forrister, D. W. Roberts, A. J. Mercer, and G. G. Gimmestad, "Infrared Lidar Measurements of Stratospheric Aerosols," *Applied Optics*, vol. 53, pp. D40–D48, 2014.

[10] D. P. Donovan et al., "Correction for Nonlinear Photon-Counting Effects in Lidar Systems," *Applied Optics*, vol. 32, pp. 6742–6753, 1993.

10 Lidar Data Analysis

Previous chapters covered statistics, the lidar and background equations, atmospheric optics, lidar optical systems, detection of optical signals, and data systems. Several types of noise were covered, as well as best practices for avoiding distortions of the signal due to the optics, aliasing, impedance mismatch, insufficient dynamic range, or electronic artifacts. If all this engineering guidance is implemented, the result will be a digitized signal with known properties, in which both the signal and the noise are understood. The digital lidar waveform shown as the last step in Figures 8.1 and 9.12 can then be faithfully modeled by the lidar and background equations, and reliable data products with good error estimates can be derived from the lidar data.

The lidar and background equations are known as *forward* models, meaning that they predict the lidar waveform using atmospheric and instrumental parameters as inputs. They are essential for designing a lidar, but the goal of lidar is to measure atmospheric parameters. This chapter is about analysis techniques called *retrievals* that are used to extract reduced atmospheric data products from lidar data. The raw digital lidar signal is uncalibrated, it usually includes an offset voltage, and it may include two or more receiver or data system channels that cover different parts of the overall dynamic range. Lidar data are often acquired in range bins smaller than the range resolution of the final data product to enable the use of digital signal processing (DSP) techniques, which were developed in tandem with the digitizer and computer hardware described at the beginning of Chapter 9. For all these reasons, the data usually require preprocessing, which is the first topic in this chapter. The second topic is analysis techniques, beginning with simple calculations made directly from the preprocessed data and progressing to actual *inversions* of the lidar equation, which enable backward calculations going from the data to the atmospheric parameters. The analyses in this chapter are for retrieving profiles of particulate β, σ, and δ from elastic backscatter lidar data. Some comments on the analysis procedures for several other types of lidar mentioned in Chapter 3 are in Chapter 11.

10.1 Preprocessing

Lidar data processing should start with a basic data quality assessment: Does the waveform have the expected level of noise? Is it saturated high or low? Does it have unexpected artifacts that cannot be removed? Is the bin number of zero range accurately

known? Does the range of full overlap appear to be as expected? The validity of the assumptions behind the intended data analysis method should also be checked at this time, if possible. If data from a photon-counting lidar requires a count rate correction, it is applied at this point, and then the data can be filtered, the background can be subtracted, and profiles from multiple receivers can be merged. The data must also be range corrected for many types of analysis.

10.1.1 Filtering

Some multi-pulse averaging has probably already been done in the data acquisition process, but further averaging of multiple records is often needed to improve SNR to the level required. After averaging, the noise level may be further decreased by *filtering* (also called smoothing), which basically averages data from multiple adjacent range bins to improve SNR at the expense of a larger range resolution. Filtering is a DSP technique enabled by acquiring data in smaller range bins than the desired resolution of the final data products, which is a common practice. For example, a GTRI tropospheric ozone lidar developed around the year 2000 acquired data in 15-m range bins but the vertical resolution of the ozone profiles was 200 m or greater [1]. The information in lidar signals is encoded in the time domain rather than the frequency domain, and the shape of a time domain signal basically shows when something occurred and what its amplitude was. Accurate time domain measurements require minimal signal shape distortion, so smoothing techniques to decrease noise must not cause time-shifting, and ideally, they should reduce noise while not smoothing out real atmospheric structure, although these latter two goals are somewhat in conflict and a compromise must be reached. Digital signal processing techniques are also known as *discrete-time signal processing* because they were developed to process digital signals recorded at discrete time intervals, so they are perfectly suited for processing digitized lidar data [2]. One appropriate and easy-to-use DSP tool for smoothing lidar data is known as the finite impulse response (FIR) filter, which restricts the spectral range of a signal's frequency by convolving the data with a sliding *window*, thereby decreasing the passband in the frequency domain by filtering in the time domain. The somewhat arcane use of the word *finite* in FIR means that the filter's response to an input value is of finite duration, in terms of range bins, as opposed to an *infinite* impulse response filter that has internal feedback and may respond indefinitely. In lidar, it is essential to avoid phase changes that would distort the signal, but this is easy to do by simply making the FIR symmetrical in shape about its center.

The FIR filters described here generate new arrays of data points from raw data arrays by finding a weighted average of the raw data points falling within the window and defining the point at the window center in the new array as the weighted average. The window is then moved one range bin and the process is repeated, until all data points in the digitized signal array have been used. The filters have an odd number of coefficients and are symmetric about the center, so they do not alter the phase of a signal. Three FIR filters commonly used to smooth lidar data are described here using the notation of [3]. If S_k is the kth raw data point, the filtered data is calculated as a moving normalized convolution integral expressed in discrete mathematics as

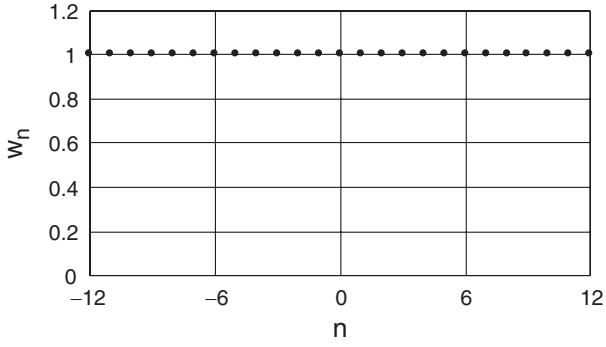


Figure 10.1 The rectangular FIR filter. The filter shown has 25 equal weights.

$$S'_k = \sum_{n=-N}^{N} c_n S_{k+n}, \tag{10.1}$$

where S'_k is the value of the filtered data at range bin k and the filter width is $2N+1$ range bins. The filter function is defined by the coefficients c_n where $-N \le n \le N$. By convention, filters are described by a set of weights w_n with a maximum value of unity, and the coefficients c_n are calculated as w_n / C, where the normalizing factor C is defined as

$$C = \sum_{n=-N}^{N} w_n. \tag{10.2}$$

The simplest such filter is rectangular (also called the boxcar filter or moving average filter), illustrated in Figure 10.1, where N is 12 and so there are 25 nonzero w_k values, all of them unity. In this case, C is 25 and all c_n values are 1/25. It is important to start smoothing at least N bins beyond the crossover range to avoid distorting the filtered data with pre-crossover data.

FIR filters are characterized by their responses to specific input functions as well as by a frequency response. The input functions are an impulse and a step. The impulse is the discrete mathematics version of the Dirac delta function; it is unity at one bin and zero everywhere else. For the symmetric FIR filters discussed here, convolving the filter with the impulse function simply reproduces the filter shape, so the filter function is the impulse response. The step function is shown in Figure 10.2. Filter functions, step functions, and responses are most often shown as continuous curves in the FIR filter literature, but they are of course discrete, so they are plotted here as discrete points.

A filter's response to a step function provides insight into how much the filter smooths out atmospheric structure. The 25-point rectangular filter's response to a step function is shown in Figure 10.3. The response is a linear ramp, as would be expected.

The rectangular filter illustrated in Figure 10.1 and defined by Eqs. (10.1) and (10.2) with an odd number of unity w_n values is very commonly used because it is

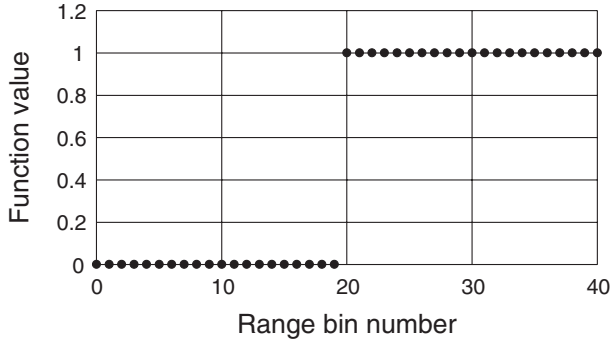


Figure 10.2 The step function.

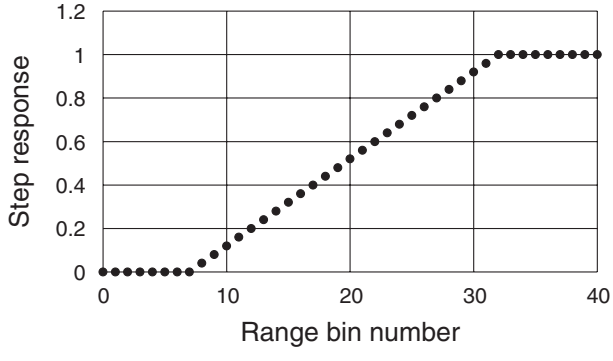


Figure 10.3 Rectangular filter step function response.

so easy to implement. If the goal of smoothing is to improve a visual presentation of lidar data, the rectangular filter may be adequate, but other FIR filters are better for preserving structure while suppressing noise. Two other common FIR filters using sliding weighted averages are the Hann window and the Hamming window [2]. The Hann window is a raised cosine defined by the weights

$$w_n = 0.5\left[1+\cos\left(\frac{\pi n}{N}\right)\right], \tag{10.3}$$

for $-N \le n \le N$. The Hann window shape is shown in Figure 10.4 for $N = 12$ and its step function response is plotted in Figure 10.5.

The Hamming window is a modified raised cosine window, defined by

$$w_k = 0.54 - 0.46\cos\left(\frac{2\pi k}{N-1}\right) \tag{10.4}$$

when N is greater than 6. The constants in Eq. (10.4) change slightly for smaller values of N, but their sum remains unity. The Hamming window function is plotted for $N = 12$ in Figure 10.6, and its step function response is plotted in Figure 10.7.

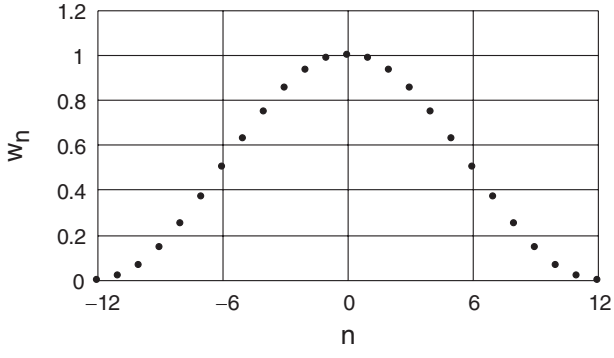


Figure 10.4 The Hann window.

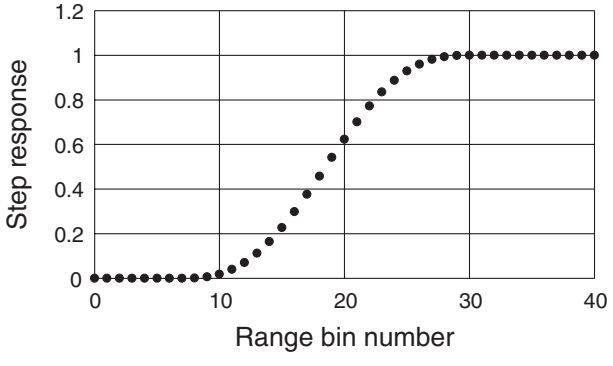


Figure 10.5 The Hann window step function response.

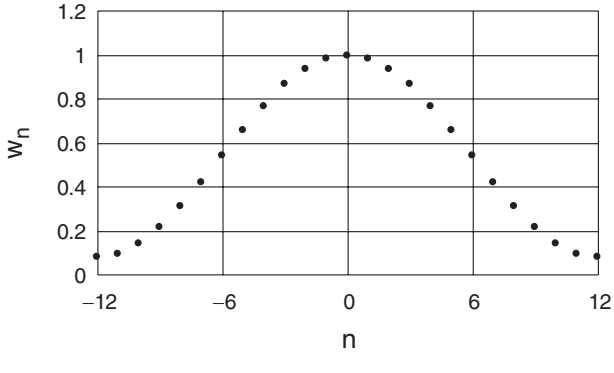


Figure 10.6 The Hamming window.

The factor by which a filter suppresses the variance in a signal is the sum of its squared coefficients, as shown by the relation [3]

$$\sigma_{S'}^2 = \sigma_S^2 \sum_{n=-N}^{N} c_n^2. \tag{10.5}$$

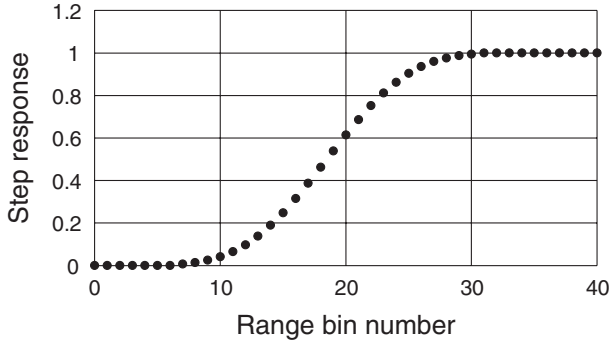


Figure 10.7 The Hamming window step function response.

Filtering suppresses all signal variance whether it is due to noise or real atmospheric structure. Suppression of real structure causes a larger range resolution, which leads to an obvious question: What is the resolution of the filtered signal, for any given filter? Historically, lidar stations have used a wide variety of data processing algorithms with various types and levels of filtering, usually implemented in software developed in-house, and researchers have used several different definitions of resolution. These facts have caused the comparison of profiles from different stations to be problematic, which is an especially important issue in networks, where such comparisons are an essential aspect of their operations. The vertical resolution of lidar data products was addressed by two networks, the NDACC and EARLINET, starting around 2010 [3, 4]. The NDACC team developed recommendations for two alternative ways to determine vertical resolution for two basic types of filter, the moving average type described here as well as the derivative type. The vertical resolution of a filter is often calculated in units of range bins, in which case the resolution in meters is the number of bins times the width of one bin. The first NDACC moving average definition is that the vertical resolution (in bins) is simply the FWHM of the filter function. As can be seen by inspecting Figures 10.4 and 10.6, the resolution of the centrally peaked filters is about one-half of their full width, whereas for the rectangular filter it is the full width, which is in fact taken as its resolution by some authors [4], although the FHWM definition only pertains to functions with central peaks. The second definition is based on the spatial frequency response of the filters, in terms of cycles per range bin. The gain G of a filter (with a range 0–1) can be shown to be [3]

$$G(f) = c_0 + 2\sum_{n=1}^{N} c_n \cos(2\pi nf), 0 \le f \le 0.5, \tag{10.6}$$

where the frequency f cannot exceed the Nyquist frequency of 0.5 cycles per bin. The second NDACC resolution definition is based on the cutoff frequency f_c at which the gain is equal to 0.5, and the resolution (in bins) is defined as $1/2f_c$. The filter gain curve for the 25-point rectangular filter for frequencies from 0 to 0.05 bin^{-1} is shown in Figure 10.8. If the sampling interval is δz, then the final resolution Δz (m) is given by the relation

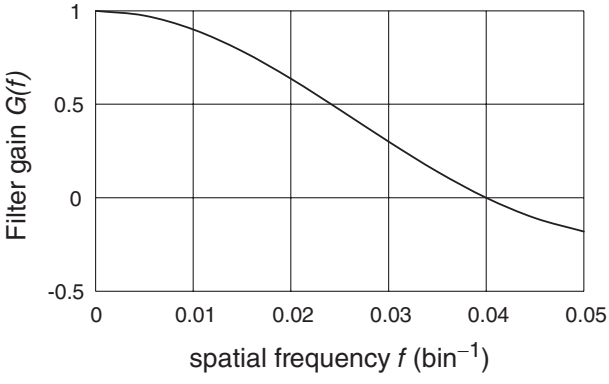


Figure 10.8 Filter gain versus spatial frequency. The plot is for a 25-point rectangular filter. The cutoff frequency, where the gain is 0.5, is 0.023 bin^{-1}.

$$\Delta z = \delta z / 2 f_c. \tag{10.7}$$

Example. The filter gain curve shown in Figure 10.8 for the 25-point rectangular filter shows that the gain is equal to 0.5 at a frequency of 0.023 bin^{-1}, so the vertical resolution is 1 / [2 × (0.023)] = 21.7 bins, which is nearly the filter's full width.

Another convenient definition of vertical resolution used mostly in Germany is based on the risetime of step function responses such as those shown in Figures 10.3, 10.5, and 10.7. Vertical resolution is defined by the number of bins required for the response to rise from 25% to 75% of the full response [3]. This definition leads to smaller numbers of bins than the other two described here.

Because lidar signals have such large dynamic ranges, it often does not make sense to use one filter width for the entire signal. For this reason, adaptive filtering is sometimes employed, in which the filter width is adjusted based on estimated raw data SNR to decrease the bandwidth with range, maintaining a reasonable SNR in the entire filtered profile. The disadvantage of this technique is that the range resolution is not constant within one profile or from one measurement to another, but the definitions of resolution summarized above can always be used to calculate it.

The properties of the three filters discussed above are summarized in Table 10.1, for both definitions of resolution. The Hann and Hamming filters are 13-point, but the rectangular filter is 9-point, so that the variance suppression factors and the vertical resolutions (by the cutoff frequency definition) are all similar. Examples of the effects of these three FIR filters on lidar data are shown in Figure 10.9. The data, which are part of the profile shown in Figure 9.14 but with less smoothing, were recorded at night on August 4, 2011 in Atlanta, Georgia, with a 1574-nm photon-counting system sampling at 40 MHz, corresponding to 3.75-m range bins [5]. The scatter in the data is signal noise plus dark count noise. The profiles smoothed

Table 10.1 Properties of three common 13-point FIR filters

Filter type	Variance factor	r.m.s factor	Vertical resolution (bins)	
			FWHM	$1/2f_c$
Rectangular	0.111	0.333	9	8.5
Hann	0.125	0.354	6.0	7.4
Hamming	0.111	0.333	6.5	8.2

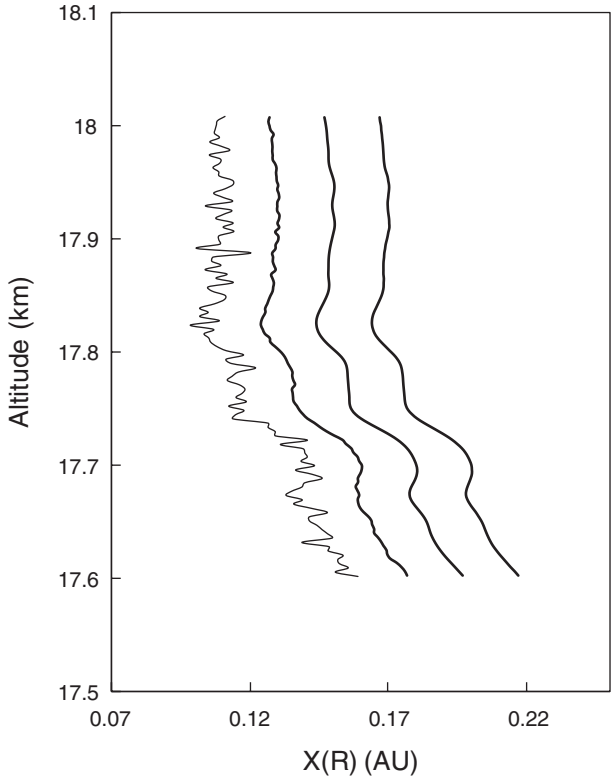


Figure 10.9 Examples of filtering. The lidar data were recorded in the stratospheric aerosol layer in 3.75-m range bins. The curve on the left is the raw data. The next one to the right was filtered with a 9-point rectangular filter, followed by the 13-point Hann and Hamming filters, each of which is offset by 0.02 units for clarity. Using the cutoff frequency definition, the filtered altitude resolutions are 32, 28, and 31 m, respectively.

with Hann and Hamming filters are almost indistinguishable because their vertical resolutions are similar, but they are better than the rectangular in two ways: They suppress high-frequency noise better, and they preserve structure better. They are not the only choices; dozens of other symmetric FIR filters have been developed and characterized [6].

10.1.2 Background Subtraction and Range Correction

The filtered raw lidar data will generally have a positive offset μ that is caused by some combination of electronic biasing, dark counts, and background light. Accurate estimation and removal of the offset, which is simply called the background in practice, is a critical first step in data analysis that enables everything that follows. Range correction is especially sensitive to background removal. Lidars usually have at least 2,000 range bins, so if the first bin is multiplied by unity, the last bin is multiplied by four million, and any background error will cause the signal to increase or decrease parabolically at long ranges.

The background must be estimated for each averaged signal profile because it may be constantly changing. The background μ is the mean value of the signal in some regions where the backscattered signal is zero. One common estimation method is to record the signal in more range bins than needed for the measurement, and then assume that the lidar signal is zero in the bins at the far end of the range. The average value in a set of those far-end bins is then taken as the background. Due to the presence of noise, the average is a statistical estimate with an uncertainty ε that depends on the standard deviation of the noise and the number of bins averaged. As shown in Chapter 2, the standard deviation of the mean is given by

$$\varepsilon = \sigma / \sqrt{n}, \tag{10.8}$$

where σ is the standard deviation of the noise and n is the number of bins in the average. However, assuming the noise is Gaussian, the measurement will only be within that limit 68% of the time, as shown in Figure 2.2, and that level of accuracy is probably not sufficient in lidar data analysis. This problem is addressed with the confidence intervals z^* also defined in Chapter 2. Adding this factor and solving for n yields the equation

$$n = \left(\frac{z^* \sigma}{\varepsilon} \right)^2. \tag{10.9}$$

Example. If the standard deviation of the data in the last bins is 0.01 and background subtraction requires an error no larger than 0.001, using $z^* = 1.96$ for 95% confidence shows that $n = [(1.96) \times (0.01) / (0.001)]^2 = 384$ bins

After the background μ is accurately estimated, it is subtracted from the data in every range bin. The data is then referred to as *background subtracted*. Depolarization, DIAL, and trace gas Raman lidar analysis algorithms use ratios of signals, in which case no further preprocessing may be required. If the transmitter has a power monitor, the signal profiles can be normalized to the same power level. The simplest method is to divide the data in each profile by the corresponding power meter reading.

For many analysis algorithms, range correction is the next step. Range correction is implemented by multiplying the value in each bin by something proportional to range squared, such as the bin number squared. It is important to know where the first

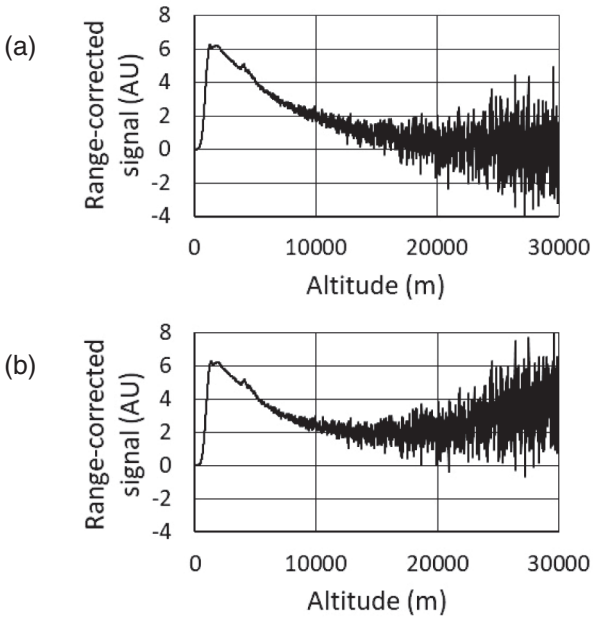


Figure 10.10 Effect of background error. (a) The background-subtracted, range-corrected signal $X(R)$ approaches zero at long ranges when the background value μ is accurately estimated and removed. (b) $X(R)$ grows quadratically at long ranges when the background value μ is underestimated.

range bin is, because otherwise the correction will not be exactly proportional to range squared, and the signal will be distorted. Quite often there is enough scattered laser light in the instrument or the laboratory to cause a spike in the digitized data when the laser pulse is transmitted, which defines zero range. If not, it is typically easy to scatter some light for this purpose. The background-subtracted, range-corrected lidar signal is often called $X(R)$ in the lidar literature, and that convention is followed here. An example of $X(R)$ for daytime EARL data is shown in Figure 10.10(a), which shows that the noise scales up with range squared along with the signal. The signal cannot be negative, but the noise certainly can be. If the signal is not distorted and the offset μ was accurately removed, the noise will have equal excursions above and below the horizontal axis. An example in which the μ estimate was artificially made too small is shown in Figure 10.10(b) for the same data, where the residual background in the data causes a quadratic growth in the range-corrected signal and consequently the mean value does not approach zero at long ranges. $X(R)$ will grow quadratically negative if the μ estimate is too large.

A background estimation error occurred in the analysis of EARL data used in [5]. The EARL data analysis software calculated the background using a few hundred range bins of data at the far end, around 30-km altitude, based on the assumption that there would be negligible backscattered signal from such a high altitude in a

micro-pulse lidar. However, when the data were averaged over 1-h periods at night, the real signal was not negligible, and a background estimation error consequently caused an error in the range-corrected data. A special-purpose algorithm was developed of necessity to correct the background error, but the problem could have been avoided with a better method of finding the background: First, many digitizers can record pre-trigger range bins, before the laser pulse is transmitted. Those bins can be used to find a reliable background, provided they are not contaminated by electromagnetic interference from the laser as it prepares to fire. Better yet, some lidar researchers trigger the data system twice for each laser pulse so that alternate profiles are nothing but background. With this method, the number of range bins can be much lower because it is not necessary to record the signal to ranges where it essentially vanishes.

10.1.3 Merging Profiles

Sometimes range profiles covering different parts of the total range must be merged to provide one continuous profile, in a process also called "stitching" and "gluing" in the lidar literature. This issue arises when a lidar has more than one receiver channel, as does EARL, and when using a hybrid data system that generates both analog and photon counting data sets. Lidar researchers have developed sophisticated data analysis tools for merging, but for clarity, only the simplest method is shown here, using the EARL data shown in Figure 10.11, which has been background-subtracted and range-corrected. Both signals are in arbitrary units and they are not on the same scale, due to differing receiver optical efficiencies and differences in detector and electronic gains.

The first step in merging is to put the signals on the same scale, which requires finding a region on the range axis where complete crossover has been reached for both receivers and both channels have reasonably good SNR. This can be accomplished by examining the *channel ratio* plot shown in Figure 10.12. EARL was designed for a short-range crossover of 550 m, so the ratio is not plotted below that altitude. The curve appears to become flat at about 2,500 m, so that value was taken as the starting point, and 4,000 m was taken somewhat arbitrarily as the end point. The accuracy of the scaling depends on the noise level and the numbers of points, like the background accuracy described by Eq. (10.6). The scale factor ρ is the average of the channel ratio in the selected region:

$$\rho = \frac{1}{N} \sum_{k=k_0}^{N+k_0-1} \frac{\left(S_{\text{long}}\right)_k}{\left(S_{\text{short}}\right)_k}, \tag{10.10}$$

where S is a signal array, k is the data array index, and N is the total number of range bins in the calculations (100 in this example). The next step is to scale the short-range profile by using the relation

$$S'_{\text{short}} = \frac{S_{\text{short}}}{\rho}, \tag{10.11}$$

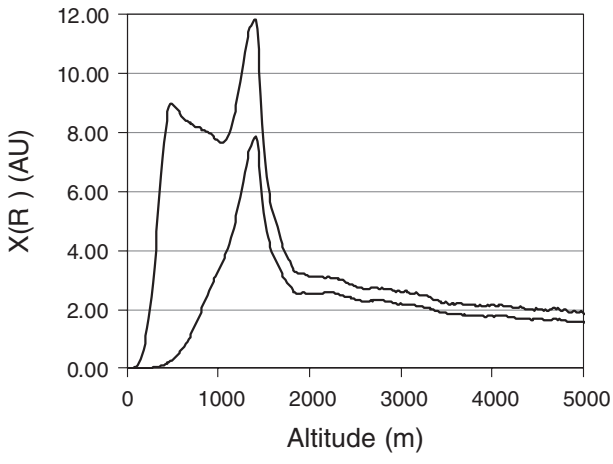


Figure 10.11 Range-corrected signals. Upper line – short-range receiver; lower line – long-range receiver.

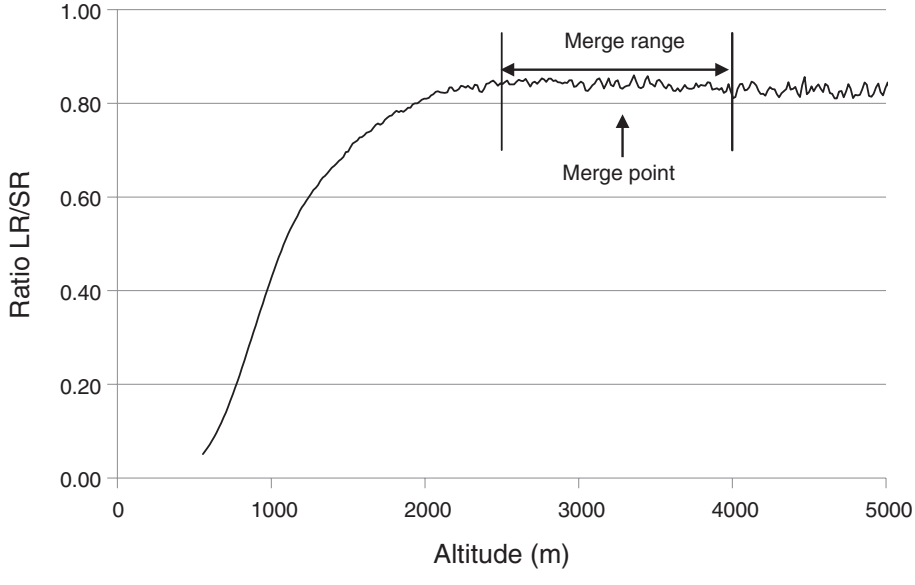


Figure 10.12 Receiver channel ratio. The curve is the ratio of the long-range signal to the short-range signal.

where S'_{short} is the scaled short-range data. The final step is to choose a range, or point, at which the merge occurs. In the example shown in Figure 10.12, it is 3250 m, in the middle of the scaling range. Below that point, the merged signal is S'_{short}; at and above that range, it is S_{long}. The merged profile is shown in Figure 10.13. Merging at one point risks having a discontinuity due to a noise spike, so it is preferable to have

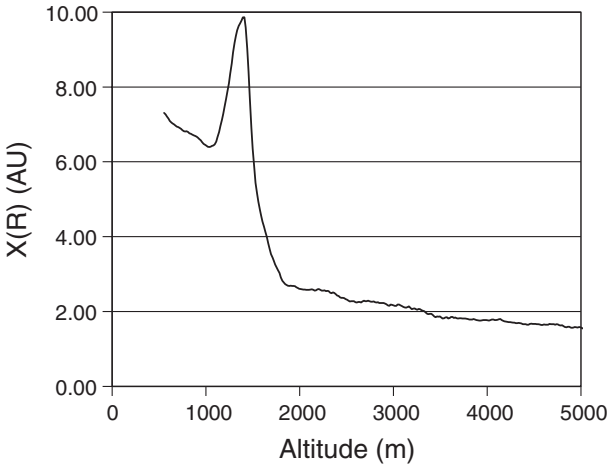


Figure 10.13 The merged signal. The curve shows the signal formed by merging the short-range and long-range signals shown in Figure 10.11.

a smooth transition from one profile to the other, and more sophisticated algorithms do use smooth transitions.

As mentioned above, the merging algorithms shown here are the simplest possible, but they illustrate the basic idea and some considerations. If the data profiles are to be merged, the quality of the short-range data is critical: it is used in the scaling factor calculation and it is part of the merged signal, so any distortions or noise in it will appear in the merged signal. The data in Figure 10.11 appear adequate for merging, but they were acquired during nighttime. An example of lower-quality short-range data is shown in Figure 10.14, which is daytime EARL data otherwise acquired with parameters identical to those of Figure 10.11. The short-range signal is quite noisy compared to the long-range signal at almost all ranges. In the region used for finding the scale factor ρ this may not be fatal because of averaging, but the short-range data is noisy at any reasonable merging range, so noise will appear in the merged data. This problem could be alleviated by using a narrower bandpass filter on the short-range receiver channel, or by diverting a greater fraction of the received signal to it. It could have been avoided by doing more thorough simulations of EARL's SNR in all intended conditions during the design phase.

After range correction and merging, the data should be plotted as a time–height diagram, as in Figure 4.6 for EARL data and the top panel of Figure 4.17 for CALIOP data. This type of plot is an excellent check on data quality because problems are revealed as discontinuities. A merging problem will result in a horizontal line or stripe, while a problem in power normalization will cause vertical bands. If the times of observation span at least one diurnal cycle, the plot will also show the differences in SNR between daytime and nighttime. Time–height plots of $X(R)$ are often used to visualize the development of the mixed layer and to estimate its thickness, as shown

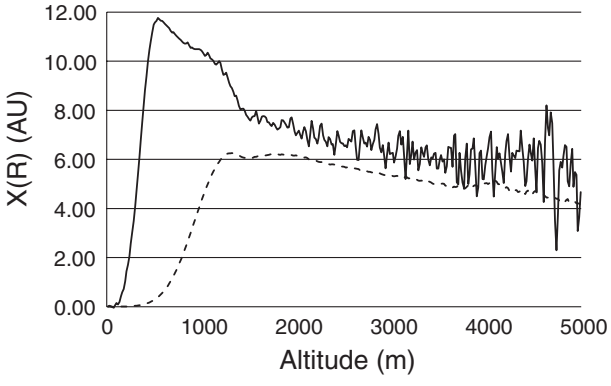


Figure 10.14 Daytime EARL data. Solid line – short-range receiver; dashed line – long-range receiver.

by the white line in Figure 4.6. Mixed layer thickness is one of the three major inputs to air quality models (the others are pollutant emission rate and wind speed), and it is typically measured only twice daily by balloon-borne sondes that may not be co-located with the polluted area. For this reason, the lidar community has made a significant effort over many years to automate the retrieval of mixed layer thickness from ground-based lidar stations, and there is a corresponding body of literature on this topic. Direct detection lidars use aerosols as tracers for thermal mixing, and several different statistical approaches have been evaluated, using the $X(R)$ profile, its gradient, and its variance [7]. Coherent lidars directly measure turbulence in the mixing airflow, which is at a maximum at the top of the layer.

10.2 Cloud and Aerosol Lidars

The lidars described in this section are the simplest type, elastic backscatter, and many use a single wavelength. The goals of cloud and aerosol lidars are to retrieve geometrical parameters, such as layer base height and thickness, and the particulate optical parameters β, σ, and δ in the layers. The lidar equation only produces one signal from those atmospheric parameters, and they have several components: β is the sum of the backscatter coefficients of molecules, aerosols, and cloud particles, and σ is the sum of the extinction coefficients of molecules, aerosols, cloud particles and trace gases. The atmospheric δ has contributions from both molecules and particles. Because there is only one signal but several unknowns, all methods of analyzing lidar data rely on combinations of simplifying assumptions and auxiliary data. The Rayleigh scattering terms can be calculated from the atmospheric density for any lidar wavelength using the methods in Chapter 3, and the atmospheric density profile is often a crucial piece of auxiliary data during lidar data analysis. One approach to resolving the ambiguity in the lidar signal is to measure two or more of something, such as polarization or pointing angle, to provide additional data and enable

a straightforward analysis, and some of the simplest algorithms are based on this approach. Another analysis method is to mathematically invert the lidar equation to create algorithms for calculating a profile, such as $\sigma(h)$, from the signal; inversions are covered in Section 10.3. The analysis algorithms used to extract atmospheric information from lidar signals must be written in discrete mathematics because the signals are digital, but as is usual in lidar, the methods are developed using continuous variables for simplicity and clarity.

10.2.1 The Scattering Ratio

Figures 10.15–10.17 illustrate a simple analysis of elastic backscatter lidar data that were acquired at night with a GTRI lidar in Albuquerque, New Mexico, on August 18, 1992. The lidar's wavelength was 511 nm, and the data were recorded with a 10 MHz, 12-bit digitizer, as 3,000-pulse averages. The site elevation was about 2,000 m. The lidar had a single receiver and a fairly long crossover range, because its only purpose was to detect cirrus at night, which is an easy task for a visible-light lidar. The raw data are plotted in Figure 10.15, which shows an electronic offset, a typical lidar signal shape, and a second small peak near 10 km.

Figure 10.16 shows the same data after background subtraction and range correction. The shape of the profile reveals a Rayleigh signal decreasing exponentially with altitude plus three kinds of scattering by particles: mixed layer aerosols below 5 km, a cirrus cloud at 10 km, and a stratospheric aerosol layer from 15 to 22 km. The stratospheric aerosol layer is prominent because of the 1991 Mt. Pinatubo eruption shown in Figure 1.7.

The next step in the analysis is to highlight the particulate signals by dividing out the molecular signal. The result is known as the attenuated scattering ratio R_{sca}, which is the total signal divided by the Rayleigh (molecular) signal, as defined in Eq. (4.27). The Rayleigh backscatter and extinction are always present and always smoothly decreasing with altitude, and the extinction can often be neglected for visible light. The scattering ratio procedure is to first find the Raleigh backscatter profile. It can be calculated from upper air data or the GMAO model (see Chapter 3), or simply approximated as an exponential decrease with an appropriate scale height, as it was for this data, using a scale height of 8,000 m. Then the profile is normalized to the range-corrected data in some chosen altitude range where the aerosol signal is believed to be negligible (8–9 km in this example) and $X(R)$ is divided by the normalized Rayleigh profile. The result of following this procedure is shown in Figure 10.17, which clearly shows the three regions of particulate backscatter. The signal is too noisy for analysis above approximately 23 km, and it also drops below unity, probably because of an inaccurate background estimate. The attenuated scattering ratio shows the boundary layer aerosol (1.5 × Rayleigh), the cirrus cloud (9 × Rayleigh), and the stratospheric aerosol (3 × Rayleigh). The signal is, of course, attenuated by the particulate matter it passes through, and by Rayleigh scattering.

Because the Rayleigh backscatter coefficient profile is known, the scattering ratio technique calibrates the particulate backscatter. Its use in finding layer optical

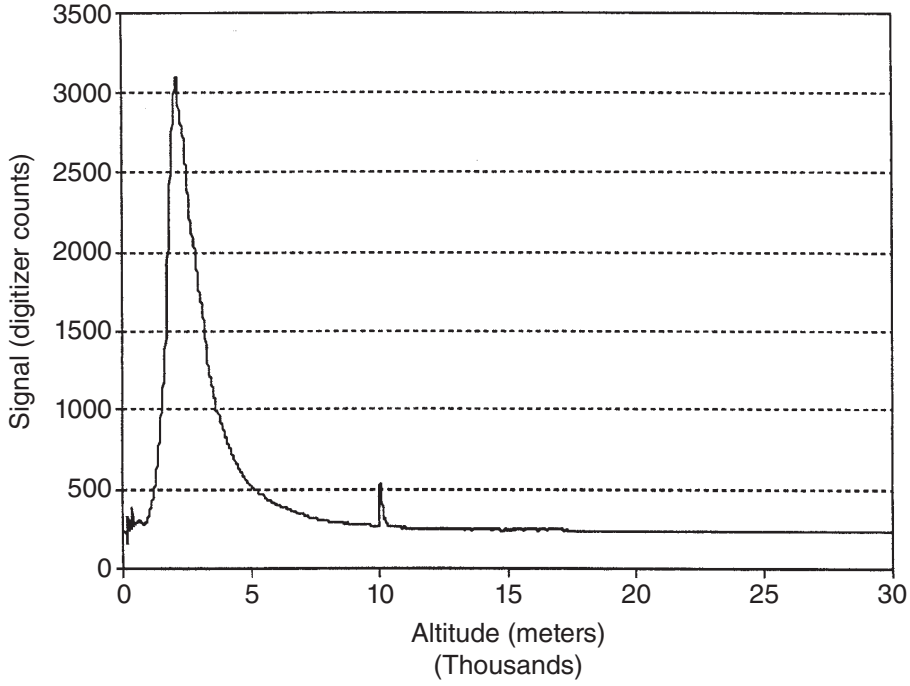


Figure 10.15 Raw data. The lidar signal is a 3,000-pulse average from a 511-nm lidar in New Mexico in 1992.

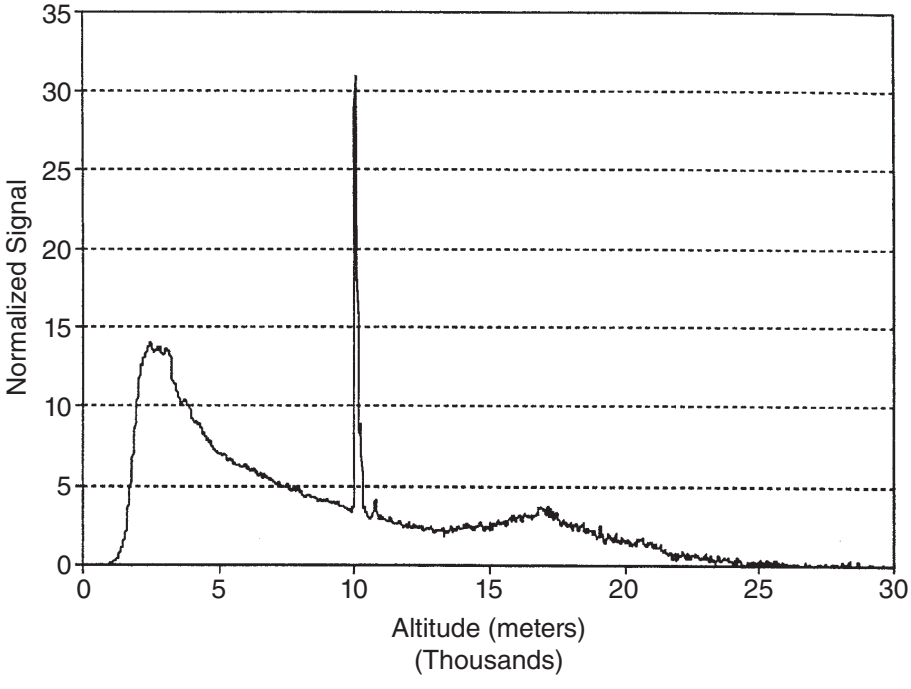


Figure 10.16 Range-corrected data. The profile shows Rayleigh scattering plus mixed layer aerosols, a cirrus cloud, and stratospheric aerosols.

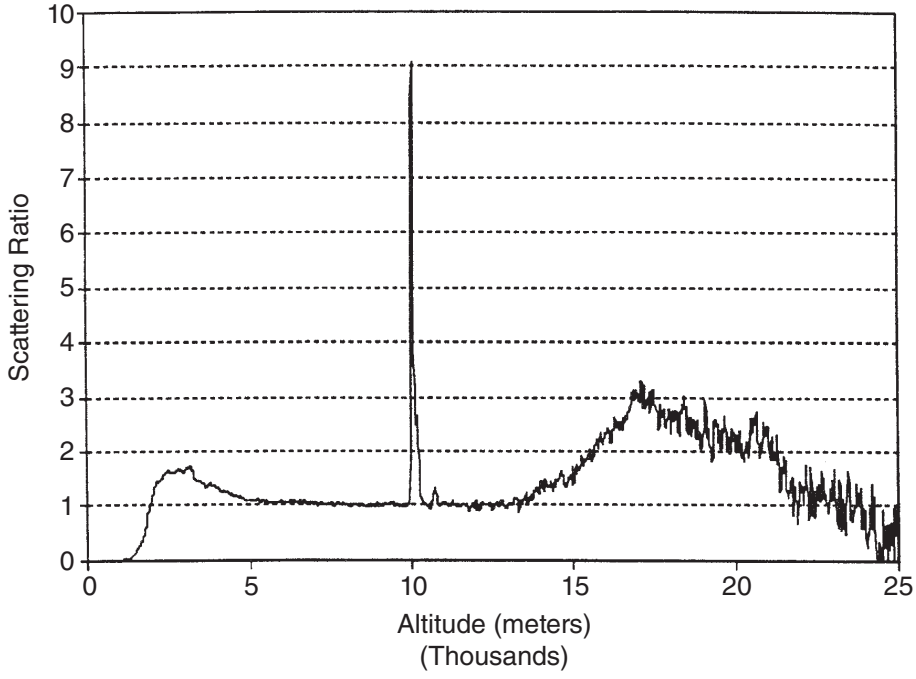


Figure 10.17 The attenuated scattering ratio. After dividing out the Rayleigh component and scaling to $X(R)$ at 8–9 km, the backscatter from mixed layer aerosols, a cirrus cloud, and the stratospheric aerosols appear as multipliers of the Rayleigh component.

properties is illustrated in Figure 10.18, which is part of a scattering ratio plot for a different cirrus layer recorded on the same night as the previous three figures. For this layer, the cloud base height was 7.7 km, its thickness was 300 m, and its peak scattering ratio was approximately 30. The Rayleigh signal from altitudes above the cloud is about 0.83. The scattering ratio method enables estimates of both geometrical and optical properties of the cloud layer.

Example. The 511-nm extinction coefficient, backscatter coefficient, and lidar ratio of the cirrus layer shown in Figure 10.18 can be estimated with simple calculations. The Rayleigh signal of about 0.83 from higher altitudes implies that the 300-m layer's two-way transmittance was 0.83, so the one-way transmittance T is $(0.83)^{1/2} = 0.91$. From Beer's law, the cloud's OD is $-\ln(0.83) = 0.094$, and so its extinction coefficient σ is $0.094/300 = 3.1 \times 10^{-4}$ m^{-1}. The backscatter coefficient is found from Eq. (3.11): $\beta = 30 \times 1.39 \times (550/511)^4 \times \exp[-(9700 / 8000)] \times 10^{-6} = 1.7 \times 10^{-5}$ m^{-1}sr^{-1}, where the atmospheric density was approximated by a scale height relation. The lidar ratio is $(3.1 \times 10^{-4}\ \text{m}^{-1}) / (1.7 \times 10^{-5}) = 18$. These results are consistent with other published values for cirrus. They are somewhat approximate, but this method of finding OD is frequently used in lidar to identify and characterize layers.

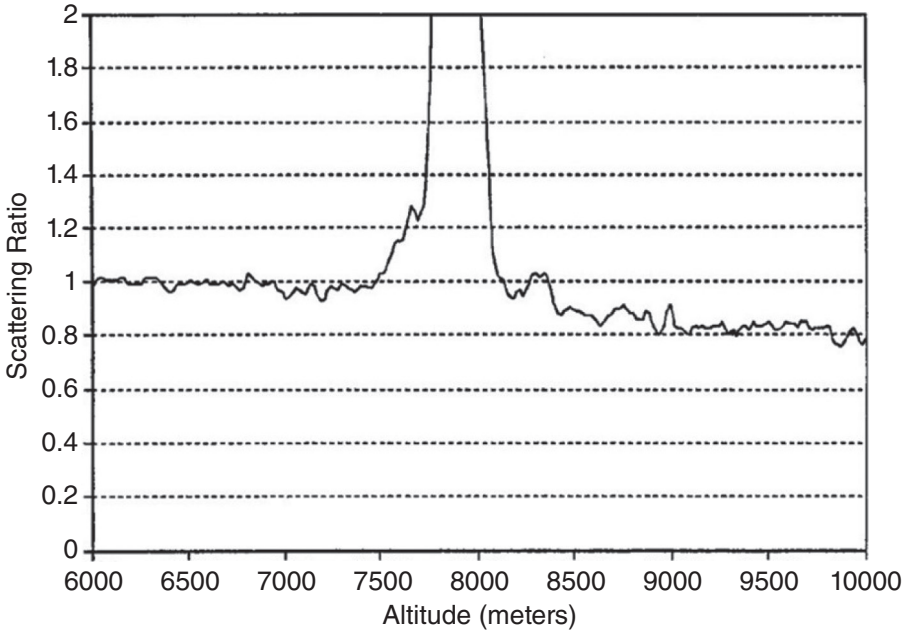


Figure 10.18 Cirrus transmittance.

10.2.2 Depolarization

Another data product that does not require inversion or even range correction is the depolarization ratio δ; as defined in Eq. (4.24) it is simply the cross-polar lidar signal divided by the co-polar signal. Depolarization is often calculated this way and presented in time–height diagrams. Technically, this is a bad practice because it is a measure of atmospheric depolarization rather than aerosol depolarization, so it mixes intrinsic and extrinsic properties. However, if the aerosol backscatter signal overwhelms the molecular signal, as it usually does at NIR and longer wavelengths, it may still be useful, and the atmospheric depolarization ratio is often part of the information used in classifying aerosol layers. If the two receiver channels have the same crossover function $G(R)$, as they do in EARL, depolarization can be calculated at all ranges that have sufficient SNR, including ranges below complete crossover.

The aerosol depolarization ratio can be factored out of the atmospheric depolarization ratio by using the scattering ratios described above, for both polarizations [8]. The measured atmospheric depolarization ratio δ_a is defined as

$$\delta_a = \frac{S_\perp}{S_\parallel}, \tag{10.12}$$

where the S values are the total lidar signals. Defining the two scattering ratios as

$$R_{\perp} = \frac{S_{a,\perp} + S_{m,\perp}}{S_{m,\perp}} \text{ and} \tag{10.13}$$

$$R_{\parallel} = \frac{S_{a,\parallel} + S_{m,\parallel}}{S_{m,\parallel}}, \tag{10.14}$$

the aerosol depolarization ratio is found from

$$\delta_a = \delta_{\text{atm}} \left[\frac{R_{\parallel}(R_{\perp} - 1)}{R_{\perp}(R_{\parallel} - 1)} \right]. \tag{10.15}$$

The proof that Eq. (10.15) yields the aerosol depolarization ratio is left as a problem for the student. See [9] for guidance on properly implementing a depolarization capability in a lidar so that accurate error estimates for δ values can be obtained.

10.2.3 The Slope Method

The simplest, and perhaps earliest, lidar method of finding the extinction coefficient σ is known as the *slope method*, which relies on the simplifying assumption that the atmosphere is homogeneous in some range interval beyond complete crossover, so that both β and σ have no range dependence. Lumping together the instrumental constants in Eq. (2.12), the lidar equation can be written as

$$P(R)R^2 = X(R) = C\beta(R)\exp[-2\int_0^R \sigma(r)dr], \tag{10.16}$$

and when β and σ are constants, the derivative of the logarithm of the range-corrected signal with respect to range is

$$\frac{d}{dR}\ln[X(R)] = -2\sigma. \tag{10.17}$$

In the simulated graphical analysis shown in Figure 10.19, the slope of a straight line fitted to $\ln[X(R)]$ vs. R is -2σ. In the simulation, the OD from 1 to 5 km is 1 and Gaussian random noise was added to the signal before it was range corrected. The natural logarithm was then calculated and plotted versus range. The slope method is conceptually simple, but it has limited utility. Practical experience by the lidar community has shown that the optical depth of the measurement region should be about unity for an accurate measurement, so that $X(R)$ decreases by a factor of $1/e^2$, but this is not always the case. The larger problem is that a linear slope in the $\ln[X(R)]$ vs. R plot does not imply that β and σ are constant; the aerosol number density may be varying along the optical path along with β and σ. The reason for mentioning the method here is to illustrate the sequence of operations: In the lidar equation, σ is the argument of an integral which is the argument of an exponential function, so the way to expose it is to find the derivative of the natural logarithm of $X(R)$. For that reason, this same sequence of mathematical operations appears in some of the algorithms in Chapter 11. For a comprehensive discussion of the slope method and its limitations, see [10].

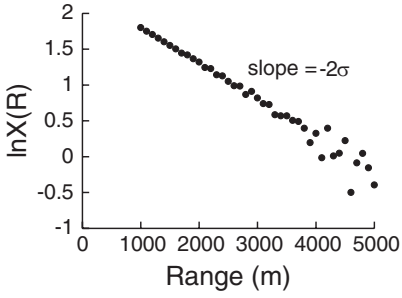


Figure 10.19 The slope method. In a uniform atmosphere, the slope of the data points is –2 times the extinction coefficient.

10.2.4 Multi-angle Lidar

The multi-angle lidar technique is like the slope method in some ways, and it does not require a mathematical inversion of the lidar equation. The method requires measurements at two or more zenith angles, as illustrated in Figure 10.20, and the most basic analysis is shown as a semi-log plot in Figure 10.21. This analysis is analogous to the Langley plot shown in Figure 4.20, so it is referred to here as the *lidar Langley plot*. The simplifying assumption behind this method is that the atmosphere is stratified (horizontally homogeneous). If this is true, the slope of a straight line fitted to the data points is the optical depth from the ground to the height h_0. The lidar system must have reached full crossover at a range less than h_0 for this analysis to be valid, but the measured optical depth includes the crossover region. One advantage of the multi-angle lidar technique is that it requires no auxiliary data; like the slope method it characterizes the total extinction from all causes, and it relies only on the measured lidar profiles.

The lidar known as ALE, mentioned in Chapter 5, was on an elevation-over-azimuth mount so that it could be pointed anywhere in the sky, with the anticipation that it would be used with the multi-angle analysis to find the optical depth of the mixed layer. When the whole lidar is to be tilted in this manner, considerations in Chapter 7 become important, because the gravitational loads change direction relative to the optical bench as the lidar is aimed off zenith. If anything in the structure flexes appreciably, $G(R)$ will change, so the entire assembly must be stiff. The other approach is to leave the lidar in a fixed position and change its aiming direction with scanning mirrors.

The lidar equation is generally written in terms of range as in Eq. (10.16), but for multi-angle lidar, it is written as a function of the backscatter height and the zenith angle using the relation $R = h/\cos(\theta) = mh$ where $m = 1/\cos(\theta)$ is the airmass, as defined in Chapter 4 in connection with sun photometry. In that case, Eq. (10.16) becomes

$$P(h,m) = \frac{C}{(mh)^2}\beta(mh)\exp[-2\int_0^{mh}\sigma(mh')dh'], \tag{10.18}$$

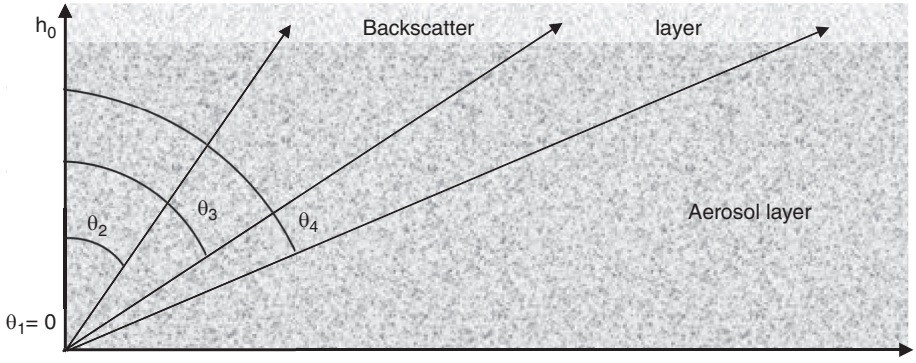


Figure 10.20 The multi-angle lidar technique. The lidar signal from altitude h_0 is measured at several zenith angles.

which is valid at any altitude above crossover. Taking the natural logarithm of the range-corrected version of Eq. (10.18) yields

$$\ln\left[X(h,m)\right] = \ln[C\beta(mh)] - 2m\int_0^{mh}\sigma(mh')dh', \tag{10.19}$$

and the integral term is just $m\tau(h)$ where $\tau(h)$ is the (vertical) optical depth from the ground to the height h. For the situation of Figure 10.20 with the backscattering region at the height h_0, a plot of $\ln\left[X(h_0,\theta\right]$ vs. $2m$ is a straight line with a slope equal to $-\tau(h_0)$, as illustrated with synthetic data in Figure 10.21. The relative errors of the measurements grow larger as $2m$ increases because the range to the scattering layer is longer and two-way transmittance is lower. It is evident from Figure 10.21 that the accuracy of the multi-angle measurement depends on having a sufficient separation of the data points on the horizontal axis in addition to the size of the uncertainties in the measurements, which is another way in which it is like the slope method. As a rule of thumb, multi-angle measurements should be made over a range of air masses from 1 to 3, as shown in the figure, which corresponds to a zenith angle range from 0 to 70.5 degrees.

A variation on the multi-angle lidar data analysis was developed at GTRI in 2007 in connection with field measurements to determine the amount of laser beam power that would reach airborne targets at various ranges and elevation angles, which depends on the transmittance on slant paths through the atmosphere. Using the definition

$$T^2(h,m) = \exp[-2m\int_0^{mh}\sigma(mh')dh'], \tag{10.20}$$

Eq. (10.18) becomes

$$P(h,m) = \frac{C}{(mh)^2}\beta(mh)T^2(h,m), \tag{10.21}$$

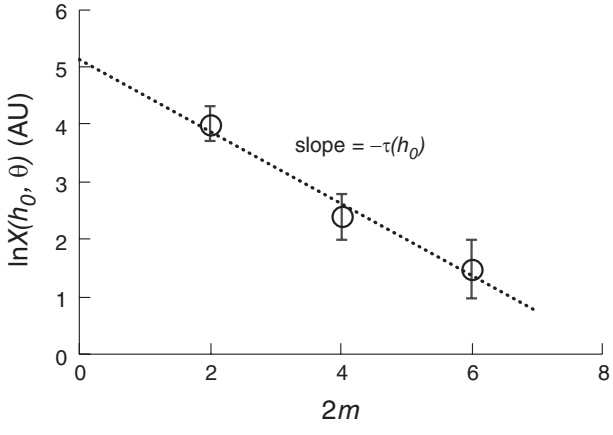


Figure 10.21 Multi-angle lidar data analysis. The slope of the fitted line is the negative of the optical depth from the surface to the altitude h_0. A wide range of airmasses is required to accurately determine the slope.

which suggests that perhaps the most natural parameter to derive with multi-angle scanning is the vertical transmittance profile $T(h)$ rather than τ. Another useful relation is that

$$T^2(h,m) = \left[T^2(h)\right]^m. \tag{10.22}$$

The transmittance approach is best elucidated using just two of the angles illustrated in Figure 10.20. Writing the lidar equations for the two angles,

$$P(h,m_1) = \frac{C}{(m_1 h)^2}\beta(h)T^2(h,m_1) \text{ and} \tag{10.23}$$

$$P(h,m_2) = \frac{C}{(m_2 h)^2}\beta(h)T^2(h,m_2). \tag{10.24}$$

To eliminate the unknowns C and β, divide Eq. (10.24) by Eq. (10.23) to get the result

$$\frac{P(h,m_2)}{P(h,m_1)} = \frac{(C/(m_2h)^2)\beta(h)[T^2(h)]^{m_2}}{(C/(m_1h)^2)\beta(h)[T^2(h)]^{m_1}} = \left(\frac{m_1}{m_2}\right)^2 [T^2(h)]^{m_2-m_1}, \tag{10.25}$$

where $m_2 > m_1$. Solving for $T(h)$ yields

$$T(h) = \left[\left(\frac{m_2}{m_1}\right)^2 \frac{P(h,m_2)}{P(h,m_1)}\right]^{\frac{1}{2(m_2-m_1)}}. \tag{10.26}$$

In cases where the m values are integers, Eq. (10.26) simplifies. For example, operating at zenith angles of 0 and 60 degrees, $m_1 = 1$ and $m_2 = 2$. Then the profile $T(h)$ is simply

$$T(h) = 2\sqrt{\frac{P(h,2)}{P(h,1)}}. \tag{10.27}$$

The foregoing derivations were all in continuous mathematics, but of course the data is discrete, and the two profiles must be sampled at the same h values to apply Eq. (10.27), so one or both profiles must be resampled before the ratio can be calculated. Resampling algorithms for this purpose have become commonplace and are now implemented in several mathematical applications. In the example illustrated here, every other range bin in the 60-degree profile is at the altitude of a bin in the vertical profile, so another option is to only use those bins in the analysis. Once $T(h)$ is known, the transmittance profile at any zenith angle is $T(h,m) = T^m(h)$ (again assuming a stratified atmosphere). Retrieving information from multi-angle lidar data in this way is not algebraically different from the lidar Langley plot method and it does not reveal any new information, but it is perhaps computationally simpler if $T(h)$ is the desired data product, and it emphasizes the fact that the lidar signal is proportional to T^2, not τ or σ.

The simple analyses described above serve to illustrate the basis for multi-angle lidar techniques, but they are rarely applicable because they require a stratified atmosphere, which the daytime mixed layer is obviously not, as shown in Figure 4.6. The analyses have other weaknesses, such as providing optical depths or transmittances when the extinction coefficient profile is usually desired. However, considerable effort has gone into making multi-angle lidar techniques less restrictive and more robust, and there is a substantial number of papers on these improvements in the lidar literature. Two-angle methods have been developed and evaluated with synthetic data that include noise and inhomogeneities [11]. These methods find the lidar's calibration constant by using data only from homogeneous altitude ranges, and then the extinction coefficient profile $\sigma(h)$ is retrieved. Strict horizontal homogeneity is not required; it is statistical only, using large sets of two-angle pairs. The multi-angle method has also been used to constrain other inversions, including the Klett inversion described in Section 10.3 [10].

10.3 Elastic Backscatter Inversions

The lidar and background equations are forward models, which predict the lidar waveform using atmospheric and instrumental inputs. Inversions are backward calculations going from the data to the atmospheric parameters. In the early days of lidar, researchers had hopes that zenith-pointing single-wavelength elastic backscatter lidars could measure the entire aerosol extinction coefficient profile from the surface to 30 km, and considerable effort was devoted to finding the enabling inversion. A 1972 paper by Fernald, Herman, and Reagan [12] is a classic example in this regard, providing an "advanced analytic solution" to the lidar solution for the situation in which both molecules and aerosols contribute to the lidar signal. Those authors also stressed the importance of supplementing lidar data with other data, such as sun photometer

optical depths. Their algebraic manipulations did not result in a closed form equation; rather, they led to a transcendental equation that was solved iteratively. The parameter S_a was assumed to be constant with altitude, and its value was determined by their analysis method.

Another approach, which was borrowed from radar, was to re-cast the lidar equation as a differential equation known as a *Bernoulli equation* (again assuming S_a was constant), which had a well-known solution. The solution requires knowledge of β (or σ) at some starting range. If $X(R)$ is known over some set of ranges, and the value of β is known in some range bin with index I, then the Bernoulli solution implies an algorithm to find β in an adjacent range bin, with index $I+1$ or $I-1$. The algorithm then steps through the $X(R)$ data bin by bin to find the profile $\beta(R)$. Early attempts at this recursive process started in a range bin near the lidar and progressed outward (again borrowing from radar techniques), and they were unsuccessful because the solution was mathematically unstable, leading to unphysical results. To address this problem, James Klett published one of the best-known lidar retrieval papers in 1981 [13]. Klett's contribution was to show that the recursive approach is mathematically stable if it starts at the far end and moves inward toward the lidar. A complete derivation of the Bernoulli solution and the corresponding algorithms is presented in the Appendix. The analysis algorithms (in discrete variables) for Klett's method were published by Fernald in 1984 [14].

An algorithm using Klett's approach is easy to derive if the atmosphere contains only one type of scatterer, in which case it is called a *one component* atmosphere. Infrared lidars approximate this condition because their signals are due to aerosol backscatter, and they are insensitive to molecular scattering. The derivation shown here begins somewhat like the one in [12], but it is more direct. Starting with the lidar equation as expressed in Eq. (10.16) and using the lidar ratio defined in Eq. (4.10) to eliminate σ yields

$$X(R) = C\beta(R)\exp[-2S_a\int_0^R \beta(r)dr], \tag{10.28}$$

which is an equation with only one variable, β, and an unknown calibration constant C. It is often remarked that β and σ enter into the lidar equation in very different ways, but in the one-component case, the lidar equation is actually in the form of a derivative, which can be seen by recalling that

$$\text{if } y(x) = \exp[f(x)], \text{ then } \frac{d}{dx}y(x) = \exp[f(x)]\frac{d}{dx}f(x). \tag{10.29}$$

Equation (10.28) can therefore be made directly integrable by writing it as

$$X(R) = \frac{-C}{2S_a}[-2S_a\beta(R)]\exp[-2S_a\int_0^R \beta(r)dr]. \tag{10.30}$$

Integrating both sides with respect to range,

$$\int_0^R X(r)dr = \frac{-C}{2S_a}\exp[-2S_a\int_0^R \beta(r)dr]\Big|_0^R = \frac{C}{2S_a}[1-T^2(R)]. \tag{10.31}$$

Rearranging,

$$T^2(R) = 1 - \frac{2S_a}{C}\int_0^R X(r)dr, \tag{10.32}$$

which is Eq. (5) of [12]. This is the relation needed for deriving the recursive inversion described above. Changing to discrete variables where I is the range bin index, consider the range-corrected signal from two adjacent bins:

$$X(I) = C\beta(I)T^2(I) \text{ and} \tag{10.33}$$

$$X(I-1) = C\beta(I-1)T^2(I-1). \tag{10.34}$$

From Eq. (10.32) we know that

$$T^2(I-1) = T^2(I) + \frac{S_a}{C}[X(I) + X(I-1)]\Delta R, \tag{10.35}$$

where we have used the middle Riemann sum, illustrated in Figure (10.22), to convert from the integral in Eq. (10.28) to a sum in discrete notation. The middle Riemann sum is just the average of the two adjacent $X(R)$ values times the range bin width. Equation (10.35) can therefore be rewritten as

$$X(I-1) = C\beta(I-1)\left[T^2(I) + \frac{S_a}{C}[X(I) + X(I-1)]\Delta R\right]. \tag{10.36}$$

Solving Eq. (10.34) for $T^2(I)$ and substituting it into Eq. (10.36) yields

$$X(I-1) = \beta(I-1)\left[\frac{X(I)}{\beta(I)} + S_a[X(I) + X(I-1)]\Delta R\right]. \tag{10.37}$$

Solving for $\beta(I-1)$, we have

$$\beta(I-1) = \frac{X(I-1)}{\frac{X(I)}{\beta(I)} + S_a[X(I) + X(I-1)]\Delta R}, \tag{10.38}$$

which is Eq. (9) of [14], also known as the Klett inversion algorithm for a one-component (aerosol) atmosphere. It is a *recurrence relation*, meaning that each value of β depends on the previous value. The lidar's calibration constant has been eliminated and the solution is calibrated by the beginning backscatter coefficient $\beta(I)$. The algorithm is mathematically stable, so it will always yield a solution profile, and Klett showed that it will converge toward the correct profile even when the initial estimate of $\beta(I)$ is incorrect and $X(R)$ includes noise. However, the algorithm for β requires knowledge of the aerosol lidar ratio S_a, which is most often not known. Using the definition of S_a, it easy to show that the retrieval algorithm for the extinction coefficient profile is

$$\alpha(I-1) = \frac{X(I-1)}{\frac{X(I)}{\alpha(I)} + [X(I) + X(I-1)]\Delta R}, \tag{10.39}$$

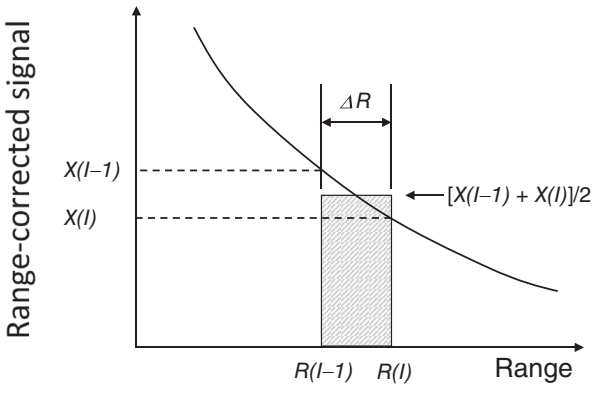


Figure 10.22 The middle Riemann sum. The sum, which corresponds to a differential of area, is shown by the shaded rectangle. It is the average of the two range-corrected signal values times the range bin width.

which does not depend on S_a. The algorithm requires the starting extinction coefficient however, so it still requires a priori knowledge. The algorithms in Eqs. (10.38) and (10.39) can only be used at ranges beyond crossover because they were derived from Eq. (10.16), which is only valid where $G(R)$ is unity. The algorithms for the two-component atmosphere are considerably more complicated [14]. Defining a new quantity A by the relation

$$A(I-1,I) = (S_a - S_m)[\beta_m(I-1) + \beta_m(I)]\Delta R, \text{ the algorithms are} \tag{10.40}$$

$$\beta_a(I-1) = \frac{X(I-1)\exp[A(I-1,I)]}{\dfrac{X(I)}{\beta_a(I)+\beta_m(I)} + S_a\left\{X(I) + X(I-1)\exp[A(I-1,I)]\right\}\Delta R} - \beta_m(I-1) \text{ and} \tag{10.41}$$

$$\sigma_a(I-1) = \frac{X(I-1)\exp[A(I-1,I)]}{\dfrac{X(I)}{\sigma_a(I)+\dfrac{S_a}{S_m}\sigma_m(I)} + \left\{X(I) + X(I-1)\exp[A(I-1,I)]\right\}\Delta R} - \frac{S_a}{S_m}\sigma_m(I-1). \tag{10.42}$$

Both algorithms depend on knowledge of S_a as well as the initial value of β_a or σ_a, and the molecular profiles. Unfortunately, the retrieved profiles depend strongly on S_a, as shown for the one-component case in Figure 10.23, where synthetic signals were generated with $S_a = 40$ sr and backscatter coefficients were retrieved using Eq. (10.38) with S_a values of 30, 40, and 50 sr. The profiles retrieved with incorrect values of S_a are badly distorted, even with relatively small errors in S_a, which ranges from 20 to 120 sr in the mixed layer and tends to vary with altitude. Retrievals of σ_a

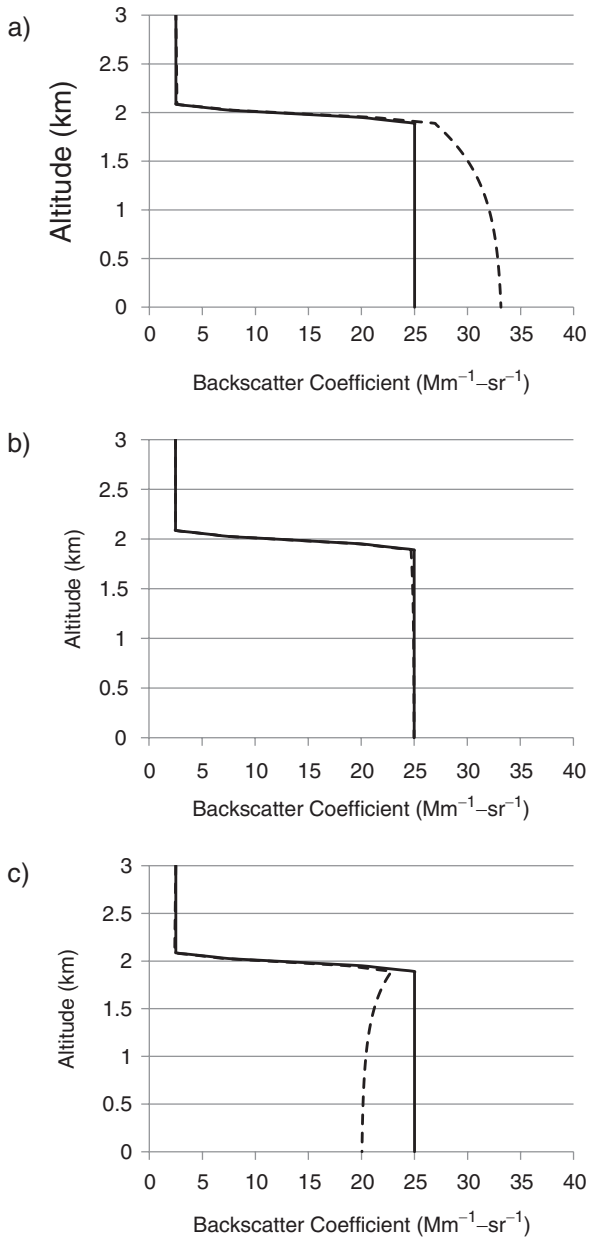


Figure 10.23 Effects of lidar ratio errors. Synthetic signals were generated with $S_a = 40$ sr and backscatter coefficient profiles were retrieved with S_a values of (a) 30 sr, (b) 40 sr, and (c) 50 sr.

for the two-component atmosphere are similarly distorted but in the opposite sense (a lower S_a causes lower retrieved values). During daytime, the Klett retrieval can be constrained by requiring that the integral of $\sigma_a(h)$ is equal to the sun photometer OD described in Chapter 4, or a multi-angle result described in Section 10.2.4 may

be used. In these methods, the value of S_a used in the retrieval is varied until the ODs match, but the profile may not be reliable because S_a often varies with altitude. Modifications to the Klett algorithms have been developed that accommodate S_a values that vary with range [15, 16], but the values are usually unknown.

Because of the problems noted above, interest in the Klett algorithms is now mainly historical. In 2012, *Applied Optics* reached its 50th year of publication, and the Optical Society of America published a list of its 50 most-cited articles [17]. Both [13] and [14] were among them. However, the quest for an accurate inversion was not successful: The lidar community has reached a consensus that a zenith-pointing elastic backscatter lidar is not capable of measuring extinction coefficient profiles in a two-component atmosphere because of the requirements for a priori knowledge, and such profiles are now measured with the more advanced HSRL or Raman aerosol lidar. The main reason for presenting the derivation in Eqs. (10.28)–(10.38) was to illustrate the process of developing an algorithm that can be implemented in software to analyze digital data, starting from a mathematical result written in continuous variables. The detailed derivations of Eqs. (10.41) and (10.42) have been relegated to the Appendix.

10.4 Further Reading

The book on digital signal processing by Oppenheim, Shafer, and Buck [2] is the standard text and it is comprehensive. It is written at the level of graduate students.

The book by Kovalev and Eichinger [10] contains a wealth of information on lidar inversions and multi-angle techniques.

10.5 Problems

10.5.1 The lidar data shown in Figure 9.14 were recorded at a sampling frequency of 40 MHz and smoothed with a 41-bin rectangular filter. (a) By what factor was the r.m.s. variation in the data reduced? (b) What is the vertical resolution of the filtered data, by the NDACC definitions? Hint – for the second definition, implement Eq. (10.4) in a spreadsheet or math application to find the cutoff frequency.

10.5.2 By inspecting Figures 10.4 and 10.6, find the resolution of the Hann and Hamming 25-point filters according to the first NDACC definition of a FIR filter's resolution. What rectangular filter width would yield the same resolution?

10.5.3 The atmospheric and aerosol depolarization ratios are not the same, because the atmospheric value depends on the aerosol concentration. (a) Show that Eq. (10.15) yields the aerosol depolarization ratio. (b) Show that when $S_{a\perp} \gg S_{m\perp}$ and $S_{a\parallel} \gg S_{m\parallel}$, $\delta_a \approx \delta_{atm}$.

References

[1] J. M. Stewart, G. G. Gimmestad, D. W. Roberts et al., "NEXLASER – an unattended tropospheric aerosol and ozone lidar," in Proceedings of SPIE, 2002, vol. 4723, pp. 172–181.

[2] A. V. Oppenheim, R. W. Shafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, 1999.

[3] T. Leblanc, R. J. Sica, J. A. E. van Gijsel et al., "Proposed Standardized Definitions for Vertical Resolution and Uncertainty in the NDACC Lidar Ozone and Temperature Algorithms – Part 1: Vertical Resolution," *Atmospheric Measurement Techniques*, vol. 9, pp. 4029–4049, 2016.

[4] M. Iarlori, F. Madonna, V. Rizi et al., "Effective Resolution Concepts for Lidar Observations," *Atmospheric Measurement Techniques*, vol. 8, pp. 5157–5176, 2015.

[5] H. N. Forrister, D. W. Roberts, A. J. Mercer et al., "Infrared Lidar Measurements of Stratospheric Aerosols," *Applied Optics*, vol. 53, pp. D40–D48, 2014.

[6] "Window function," in Wikipedia. [Online]. Available: https://wikipedia.org/wiki/Window_function [Accessed December 17, 2021].

[7] M. Haeffelin, F. Angelini, Y. Morille et al., "Evaluation of Mixing-Height Retrievals from Automatic Profiling Lidars and Ceilometers in View of Future Integrated Networks in Europe," *Boundary-Layer Meteorology*, vol. 143, pp. 49–75, 2012.

[8] E. V. Browell, C. F. Butler, S. Ismail et al., "Airborne Lidar Observations in the Wintertime Arctic Stratosphere: Polar Stratospheric Clouds," *Geophysical Research Letters*, vol. 17, pp. 385–388, 1990.

[9] V. Freudenthaler, "About the Effects of Polarising Optics on Lidar Signals and the Δ90 Calibration," *Atmospheric Measurement Techniques* vol. 9, pp. 4181–4255, 2016. See also supplement doi:10.5194/amt-9-4029-2016-supplement (45 pp).

[10] V. A. Kovalev and W. E. Eichinger, *Elastic Lidar – Theory, Practice, and Analysis Methods*. Hoboken, New Jersey: John Wiley & Sons, 2004.

[11] M. Pahlow and V. A. Kovalev, "Calibration Method for Multiangle Lidar Measurements," *Applied Optics*, vol. 43, pp. 2948–2956, 2004.

[12] F. G. Fernald, B. M. Herman, and J. A. Reagan, "Determination of Aerosol Height Distributions by Lidar," *Journal of Applied Meteorology*, vol. 11, pp. 482–489, 1972.

[13] J. D. Klett, "Stable Analytical Inversion Solution for Processing Lidar Returns," *Applied Optics*, vol. 20, pp. 211–220, 1981.

[14] F. G. Fernald, "Analysis of Atmospheric Lidar Observations: Some Comments," *Applied Optics*, vol. 23, pp. 652–653, 1984.

[15] J. D. Klett, "Lidar Inversion with Variable Backscatter/Extinction Ratios," *Applied Optics*, vol. 24, pp. 1638–1643, 1985.

[16] Y. Sasano, E. V. Browell, and S. Ismail, "Error Caused by Using a Constant Extinction/Backscattering Ratio in the Lidar Solution," *Applied Optics*, vol. 24, pp.3929–3932, 1985.

[17] "50 Most cited articles," in Optica publishing group. [Online]. Available: 50 Most Cited Articles (osapublishing.org) [Accessed: November 11, 2021].

11 Applications

The engineering principles described in the previous 10 chapters are applied in all types of atmospheric lidars. Most of the examples presented so far have involved elastic backscatter aerosol lidars, so this chapter includes applications to many of the other types described in Chapter 3: wind lidars, Rayleigh lidar, DIAL, Raman lidar, HSRL, and resonance fluorescence lidar. Some comments are included on the engineering challenges of these various types of lidars, and on their data analysis algorithms. These lidar techniques have all been described in other books and papers, so only brief descriptions are given here, with appropriate references. The engineering principles also apply to several other types of lidar that are not covered here.

11.1 Cloud-Aerosol Lidar with Orthogonal Polarization

The CALIOP lidar, as described in Section 4.5.3, was a calibrated elastic backscatter lidar operating at 532 and 1064 nm, with parallel and perpendicular polarization channels in the 532-nm receiver. In view of the comments in the previous chapter, it may seem strange that such a limited instrument was used, but CALIOP was an early spaceborne lidar and its simplicity was no doubt a key reason that it provided continuous global coverage of cloud and aerosol layers for well over a decade. The CALIOP data analysis algorithms were developed in the best tradition of geoscience, which is to extract the maximum amount of information from the data while also providing reliable estimates of the uncertainty in the data products.

The CALIOP lidar is described in [1], where the table of design parameters will look familiar to anyone who has read the previous chapters of this book. The need for a large dynamic range in lidar receivers and data systems is especially acute for CALIOP, because it is calibrated with the weak molecular scattering from altitudes above 30 km and it also experiences very bright returns from the surface. As mentioned in Chapter 9, the solution was to split each TIA output in two, with one output going to a low gain voltage amplifier and the other to a high gain amplifier. The two signals then went to two 14-bit ADCs, with some overlap, resulting in a 22-bit single-pulse dynamic range. The background in the recorded signal is assessed using range bins from 112 to 97 km, where the real signal is assumed to be zero. The lidar is pointed 3 degrees off nadir in the along-track direction to avoid the problem with oriented plates in cirrus clouds mentioned in Chapter 4 and to avoid potential damage from specular reflections off still ponds.

Fully automated data analysis was implemented pre-launch, and the algorithms have gone through several revisions over the years, with Version 4 being released in 2019. The CAD is described in [2], the extinction retrieval updates are in [3], and the algorithms are explained in [4]. As illustrated in Figure 4.17, analysis starts with a curtain plot, which is called a *scene*. Features, which are regions with backscatter greater than clear air, are located by a process call selective iterated boundary locator (SIBYL). The tasks of determining the type of each feature, finding its optical depth, and finding profiles of particulate backscatter and extinction are then performed by hybrid extinction retrieval algorithm (HERA). Cloud-aerosol lidar with orthogonal polarization algorithms operate from the top down, stopping when they encounter either an opaque cloud or the surface. When semitransparent layers are encountered, their optical depths are calculated using the scattering ratio method shown in Figure 10.18. Finding profiles of optical properties poses unique challenges: far-end solutions such as the Klett method could not be used because finding an aerosol-free region under a feature in which to initialize the retrieval is unlikely, and in any event, the far end of the PBL is the surface. Another problem was that the algorithms due to Fernald and Klett did not lend themselves to multiple scattering corrections. Cloud-aerosol lidar with orthogonal polarization data generally include multiple scattering contributions, because the great range from the lidar to the atmosphere causes the FOV to cover a large area on the atmosphere that includes both singly and multiply scattered photons, so a correction factor was required in the algorithm. For these reasons, the solution embodied in HERA, the *linear iterative method*, was adapted from pre-lidar aerosol studies using searchlights. This method is not based on a closed-form solution of the lidar equation; instead it uses iteration to solve a transcendental equation, as mentioned in Section 10.3. As a familiar example in lidar, assuming that $\sigma(h) = S_a\beta(h)$ in a one-component atmosphere, the lidar equation can be solved for $\beta(h)$ in terms of $X(h)$, C and $T^2(h)$, but the product $S_a\beta$ is in the argument of the exponential function in $T^2(h)$, so the equation does not have a closed-form solution. The iteration in HERA uses the Newton–Raphson method, which starts with phrasing the problem as finding the root of an equation in the form $f(x) = 0$. The method starts with an initial guess at the solution called x_1, and the algorithm is

$$x_{k+1} = x_k - \left(\frac{f(x_k)}{f'(x_k)} \right), \tag{11.1}$$

where $f'(x)$ is the first derivative of the function with respect to x. A demonstration of the rapid convergence of this method is left as a problem for the student. It is said to converge to a sufficient accuracy in HERA calculations within two or three iterations. The HERA algorithm operates from the top down, it accommodates multiple scattering corrections, and it operates in two modes: constrained, when the layer optical depth is known, and unconstrained, when it is not. The HERA profile retriever requires the lidar ratio S_a for each layer, and the Version 4 values of S_a used in 532-nm retrievals of CALIOP data are listed in Table 11.1. CALIOP uses a model-matching scheme to select S_a that is embodied in flowcharts like Figure 4.16 but more complicated. In addition to lidar observables (mean attenuated 532-nm

Table 11.1 S_a values for CALIOP retrievals

Aerosol type	S_a (532 nm)
1. Tropospheric	
Clean marine	23 ± 5
Dust	44 ± 9
Polluted continental/smoke	70 ± 25
Clean continental	53 ± 24
Polluted dust	55 ± 22
Elevated smoke	70 ± 16
Dusty marine	37 ± 15
2. Stratospheric	
Polar stratospheric aerosol	50 ± 20
Volcanic ash	44 ± 9
Sulfate/other	50 ± 18
Smoke	70 ± 16

backscatter coefficient, depolarization ratio, and color ratio), latitude, longitude, and seasonal characteristics are used in decision points. The troposphere and stratosphere are considered separately. The first set of CALIOP aerosol types with corresponding S_a values came from analysis of an extensive multiyear AERONET data set in 2005 [5]. That analysis identified six aerosol types, and more have been added over the years. AERONET has also been used for validation of CALIOP data products.

11.2 Wind Lidars

Henderson et al. have provided an encyclopedic review of lidar wind sensing methods including both direct and coherent detection, as a chapter in a laser remote sensing book [6]. As those authors pointed out, there are two basic types of direct detection wind lidar: those that measure the Doppler shift in backscattered laser light and those that track the motion of tracers such as aerosols or clouds.

11.2.1 Tracer Techniques

Spaceborne imagers often record cloud motions as changes of positions in successive sets of images, and those motions are used to produce atmospheric motion vectors, which are related to winds. Some ground-based lidar stations equipped with imagers use a similar technique to measure winds aloft. The wind vector can be calculated by correlating cloud positions in sequential images and using the imager's FOV to obtain the direction and rate of angular motion, after which the cloud base height measured by the lidar is used to convert angular motion to linear motion (m/s). The usefulness of this technique is limited because wind measurements are only available at altitudes where clouds happen to pass over the station. Aerosols provide a more robust method

in the boundary layer because they are always present. Aerosols are not homogeneous; they tend to occur with regions of higher concentrations known as *eddies*, with lower concentrations between the eddies. A lidar scanning repeatedly in a plane will observe the eddies as regions of higher backscatter, and successive scans will reveal the motions of the eddies. Because the eddies are advected by the wind, those motion vectors are interpreted as the wind field. The technique of measuring winds in this way is known as *eddy correlation*. Scanning requires a powerful lidar because it requires single-pulse signals with good SNR out to some useful range.

Pioneering work in eddy correlation was conducted at the University of Wisconsin with a two-axis scanning lidar known as the volume imaging lidar (VIL), starting around 1990 [7]. Volume imaging lidar employed a 1,064-nm laser with 30-Hz PRF and an average power of 20 W (the pulse energy was 667 mJ) and a two-axis scanner with a maximum scan rate of 20 degrees per second. A more detailed description of VIL as well as movies and 3D images made from VIL data on both aerosols and clouds can be seen at the University of Wisconsin Lidar Group website [8]. A main goal of this early work was visualization of flow fields, so the early data analysis was aimed at video presentations. The contrast between the higher and lower backscatter regions was often very low, so individual profiles were range-corrected and then corrected for estimated extinction profiles so that they did not fade to black with increasing range. Digital imaging was becoming common at the time, along with image processing algorithms, but conventional images are in the Cartesian coordinate system (rows and columns of pixels) whereas lidar data are acquired in polar coordinates (angle and range), so commercial software for image display and analysis was not directly applicable and the processing software was developed in-house. In addition to producing the movies showing 2D flow fields, the Wisconsin group also analyzed VIL data to find vertical profiles of horizontal winds [9].

Volume imaging lidar demonstrated the use of scanning lidar to measure winds, but the 1,064-nm laser beam presented an eye hazard that limited its applications. The next step was to develop a safe version. An eye safe aerosol lidar had been demonstrated in 1989 by the GTRI lidar team; it was based on Raman-shifting 1064 nm to 1543 nm in a high-pressure methane cell as shown in Figure 5.16 [10]. The SWIR wavelength also benefits from a low sky radiance, as shown in Figure 3.7. The eye safe technology was adapted and taken to a new level in a scanning lidar known as Raman-shifted Eye safe Aerosol Lidar (REAL) that was reported in 2004 by Mayor and Spuler [11]. REAL, developed between 2001 and 2007 at the National Center for Atmospheric Research (NCAR), transmits 170-mJ pulses at 1543 nm with 10 Hz PRF. The receiver diameter is 40 cm, and the 14-bit digitizer has a 100 MHz sampling frequency (1.5-m range bins). It is presently maintained and operated at the California State University, Chico; see the website [12] with many example movies. The movies, which are time sequences of range-corrected signals in planar scans, provide excellent qualitative visualizations of aerosol motions. To obtain a field of wind vectors, it was necessary to perform correlations between the individual scans. The data from a planar lidar scan is generally plotted in the form of a 2D image (although it is measured in polar coordinates, as mentioned above) and by the time REAL was developed, a great deal of relevant image

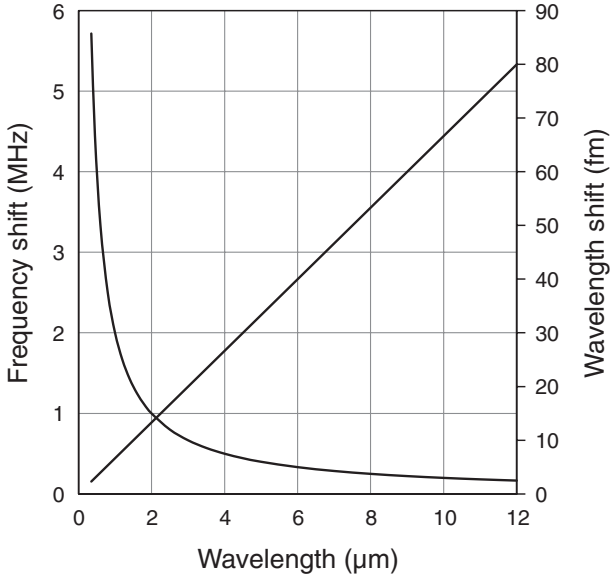


Figure 11.1 Doppler shifts for 1 m/s wind. Curved line – frequency shift; straight line – wavelength shift. Absolute values are plotted because the shifts have opposite signs.

analysis work had been done in fluid mechanics for another laser sensing technique, particle image velocimetry. Early algorithms were based on cross-correlation between image pairs, but more recently, a class of algorithms known as optical flow has become popular. Raman-shifted Eye safe Aerosol Lidar data are analyzed with software that provides a two-component displacement vector at every grid point. Examples of wind vectors overlaid on the lidar data visualizations can be seen at [12].

11.2.2 Doppler Techniques

Laser Doppler techniques measure the component of the wind vector along the line of sight by measuring the Doppler shift in backscattered laser light. The frequency shift and the wavelength shift are given by

$$\Delta\nu = -2\mathrm{v}/\lambda \text{ and} \tag{11.2}$$

$$\Delta\lambda = 2\frac{\mathrm{v}}{c}\lambda, \tag{11.3}$$

where v is the wind component (which is positive for motion away from the receiver). The absolute values of these shifts are shown in Figure 11.1 for a 1 m/s wind. The frequency shift rises rapidly at shorter wavelengths, and it is about 1 MHz per m/s in the SWIR wavelength range. Frequency shifts of this size are easily measurable with coherent lidar. The SWIR range is eye safe and solid-state lasers are available, and these facts have enabled very successful commercial lidars, including the Leosphere Windcube series [13] and the Lockheed Martin WindTracer [14], which are widely deployed at wind farms and airfields. See [6] for an extensive discussion of coherent Doppler lidar wind sensing.

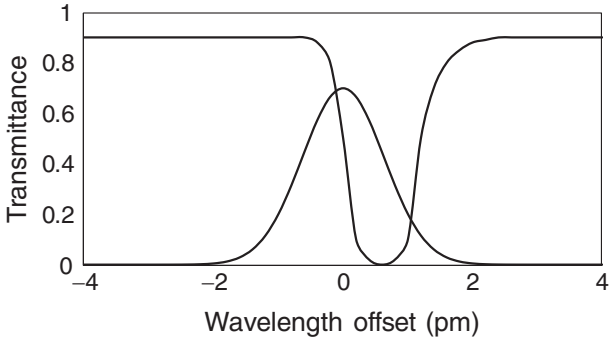


Figure 11.2 The edge technique. The Gaussian curve is the Doppler-broadened molecular lidar signal; the other curve is the I_2 filter transmittance.

The other types of Doppler wind sensing measure wavelength shifts using direct detection. In contrast to the large Doppler frequency shifts shown in Figure 11.1, wavelength shifts are tiny, less than 0.1 pm for visible light with a 1 m/s wind. For this reason, lidars that measure Doppler shifts optically require extremely high spectral resolution, which can only be obtained by exploiting atomic or molecular absorption lines or with interferometers. An early approach called the *edge technique* is illustrated schematically in Figure 11.2, where the Doppler-broadened molecular lidar signal is shown by the Gaussian curve and an iodine vapor absorption line is shown by the other curve. The basic principle is that a steep edge in the filter response means a small wavelength change causes a large transmittance change. The method does not require spectral scanning. This idea was thoroughly explored through analysis and modeling around 1990 [15], using a differential technique that made it insensitive to the laser line width and frequency jitter. It can be applied to either the aerosol peak or the Doppler-broadened spectrum shown in Figure 3.10, where the atmospheric Doppler spectrum spans about ±2 pm for 532-nm laser light. The received signal is split into two parts, with one going directly to a detector and the other through the filter. If the Doppler-broadened signal shifts due to wind, the ratio of the two signals will change. The results of the study were very promising, and the edge technique has been deployed, notably in the ground-based Rayleigh wind lidar at ALOMAR (discussed later in this Section).

The double-edge technique illustrated in Figure 3.11 uses two filter passbands spaced symmetrically about the laser wavelength. These passbands are usually created with a Fabry–Perot etalon. A wavelength shift causes a change in the relative signals through the two receiver channels. As pointed out in [6], the technique should really be called two-channel because edges are not involved, but the conventional terminology is double-edge. The Fabry–Perot etalon is perhaps best introduced by considering a single element, the etalon, which was discussed in Section 5.2.6 as a means of limiting the free spectral range (FSR) of a laser cavity so that there is only one longitudinal mode under the gain curve. That type of etalon is illustrated in Figure 11.3, which shows some of the reflected and transmitted rays that result from an incident ray on a

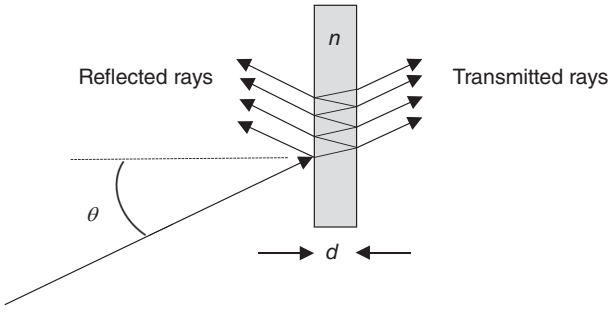


Figure 11.3 The etalon. Several reflected and transmitted rays are shown. The front and back surfaces are both partially reflective.

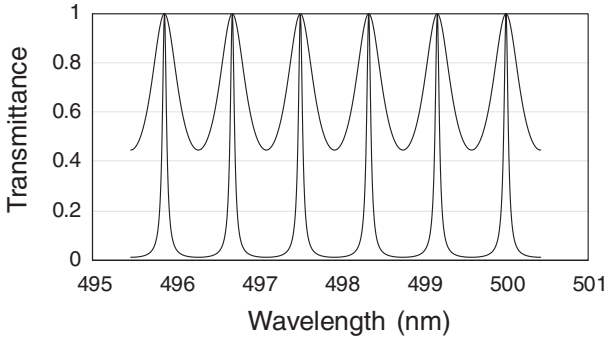


Figure 11.4 Transmittance spectrum of an etalon. The lower curve is for a coating reflectance of 0.8 and the upper curve is for a coating reflectance of 0.2.

transparent planar material in air with thickness d and refractive index n. Both front and back surfaces have partially reflective coatings. The angle of incidence θ is exaggerated for clarity; at normal incidence all reflected and transmitted rays are colinear with the incident ray. If the optical path $nd\cos(\theta)$ inside the etalon is a certain number of wavelengths, the reflected rays will suffer complete destructive interference. In that case, the etalon reflectance R_e becomes zero and its transmittance T_e becomes unity, because $R+T=1$ in a non-absorbing medium. This phenomenon is shown in Figure 11.4, which is a portion of the transmittance spectrum of an etalon with a thickness d of 100 μm and refractive index n of 1.5. Normal incidence is assumed. The spacing between the peaks, about 0.83 nm, corresponds to the FSR of the etalon at 500 nm (FSR is defined as $c/2nd$ and it is in frequency units). The transmittance spectrum in Figure 11.4 was calculated with the formula

$$T_e = \frac{1}{1+F\sin^2(2\pi nd/\lambda)}, \tag{11.4}$$

where

$$F = \frac{4R}{(1-R)^2}. \tag{11.5}$$

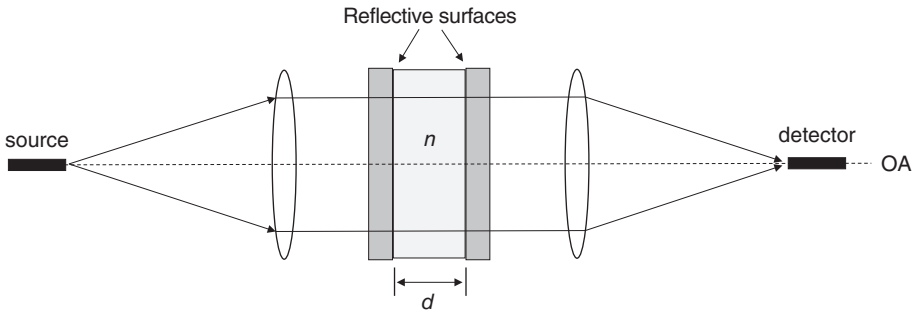


Figure 11.5 The Fabry–Perot etalon. This configuration is used when the etalon is a bandpass filter. In lidars, the source is often an optical fiber from the receiver telescope.

The parameter F is known as the *coefficient of finesse* of the etalon, and it is the ratio of the spacing between transmittance peaks to the width of each peak (FWHM) [16]. The etalon transmittance becomes unity at wavelengths where the argument of the sine function is an integer multiple of π. The parameter R is the reflectance of each surface. Etalons with high finesse show sharper transmittance peaks with lower minimum transmittance coefficients. The phenomena of constructive and destructive interference caused by pathlength-induced phase shifts is the basis for AR coatings, and it is also the fundamental principle used in the interference filters described in Chapter 6, although those filters always have multiple layers and their designs are optimized with sophisticated models.

The Fabry–Perot etalon has two optical elements with reflective coatings and a gap between them filled with a lower-index medium (often air), and the gap size is sometimes adjustable, by means of piezo-electric actuators or by changing the temperature of the etalon structure. The other tuning method is pressure tuning, which changes n by changing the gas density in the gap. If the etalon is designed such that one element can be scanned in position, the device is called an interferometer. A Fabry–Perot etalon used as a bandpass filter is illustrated in Figure 11.5. In practice, the two optical elements are slightly wedged to avoid spurious spectra caused by reflections within them. The bandpass width is again governed by the finesse [16]. Multiple bandpasses can be obtained from one Fabry–Perot etalon by depositing layers of silica in sub-apertures of an etalon element to make d slightly smaller in those areas.

The lidar called Goddard Lidar Observatory for Winds (GLOW) was a good example of the double-edge technique with multiple bandpasses [17]. Goddard Lidar Observatory for Winds was a NASA prototype for a proposed space shuttle lidar. It used a tripled Nd:YAG laser transmitting 70 mJ at 355 nm and a 45-cm receiver telescope. The Fabry–Perot etalon had three sub-apertures, one of which was used to sample the outgoing laser beam as a reference, because any frequency drift in the laser or the etalon would cause a systematic error in the retrieved wind speeds. Goddard Lidar Observatory for Winds was designed to have equal sensitivities to Doppler shifts in the aerosol and molecular spectra, and the two Doppler channels used photon counting PMTs. The lidar was validated with rawinsonde wind data while measuring wind profiles from 1.8 to 35 km, but the dynamic range of the signal levels was so

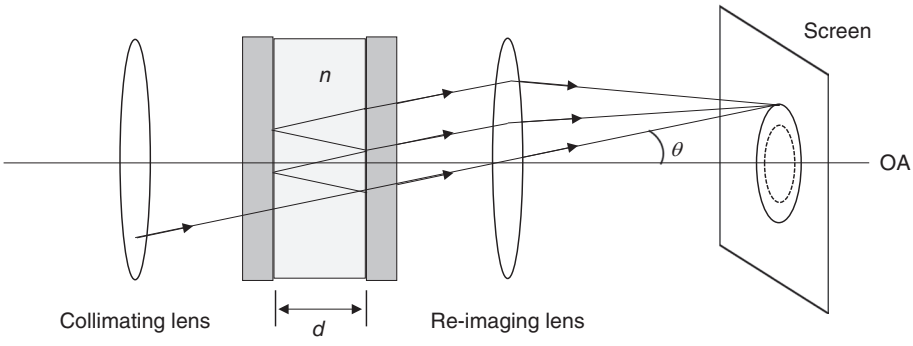


Figure 11.6 Fabry–Perot etalon fringes. Because of axial symmetry, rays at angles causing constructive interference will form bright circular fringes on the screen.

large that it could not be accommodated due to a maximum count rate restriction. The solution was to split the altitude range into two parts, 1.8 to 7.0 km and 7.0 to 35 km, and decrease the pulse energy to 0.4 mJ for the lower part. Raw data were recorded at 45-degree elevation for four compass points using 250-m range bins and 30-s averaging. The data were analyzed by ratioing the signals from the two filters and converting the ratios to projected wind speeds by using a lookup table. The lookup table was calculated with models of the filter spectra and the signal spectra.

Fabry–Perot etalons in lidar receivers tend to be very expensive, so they are only used in major facilities or in spaceborne lidars. One reason they are expensive is their size: They are sensitive to incidence angle, and as shown by Eq. (6.3) and illustrated in Figure 6.3, minimizing the divergence of rays on the receiver filter can require making the filter large. Another reason is the requirements for flatness and parallelism of the optical elements. Henderson et al. provided a table of requirements for 355-nm wind lidar etalons that include sub-aperture diameters of 40 mm, plate flatness of $\lambda/120$ r.m.s., and plate relative alignment error less than 0.1 μrad [6]. Such optical specifications can be met, but at considerable cost.

A third alternative for direct detection Doppler wind lidar is called fringe imaging. This technique is based on using a high-resolution spectrometer to form a spatial irradiance distribution that corresponds to the received signal spectrum. The spatial distribution is recorded with an array detector, and changes in the fringe images are analyzed to find wind speeds. This technique requires range-gated imaging detectors. The formation of *fringes* is illustrated in Figure 11.6. If collimating and re-imaging lenses are placed before and after a Fabry–Perot etalon, and the light through the etalon is slightly diverging, an axially symmetric pattern of light and dark circles will be seen on a screen in the focal plane of the re-imaging lens. The etalon transmittance is unity when $nd\cos(\theta)$ is an integer number of wavelengths, so when the number of wavelengths changes, θ also changes. A monochromatic spectrum is thereby converted to a set of concentric bright circles that depend on the wavelength, and in a general sense, if the spectrum changes, the fringe pattern changes. Direct detection Doppler wind lidars record the fringe pattern with range-gated imagers and convert

the patterns to wind speed. Other types of interferometers are sometimes used to create other fringe patterns. The laser frequency must be very stable because a change in transmitted wavelength will also cause a change in the fringe pattern.

The very large ground-based direct-detection RMR wind lidar at ALOMAR initially employed a Fabry–Perot etalon and fringe imaging, using a special type of PMT with 24 separate channels in concentric rings [6]. The RMR lidar is based on two 1.8-m diameter telescopes that can be independently tilted up to 30° off the zenith to measure the component of horizontal winds projected onto the lidar's line of sight. The telescopes are independent lidars, and each one has a 532-nm laser transmitting 470-mJ pulses at 30-Hz PRF. The injection-seeded lasers have a linewidth of less than 70 MHz and their wavelengths are continuously monitored. The receivers are coupled to the spectrometer with optical fibers, and the two receivers are multiplexed into one spectrometer by pulsing the lasers alternately. The wind vector can be measured by tilting the lidars at orthogonal angles. Around 2010, the RMR receiver for winds was converted to the edge technique illustrated in Figure 11.2, using an iodine filter [18]. The received beam diameter is 36 mm and the cell diameter is 50 mm. For daytime operation, the beam also goes through a double etalon filter with an optical bandwidth B_{opt} of ~4 pm. The system was developed with PMTs as the detectors, but they were replaced with APDs for higher quantum efficiency. The developers claimed single-photon sensitivity and a relative resolution ($\Delta\lambda / \lambda$) of 10^{-8}.

Example. At 532 nm, a relative resolution of $\Delta\lambda / \lambda = 10^{-8}$ corresponds to a wavelength shift of $(5.32 \times 10^{-7}) \times (10^{-8}) = 5.32$ fm. This is comparable to the Doppler shift from a 1 m/s wind. Using Eq. (11.3) with a 1 m/s wind speed, the Doppler shift is $2 \times 1 \times (5.32 \times 10^{-7}) / (3 \times 10^{8}) = 3.55$ fm.

After a great deal of care in frequency stabilization and monitoring, the RMR wind lidar data analysis is simple: There are only two signals, one through the iodine cell and one without it. The ratio of these becomes the single parameter from which the projected wind speed is calculated. Values of the ratio were modeled for wind speeds from −200 to +200 m/s, and a lookup table was generated. For a 2-km altitude resolution and 2-hour time resolution, the r.m.s errors of the lidar were shown to be ±0.6 m/s at 49 km and ±10 m/s at 80 km for the projected wind speeds (the horizontal wind speed errors are twice those values), and those errors are due only to photon statistics. The lidars can also be pointed vertically, and vertical winds are significant in the mesosphere. The unfiltered signal is also used to find temperature profiles up to 90 km, using the Rayleigh lidar technique described in the next section.

The first spaceborne wind lidar was ESA's Aeolus, launched in August 2018. The primary mission objective was to demonstrate the Doppler wind lidar technique for measuring wind profiles from space, which were intended for assimilation in numerical weather prediction (NWP) models. Conventional wind profile measurements are mainly over land in the Northern Hemisphere, and the lack of global wind data has long been recognized as a major limiting factor in NWP improvement. Aeolus was pointed 35 degrees off nadir in a cross-track direction, so that it measured horizontally

projected line-of-sight (HLOS) winds (one component of the wind vectors). Data were recorded in 24 programmable altitude bins with resolutions from 0.25 to 2 km. Data were averaged for 3 km along-track onboard and were further averaged on the ground. The mission requirements included a maximum HLOS wind observation random error of 1 m/s in the PBL, 2.5 m/s in the free troposphere, and 3–5 m/s in the stratosphere, with a maximum bias error of 0.7 m/s. The data products also included profiles of particle and molecular parallel-polarized backscatter and extinction coefficients, scattering ratios, and backscatter-to-extinction ratios [19].

Aeolus was similar to GLOW in that it operated at 355 nm and it used the double-edge technique for the Rayleigh return, implemented with a Fabry–Perot etalon. However, it used a Fizeau interferometer and fringe imaging for the Mie return. A Fizeau interferometer creates straight-line fringes rather than rings, and wind speeds were indicated by the position of a single fringe on the imaging detector. Both receivers used accumulation CCDs as detectors. These devices had several zones: image, transfer, storage, and finally, the readout register. The image zone was 16 rows by 16 columns and the memory zone was 25 rows by 32 alternating transfer and storage columns. Both ACCDs were silicon CCDs with quantum efficiencies greater than 83% and they were cooled to −30°C with thermo-electric coolers to limit dark-current noise. The image zone could be read out in 1 μs, when all charges were binned into one line that was stored in the memory area, where each row corresponded to a height bin. The signals were accumulated on the CCD chip over several laser shots, and the charges became sufficiently large that the noise contribution of the preamplifier became negligible compared to the shot noise from the detected signal itself, and the SNR approached the shot-noise detection limit. The ACCDs were used differently for the two receiver channels: for the Mie channel, the detector recorded the position of the fringe in each height bin, and for the Rayleigh channel, the two Fabry–Perot outputs were imaged onto the detector as two circular spots and differences in intensity between the two spots were recoded as different amounts of charge [20].

Resonance fluorescence lidars provide another way to measure winds, by pointing off-zenith and tuning the laser wavelength to perform high-resolution spectroscopy on layers of metal atoms in the mesosphere. Figure 3.22 illustrates a method for measuring temperatures by means of measurements at a few selected wavelengths; a similar method detects Doppler shifts of the whole spectrum, thereby measuring the projected wind speed at the location of the layer [21]. Resonance fluorescence lidar instrumentation is described in Section 11.6.

In summary, lidar researchers have developed four ways to remotely sense winds, with applications from the surface up to ~100 km: coherent Doppler lidar; eddy correlation; direct detection Doppler (the edge, double edge, and fringe imaging techniques); and resonance fluorescence wind lidar.

11.3 Rayleigh Lidar

Rayleigh lidar is an elastic backscatter technique. Its name comes from its main simplifying assumption: the signal is caused only by molecular Rayleigh scattering,

because the measurements are made in a region of the atmosphere that is free of aerosols, taken to be above 25 or 30 km altitude. Two more assumptions are then invoked: the atmosphere obeys the ideal gas law (which is quite accurate in the upper atmosphere), and it is in hydrostatic equilibrium, which means that the pressure at any altitude it due to the weight of the air above that altitude. If hydrostatic equilibrium applies, the differential of pressure P due to a differential of height is expressed as

$$dP = -N(h)mg(h)dh, \tag{11.6}$$

where $N(h)$ is the number density of air molecules (m^{-3}), m is the mass of an air molecule (4.8×10^{-26} kg), and $g(h)$ is the acceleration due to gravity (ms^{-2}). The last term varies according to the inverse square law of gravitation. At a midlatitude sea level location, g is 9.80 ms^{-2}. As a function of altitude, $g(h) = 9.80[6{,}371 / (6{,}371 + h)]^2$ where h is in km and 6,371 km is the radius of Earth. The parameter m also varies slightly with altitude above about 90 km, but that effect is ignored here. Integrating Eq. (11.6) over h,

$$P(h_{\max}) - P(h) = -\int_h^{h_{\max}} N(h')mg(h')dh'. \tag{11.7}$$

where $h_{\max}$ is the maximum height and starting point of the analysis, determined by the SNR of the lidar signal. The retrieval of the temperature profile proceeds step by step from $h_{\max}$ downward. Substituting the ideal gas law expression for pressure $P(h) = n(h)kT(h)$ into Eq. (11.7) and rearranging yields the relation

$$T(h) = T(h_{\max})\frac{N(h_{\max})}{N(h)} + \frac{1}{k}\int_h^{h_{\max}} \frac{N(h')}{N(h)} mg(h')dh'. \tag{11.8}$$

Because the Rayleigh $\beta(h)$ is proportional to the number density of air molecules $N(h)$, so is the range-corrected lidar signal $X(h)$ and the number densities appear only in ratios, so the $X(h)$ values can be used directly in the analysis as proxies for the densities, without calibration. However, an initial guess of the temperature $T(h_{\max})$ is required.

Equation (11.8) often appears in the literature with the factor before the integral as $1/R$ where R is the gas constant, 8.314 J/mol-K, rather than the inverse of the Boltzmann constant k. The difference is in the definition of the mass m. In Eq. (11.8) it is the mass of one molecule, whereas when R is used, it is the mass of one mole of air, which is 2.896×10^{-2} kg. The gas constant R is the product of the Avogadro constant (6.022×10^{23} mol^{-1}) and the Boltzmann constant. Equation (11.8) is often referred to as the Chanin–Hauchecorne (or CH) Rayleigh lidar temperature, because of a classic 1980 paper [22]. As is usual in lidar, the foregoing derivation was performed in the mathematical language of continuous variables, but the actual analysis algorithms must be in discrete variables. In this case, the algorithm moves from the highest altitude downward assuming isothermal layers of thickness Δh, and the measured $X(h)$ values represent the layer averages of density. The algorithm then becomes

$$T_i = \frac{N_o}{N_i}T_0 + \frac{1}{N_i k}\sum_{k=0}^{i} \bar{N}_k mg_k \Delta h. \tag{11.9}$$

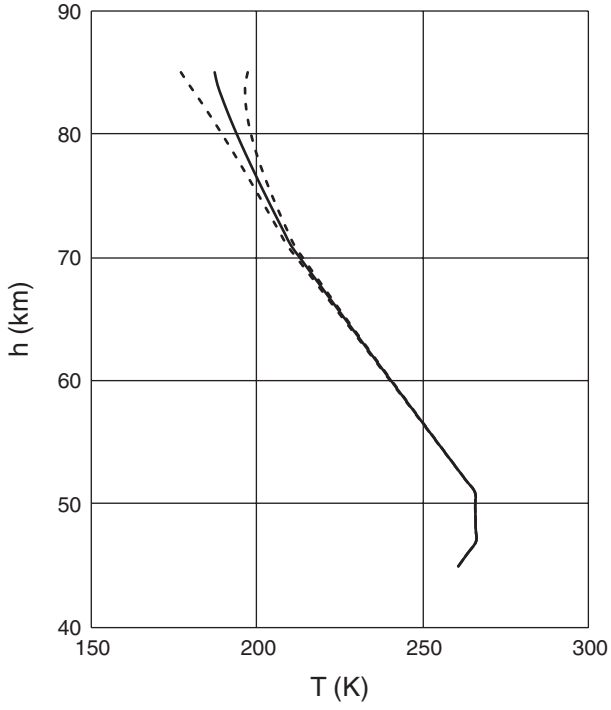


Figure 11.7 Rayleigh temperature profile retrievals. The initial guess for the solid line was the correct temperature; for the dashed lines, the initial guesses were 10 K lower and higher.

There is a subtlety in this algorithm, because N decreases exponentially with altitude and linear averages will cause serious errors. As shown by Behrendt [23], the appropriate average is

$$\bar{N}_k = \frac{N_k - N_{k-1}}{\ln[N_k / N_{k-1}]}. \tag{11.10}$$

Stepping downward in altitude, the first term in Eq. (11.9) rapidly decreases as the second increases. This fact causes the retrieved temperature profile to converge to the true profile rapidly when the initial guess $T(h_{\max})$ is wrong. This feature can be easily illustrated with the Standard Atmosphere profiles plotted in Figures 1.2 and 1.3. Figure 11.7 was generated by implementing Eqs. (11.9) and (11.10) in a spreadsheet, with the highest altitude at the top. The actual densities (kg/m^3) were used for the N values, and retrievals were generated for the three initial guesses at the top temperature: the correct value and 10 K above and below it.

Rayleigh lidars are most often used at night, when their main measure of merit is the power-aperture product, so they use powerful lasers and large receiver telescopes. The temperature profile in the upper regions of the atmosphere is slowly varying, so integration times are usually measured in hours. GTRI's first successful lidar, known as Megalidar, used a receiver telescope with a diameter of 2.54 m, which was the largest reported lidar receiver at the time. The lidar transmitted 320 mJ pulses of 532-nm light at 16.7 Hz, for a power-aperture product of 27 W-m^2 [24].

The CH analysis method was used with Megalidar data, and during the first two weeks of operation, the investigators observed a sudden stratospheric warming, also known as a *stratwarm*, and the retrieved lidar temperature profiles compared well with data from spaceborne microwave temperature sounders. The CH algorithm is the oldest and best known for analyzing Rayleigh temperature lidar data, but other more sophisticated algorithms have been developed with desirable features such as not requiring isothermal layers [25, 26].

11.4 Differential Absorption Lidar

Differential absorption lidar is another elastic backscatter technique, in which the signals at the two wavelengths are caused by backscatter from molecules and aerosols. DIAL's basis is illustrated in Figure 3.16 and a specific application to ozone is illustrated in Figure 3.23. For addressing the anthropogenic problems in Earth's atmosphere described in Chapter 1, DIAL is an essential lidar technique. In principle, it can be applied to water vapor or any trace gas that has an absorption feature in a window region, although there are complicating factors. GHGs with absorptions in windows include CO_2, CH_4, and N_2O, and pollutants include SO_2 and Hg vapor. Tropospheric O_3 is an urban pollutant, but stratospheric O_3 enables life forms on Earth by blocking the intense UV solar radiation, and both regions are monitored with DIAL. Near-real-time water vapor profiles would be invaluable for forecasting the local summertime convective storms that often disrupt commercial aviation, and DIAL is a promising technique for affordable, eye safe, day and night monitoring of them.

The DIAL algorithm is easily derived. The lidar equations for the "on" and "off" wavelengths are

$$P_{\text{on}}(R) = \frac{C_{\text{on}}}{R^2}\beta(r)\exp\left[-2\int_0^R \alpha_{\text{on}}(r)dr\right] \text{ and} \tag{11.11}$$

$$P_{\text{off}}(R) = \frac{C_{\text{off}}}{R^2}\beta(r)\exp\left[-2\int_0^R \alpha_{\text{off}}(r)dr\right]. \tag{11.12}$$

Note that the symbol α (m^{-1}) is used here for the extinction coefficient, because σ (m^2) is universally used in spectroscopy for the molecular absorption cross section. Finding the ratio of Eq. (11.12) to Eq. (11.11) yields

$$\frac{P_{\text{off}}(R)}{P_{\text{on}}(R)} = \frac{C_{\text{off}}}{C_{\text{on}}}\exp\left[2\int_0^R [\alpha_{\text{on}}(r) - \alpha_{\text{off}}(r)]dr\right], \tag{11.13}$$

assuming the atmospheric backscatter coefficient is the same at the two wavelengths. As pointed out in Chapter 10 in connection with the slope method, the extinction coefficient is exposed by finding the range derivative of the natural logarithm, which results in the relation

$$\Delta\alpha(R) = \frac{1}{2}\left[\frac{d}{R}\ln\left(\frac{P_{\text{off}}(R)}{P_{\text{on}}(R)}\right)\right], \tag{11.14}$$

where $\Delta\alpha = \alpha_{on} - \alpha_{off}$. The number density N (molecules/m^3) of a gas is often the desired quantity, and if the absorption cross section is denoted by σ (m^2) then the extinction coefficient α is $N\sigma$, and the number density is

$$N(R) = \frac{1}{2\Delta\sigma}\left[\frac{d}{dR}\ln\left(\frac{P_{off}(R)}{P_{on}(R)}\right)\right]. \tag{11.15}$$

Equation (11.14) shows that DIAL is a self-calibrating technique: All instrument constants have been removed by the sequential operations of finding the logarithm and differentiating with respect to range, as they were in the slope method. Note also that range correction is not necessary because the ratio of the signals is used. In discrete mathematics, Eq. (11.15) becomes

$$N(R) = \frac{1}{2\Delta\sigma\Delta R}\ln\left[\frac{P_{off}(R+\Delta R)}{P_{off}(R)}\cdot\frac{P_{on}(R)}{P_{on}(R+\Delta R)}\right]. \tag{11.16}$$

The derivation of Eq. (11.16) was simplified for clarity by ignoring the facts that the atmospheric backscatter coefficients at the two wavelengths may be different, and that differential extinction may be present due to air molecules, aerosols, or interfering gases. In total, there are four DIAL correction terms that may be important [27].

The number density of trace gas molecules is sometimes not the desired measure of concentration. Other units are mass density and mixing ratio. Mass density can be found by multiplying the number density by the gas's molecular weight in AMU and the weight of 1 AMU, which is 1.6605×10^{-27} kg. Mass density is usually expressed as μg/m^3. Mixing ratios have the advantage that they are unchanged when temperature and pressure change. By weight, mixing ratios are usually kg/kg or g/kg. By volume, the units are usually parts per million by volume (ppmv) or parts per billion by volume (ppbv). These are calculated from number density as N/N_{air}, where N_{air} at STP is 2.55×10^{25} molecules/m^3 with STP defined as the International Standard Metric Conditions that are used throughout this book. However, DIAL is sensitive to N, not mixing ratio, so the conversion relies on having a profile of atmospheric density.

Equation (11.16) shows that DIAL has an unusual error propagation feature: A fractional error in the power measurements becomes an absolute error in the concentration because if $y = a\ln(x)$, then $\Delta y = a\Delta x / x$. The accuracy of the quantity in square brackets in Eq. (11.16) sets the *limit of detection* for the gas of interest, and it may be rather large because $\Delta\sigma$ is usually very small. The resolution ΔR is often chosen to be large for this reason, and tropospheric ozone lidar data are generally filtered to a resolution of hundreds of meters before analysis. An example calculation for an ozone lidar is left as a problem for the student.

Optimizing the choice of "on" and "off" DIAL wavelengths involves tradeoffs. The absorption cross section at λ_{on} must be low enough so that the lidar can reach its maximum range requirement, but the cross-sectional difference must be large enough to meet the range resolution requirement. The two wavelengths should be close together to minimize the correction terms mentioned after Eq. (11.16), and there should ideally be no interfering gas spectra at either wavelength (or the same absorption at both

wavelengths). Molecular absorption lines tend to have a temperature dependence due to the Boltzmann distribution of energy level populations, so λ_{on} should be selected as a line with minimal temperature dependence, if possible.

Once the two DIAL wavelengths have been selected, a major challenge in implementing the technique has always been in generating laser pulses with appropriate wavelengths, pulse energies, divergences, and PRF. Tunable dye lasers were used in many of the initial demonstrations, but they have high maintenance requirements and have fallen out of favor in lidar. SRS cells pumped by harmonics of Nd:YAG or Nd:YLF lasers are commonly used in ozone lidars, where the wavelengths are not critical because the absorption bands are so broad. Such cells are very reliable, and once they are optimized, they are treated as passive optical elements in lidar transmitters. Another option is to use a Ce:LiSAF laser, which is tunable from 284 to 299 nm. Optical parametric oscillators can be used to build tunable all-solid-state transmitters covering a wide range of lidar wavelengths, so they are the optimum solution in that sense, but engineering them is a sophisticated task. For example, OPO cavities are often illustrated as in Figure 5.18, but that configuration generates a beam with different divergences in plane and out of plane, one of which is very large. The RISTRA mentioned in Chapter 5 alleviates that problem.

In meteorology, the need for an affordable lidar water vapor profiler for deployment in national networks has long been recognized, and several approaches have been attempted, including both DIAL and Raman lidar. Water vapor DIAL in the 700–950 nm range has a long history, but giant-pulse lasers in that spectral region are eye hazards, and background light can be a problem because of the significant daytime sky radiance. The only eye safe solution in the NIR is the micro-pulse lidar, and researchers at several organizations worked over a period of decades to overcome affordability, reliability, and background issues with it. In 2015, they reported a fourth-generation design that was implemented in five identical field units [28]. The technology was micro-pulse DIAL based on NIR laser diodes operating at 828.2 nm (λ_{on}) and 828.3 nm (λ_{off}). The transmitter alternated wavelengths 60 times per second. The transmitted beam was shaped into an annulus by axicons and expanded by a 406-mm telescope (as shown in Figure 5.4) to yield a beam divergence of 56 μrad. The pulse energy was 5 μJ, and with the effective 114 mm diameter, the beam met the ANSI eye safety criteria at zero range. A high PRF (9 kHz) was used to minimize measurement times. To minimize alignment issues and overall size, the transmitter–receiver configuration was common optics, in which the transmitted beam used the inner part of the primary mirror and the received light used an outer annular area. The data system used photon counting to approach the statistical limit of SNR. To cover an altitude range from 0.5 to 3 km without exceeding the maximum count rate (5 MHz), two receiver channels were used. The near channel, with an FOV of 451 μrad, had an optical efficiency of 0.04, while the far channel had an FOV of 115 μrad and an optical efficiency of 0.23. The detectors were APDs with quantum efficiencies of 0.45. Background counts were minimized with a combination of interference filters and an etalon, which yielded optical bandwidths of 20 pm for the near channel and 14 pm for the far channel. The system was fiber coupled as much as possible to avoid alignment

issues, and it used commercial off-the-shelf parts to minimize costs. The laser pulse width was unusually long at 900 ns, and 150-m range bins were used so that they would be longer than the pulses. With averaging times of 1–10 minutes, the lidars obtained high-quality profiles to 3 km in a comparison test with other instruments. Meteorologists desire profiles to 10 km, and studies (including one at GTRI) have shown that an eye safe water vapor DIAL capability could be developed in the SWIR region around 1.6 μm, using Nd:YAG-pumped OPOs, but not at low cost.

Molecular backscatter is quite small at wavelengths in the SWIR–MWIR region, and aerosol backscatter generally decreases with wavelength, too. This is unfortunate for DIAL because many gases of interest, both GHGs and pollutants, have infrared absorption lines. One possible enhancement is to use coherent detection with DIAL, which has been demonstrated but is not commonplace. Another option is to use the integrated path differential absorption (IPDA) technique. Integrated path differential absorption is a remote sensing technique for gases that is based on pulsed lasers at "on" and "off" wavelengths, so it has much in common with DIAL. The major difference is that IPDA uses the backscatter from a hard target at the far end of its range for its signals, so it provides only a path-integrated measurement of the gas. For monitoring many GHGs from orbit with laser remote sensing, it is the only option. The German and French space agencies CNES and DLR have planned a 2024 launch of a small satellite using IPDA to globally monitor methane, the second most important anthropogenic GHG. The satellite has been dubbed MERLIN, which is an acronym formed from the name methane remote sensing lidar mission [29]. Methane remote sensing lidar mission was designed to measure the spatial and temporal gradients of atmospheric methane accurately enough to constrain fluxes significantly better than the existing observation network, in all seasons at all latitudes.

A sensitivity analysis for MERLIN was reported in 2011 [30]. The SWIR IPDA instrument design used a Nd:YAG-pumped OPO to generate an "off" line at 1645.846 nm and an "on" line at 1645.552 nm, transmitting 9-mJ pulses at 50-Hz PRF (25 Hz for the two-wavelength pair). With a 0.55-m diameter telescope, 506-km orbit altitude, and 50-km along-track averaging, MERLIN was expected to provide weighted average dry air volume mixing ratios of CH_4 with accuracies and precisions of 1%. The wavelength pair was chosen using line-by-line HITRAN calculations for an optimal optical depth and to avoid sensitivities to temperature and interfering gases. The "on" line was in a trough between two closely spaced methane lines, to reduce the laser stability requirements. The detector was an InGaAs APD with a quantum efficiency of 0.6, a gain of 10, an excess noise factor of 3.2, and an NEP value of 43 fW/Hz$^{1/2}$ for the detector–TIA combination. The expected SNR was calculated with Eq. (8.58). When there are no interfering gases, the IPDA equation for power on the detector, analogous to the lidar equation, can be written as

$$P = \frac{E_{\text{pulse}}}{\tau_{\text{pulse}}} k_T k_R \left(\frac{A}{R^2} \right) \rho \exp[-2(OD_0 + OD_{\text{gas}})], \tag{11.17}$$

where τ_{pulse} is an effective pulse width that includes the transmitted pulse width, the pulse stretching due to uneven terrain, and the detector–TIA impulse response time of

111 ns; the terrain reflectance ρ (sr^{-1}) is defined such that the albedo of a Lambertian surface is $\rho\pi$; and OD_0 is the optical depth due to aerosols, molecules, and clouds. Writing Eq. (11.17) for the "off" and "on" wavelengths, dividing P_{off} by P_{on}, and finding the natural logarithm yields the equation for the IPDA measured quantity, which is the differential atmospheric optical depth (DAOD), defined by

$$\mathrm{DAOD} = \frac{1}{2}\ln\left[\frac{P_{off}E_{on}}{P_{on}E_{off}}\right], \tag{11.18}$$

where the P values are the received powers at the two wavelengths and the E values are the corresponding transmitted energies, which are recorded for each outgoing pulse. The two wavelengths were carefully chosen so that only CH_4 contributes to the DAOD, but the quantity of scientific interest is not the DAOD but rather the weighted average dry air volume mixing ratio of CH_4, and there is one more step in data analysis to get it. Molecular spectra change with temperature and pressure, which vary with height in the atmosphere over wide ranges. The temperature sensitivity is due to the Boltzmann distribution of energy level populations, and the pressure sensitivity is due to the phenomenon of pressure broadening, which determines the widths of spectral lines in the lower atmosphere. For this reason, the differential absorption coefficient between the two lines (called $\Delta\alpha$ in the DIAL discussion) is not constant with altitude, hence the sensitivity of IPDA measurements has an altitude dependence described by a *weighting function*, and the DAOD values must be divided by the weighting function to obtain the weighted average dry air volume mixing ratio of CH_4. The shape of the weighting function depends on the choices of "on" and "off" wavelengths, and for MERLIN, the two wavelengths were chosen for optimum sensitivity in the lower troposphere [30].

11.5 Raman Lidar and HSRL

Raman lidar techniques are inelastic and they are all challenging, especially for daytime operation, because the Raman backscatter coefficients are orders of magnitude smaller than Rayleigh. However, great progress has been made with Raman lidar instrumentation over the years, and three standard Raman techniques have been developed: trace gas Raman, aerosol Raman, and temperature Raman. Because of their small signals, Raman lidars tend to have lasers with large pulse energies and receivers with large telescopes, narrow FOVs, and sophisticated optical filtering. Because Raman backscatter coefficients increase with the fourth power of the laser frequency, all Raman lidars operate in the UV-VIS region.

11.5.1 Trace Gas Raman Lidar

Trace gas Raman lidars transmit one wavelength, generally 532 nm or 355 nm, and receive Raman-scattered light from nitrogen and the trace gas, which is usually water vapor. The mixing ratio profile of the trace gas is found from the ratio of the two signals. This type of lidar was described in detail by Wandinger [31]. The lidar equation

requires two modifications for trace gas Raman: (1) the backscatter coefficient $\beta(h)$ becomes $N(h)(d\sigma / d\Omega)_\pi$, where N is the number density of the Raman scattering gas and $(d\sigma / d\Omega)_\pi$ is the Raman *differential scattering cross section* in the backward direction in m^2/sr; and (2) the extinction coefficient on the downward paths is at the Raman-shifted wavelength. The calculations of $(d\sigma / d\Omega)_\pi$ for the various Raman spectral lines are described in [31]. The mixing ratio $m(h)$ of water vapor relative to dry air is found from

$$m(h) = C \frac{P_{H_2O}(h) \exp\left[-\int_0^h \sigma_{N_2}(h')dh'\right]}{P_{N_2}(h) \exp\left[-\int_0^h \sigma_{H_2O}(h')dh'\right]}, \tag{11.19}$$

where the subscripts H_2O and N_2 refer the wavelengths of the respective Raman lines. The derivation of Eq. (11.19) is left as a problem for the student. A main difference in the two transmittance terms is caused by the wavelength dependence of Rayleigh scattering, so it must be corrected for by using the atmospheric density profile. The difference in aerosol extinction at the two wavelengths is often neglected. The calibration constant C accounts for differences in the optical efficiencies, quantum efficiencies, and detector and amplifier gains between the two channels, as well as the effective Raman cross sections of the two gases, which depend on the optical bandwidths of the channels because Raman spectra consist of bands of lines, as shown in Figure 3.18. The value of C is most often found by comparison with balloon-borne sondes, which makes the accuracy of the lidar dependent on other instruments. The stability of the calibration is another issue because the transmittance spectra of filters change with age, as illustrated in Chapter 6, and electronic gains may also change. The calibration errors are estimated to be on the order of 5%.

In Raman trace gas lidars, the relevant signal increases as the trace gas concentration increases. This is the opposite of DIAL, in which the "on" line signal decreases as the trace gas concentration increases. The atmospheric water vapor concentration tends to decrease rapidly with altitude and almost all of it is contained in the troposphere, which limits the maximum measurement altitudes of water vapor Raman lidar measurements to about 10 km. The daytime maximum altitude is usually lower because the sky background decreases the SNR.

11.5.2 Aerosol Lidar Techniques

The slope method, described in Section 10.2.3, addresses the age-old lidar problem of having one equation but several unknowns. It does not yield a range profile of the atmospheric extinction coefficient; it only finds a value in some range interval, where the simplifying assumption is that β and σ are constants in that interval. The reason for assuming that β is constant is to eliminate it with the mathematical operation of finding the range derivative, thereby reducing the number of unknowns to one, the atmospheric extinction coefficient σ. A more general version of Eq. (10.17) is

$$\frac{d}{dR}\ln[X(R)] = \frac{d}{dR}\ln[\beta(R)] - 2\sigma(R). \tag{11.20}$$

If $\beta(R)$ is known, in the sense that it can be accurately modeled, Eq. (11.20) can be solved for $\sigma(R)$. The coefficients β and σ both have two components, molecular and aerosol, so there are four unknowns. The molecular terms can be modeled but the aerosol terms cannot, so Eq. (11.20) cannot be solved for a simple elastic backscatter lidar. However, if $\beta(R)$ is only due to the molecular atmosphere (no aerosol contribution), then it can be modeled, and the molecular component of $\sigma(R)$ can also be modeled, reducing the number of unknowns to one. Two types of lidar have been developed in which one of the receiver channels records a signal that is caused by the molecular atmosphere alone: Raman aerosol lidar and HSRL.

A Raman aerosol lidar has receiver channels at two wavelengths, the elastic transmitted wavelength and the inelastic Raman wavelength. The Raman lidar signal is only due to air molecules (usually nitrogen) so it is proportional to the air density, and the Raman backscatter coefficient profile can be accurately modeled using the air density profile. The HSRL also has two receiver channels. One channel blocks the aerosol signal with a very narrow filter to record a signal due only to molecules, which again can be accurately modeled using the air density profile. High spectral resolution lidar also records the total signal without the blocking filter, so subtracting one signal from the other (with the proper scale factor) yields the aerosol signal. Because the two techniques are based on similar principles, the solution of Eq. (11.20) is nearly the same for both:

$$\sigma_a(R,\lambda_0) = \frac{\dfrac{d}{dR}\ln\dfrac{N_{\mathrm{Ra}}(R)}{X(R,\lambda_{\mathrm{Ra}})} - \sigma_m(R,\lambda_0) - \sigma_m(R,\lambda_{\mathrm{Ra}})}{1+\left(\dfrac{\lambda_0}{\lambda_{\mathrm{Ra}}}\right)^{a(R)}}, \tag{11.21}$$

where λ_0 is the transmitted wavelength, $N_{\mathrm{Ra}}(R)$ is the number density profile, the subscript *Ra* refers to Raman or Rayleigh depending on the type of lidar, and $a(R)$ is the Angstrom parameter profile [32]. For HSRL, there is only one wavelength, so the denominator becomes the number 2, and the two σ_m values are the same. For Raman lidar, the particle backscatter coefficient profile $\beta_a(R)$ can be found from a ratio of signals at λ_0 and λ_{Ra} (although an initial estimate is required at the starting altitude of the retrieval) and then the lidar ratio profile can be calculated as $S_a(R) = \sigma_a(R) / \beta_a(R)$. See [32] for more detailed derivations. For HSRL, the scattering ratio profile R_{sca} can be found directly as the ratio of total backscatter profile to the molecular backscatter profile.

The implementations of the two aerosol lidar concepts are quite different. The Arctic HSRL, shown in Figure 7.19, was an eye safe micropulse lidar operating at 532 nm. It used the common optics configuration, transmitting a circularly polarized beam for T/R switching, with a pulse energy of 150 μJ and 4-kHz PRF. The beam was expanded to 400 mm. With a receiver FOV of 45 μrad and an optical bandwidth of 8 GHz, it was essentially immune to the sky background. The HSRL had a photon counting data system and the crossover function $G(R)$ was engineered to discriminate against the nearby signal and allow photon counting to altitudes as low as 50 m (the system operated in the crossover region because there was only one received beam)

[33]. As mentioned in Chapter 7, HSRLs have operated unattended as Internet appliances for up to three years.

As an example of Raman aerosol lidars, the Polly series of lidars was developed by the Leibniz Institute for Tropospheric Research (TROPOS), and dozens of identical systems have been fielded in a network that features central data processing, as AERONET does [34]. The characteristics of an advanced generation, the Polly^{XT}, were described by Engelmann et al. in 2016 [35]. Raman lidars must use large pulse energies because the backscatter is weak, so they are not eye safe. The Polly^{XT} transmits 180, 110, and 60 mJ at 1,064, 532, and 355 nm, respectively, and the transmit beam diameter is 45 mm, so it is hazardous. The lidar can be interfaced to a safety radar for this reason. The biaxial receivers have six wavelengths, the elastic ones plus nitrogen Raman at 607 nm, water vapor Raman at 407 nm, and nitrogen Raman at 387 nm. The receiver optical bandwidths are all either 0.3 or 1.0 nm. Measurements are made down to 100-m altitude by using two receiver telescopes, called Far, with 300-mm diameter and 1.0-mrad FOV and Near, with 50-mm diameter and 2.2 mrad FOV. Both receivers use Fabry lenses to image their telescope objectives onto the detectors. The optical components, which include a fixture for using the Δ90 depolarization calibration method, are mounted on a custom carbon fiber optical bench that aims the FOVs 5 degree off the zenith to avoid the problem with oriented ice crystals. Polly lidars are used for aerosol classification based on measurements of the lidar ratio S_a, the Angstrom exponent a, and the depolarization ratio δ. For Raman lidars, Polly receivers have unusually large bandpasses and FOVs, which makes them susceptible to the sky background, especially in the Raman channels, and in fact the 407-nm channel is switched off during daytime for that reason. The detectors are all PMTs, and the lidar uses a paralyzable photon counting system. To achieve high count rates, the developers measured the count rate correction in the laboratory, and they allowed photocount rates as high as 60 Mcps. Calculation of a Polly background count rate is left as a problem for the student.

11.5.3 Raman Temperature Lidar

As mentioned in Section 11.3, Rayleigh temperature lidar is restricted to the aerosol-free region of the atmosphere above 30 km. For this reason, there was a need for another lidar technique that could operate lower down in the presence of aerosols, and Raman temperature lidar fulfilled that need. It is a Boltzmann technique, based on the temperature dependence of line strengths in the rotational Raman spectrum. Behrendt [23] has provided a concise explanation of the theory, analysis of errors, wavelength optimization, and engineering considerations for the technique, along with an example of an outstanding system that combined Rayleigh and Raman temperature measurement techniques. As shown in Figure 3.18, there is a cluster of rotational lines around the elastic backscatter line at 532 nm. Those lines are due to both N_2 and O_2, and they span a wavelength region several nm wide. The technique is based on selecting two wavelength regions within the rotational spectrum with interference filters, such that the signal through one filter increases with temperature and the other decreases. The ratio of the two signals is then related to the temperature, although a

calibration against a radiosonde is also required. This technique was difficult to implement because of stringent requirements on the optical filters, but filter technology eventually improved to meet the need, and Raman temperature lidars became capable of tropospheric measurements with uncertainties of ±1 K with a few minutes of averaging and 100-m vertical resolution. The elastic backscatter line is several orders of magnitude greater than the rotational Raman lines, so a blocking factor of 10^{-7} is required to keep it from corrupting the Raman receiver channels. The two spectral regions are usually chosen in the anti-Stokes part of the spectrum so that tilt tuning can be used to optimize the filter center wavelengths, as explained in Chapter 6.

11.6 Resonance Fluorescence Lidar

Other than DIAL and IPDA, the techniques described above implement spectral sensitivity in the receiver, while the transmitter has a single wavelength. Resonance fluorescence lidars are the opposite, where the transmitted wavelengths must span an atomic spectral feature, either with a few discrete wavelengths or by scanning, and the receiver has a bandwidth broad enough to include the spectral range of those wavelengths. This difference leads to stringent spectral requirements on the laser transmitters. In addition, the atoms of interest are all at high altitudes, 80 km and above, so the lidars must have a high power-aperture product, and daytime operation is challenging because the signals are weak and the fluorescence wavelengths are in the UV-VIS region. For these reasons, resonance fluorescence lidars make use of all the engineering principles and tradeoffs in the previous chapters for optimizing SNR by maximizing the signal while minimizing the background. They also use photon counting, to get as close to the statistical limit of SNR as possible.

Resonance fluorescence lidar engineers have made steady progress over the decades, and their technology is now generally more sophisticated than in other types of lidar. In addition, it has become rugged enough to enable transportable lidar systems in shelters that can operate in hostile conditions such as the polar regions. Progress in daytime operation has required very small FOVs to minimize background photocounts, so the necessity to maintain transmitter–receiver alignment has put stringent stability requirements on the mechanical systems. A chronology of progress in resonance fluorescence lidar technology, with example systems, is presented in [21], and some of that technology is briefly described in the paragraphs that follow. The resonance fluorescence lidar equation is also more complicated than the equation for Rayleigh and Mie scattering, largely because the scattering process is not instantaneous. If an atom that has been excited by absorption of a laser photon is de-excited by a collision, it will not emit a signal photon. If the lifetime of an excited state is too long relative to the laser pulse, saturation may occur, because a significant fraction of the atoms may already be pumped up to the excited state. Because of such phenomena, extra factors may have to be considered, including the laser spectral shape and linewidth; the laser and signal polarizations; extinction in the layer of atoms being probed; the finite lifetime of the excited state of the atoms; and the temporal shape of the laser pulse [21].

The temperature Doppler technique using the hyperfine spectrum of the sodium (Na) D_2 line, as illustrated in Figure 3.22, requires at least three laser lines. The whole spectral feature is only about 5 GHz wide, so the laser lines must be narrow and at precisely controlled wavelengths. The laser transmitter is usually locked to an absorption feature in a sodium vapor cell to provide an absolute reference wavelength, which is necessary if winds are to be measured by tilting the lidar off zenith. As an example of the instrumentation, a large sodium layer wind and temperature lidar at the University of Illinois was based on a CW ring dye laser pumped by a CW Nd:YVO4 laser with its output locked to a sodium vapor cell, an acousto-optic modulator to add shifts to two nearby wavelengths, and finally, a pulsed dye amplifier pumped by a pulsed Nd:YAG laser. The transmitter produced 6 ns, 300-mJ pulses at 30-Hz PRF, and the laser line width was 60 MHz. The researchers also developed a daytime measurement capability by using a Fabry–Perot etalon to narrow the receiver's optical bandwidth. This type of lidar achieved ±1 K temperature accuracy in the Na layer, with 1-km altitude resolution. The system required a stable laboratory environment for all the laser equipment. Eventually, solid-state transmitters for Na wind and temperature lidars were realized, by using a fortuitous feature of the Nd:YAG energy levels: There are two possible laser transitions at 1,064 and 1,319 nm, and mixing them produces 589 nm. Using this concept, a transportable narrowband Nd:YAG lidar transmitter was developed that was tunable and locked to an Na vapor cell, and the lidar was deployed at ALOMAR.

A completely different temperature measurement technique was developed around the year 2000 by exploiting the Boltzmann temperature dependence of energy level populations in iron (Fe) atoms. The Fe Boltzmann lidar had all solid-state transmitters and less stringent laser linewidth constraints, so it was more rugged than the Na lidars of that time. It was successfully deployed on aircraft and in the polar regions, and it was capable of daytime operation. The Fe spectral lines used were in the UV, at about 372 nm and 374 nm. The lidar had two transmitters and two receivers. The transmitters used seed lasers and Alexandrite lasers to generate pulses at 744 nm and 788 nm that were then doubled. The beams were expanded to reduce their divergence to 350 μrad. Sufficiently accurate wavelength control was achieved with commercial wavemeters. The lasers were multi-mode and had widths of about 800 MHz. The photon-counting receivers each had a 0.4-m telescope, an interference filter, and a pressure-tuned Fabry–Perot etalon to enable daytime operation. Accuracies of ±1 K were easily achieved at night, although they took hours of integration during daytime. The lidar had a clever feature: It used a higher optical bandwidth during nighttime to achieve a higher receiver optical efficiency by removing a filter. Another feature of the Fe Boltzmann lidar was that its data were analyzed with the Rayleigh temperature lidar algorithm in the 30–75 km region, using the Fe temperature to initialize the retrieval. The two techniques together provided temperature profiles from 30 to 110 km.

Another approach to building a robust, transportable Doppler temperature lidar was to exploit the potassium (K) spectrum rather than Na, because the D_1 and D_2 lines of K are at 770 and 767 nm, which are within the tuning ranges of several tunable solid-state lasers. This technique was realized with an Alexandrite ring laser, locked to a tuning and scanning external cavity diode laser. The transmitter produces

single-mode laser pulses with <20 MHz bandwidth, a temporal pulse width of 250 ns, a pulse energy of 150 mJ, and a PRF of 33 Hz. The transmitter has a 50 pm tuning range, and the pulse width and wavelength are monitored with a Fabry–Perot etalon using fringe imaging. The receiver diameter is 80 cm. With 3 hours of integration time at night, an uncertainty of ±2 K was achieved with 200-m vertical resolution. Daytime operation was also possible, by using a Faraday anomalous dispersion optical filter (FADOF) with a 10-pm width, along with an FOV of 192 urad. An APD was used as the detector, for higher QE than a PMT.

11.7 Further Reading

C. Weitkamp, Ed., *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere*. New York: Springer, 2005.

This book has 14 chapters covering basic scattering phenomena (aerosol scattering, depolarization, visibility and clouds, and multiple scattering) as well as chapters on Doppler wind lidar; HSRL; DIAL techniques for ozone, pollutants, and water vapor; Raman lidar including aerosol extinction; resonance fluorescence lidar; and temperature profiling techniques.

T. Fujii and T. Fukuchi, Eds., *Laser Remote Sensing*. New York: Taylor & Francis, 2005.

This book includes chapters on elastic backscatter lidar, trace gas monitoring, resonance fluorescence, and wind lidar. The chapters on resonance fluorescence lidar and wind lidar are comprehensive. The introduction, by Claus Weitcamp, includes a taxonomy of atmospheric scattering processes and nomenclature, with definitions of every scattering, absorption, and fluorescence term that a lidar researcher will encounter.

11.8 Problems

11.8.1 A classic example of a transcendental equation is $\theta = \cos(\theta)$. Solve this equation with an error of ±1% or less by using the Newton–Raphson method, with an initial guess that $\theta = 0.6$ radian.

11.8.2 One of the transmittance peaks in Figure 11.4 is at 500 nm. Where are the others?

11.8.3 Two differential absorption techniques are used in laser remote sensing, DIAL and IPDA. Elucidate their basic data analysis relations by (a) deriving Eq. (11.16) from Eq. (11.15); and (b) deriving Eq. (11.18) from Eq. (11.17).

11.8.4 Tropospheric ozone lidars often operate at 288.9 nm and 299.1 nm because these wavelengths can be generated in SRS cells with D2 and H2 pumped by the

fourth harmonic of the Nd:YAG laser at 266 nm. GTRI's ozone lidar known as NEXLASER was designed for a vertical resolution of 300 m and an ozone uncertainty of 10 ppbv. What was the maximum relative error for the term in square brackets in Eq. (11.16)? The ozone absorption cross sections are 1.59×10^{-19} at 289 nm and 4.55×10^{-19} at 299 nm, so $\Delta\sigma$ is 1.14×10^{-18} cm^2.

11.8.5 What is the bandwidth of MERLIN's detector–TIA combination? What is its NEP in watts? The detector's impulse response time is 111 ns and the spectral NEP is 43 fWHz-1/2.

11.8.6 Derive Eq. (11.19) for water vapor Raman lidar.

11.8.7 The lidars known as Polly^{XT} have photon counting data systems with paralyzable counters. The maximum photon count rate is 60 Mcps, which is said to result in correction factors less than 1.3. Assume the dead time τ_d is 4 ns. Are these numbers consistent with Eq. (9.19)? Are such large count rates consistent with the design parameters? Use Eq. (2.14) with the parameters for the 532 nm Near receiver listed in the following table, which were taken from the literature. Assume a sky radiance of 40 W/m^2-μm-sr, a receiver efficiency k_R of 0.3, and a detector quantum efficiency η of 0.2 to calculate the daytime background count rate.

Polly^{XT} 532 Near receiver design parameters

Parameter	Value	Units
Diameter	0.05	m
FOV	2.2	mrad
Optical bandwidth	1.0	nm

11.8.8 Derive Eq. (11.21) from Eq. (11.20) for the HSRL.

References

[1] W. H. Hunt, D. M. Winker, M. A. Vaughan et al., "CALIPSO Lidar Description and Performance Assessment," *Journal of Atmospheric and Oceanic Technology*, vol. 26, pp. 1214–1228, 2009.

[2] Z. Liu, J. Kar, S. Zeng et al., "Discriminating between Clouds and Aerosols in the CALIOP Version 4.1 Data Products," *Atmospheric Measurement Techniques*, vol. 12, pp. 703–734, 2019.

[3] S. A. Young, M. A. Vaughan, A. Garnier et al., "Extinction and Optical Depth Retrievals for CALIPSO's Version 4 Data Release," *Atmospheric Measurement Techniques*, vol. 11, pp. 5701–5727, 2018.

[4] S. A. Young and M. A. Vaughan, "The retrieval of Profiles of Particulate Extinction from Cloud Aerosol Lidar Infrared Pathfinder Satellite Observations (CALIPSO) Data: Algorithm Description," *Journal of Atmospheric and Oceanic Technology*, vol. 26, pp. 1105–1119, 2009.

[5] A. H. Omar, J. G. Won, D. M. Winker et al., "Development of Global Aerosol Models Using Cluster Analysis of Aerosol Robotic Network (AERONET) Measurements," *Journal of Geophysical Research*, vol. 110, D10S14 (14 pp.), 2005.
[6] S. W. Henderson, P. Gatt, D. Rees, and M. Huffaker, "Wind Lidar," in *Laser Remote Sensing*, T. Fujii and T. Fukuchi, Eds. Boca Raton: Taylor & Francis, 2005, pp. 469–722.
[7] E. W. Eloranta and D. K. Forrest, "Volume Imaging Lidar Observation of the Convective Structure Surrounding the Flight Path of an Instrumented Aircraft," *Journal of Geophysical Research*, vol. 97, pp. 18383–18394, 1992.
[8] University of Wisconsin Lidar Group. [Online]. Available: http://lidar.ssec.wisc.edu/index.htm. 1/10/22. [Accessed October 1, 2022].
[9] J. L. Schols and E. W Eloranta, "Calculation of Area-Averaged Vertical Profiles of the Horizontal Wind Velocity from Volume-Imaging Lidar Data," *Journal of Geophysical Research*, vol. 97, pp. 18395–18407, 1992.
[10] E. M. Patterson, D. W. Roberts, and G. G. Gimmestad, "Initial Measurements Using a 1.54 Micron Eyesafe Raman Shifted Lidar," Letters to the Editor, *Applied Optics*, vol. 28, pp. 4978–4981, 1989.
[11] S. D. Mayor and S. M. Spuler, "Raman-Shifted Eye-Safe Aerosol Lidar," *Applied Optics*, vol. 43, pp. 3915–3924, 2004.
[12] Atmospheric lidar research Group. [Online]. Available: https://physics.csuchico.edu/lidar/marine/. [Accessed March 9, 2022].
[13] Vaisala WindCube. [Online]. Available: www.vaisala.com/en/wind-lidars/wind-energy/windcube. [Accessed October 1, 2022].
[14] Lockheed Martin WindTracer. [Online]. Available: www.lockheedmartin.com/en-us/products/windtracer.html. [Accessed October 1, 2022].
[15] C. L. Korb, B. M. Gentry, and C. Y. Weng, "Edge Technique: Theory and Application to the Lidar Measurement of Atmospheric Wind," *Applied Optics*, vol. 31, pp. 4202–4213, 1992.
[16] E. Hecht, *Optics*, 5th ed., London: Pearson, 2017.
[17] B. M. Gentry, H. Chen, and S. X. Li, "Wind Measurements with 355-nm Molecular Doppler Lidar," *Optics Letters*, vol. 25, pp. 1231–1233, 2000.
[18] G. Baumgarten, "Doppler Rayleigh/Mie/Raman Lidar for Wind and Temperature Measurements in the Middle Atmosphere up to 80 km," *Atmospheric Measurement Techniques*, vol. 3, pp. 1509–1518, 2010.
[19] A. G. Straume et al., "ESA's Space-Based Doppler Wind Lidar Mission Aeolus – First Wind and Aerosol Product Assessment Results," 29th International Laser Radar Conference, 2020. [Online]. Available: www.epj-conferences.org/articles/epjconf/abs/2020/13/epjconf_ilrc292020_01007/epjconf_ilrc292020_01007.html. [Accessed February 2, 2022].
[20] ESA ADM-Aeolus Science Report SP-1311, 122 pp., April 2008. [Online]. Available: https://earth.esa.int/pi/esa?id=3409&sideExpandedNavigationBoxId=Aos&cmd=image&topSelectedNavigationNodeId=AOS&targetIFramePage=/web/guest/pi-community/apply-for-data/ao-s&ts=1496439496255&type=file&colorTheme=03&sideNavigationType=AO&table=aotarget. [Accessed: February 2, 2022].
[21] X. Chu and G. C. Papen, "Resonance Fluorescence Lidar for Measurements of the Middle and Upper Atmosphere," in *Laser Remote Sensing*, T. Fujii and T. Fukuchi, Eds. New York: Taylor and Francis, 2005, pp. 179–432.
[22] A. Hauchecorne and M. Chanin, "Density and Temperature Profiles Obtained by Lidar between 35 and 70 km," *Geophysics Research Letters*, vol. 7, pp. 565–568, 1980.

[23] A. Behrent, Temperature Measurements with Lidar, in *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere*, C. Weitcamp, Ed. New York: Springer, 2005, pp. 273–305.

[24] D. W. Roberts, G. G. Gimmestad, A. K. Garrison et al., "Design and Performance of a 100-Inch Lidar Facility," *Optical Engineering*, vol. 30, pp. 79–87, 1991.

[25] J. Khanna, J. Bandoro, R. J. Sica, and C. T. McElroy, "New Technique for Retrieval of Atmospheric Temperature Profiles from Rayleigh-Scatter Lidar Measurements Using Nonlinear Inversion," *Applied Optics*, vol. 51, pp. 7945–7952, 2012.

[26] R. J. Sica and A. Haefele, "Retrieval of Temperature from a Multiple-Channel Rayleigh-Scatter Lidar Using an Optimal Estimation Method," *Applied Optics*, vol. 54, pp. 1872–1889, 2015.

[27] G. G. Gimmestad, "Differential-Absorption Lidar for Ozone and Industrial Emissions," in *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere*, C. Weitcamp, Ed. New York: Springer, 2005.

[28] S. M. Spuler, K. S. Repasky, B. Morley, et al., "Field-Deployable Diode-Laser-based Differential Absorption Lidar (DIAL) for Profiling Water Vapor," *Atmospheric Measurement Techniques*, vol. 8, pp. 1073–1087, 2015.

[29] MERLIN (Methane Remote Sensing Lidar MIssion): an overview. [Online]. Available: www.researchgate.net/publication/280090102_MERLIN_Methane_Remote_Sensing_Lidar_MIssion_an_overview. [Accessed February 6, 2022].

[30] C. Kiemle, M. Quatrevalet, G. Ehret, et al., "Sensitivity Studies for a Space-Based Methane Lidar Mission," *Atmospheric Measurement Techniques*, vol. 4, pp. 2195–2211, 2011.

[31] U. Wandinger, "Raman Lidar", in *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere*, C. Weitcamp, Ed. New York: Springer, 2005, pp. 241–271.

[32] A. Ansmann and D. Muller, "Lidar and Atmospheric Aerosol Particles", in *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere*, C. Weitcamp, Ed. New York: Springer, 2005, pp. 105–141.

[33] E. W. Eloranta, "High Spectral Resolution Lidar," in *Lidar: Range-Resolved Optical Remote Sensing of the Atmosphere*, C. Weitcamp, Ed. New York: Springer, 2005, pp. 143–163.

[34] Pollynet. [Online]. Available: https://polly.tropos.de. [Accessed February 11, 2022].

[35] R. Engelmann, T. Kanitz, H. Baars, et al., "The Automated Multiwavelength Raman Polarization and Water-Vapor Lidar PollyXT: The neXT Generation," *Atmospheric Measurement Techniques*, vol. 9, pp. 1767–1784, 2016.

Appendix A The Klett Retrieval

In the absence of absorbing gases, the instantaneous received power $P(R)$ from scatterers located at a range R from the lidar receiver can be calculated using the single-scattering elastic backscatter lidar equation, which is

$$P(R) = \frac{P_o C}{R^2}[\beta_m(R) + \beta_a(R)]\exp\left[-2\int_0^R [\sigma_m(r) + \sigma_a(r)dr\right], \tag{A.1}$$

where

P_0 is the transmitted power,
C is the calibration constant of the lidar,
$\beta_m(R)$ is the molecular volume backscatter coefficient (1/m·sr),
$\beta_a(R)$ is the aerosol volume backscatter coefficient (1/m·sr),
$\sigma_m(R)$ is the molecular extinction coefficient (1/m), and
$\sigma_a(R)$ is the aerosol extinction coefficient (1/m).

The age-old problem in lidar is that four variables contribute to one signal. Many approaches to solving Eq. (A.1) had been developed over the decades, including making restrictive assumptions about the atmosphere, choosing special measurement scenarios, and using lidar techniques in which some variables are eliminated. The variables β_m and σ_m are generally assumed to be known because they only depend on wavelength and the atmospheric molecular density, which can be calculated from upper air data or models, but there are still two remaining unknowns, σ_a and β_a.

In the early days of lidar, much effort was expended toward finding a general solution to Eq. (A.1) in the hope that a ground-based elastic backscatter lidar could be used to measure the entire profile of the aerosol extinction coefficient $\sigma_a(R)$ from the surface to 30-km altitude. Several classic papers on this effort were written during the 1972–1985 period, and two of them were among the most-cited papers in the first 50 years of *Applied Optics*, but taken as a whole, they do not tell a coherent story and they are somewhat baffling to the uninitiated. For this reason, one of us (David Roberts) re-derived all the equations and algorithms in those classic papers. The purpose of this appendix is to explain the so-called Klett retrieval (also called an inversion or a solution to the lidar equation) in full mathematical detail while also clarifying the contributions of the other authors. The chronology of relevant papers is as follows:

(1) In 1972, in the days of ruby lasers at 694 nm and data recorded as photographs of oscilloscope traces, Fernald, Herman, and Reagan published a paper describing what they called "a new analytic solution" to the lidar equation that enabled them to derive vertical profiles of the aerosol extinction coefficient, provided that the optical depth of the aerosol layer was known from sun photometer data [1]. They first derived useful relationships for the one-component case (aerosols) by mathematical analyses. They continued their analyses for the two-component case (molecules and aerosols) and derived a transcendental equation that they solved iteratively. Their method relied on a relationship between the aerosol extinction and backscatter coefficients defined by the lidar ratio, $S_a = \sigma_a/\beta_a$, which was assumed to be independent of altitude. The lidar ratio was not known a priori, but the method yielded a value for it. The method was illustrated with examples in both clear and dusty conditions.

(2) In 1981, Klett published his paper on a stable analytical solution for analyzing lidar data [2]. The paper dealt with a known solution to a Bernoulli equation, in which the extinction and backscatter coefficients were related by a power law. Klett's contribution was to point out that, whereas previous unsuccessful efforts to implement the recursive solution started at the lidar and moved outward, the method was mathematically stable if one started at the far end and moved inward. The paper is somewhat difficult to read, partly because of a cumbersome notation, which uses the variable $S(R) = \ln[X(R)]$, where $X(R)$ is the range-corrected signal, and it is written in continuous variables, with no explicit algorithms. Only "turbid" atmospheres were considered, meaning one-component atmospheres. Klett used the results of numerous simulations to demonstrate stability, but the method required a priori knowledge of the extinction-to-backscatter relationship as well as the value of the extinction coefficient at the starting range.

(3) Apparently inspired by Klett's paper, in 1984 Fernald published what he called a "restatement" of his earlier paper [3]. It is not actually a restatement; rather it defines the actual algorithms, in discrete math, that correspond to Klett's equations, for the two-component case. Fernald also showed how the algorithms reduced to simpler forms for a one-component atmosphere.

(4) Klett updated his results in 1985 by including a range-dependent extinction-to-backscatter relationship and "the effect of a background of Rayleigh scatterers," meaning a two-component atmosphere [4].

(5) In 1985, Sasano, Browell, and Ismail also derived modified algorithms in which the lidar ratio is a function of range [5]. Their derivation was in continuous variables, but they translated their results into discrete variables so that they are modifications of the algorithms presented by Fernald in 1984. They also provided some useful detail on re-casting the lidar equation as a Bernoulli equation.

Over the years, the method explained here has become known as the Klett retrieval even though the lidar community benefitted from the contributions of several other authors in developing the data analysis algorithms. The derivations of analytic solutions for the profiles $\sigma_a(R)$ and $\beta_a(R)$ presented here do not precisely follow the

derivations in the original papers, because it is the combination of the contributions that is most useful. The gory mathematical details are presented in full because they are important for understanding where the algorithms come from. This level of detail is not in the scientific journals, so the effort of filling in the intermediate steps was left to the reader, who often did not have the time to ferret out the details on his own. The derivation proceeds in the following four steps:

(1) Elimination of one unknown from the lidar equation by assuming a functional relationship between σ_a and β_a and substituting it into the equation.
(2) Transformation of the lidar equation to a nonlinear differential equation followed by conversion of that equation to an ordinary linear differential equation that can be solved using the method of integrating factors.
(3) Determination of a constant of integration by using a priori information about, or assumption of, a boundary condition.
(4) Development of an efficient computer algorithm for calculating σ_a or β_a as a function of range.

A.1 Elimination of an Unknown from the Lidar Equation

If a functional form is assumed to exist between σ_a and β_a, then one of the unknowns can be eliminated from the lidar equation. Klett assumed that

$$\beta = \text{const} \cdot \sigma^{k}, \tag{A.2}$$

while Fernald assumed that

$$\beta_a = \sigma_a / S_a, \tag{A.3}$$

which is just Klett's assumption with $k = 1$ and the constant equal to $1/S_a$, where S_a is the aerosol extinction-to-backscatter ratio, more commonly called the lidar ratio, in units of sr. Fernald's notation is used here because it is the convention most often used by the lidar community. The lidar ratio S_a is initially assumed to be constant with range. Fernald described this assumption as being "not exceedingly restrictive," but in the convective boundary layer it is, because aerosols grow rapidly with the high relative humidity that often occurs near the top of the mixed layer, and that growth changes their optical properties. The spatial variability of S_a will be dealt with later. Measured values of S_a range from approximately 20 sr to 120 sr, with higher values signifying increased aerosol extinction relative to backscattering, caused by more highly absorbing components in the aerosol particles. In contrast to the wide range of S_a values, which depend on the aerosol size distribution, shape, particle composition, and relative humidity, the molecular extinction to backscatter ratio S_m is $8\pi/3$ sr, or 8.38 sr, to first order (it actually has a slight wavelength dependence).

The extinction-to-backscatter ratios can be used to find an expression for the backscatter coefficient in terms of the aerosol and molecular extinction coefficients, to eliminate an unknown from the lidar equation. The sum of backscatter coefficients is just

$$\beta_a(R)+\beta_m(R)=\frac{\sigma_a(R)}{S_a}+\frac{\sigma_m(R)}{S_m}=\frac{1}{S_a}\left(\sigma_a(R)+\frac{S_a}{S_m}\sigma_m(R)\right)=\frac{1}{S_a}y(\boldsymbol{R}), \quad \text{(A.4)}$$

where

$$y(R)=\sigma_a(R)+\frac{S_a}{S_m}\sigma_m(R). \quad \text{(A.5)}$$

The quantity $y(R)$ is referred to as the "normalized total extinction coefficient" by Sasano et al. [5]. This is a useful quantity that will be referred to later. Making the substitution in Eq. (A.4) in the range-corrected lidar equation results in

$$X(R)=\frac{C}{S_a}\left(\sigma_a(R)+\frac{S_a}{S_m}\sigma_m(R)\right)\exp\left[-2\int_0^R\left(\sigma_a(r)+\sigma_m(r)\right)dr\right], \quad \text{(A.6)}$$

where the transmitted power P_0 has been included in the calibration constant C.

A.2 Transformation of the Lidar Equation to a Differential Equation

Equation (A.6) can be transformed into a differential equation by taking the natural logarithm of both sides to get rid of the exponential terms and then taking the derivative with respect to the range to eliminate the integral. But first, some rearranging is helpful. Because S_a is assumed to be constant and the number of unknowns should be as few as possible, it would be convenient to eliminate $\sigma_a(R)$ from the exponential term of the range-corrected lidar equation. Rearranging Eq. (A.5) to get an expression for $\sigma_a(R)+\sigma_m(R)$ yields the relation

$$\sigma_a(R)+\sigma_m(R)=y(R)+\left(1-\frac{S_a}{S_m}\right)\sigma_m(R), \quad \text{(A.7)}$$

and substituting that result into Eq. (A.6) results in the equation

$$X(R)=\frac{C}{S_a}y(R)\exp\left[-2\int_0^R\left(y(r)+\left(1-\frac{S_a}{S_m}\right)\sigma_m(r)\right)dr\right]. \quad \text{(A.8)}$$

This equation has a mixture of knowns (S_a, S_m, and $\sigma_m(R)$) and unknowns (C and $y(R)$) on the right-hand side. To get the knowns on the right-hand side and the unknowns on the left-hand side, the exponential term can be split into the product of two exponentials, one with knowns and one with unknowns. Doing this and rearranging yields the equation

$$Cy(R)\exp\left[-2\int_0^R y(r)dr\right]=S_aX(R)\exp\left[-2\int_0^R\left(\frac{S_a}{S_m}-1\right)\sigma_m(r)dr\right], \quad \text{(A.9)}$$

where the sign of the quantity in parentheses in the exponential on the right-hand side was changed to make it look more like the exponential term on the left-hand side. Now the exponentials and the integrals can be eliminated. Taking the natural logarithm of both sides yields the relation

$$\ln(Cy(R)) - 2\int_0^R y(r)dr = \ln\left\{S_a X(R)\exp\left[-2\int_0^R\left(\frac{S_a}{S_m}-1\right)\sigma_m(r)dr\right]\right\}, \quad \text{(A.10)}$$

and then taking the derivative with respect to range results in the differential equation

$$\frac{1}{y(R)}\frac{dy(R)}{dR} - 2y(R) = \frac{d}{dr}\ln\left\{S_a X(R)\exp\left[-2\int_0^R\left(\frac{S_a}{S_m}-1\right)\sigma_m(r)dr\right]\right\}. \quad \text{(A.11)}$$

The lidar constant C has been eliminated by taking the derivative. Now the problem is that Eq. (A.11) is a nonlinear differential equation, because of the factor $1/y(R)$. To make it linear and to simplify the notation, represent the right-hand side by $f(R)$ and multiply both sides by $y(R)$. The equation then becomes

$$\frac{dy(R)}{dR} - 2y(R)^2 = f(R)y(R). \quad \text{(A.12)}$$

Equation (A.12) is a standard form that can be found in most books on differential equations or engineering mathematics: it is the Bernoulli equation (named for Swiss mathematician Jacob Bernoulli, 1654–1705), which is typically written in textbooks as

$$\frac{dy}{dx} + P(x)y = Q(x)y^k. \quad \text{(A.13)}$$

In the lidar case, $Q = 2$, $k = 2$, $P(x) = -f(R)$, and the range R corresponds to the variable x.

At this point, a digression into the solution of the Bernoulli equation is required. At first glance, our Bernoulli equation is still nonlinear, but it can be transformed into a linear ordinary differential equation by substituting a new variable for y. This is easiest to see if we rearrange the equation a bit first. Multiplying both sides by $1/y^k$ results in the equation

$$\frac{1}{y^k}\frac{dy}{dx} + P(x)\frac{1}{y^{1-k}} = Q(x). \quad \text{(A.14)}$$

The new variable υ is defined as

$$\upsilon = y^{1-k}, \quad \text{(A.15)}$$

which implies that

$$y^k = \upsilon^{\frac{k}{1-k}} \quad \text{(A.16)}$$

and

$$\frac{d\upsilon}{dx} = (1-k)y^{-k}\frac{dy}{dx}. \quad \text{(A.17)}$$

Putting these results into the Bernoulli equation, we get

$$\frac{1}{1-k}\upsilon^{\frac{k}{1-k}}\frac{d\upsilon}{dx}+P(x)\upsilon^{\frac{1}{1-k}}=Q(x)\upsilon^{\frac{k}{1-k}}, \tag{A.18}$$

which is still pretty opaque. Cleaning things up by multiplying through by the factor in front of the derivative on the left side and cancelling out some terms results in the simpler equation

$$\frac{d\upsilon}{dx}+(1-k)P(x)\upsilon=(1-k)Q(x), \tag{A.19}$$

which is a linear ordinary differential equation. It is "ordinary" because it depends on only one variable, x, as opposed to "partial" where it would depend on more than one variable. This type of differential equation can be solved with the method of *integrating factors*. The integrating factor method works by turning the left-hand side of Eq. (A.19) into the derivative of a product of two functions. To see how this works, multiply both sides of Eq. (A.19) by some function $M(x)$ (which will be figured out later):

$$M(x)\frac{d\upsilon}{dx}+M(x)(1-k)P(x)\upsilon=M(x)(1-k)Q(x). \tag{A.20}$$

By assuming that the left side is indeed the derivative of a product, we can write

$$\frac{d}{dx}M(x)\upsilon(x)=M(x)(1-k)Q(x). \tag{A.21}$$

Integrating both sides and solving for (x) gives

$$\upsilon(x)=\frac{\int M(x)(1-k)Q(x)dx+\text{constant}}{M(x)}. \tag{A.22}$$

The constant in Eq. (A.22) is an arbitrary constant of integration and it plays a very important role. The variable $\upsilon(x)$, and hence $y(x)$, can now be found if we can just figure out what $M(x)$ is. By expanding the derivative in Eq. (A.21) using the product rule for differentiation,

$$\frac{d}{dx}M(x)\upsilon(x)=\upsilon(x)\frac{d}{dx}M(x)+M(x)\frac{d}{dx}\upsilon(x)=M(x)(1-k)Q(x), \tag{A.23}$$

and comparing the terms in Eq. (A.23) to Eq. (A.20), we see that

$$\frac{dM(x)}{dx}=M(x)(1-k)P(x). \tag{A.24}$$

Rearranging slightly, we see that we have something that looks like the derivative of a logarithm:

$$\frac{1}{M(x)}\frac{dM(x)}{dx}=\frac{d}{dx}\ln M(x)=(1-k)P(x). \tag{A.25}$$

Knowing that fact, we can solve for $M(x)$, which is the integrating factor for the linear ordinary differential equation that we got by performing that convenient substitution of variables into the Bernoulli equation. Integrating both sides and solving for $M(x)$ yields the expressions

$$\ln M(x) = (1-k)\int P(x)dx \text{ and} \tag{A.26}$$

$$M(x) = \exp\left[(1-k)\int P(x)dx\right]. \tag{A.27}$$

We can now use Eq. (A.27) in Eq. (A.22) to get the expression for (x),

$$\upsilon(x) = \frac{\int \exp\left[(1-k)\int P(x)dx\right](1-k)Q(x)dx + \text{constant}}{\exp\left[(1-k)\int P(x)dx\right]}, \tag{A.28}$$

which is related to $y(x)$, the quantity we are trying to find, by the substitution of variables given in Eq. (A.15) that made it possible to get this far. Going back to the lidar equation, and remembering that $Q = 2$, $k = 2$, and $-f(R)$ corresponds to $P(x)$, substituting those quantities in Eq. (A.28) yields

$$\upsilon(R) = \frac{-2\int \exp\left[\int f(r)dr\right] dr + \text{constant}}{\exp\left[\int f(r)dr\right]}. \tag{A.29}$$

Back in Eq. (A.12) we substituted the notation $f(R)$ for a rather unwieldy expression, and now is the time to call that expression into use. The unwieldy expression is

$$f(R) = \frac{d}{dr}\ln\left\{S_a X(R)\exp\left[-2\int_0^R\left(\frac{S_a}{S_m}-1\right)\sigma_m(r)dr\right]\right\}. \tag{A.30}$$

The exponentials in Eq. (A.29) are then equal to

$$\exp[\int f(r)dr] = \exp\int\frac{d}{dr}\ln\left\{S_a X(R)\exp\left[-2\int_0^R\left(\frac{S_a}{S_m}-1\right)\sigma_m(r)dr\right]dr\right\}, \tag{A.31}$$

which reduces to

$$\exp[\int f(r)dr] = S_a X(R)\exp\left[-2\int_0^R\left(\frac{S_a}{S_m}-1\right)\sigma_m(r)dr\right]. \tag{A.32}$$

Because $k = 2$, $\upsilon(R) = 1/y(R)$, and $y(R)$ is what we want to find because it includes either $\sigma_a(R)$ or $\beta_a(R)$. It is time to leave the variable $\upsilon(R)$, which made it possible to deal with the Bernoulli equation, by using $y(R) = 1/\upsilon(R)$. Doing that and substituting Eq. (A.32) into the resulting expression results in the equation

$$y(R) = \frac{S_a X(R)\exp\left[-2\int_0^R\left(\frac{S_a}{S_m}-1\right)\sigma_m(r)dr\right]}{\text{constant} - 2S_a\int_0^R X(r)\exp\left[-2\int_0^r\left(\frac{S_a}{S_m}-1\right)\sigma_m(r')dr'\right]dr}. \tag{A.33}$$

At this point, either $\sigma_a(R)$ or $\beta_a(R)$ can be found by using one of the expressions in Eq. (A.4). The derivation is continued using backscatter coefficients to derive the solution that Fernald provided in 1984. Noting that

$$\left(\frac{S_a}{S_m}-1\right)\sigma_m(R) = \left(S_a - S_m\right)\beta_m(R), \tag{A.34}$$

and from Eq. (A.4) that

$$y(R) = S_a\left(\beta_a(R) + \beta_m(R)\right), \tag{A.35}$$

we find that

$$\beta_a(R) = \frac{X(R)\exp\left[-2\left(S_a - S_m\right)\int_0^R \beta_m(r)dr\right]}{\text{constant} - 2S_a\int_0^R X(r)\exp\left[-2\left(S_a - S_m\right)\int_0^r \beta_m(r')dr'\right]dr} - \beta_m(R), \tag{A.36}$$

which is Eq. (2) of [3]. Notice that the integrals in the exponentials involve only the molecular backscatter coefficient, which can be calculated from meteorological data or from a standard atmospheric profile adjusted for the ground-level temperature and pressure at the lidar location. The solution for $\beta_a(R)$ therefore depends only on the measured lidar signal values and known parameters, except for the unknown constant of integration.

A.3 Solving for the Constant of Integration

At this point, we could find the aerosol volume backscatter coefficient from our measurement of $X(R)$, our knowledge of the molecular scattering parameters, and our assumption of S_a if we knew what the arbitrary constant of integration was. It can be found by solving Eq. (A.36) for the constant's value at a boundary condition range R_c where we know (or assume to know) the atmospheric characteristics and then substituting that result back into Eq. (A.36). Solving for the constant, we find that it is

$$\begin{aligned}\text{constant} = {} & \frac{S_a X(R_c)\exp\left[-2\left(S_a - S_m\right)\int_0^{R_c} \beta_m(r)dr\right]}{S_a\left(\beta_a(R_c) + \beta_m(R_c)\right)} \\ & + 2S_a\int_0^{R_c} X(r)\exp\left[-2\left(S_a - S_m\right)\int_0^r \beta_m(r')dr'\right]dr.\end{aligned} \tag{A.37}$$

Substituting the expression for the constant back into Eq. (A.36) gives a rather complicated expression for the aerosol volume backscatter coefficient,

$$\beta_a(R) = \frac{X(R)\exp\left[-2\left(S_a - S_m\right)\int_0^R \beta_m(r)dr\right]}{\left\{\begin{array}{l}\dfrac{X(R_c)}{\beta_a(R_c) + \beta_m(R_c)}\exp\left[-2\left(S_a - S_m\right)\displaystyle\int_0^{R_c} \beta_m(r)dr\right] \\ +2S_a\displaystyle\int_0^{R_c} X(r)\exp\left[-2\left(S_a - S_m\right)\int_0^r \beta_m(r')dr'\right]dr \\ -2S_a\displaystyle\int_0^R X(r)\exp\left[-2\left(S_a - S_m\right)\int_0^r \beta_m(r')dr'\right]dr\end{array}\right\}} - \beta_m(R). \tag{A.38}$$

This expression can be simplified by taking advantage of the integration limits in each integral. Note that the second term in the denominator looks just like the third term, but its integration runs from 0 to R_c while the third runs from 0 to R. The second term therefore can be used to cancel the portion of the third term's integration that runs from 0 to R_c. Likewise, the integral of the exponential in the third term can be broken into two parts: one that runs from 0 to R_c and one that runs from R_c to R. The part that runs from 0 to R_c is not affected by the third term's integration, which now runs from R_c to R, so it is a constant that can be pulled outside of the integral, giving

$$\beta_a(R) = \frac{X(R)\exp\left[-2\left(S_a - S_m\right)\int_0^R \beta_m(r)dr\right]}{\left\{\begin{array}{l}\dfrac{X(R_c)}{\beta_a(R_c)+\beta_m(R_c)}\exp\left[-2\left(S_a - S_m\right)\displaystyle\int_0^{R_c}\beta_m(r)dr\right]+ \\ -2S_a\exp\left[-2\left(S_a - S_m\right)\displaystyle\int_0^{R_c}\beta_m(r')dr'\right]\displaystyle\int_{R_c}^{R} X(r)\exp\left[-2\left(S_a - S_m\right)\int_{R_c}^{r}\beta_m(r')dr'\right]dr\end{array}\right\}} - \beta_m(R). \tag{A.39}$$

The exponential terms with integrals from 0 to R_c occur in the numerator and denominator, so they cancel, to arrive at

$$\beta_a(R) = \frac{X(R)\exp\left[-2\left(S_a - S_m\right)\displaystyle\int_{R_c}^{R}\beta_m(r)dr\right]}{\dfrac{X(R_c)}{\beta_a(R_c)+\beta_m(R_c)} - 2S_a\displaystyle\int_{R_c}^{R} X(r)\exp\left[-2\left(S_a - S_m\right)\int_{R_c}^{r}\beta_m(r')dr'\right]dr} - \beta_m(R), \tag{A.40}$$

which is equation (3) of [3]. The need to know the integration coefficient forces us to assume a value for $\beta_a(R_c)$. If R_c lies in an aerosol-free region of the atmosphere, then $\beta_a(R_c)$ is zero and we only have to assume a value for S_a.

This is not a new equation in the literature of remote sensing. The first known appearance of one like it was in a 1954 paper by Hitschfeld and Bordan [6] concerning rain intensity measurements by radar. The problem is that under very turbid atmospheric conditions the two terms in the denominator are of nearly equal size, both involve the signal that has been contaminated with noise, and we are taking the difference of the two. Integrating outward from the calibration range leads to numerical instabilities that often produce unphysical results, and one would like an algorithm that (ideally) works no matter what. This is where Klett's contribution comes in. He found that by integrating from the calibration range *inward* toward the lidar, not only was the analysis much more stable, but that the integration process rapidly lost its dependence on the initial guess for the aerosol scattering cross section $\beta_a(R_c)$ at the calibration range. To use this insight, the modification to Eq. (A.36) is trivial:

$$\beta_a(R) = \frac{X(R)\exp\left[+2\left(S_a - S_m\right)\int_R^{R_c}\beta_m(r)dr\right]}{\frac{X(R_c)}{\beta_a(R_c)+\beta_m(R_c)} + 2S_a\int_R^{R_c}X(r)\exp\left[+2\left(S_a - S_m\right)\int_r^{R_c}\beta_m(r')dr'\right]dr} - \beta_m(R). \tag{A.41}$$

Notice the reversal of signs and the reversal of the integration limits for the downward integration. That simple change is what creates the numerical stability of the Klett retrieval. The corresponding formula in terms of extinction coefficients can be found by substituting Eq. (A.4) and Eq. (A.34) into Eq. (A.41) yielding

$$\sigma_a(R) = \frac{X(R)\exp\left[+2\left(\frac{S_a}{S_m} - 1\right)\int_R^{R_c}\sigma_m(r)dr\right]}{\frac{X(R_c)}{\sigma_a(R_c)+\frac{S_a}{S_m}\sigma_m(R_c)} + 2\int_R^{R_c}X(r)\exp\left[+2\left(\frac{S_a}{S_m} - 1\right)\int_r^{R_c}\sigma_m(r')dr'\right]dr} - \frac{S_a}{S_m}\sigma_m(R). \tag{A.42}$$

If a calibration range can be found that is essentially aerosol free, then the extinction or backscatter terms in the calibration constant are due to molecules only and can be calculated from standard atmospheric profiles. But this is a big if. To take advantage of that simplification, the lidar receiver has to generate an output signal of reasonably good quality from that range. Lidars operating in the infrared where the combination of low values for $\beta_a(R_c)$ and low system SNR caused by noisy photodetectors can be subject to radar-like calibration problems if they do not have adequate light collection ability. In this case, we must use some estimation for β_a at a calibration range R_c that is sufficiently laden with aerosols so that it provides a measurable signal with acceptable SNR. In reasonably uniform atmospheres that value might be estimated using the slope method, but even that approach is fraught with peril because the atmospheric conditions must be right for the slope method to work.

A.4 Algorithms for Data Analysis

Equations (A.41) and (A.42) are the jumping-off points for the development of the computer algorithm that many lidar researchers used in their data analysis before there were alternatives to elastic backscatter lidar. Fernald presented the algorithm in [3], but he did not explain its origins as clearly as they are explained here. In discrete mathematics, ranges become $R(I)$ where I is the range index. The range-corrected signal $X(I)$ is the value of $X(R)$ at the range $R(I)$, and the difference between adjacent ranges is just the range bin width ΔR. Equations (A.41) and (A.42) include integrals, so these must be converted to sums in which the differential of range dr is replaced by ΔR. An integral of a function is just the area of the region bounded by the

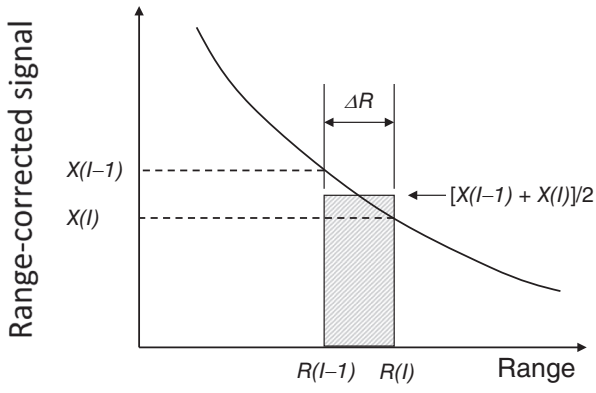


Figure A.1 The middle Riemann sum. The sum, which corresponds to a differential of area, is shown by the shaded rectangle. It is the average of the two range-corrected signal values times the range bin width.

graph of the function, the x-axis, and the vertical lines at the limits of the integral's range. The simplest way to compute an integral in discrete math is the Riemann sum, illustrated in Figure A.1 for the integral of the range-corrected signal over one range bin. The area shown is known as the middle Riemann sum, where the height of the rectangle assumes the function value at the middle of the interval, and it is calculated from an average of the function values at adjacent values of range. This integration procedure can be applied to the integrals in Eq. (A.41). We are interested in integrating downward from the calibration range, so we will first develop that algorithm. Following the derivation in [3], the value of the integral in the exponentials between range $R(I)$ and the range element just before it, $R(I-1)$, is

$$A(I-1,I) = (S_a - S_m)(\beta_m(I-1) + \beta_m(I))\Delta R. \tag{A.43}$$

The factor of ½ in the formula for the area of the Riemann sum rectangle cancels the factor of 2 in the exponential. The first integral in the second term of the denominator is treated the same way, but this time there is the added complication of the exponential weighting factor for $X(r)$. On the right side of the Riemann rectangle $r = R_c$. Letting R_c correspond to the index I, the integrand is

$$X(R_c)\exp\left[+2\left(S_a - S_m\right)\int_{R_c}^{R_c}\beta_m(r')dr'\right] = X(I). \tag{A.44}$$

On the left side of the Riemann rectangle, $r = R$. The range R corresponds to the array index $I-1$ and the integrand is

$$X(R)\exp\left[+2\left(S_a - S_m\right)\int_{R}^{R_c}\beta_m(r')dr'\right] = X(I-1)\exp\left[+A(I-1,I)\right]. \tag{A.45}$$

The area of the Riemann rectangle is then just

$$a_{\text{Riemann}} = \frac{X(I) + X(I-1)\exp\left[+A(I-1,I)\right]}{2}\Delta R. \tag{A.46}$$

Substituting these expressions into Eq. (A.41) gives the formula for calculating the value of $\beta_a(I-1)$ given the value of $\beta_a(I)$ one range step in front of it at the calibration range,

$$\beta_a(I-1)=\frac{X(I-1)\exp\left[+A(I-1,I)\right]}{\dfrac{X(I)}{\beta_a(I)+\beta_m(I)}+S_a\left\{X(I)+X(I-1)\exp\left[+A(I-1,I)\right]\right\}\Delta R}-\beta_m(I-1). \tag{A.47}$$

In terms of extinction coefficients,

$$\sigma_a(I-1)=\frac{X(I-1)\exp\left[+A(I-1,I)\right]}{\dfrac{X(I)}{\sigma_a(I)+\dfrac{S_a}{S_m}\sigma_m(I)}+\left\{X(I)+X(I-1)\exp\left[+A(I-1,I)\right]\right\}\Delta R}-\frac{S_a}{S_m}\sigma_m(I-1). \tag{A.48}$$

If the value of the calibration constant (the first term in the denominator) is fixed at the value for range R_c during the entire downward integration, then the other integral contributions become sums of Riemann rectangle areas because those integrals are over the path from R_c down to the current range of interest at R. *However*, the atmospheric parameters that have been calculated at range step I can be used to calculate the value of the calibration constant for the current range step $I-1$. This approach allows the sum to be avoided and the algorithm is computationally efficient.

A.4.1 Refining the Estimate for S_a

If other information about the extinction or backscattering properties of the atmosphere is available, then the initial estimate for S_a can be refined. One possibility is to use sun photometer measurements of total column extinction as a boundary value in the retrieval with S_a as a free parameter. An initial estimate is made for S_a and the retrieval is performed to calculate $\sigma_a(R)$, which is then integrated to provide a column optical depth that is compared to the sun photometer measurement. If the two do not match, the value for S_a is adjusted and the retrieval is performed again. This procedure is performed iteratively until the lidar-derived optical depth matches the sun photometer-derived value within the desired tolerance. In addition to refining the value of S_a, it has the added value of improving the accuracy of the extinction profile, but unfortunately S_a is often not constant, and when it is not, neither the value of S_a nor the profile $\sigma_a(R)$ derived by this method will be reliable.

A.5 Spatially Variable Lidar Ratio

Klett later extended his algorithm to the case of a spatially variable aerosol extinction-to-backscatter ratio [4]. Shortly afterward, Sasano, Browell, and Ismail reformulated

the algorithms of Fernald [3] to include spatial variation and investigated the errors caused by assuming a constant value of S_a [5]. Those authors acknowledged that *a priori* knowledge of the range dependence of S_a is often not available, but they suggest that estimates might be made for how it changes with altitude. Taking the lead of [5], the equivalent of Eq. (A.42) for a spatially variant $S_a(R)$ is found quite easily by substituting $S_a(R)$ for S_a and moving the appropriate range-dependent terms inside the integrals:

$$\sigma_a(R)=\frac{S_a(R)X(R)\exp\left[+2\int_R^{R_c}\left(\frac{S_a(r)}{S_m}-1\right)\sigma_m(r)dr\right]}{\frac{S_a(R_c)X(R_c)}{\sigma_a(R_c)+\frac{S_a(R_c)}{S_m}\sigma_m(R_c)}+2\int_R^{R_c}S_a(r)X(r)\exp\left[+2\int_r^{R_c}\left(\frac{S_a(r')}{S_m}-1\right)\sigma_m(r')dr'\right]dr}-\frac{S_a(R)}{S_m}\sigma_m(R). \tag{A.49}$$

The parameter S_a, which didn't enter into the equation for $\sigma_a(R)$ before now makes an appearance in both the numerator and denominator due to its new range dependency. When it was a constant, it canceled out between the numerator and denominator and did not appear anywhere except in the ratio of the aerosol to the molecular extinction-to-backscatter coefficients. An equivalent equation can be found in terms of the backscatter coefficients,

$$\beta_a(R)=\frac{X(R)\exp\left[+2\int_R^{R_c}\left(S_a(r)-S_m\right)\beta_m(r)dr\right]}{\frac{X(R_c)}{\beta_a(R_c)+\beta_m(R_c)}+2\int_R^{R_c}S_a(r)X(r)\exp\left[+2\int_r^{R_c}\left(S_a(r')-S_m\right)\beta_m(r')dr'\right]dr}-\beta_m(R). \tag{A.50}$$

As before, S_a must be known at the calibration range R_c in order to initialize the retrieval, but now the profile of $S_a(R)$ must also be known.

Fernald's algorithms must also be modified to accommodate the range dependency of $S_a(R)$. The exponential term in this case becomes

$$A(I-1,I)=\left\{\left[S_a(I-1)-S_m\right]\beta_m(I-1)+\left[S_a(I)-S_m\right]\beta_m(I)\right\}\Delta R. \tag{A.51}$$

The backscatter coefficient algorithm is

$$\beta_a(I-1)=\frac{X(I-1)\exp\left[+A(I-1,I)\right]}{\frac{X(I)}{\beta_a(I)+\beta_m(I)}+\left\{S_a(I)X(I)+S_a(I-1)X(I-1)\exp\left[+A(I-1,I)\right]\right\}\Delta R}-\beta_m(I-1), \tag{A.52}$$

and the extinction coefficient algorithm is

$$\sigma_{\mathrm{a}}(I-1)=\frac{S_{\mathrm{a}}(I-1)X(I-1)\exp\left[+A(I-1,I)\right]}{\dfrac{S_{\mathrm{a}}(I)X(I)}{\sigma_{\mathrm{a}}(I)+\dfrac{S_{\mathrm{a}}(I)}{S_{\mathrm{m}}}\sigma_{\mathbf{m}}(I)}+\left\{S_{\mathrm{a}}(I)X(I)+S_{\mathrm{a}}(I-1)X(I-1)\exp\left[+A(I-1,I)\right]\right\}\Delta R}$$
$$-\frac{S_{\mathrm{a}}(I-1)}{S_{\mathrm{m}}}\sigma_{\mathrm{m}}(I-1) \tag{A.53}$$

The derivations of the equations and data analysis algorithms in [1–5] have been presented in their full mathematical detail. The algorithms can be implemented in computer software or in applications such as spreadsheets. However, the lidar community reached a consensus by the mid-1990s that a simple elastic backscatter lidar measures only geometrical properties, such as the base height and thickness of an elevated aerosol layer (although optical depths are sometimes estimated by scattering ratios). Such a lidar is now considered to be incapable of measuring a reliable $\sigma_{\mathrm{a}}(R)$ profile. Fortunately, two reliable lidar techniques for this purpose have been developed, HSRL and aerosol Raman lidar, both of which date to 1990.

References

[1] F. G. Fernald, B. M. Herman, and J. A. Reagan, "Determination of Aerosol Height Distributions by Lidar," *Journal of Applied Meteorology*, vol. 11, pp. 482–489, 1972.

[2] J. D. Klett, "Stable Analytical Inversion Solution for Processing Lidar Returns," *Applied Optics*, vol. 20, pp. 211–220, 1981.

[3] F. G. Fernald, "Analysis of Atmospheric Lidar Observations: Some Comments," *Applied Optics*, vol. 23, pp. 652–653, 1984.

[4] J.D. Klett, "Lidar Inversion with Variable Backscatter/Extinction Ratios," *Applied Optics*, vol. 24, pp. 1638–1643, 1985.

[5] Y. Sasano, E. V. Browell, and S. Ismail, "Error Caused by Using a Constant Extinction/Backscattering Ratio in the Lidar Solution," *Applied Optics*, vol. 24, pp. 3929–3932, 1985.

[6] W. Hitschfeld and J. Bordan, "Errors Inherent in the Radar Measurement of Rainfall at Attenuating Wavelengths," *Journal of Meteorology*, vol. 11, pp. 58–67, 1954.

Index