CHAPTER 2

Sources and Illuminators

And God saw the light, that *it* was good: and God divided the light from the darkness.
—Genesis 1:4 (Authorized Version)

2.1 INTRODUCTION

In order to make an optical measurement, we need a light source. In some systems, the only source required is ambient illumination or the luminosity of the object itself. More often, though, the instrument must supply the light. Instrument designers are less likely to neglect the light source than the detector, but still it is often chosen without proper consideration of the pain and suffering its deficiencies may cause one, or without regard to the special attributes of an alternative.

This chapter deals with light sources and illumination systems, dwelling on their strengths and weaknesses in applications, rather than on the physics of how they work. We stick with the mainstream choices: gas, solid state, and diode lasers, tungsten bulbs, arc lamps, and LEDs. There are other interesting sources, useful in special cases, many of which are discussed in the OSA Handbook.

2.2 THE SPECTRUM

Dye laser catalogs usually have a nice chart showing the electromagnetic spectrum from the infrared (IR), through the surprisingly narrow visible range, to the ultraviolet (UV), short UV, and vacuum UV (VUV). Electrodynamics doesn't alter its character between regions, but the interaction of light and matter does change systematically with wavelength.

2.2.1 Visible Light

The visible part of the spectrum is conventionally taken to be the octave from 400 to 800 nm. Any measurement that can be done with visible light should be.[†] A visible light

[†]There are rare exceptions, such as IR laser based intraocular measurements, where the use of visible light would be very uncomfortable, or solar blind UV photomultipliers for seeing dim UV sources in normal daylight.

Building Electro-Optical Systems, Making It All Work, Second Edition, By Philip C. D. Hobbs

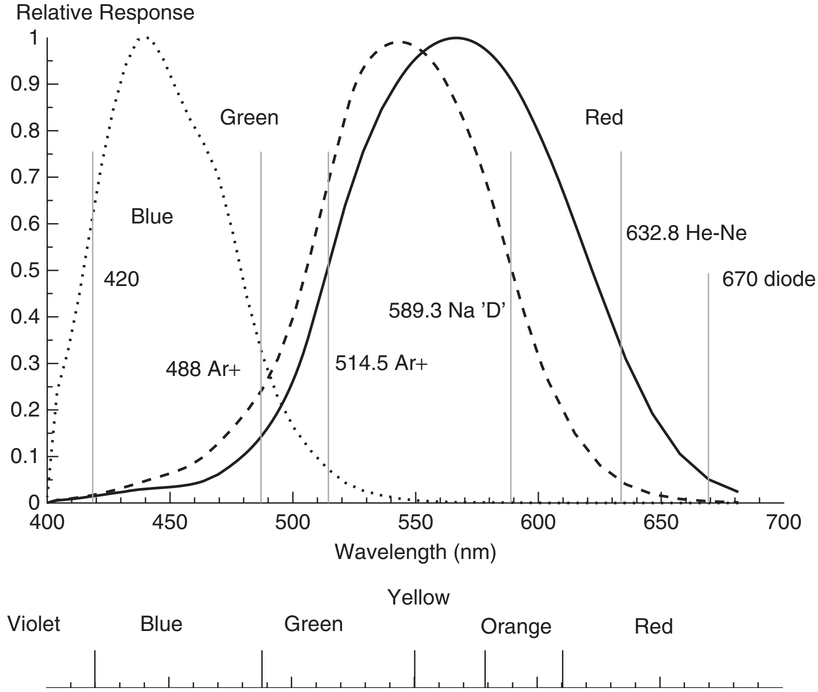


Figure 2.1. Relative response function of human cone cells versus wavelength. Note the almost complete overlap of the red and green pigments, leading to good wavelength discrimination in the yellow and less elsewhere. Note: The blue channel has been normalized to the same height as the others, but it's much smaller in reality.

system is enormously easier to align, use, and find parts for than any other kind. If lasers are involved, it's also somewhat safer for your eyes, because of the blink reflex and because you can see the stray beams and block them all before they leave your setup.

The wavelength resolution of human color perception is wildly variable, due to the pronounced overlap of the sensitivities of the retinal pigments, as shown in Figure 2.1.†

2.2.2 Ultraviolet

The UV begins at about 400 nm and extends to the border of X-rays at a few nanometers. Near UV (320–400 nm) is generally safe and convenient, because lenses still work. Deep UV (180–320 nm) is less safe, and lens materials are generally scarce there and may be darkened by exposure. The most miserable is vacuum UV (below about 180 nm), where O_2 dissociates to produce ozone, and room air and everything else becomes strongly absorbing. Beware of insidious skin and corneal burns from bright deep-UV sources.

Ultraviolet photons are energetic enough to cause ground state electronic transitions in gases and solids, so every material known becomes strongly absorbing somewhere in the UV. Finding optical materials becomes very difficult around 200 nm, another reason

†The blue channel is about 10 times less sensitive, so the light-adapted eye has its peak sensitivity at 550 nm, with its 50% points at 510 and 610 nm.

for using vacuum systems. This same property makes solid state UV detectors difficult to build; photocathodes are the usual choice in the VUV.

2.2.3 Infrared

The IR starts at about 800 nm, although since the human eye is nearly blind beyond 700, the point at which light becomes invisible depends on its brightness. Out to about 2.5 μm is the near infrared, defined roughly as the region in which glass lenses work and decent photodiodes are available. From there to about 10 μm is the mid-IR, where radiation from room temperature objects is still not a critical limitation, and beyond 10 μm almost to 1 mm is the far-IR, where everything is a light bulb, including the detector itself.

The infrared contains the fundamental vibrational frequencies of molecular bonds (e.g., C—H), so most of the interesting spectroscopy happens there. Infrared spectral lines are often combinations of rotational and vibrational transitions, whose frequencies add and subtract; it is more convenient to quote such systems in terms of frequency than wavelength. The usual frequency unit is not hertz, but wave numbers,[†] that is, how many waves fit in one centimeter. Fortunately, the speed of light is a round number to three decimal places, so the rough conversion that 1 $cm^{-1} \approx 30.0$ GHz is convenient for rapid calculation. Interesting spectroscopy happens between 100 cm^{-1} and about 5000 cm^{-1} (100 μm to 2 μm). Note the near-total failure of interesting spectroscopy and good detectors to overlap. Lots of work has gone into that problem.

2.3 RADIOMETRY

In order to be able to choose between illuminators, we need some way to compare them. Measurement and comparison of illumination conditions is the aim of radiometry. Radiometry is an almost completely isolated subfield of optics. This is a bit strange, since all designers need to know how much power crosses a given surface element into some solid angle, in some frequency interval, and most of the time they do. It sounds simple. It would *be* simple, too, if it weren't for the names.

Radiometric nomenclature is important in fields such as remote sensing and architecture, but because it's such a horrible mishmash, it remains a foreign language to most instrument designers, engineers as well as physicists. For one thing, the names have no mnemonic value of their own, and preexisting terms that everybody understands (*intensity* and *brightness*) have been redefined to mean something completely different; for another, there's an impenetrable thicket of redundant units in various subfields. Ideally, technical language provides a concise method of communicating difficult concepts precisely, and radiometric and photometric nomenclature is an excellent example of how *not* to do that.

We'll eschew the footcandles per fortnight and stick with something the author can cope with: $L_{\overline{\nu}}(\nu, \theta)$, the power emitted per square meter per steradian per hertz.[‡] This is the most natural description of a low coherence source such as a light bulb. Its official name is something out of Edgar Allan Poe—*spectral radiance*—and it has a welter of

[†]The official name for the inverse centimeter is the kayser, which nobody uses.

[‡]The subscript Greek *nu* is easily confused with subscript v, which means something else, so we write it with an overbar, $\overline{\nu}$, instead.

other names and symbols. Everything else is an integral of the spectral radiance over one or more of the quantities in the denominator: frequency, solid angle, area. You can find charts and tables and explanations of this in several sources. One thing to remember is that radiometrists like to do things in per-wavelength units instead of the per-frequency units that most physicists and engineers tend to prefer, leading to mysterious factors of ν^2 turning up on account of the chain rule.

Unless you need to communicate with diehard radiometrists, and don't care that few others will understand, stick with SI units written out in full. Use flux for total power through a surface, flux density for power per unit area, and brightness or radiance for flux density into unit solid angle. Watts per square meter per steradian per hertz are widely understood, internationally standardized, and won't get you into trouble, even if you can't tell a metrelambert from an apostilb.[†]

Photometry is similar, except squishier: the aim there is to compare illumination conditions not in W/Hz, but by what a standardized human eye defines as brightness. Since human eyes have different spectral sensitivities under different lighting conditions (rods vs. cones), this is a bit fraught, so it is solved by fiat: there are two published curves for the scotopic (dark adapted) and photopic (light adapted) response of the eye, and you just multiply $L_{\overline{\nu}}$ by that and integrate over ν to get the visual brightness or *luminance*, measured in lumens/m^2/sr. The photopic curve peaks at 550 nm, and is down 50% at 510 and 610 nm. The unit of "photometric power" is the lumen, corresponding roughly with the watt; for historical reasons 1 W at 552 nm (540 THz) is equivalent to 683 lumens, probably the largest prime number ever used for unit conversion.

The reason we can't completely ignore all this is that making accurate measurements of optical power, flux, fluence, and so on is *hard*, which is why there are people who specialize in it. A decent DVM costing $200 will measure voltage and current to an accuracy of 0.1% or so over a dynamic range of 120 dB. Nothing remotely comparable exists for optical power measurements: $\pm 1\%$ is quite respectable even for narrow signal level ranges. The skill of radiometry is in making decent measurements, not in expressing them in units.[‡]

2.4 CONTINUUM SOURCES

A basic distinction between sources is whether their outputs are predominantly in a spectral continuum or in isolated spectral lines. Lasers and low pressure gas discharge sources have narrow line spectra, high pressure discharge tubes (such as arc lamps and flashlamps) produce a mix of broadened lines and continuum, and incandescent objects produce continuum only. Within each class, there are distinctions as to how wide the spectral features are. The groupings are somewhat rough-and-ready, but nonetheless useful.

In practical applications, the other main distinction is spatial coherence—basically how much of your source light you can corral into your optical system, and how small a spot you can focus it into. Piping laser beams around is easy, because the beam is so well behaved; all the light can be pointed in the same direction. Other sources are not so tractable. Large diameter thermal sources cannot be focused down as tightly as

[†]They're the same thing.

[‡]One of the most heroic feats of radiometric calibration to date is the Cosmic Background Explorer (COBE) satellite—see Kogut et al., *Astrophysical Journal* **470**(2), 653–673 (1996).

small diameter ones. Since we're always trying either to get more photons or to reduce the source power (to save cost, weight, and electric power), extremely low coherence sources such as large arc lamps or tungsten bulbs present problems in illuminator design.

2.4.1 Black Body Radiators

Electromagnetic radiation from an object in thermal equilibrium with its surroundings is called *black body radiation*. Since the closest thing to pure black body radiation is generated experimentally by putting a small hole into a hot cavity, the name may seem odd; however, it is the same as the radiation emitted by a purely theoretical object whose coupling to the electromagnetic field is perfect—every photon hitting it is absorbed. More subtly, by the second law of thermodynamics, it must have the maximum emittance as well. If this were not so, a black body would spontaneously heat up or cool down in isothermal surroundings, since it would absorb every photon incident but emit more or fewer. Thus a black body forming one wall of an isothermal cavity must emit a radiation spectrum exactly the same as that which fills the cavity. Since resonator effects in a large cavity are slight, they can modify the black body's spectrum only slightly (asymptotically, not at all), so a black body radiates cavity radiation regardless of where it is placed. This is a remarkable result.

Such an object would indeed appear black when cold, and many objects have emission spectra that are reasonably well approximated by that of a black body. Among these approximate black bodies are the Sun and tungsten bulbs. The ratio of the total power emitted by some real surface at a given wavelength to that predicted from an ideal black body is known as its *spectral emissivity*, or emissivity for short.† It depends strongly on surface composition, angle of incidence, finish, and coatings, but tends to vary slowly with frequency. The ratio of total power absorbed to total power incident is the absorptivity, but this term is almost never used, since it is identical to the emissivity, by the same thermodynamic argument we used before. Landau and Lifshitz, *Statistical Physics Part I*, present an excellent discussion of this topic. We're going to do everything in per-frequency units, for conceptual unity and calculating convenience.

The power spectral density curve of black body radiation peaks at a frequency given by the Wien displacement law,

$$h\nu_{\text{peak}} = 2.8214 k_B T, \tag{2.1}$$

where, as usual, h is Planck's constant, ν is frequency, k_B is Boltzmann's constant,‡ and T is absolute temperature. It has a FWHM of about two octaves, in round figures. If the spectral density is quoted in per-wavelength units, the peak is slightly shifted toward the blue. (Why?) Although the position of the peak moves, black body curves from different temperatures never cross; heating up a black body makes it glow more brightly at all wavelengths (this is also obvious from the second law, if reflective colored filters are admitted).

The formulas for black body radiation tend to be a bit confusing, since different authors quote formulas for different things: exitance from a semi-infinite black body,

†Emissivity is formally the average value over some spectral range, but since the emissivity changes slowly with frequency and is anyway a function of temperature, the distinction is not usually emphasized in real life.

‡This is one of the few places in this book where Boltzmann's constant k can be confused with the propagation constant k, so we need the inelegant subscript B.

cavity energy density per unit frequency, or per unit wavelength, and so on. The most fundamental is the energy density per hertz of the cavity radiation,

$$e_0(\nu, T) = \frac{2hn^3\nu^3}{c^3\{\exp[h\nu/(k_B T)] - 1\}}, \tag{2.2}$$

where n is the refractive index of the (uniform and isotropic) medium filling the cavity.

There is no special direction inside the cavity, so the energy is equally distributed over all propagation directions. The rate at which a component with wave vector $\mathbf{k}$ leaves the cavity through a patch of area dA is $c\mathbf{k}\cdot\mathbf{dA}/k$, where the vector $\mathbf{dA}$ is dA times the outward directed unit surface normal $\hat{\mathbf{n}}$. Thus there is a cosine dependence on the angle of incidence θ, and the spectral radiance leaving the cavity is

$$L_{\overline{\nu}}(\nu, T) = \frac{c}{n}\cos\theta e_0(\nu, T) = \frac{2hn^2\nu^3\cos\theta}{c^2\{\exp[h\nu/(k_B T)] - 1\}} \tag{2.3}$$

which is the most useful all around black body formula. Integrated over all ν, this yields the Stefan–Boltzmann formula for the total power emitted from unit surface area of a black body into unit solid angle,

$$P_{\text{tot}}(T) = \sigma n^2 T^4 \cos\theta, \tag{2.4}$$

where σ is Stefan's constant, approximately 1.8047×10^{-8} W/m^2/sr/K^4. A real object cannot (as we saw) radiate any more than this, and most will radiate quite a bit less; not everything is black, after all.

The cosine dependence of black body radiation arose, remember, because there was no special direction. A source with this cosine dependence is said to be *Lambertian*, and broad area sources with Lambertian or near-Lambertian angular dependence are said to be *diffuse*.

2.4.2 Radiance Conservation and the Second Law of Thermodynamics

In imaging with thermal light, it's important to keep in mind a fundamental limitation: *no passive system can form an output whose surface radiance is higher than that of the source*. We've seen this as conservation of phase space volume. This can be shown rigorously from Maxwell's equations, and for paraxial optics it follows from the *ABCD* matrix for an imaging system: the image-side NA goes as $1/M$. For our purposes we'll use a thought experiment and the second law of thermodynamics.

Consider two reservoirs, initially at the same temperature, and well insulated except for a sufficiently small heat engine (SSHE) running off any temperature difference from R_{hot} to R_{cold}. Drill a hole in each reservoir's insulation, and place a conceptual highly asymmetric optical system (CHAOS) in between (also suitably insulated). Add well-silvered baffles so that only light that will make it through the CHAOS gets out of the reservoir insulation.

Call the $A\Omega'$ products looking into each end of the CHAOS $A\Omega'_{\text{hot}}$ and $A\Omega'_{\text{cold}}$. Then the net radiative energy transfer from hot to cold will be

$$\dot{Q} = \frac{\sigma}{\pi}(T^4_{\text{hot}}A\Omega'_{\text{hot}} - T^4_{\text{cold}}A\Omega'_{\text{cold}}), \tag{2.5}$$

assuming Lambertian radiation (which should be right, since it's cavity radiation we're considering).

If $A\Omega_{hot} < A\Omega_{cold}$, the hot reservoir will spontaneously heat up until $\dot{Q} = 0$, which will happen when $T_{hot}/T_{cold} = (A\Omega_{hot}/A\Omega_{cold})^{0.25}$, and the SSHE will run forever as a perpetual motion machine. If we allow the CHAOS to contain lossless reflective optical filters and amend (2.5) appropriately, the same argument shows that radiance conservation applies for each wavelength independently, not just in aggregate. If the two ends are immersed in different media, then there's a factor of n^2 that has to be added to Stefan's law on each end, and the math comes out the same: there's no temperature drop created.

2.4.3 Tungsten Bulbs

Tungsten bulbs are excellent optical sources for many purposes. They are quiet, stable, cheap, broadband, and reasonably bright. Although their radiance (in W/m^2/sr) is far lower than a laser's, they can put out a lot of photons cheaply. Their electrical to optical conversion efficiency is excellent—75% or better—provided that the application does not require throwing away major fractions of their output.

The primary problem with tungsten bulbs is their relatively short operating life, which is limited by the evaporation of the tungsten filament, with the subsequent formation of hot spots, leading to filament breakage. The rate of tungsten evaporation goes as $e^{-10500/T}$, while the brightness of the bulb is proportional to T^4, which makes the trade-off between lifetime and brightness pretty sharp; you really can't push tungsten beyond about 3300 K. This isn't quite as bad as it sounds, because hot tungsten's emissivity is around 0.45 in the visible but only 0.15–0.2 in the IR, so that you get proportionately more visible photons than you expect, and the bulb color looks about 100 K hotter than it is (maximum color temperature $\approx$ 3400 K).

The trade-off between lifetime and brightness can be exploited in appropriate circumstances: a tungsten bulb run at 10% higher than rated voltage will have three times shorter life and about a third higher output. Alternately, running it at 10% lower than rated voltage will give five times longer life at about 30% less output.

Evaporated tungsten collects on the bulb envelope, which produces progressive dimming with time, as well as some spectral shift. The evaporation rate can be reduced by reducing the bulb temperature or by increasing the pressure of the gas filling the envelope, which retards the diffusion of tungsten vapor away from its source. Thermal losses and safety considerations put a practical limit on the gas pressure.

Tungsten–halogen bulbs offer a partial solution to the tungsten loss problem, through a clever regenerative mechanism that keeps the envelope clean. A bit of iodine is introduced into a small, very strong quartz envelope around the filament. Evaporating tungsten combines with the iodine to form tungsten iodide, which has a low boiling point. The envelope is hot enough to prevent tungsten iodide from condensing on it, so that the halide stays in the gas phase until it encounters the filament and is redeposited as metallic tungsten. Unfortunately, the redeposition does not take place selectively on the hot spots, so the filament does not heal itself. The longer life of halogen bulbs is due to the high gas pressure inside, which by slowing vapor diffusion *does* cause selective redeposition.

Running a bulb on AC helps its long-term stability, since electromigration of the filament metal partly cancels out when the current changes direction frequently.

Temperature variations over a cycle give rise to intensity variations. These aren't that small—asymptotically, their power spectrum goes as $1/f^2$ (1 pole). Square wave

AC drive helps by keeping the power dissipation constant over a cycle. Ripple in the dissipated power will cause thermal forcing of mechanical oscillations of the filament. You can't readily simulate this, so be sure to test it if you're planning anything fancy.

The other major problem with tungsten bulbs (as with other low coherence sources) is how to keep the source radiance as high as possible as the light passes through the bulb envelope and condenser. More on that soon.

2.4.4 Glow Bulbs and Globars

Occasionally one runs across these odd ceramic light bulbs. They provide nice wide, uniform diffuse thermal emission in the IR, but are not very bright as the ceramic seldom goes much above 1000 K. They are primarily useful in FTIR spectrometers. Good ones have low thermal mass so you can do chopping measurements by pulsing the light source (see Section 10.9.1).

2.5 INTERLUDE: COHERENCE

In various places we've already encountered the idea of coherence, which is basically a measure of how well different parts of an electromagnetic field stay in phase with each other. It's time to be a bit more systematic about what we mean by that. It's a big subject, though, and is explained well in J. W. Goodman's *Statistical Optics*. We're just going to dip our toes here.[†]

Coherence theory is easiest to understand in the context of two famous interferometers: Young's slits (actually pinholes) for spatial coherence, and the Michelson interferometer for temporal. Both of these experiments involve sampling the fields at different points in space-time, and looking at the time-averaged interference term. The result of the experiments is a fringe pattern; the visibility of the fringes expresses the degree of coherence between the fields at the two points. This is an appealingly intuitive definition of coherence, and what's more, it is powerful enough for all optical purposes, provided the averaging time is a small fraction of a cycle of the highest signal frequency we care about.

Thermal sources such as tungsten bulbs are said to be *spatially incoherent*; the phase and amplitude of the light emitted from a point on the surface are independent of that from points more than a wavelength or so away. As a result, the intensity of the light at any point in the system can be calculated by finding the intensity due to a single arbitrary source point, and integrating over the source, without regard for the interference of light from different points on the source.

The light from a tungsten bulb is also *temporally incoherent*, meaning that the phase and amplitude of light from the same source point at some time t is independent of what it was at $t - \tau$, for sufficiently large τ (several femtoseconds for a 3000 K bulb).

These properties may be substantially modified by spatial- and frequency-selective filters, condensers, and even free-space propagation. For example, starlight is temporally as incoherent as tungsten light, but spatially highly coherent, due to the small angular size of the star as seen from Earth. A nearby star would have an angular size of a few

[†]This discussion assumes some familiarity with interferometers. Have a look at Section 1.6, or for more basic stuff, try Hecht and Zajac or Klein and Furtak.

nanoradians. The angular size of an incoherent (thermal) source determines its spatial coherence; the divergence of the beam is equal to the angular size of the source. (Why?)

The basic quantity in coherence theory is the *complex degree of coherence* γ, which is the normalized statistical cross-correlation of the optical field at two points ($\mathbf{P}_1$, t) and ($\mathbf{P}_2$, $t+\tau$),

$$\gamma_{12}(\tau) \equiv \frac{\Gamma_{12}(\tau)}{\sqrt{\Gamma_{11}(0)\Gamma_{22}(0)}}, \tag{2.6}$$

where Γ_{12} is the usual ensemble-averaged statistical cross-correlation (see Section 13.5),

$$\Gamma_{12}(\tau) \equiv \langle \psi(\mathbf{P}_1, t)\psi(\mathbf{P}_2, t+\tau)\rangle. \tag{2.7}$$

A pure temporal coherence discussion sets $\mathbf{P}_1 = \mathbf{P}_2$ and usually drops the subscripts.

Given a screen with unresolved pinholes at $\mathbf{P}_1$ and $\mathbf{P}_2$, the fringe pattern at some point $\mathbf{x}$ on the far side of the screen from the source is

$$I(\mathbf{x}) = I_1(\mathbf{x}) + I_2(\mathbf{x}) + 2\sqrt{I_1(\mathbf{x})I_2(\mathbf{x})}\mathrm{Re}\left\{\gamma_{12}\left[\frac{|\mathbf{x}-\mathbf{P}_1| - |\mathbf{x}-\mathbf{P}_2|}{c}\right]\right\}. \tag{2.8}$$

Temporal coherence can be increased by a narrowband filter, at the price of throwing most of the light away; you can make fringes using light with a frequency range $\Delta\nu$ if the time delay τ between components is less than $1/\Delta\nu$ (if $\tau > 1/\Delta\nu$ you get fringes in the spectrum instead). The envelope of the fringes in a Michelson interferometer will decay as the path difference is increased, so that we can define coherence time more rigorously as the equivalent width of the squared modulus of the fringe envelope,

$$\tau_c \equiv \int_{-\infty}^{\infty} |\gamma(\tau)|^2 d\tau. \tag{2.9}$$

The *coherence length* sounds like a spatial parameter but is in fact defined as $c\tau_c$. The real spatial parameter is the *coherence area*, A_c, the area over which a given field's phase stays highly correlated. It is analogously defined as the two-dimensional (2D) equivalent width in the $\xi\eta$ plane, where $\mathbf{P}_2 - \mathbf{P}_1 = (\xi, \eta, 0)$.

$$A_c \equiv \iint_{-\infty}^{\infty} |\gamma_{12}(0)|^2 d\xi d\eta. \tag{2.10}$$

Young's fringes also fall off in strength as the pinholes are moved apart, but measurements must be made in the plane of zero path length difference in order that temporal coherence effects not enter. In practice, $\Delta\nu$ must be small enough that at least a few fringes are visible for us to be able to measure the fringe visibility as a function of pinhole separation, but mathematically this can be swept under the rug.

Coherence theory predicts how these properties change and what their effects are. It also provides a fairly rigorous conceptual framework in which to describe them. It is a huge topic, which is really beyond our scope. It is treated well in Goodman and in Born and Wolf.

2.5.1 Speckle

Light scattered from a rough surface undergoes wholesale modification of its plane wave spectrum. A single plane wave gets scattered into a range of angles that is characteristic of the surface. At each angle, the light from different locations interferes together, producing strongly modulated, random-looking fringes called *speckle*. Speckle is a three-dimensional (3D) interference pattern, and the characteristic size of the speckles in radians gives the characteristic length scale of the roughness in wavelengths. Speckle is a particular problem in laser-based full-field measurements and when using diffusers (see Section 5.7.11). Although it's usually a nuisance, speckle does contain information about the position and shape of the surface; this information can be extracted by electronic speckle pattern interferometry (ESPI, frequently called TV interferometry).

When low coherence light is scattered from a rough surface, its angular spectrum is modified in a fairly simple way; usually it just spreads out some more, and the scattered radiance remains reasonably smooth in both position and angle. The lack of speckle is a consequence of both spatial and temporal incoherence; in each propagation direction, temporal incoherence (i.e., wide optical bandwidth with no special phase relationships, as in short pulses) smears the fringes out in the space domain (due to the path differences), and spatial incoherence makes the phase of the interference pattern at each wavelength vary at optical frequencies, so that no fringe contrast can be observed in any reasonable measurement.

2.5.2 Imaging Calculations with Partially Coherent Light

Besides the two limiting cases of fully coherent and incoherent sources, there's a broad range of partially coherent cases, where the source linewidth and angular subtense are big enough to notice but too small to dominate. Most of the time, partially coherent imaging is done by starting with an incoherent source such as a light bulb. There are lots of books and papers telling how to predict the results of a partially coherent imaging system, but for most of us, getting through that mathematical machinery is a big time investment. Another way, less general but conceptually much easier, is to replace the incoherent source by a collection of monochromatic point sources, calculate the (coherent) image from each one, and then add all the resulting intensities by integrating over the source distribution in space and frequency. Integrating the intensities expresses the fact that the expectation value of the interference term between two incoherent sources (or two parts of a single one) is 0. Even from a single point, light of two different frequencies produces fringes that move at optical frequency, averaging to 0 (but see below). The interference term in (1.67) is thus 0, and the total photocurrent from the detector is the sum of those from each source point.

If your source is inherently partially coherent (e.g., starlight coming through a turbulent atmosphere), this procedure breaks down and you have to use the big mathematical guns.

2.5.3 Gotcha: Coherence Fluctuations at Finite Bandwidth

One coherence effect that we're going to run into again and again is intensity fluctuations due to the interference of light with a delayed copy of itself. This effect is often overlooked, but it limits the SNR of measurements a lot of the time; it's normally what sets the ultimate SNR of laser-based fiber sensors, for example, and Section 19.1.1 has a very sad story about what can happen when it bites. It's true that the DC photocurrent from an interferometer whose path difference is $\gg 1/\Delta\nu$ will be the sum of the intensities of the two beams, but that doesn't mean that the instantaneous noise current is just the sum of their noises.

The autocorrelation is an ensemble- or long-time-averaged quantity, whereas we're actually detecting the instantaneous intensity, including the instantaneous interference term, with only the short-time averaging due to our finite measurement bandwidth.

Whenever we have two beams adding together, they do interfere, because at any instant the total energy arriving at the detector goes as the integral of E^2; all that happens at τ_C is that the *DC component* of the interference becomes small. So where does the interference signal go?

Temporal coherence limits usually arise from the gradual building up of phase fluctuations as Δt grows. When this happens, the interference contribution from each point doesn't go away, it just gets turned into noise spread out over a wide bandwidth, about $2/\tau_C$. A light bulb has a bandwidth of two octaves, centered around 700 nm, so its temporal bandwidth is 600 THz and its coherence time is 2 fs. It takes a lot of noise power to fill that up.

Furthermore, since these AC fluctuations are uncorrelated across the surface of an incoherent source, they average to zero pretty well too if the source and detector are big enough. On the other hand, if you build an interferometer using a spatially coherent source of limited temporal coherence, it will give you fits, and this is true whether you meant it to be an interferometer, or just have a stray fringe or two.

In a single-mode fiber measurement, for example, we get no spatial averaging whatever, so we have to cope with the whole interference term, changed into random noise, thundering into our detector. If there is no delayed wave to interfere with, it's no problem, but since the effect is gigantic and the intrinsic SNR of optical measurements is so high, it doesn't take much delayed light to reduce the SNR catastrophically. The detection problem for FM noise in a system with multiple paths of different delays is treated in Section 15.5.12; here we'll just wave our arms a bit.

Consider a Michelson interferometer using a well-collimated, 820 nm diode laser. It is VHF modulated (see Section 2.14.3) so that $\Delta\nu \approx 1$ THz ($\Delta\lambda \approx 2.2$ nm), but has a single transverse mode. Its spatial coherence is then nearly perfect. Recall from Section 1.5 that the instantaneous intensity reaching the photodetector varies as $I_1 + I_2 \pm 2\cos\phi\sqrt{(I_1 I_2)}$ (see Section 1.6), where ϕ is the instantaneous phase difference between the two beams. Because ϕ is constant across the detector, spatial averaging of the fringes does not reduce the fringe contrast in the detected signal. For $\Delta t \gg 1/\Delta\nu$, the variance of $\phi\langle|\phi - \langle\phi\rangle|^2\rangle \gg (2\pi)^2$, so ϕ modulo 2π is a uniformly distributed random variable. The noise probability density is therefore that of the cosine (peaked at the edges). The noise current variance is then $2I_1 I_2$, spread out over a frequency range from 0 to about $\Delta\nu$. The resulting 1 Hz current noise i_n as a fraction of the peak interference term is on the order of

$$\frac{i_n}{\sqrt{i_1 i_2}} \sim \frac{2}{\sqrt{\Delta\nu}}, \tag{2.11}$$

which can easily dwarf the shot noise. Using our 1 THz wide diode laser example, two 1 mW beams at 820 nm will generate fluctuation noise on the order of 2 nW/Hz$^{1/2}$ (optical), so that the detected noise will be nearly 40 dB (electrical) greater than the shot noise. This can be a pretty big shock when you're expecting the interference signal to go away completely. The same diode laser with the VHF modulation turned off might have $\Delta\nu \approx 10$ MHz, in which case the measurement will be much quieter for small path differences, but much noisier for large ones; for $\Delta t > 100$ ns, i_n will be more like 0.4 μW/Hz$^{1/2}$, an absolutely grotesque figure—the photocurrent noise will be 86 dB over the shot noise.

Even for path differences well within $1/\Delta\nu$, this effect can be very large; at 1% of the coherence length (3 μm path difference here), the RMS phase wobble is on the order of 0.01 radian, which in the 1 THz instance would bring the fluctuation noise near but not below the shot noise level. Note that this noise current grows as i, as signal does, not $\sqrt{i}$ as shot noise does; increasing the laser power will make it relatively worse. We see that even if $\Delta\nu$ is huge compared with the measurement bandwidth, these seemingly unimportant fluctuations may be the dominant noise source. How's that for an insidious gotcha?

One final addendum: it is not unusual for a very small-amplitude side mode in a diode laser to have very significant mode partition noise, and it gets out of phase with the carrier really easily in the presence of dispersion.

2.5.4 Measuring Laser Noise in Practice

Okay, so there are lots of noise sources to worry about. How do we measure the actual noise? In two steps: first, calibrate the detection setup; and second, measure the noise.

For calibration, take whatever setup you have and shine a battery-powered, incandescent-bulb flashlight on it. Move the flashlight in and out to make the photocurrent roughly the same as in your laser measurement. This will give you a beautifully calibrated white noise source (the shot noise of the photocurrent) of spectral density $i_N(1 \text{ Hz}) = \sqrt{2eI_{DC}}$, delivered right to your photodiode. Measuring the noise level in the dark (circuit noise) and with the flashlight (circuit noise + shot noise) will give you the gain versus frequency and noise versus frequency of your measurement system. This will allow you to compare the noise of the laser to the shot noise.

To do the actual measurement, the best bet is a spectrum analyzer. You can do this with a fancy digitizing scope, by taking the discrete Fourier transform of the output, squaring it, and averaging many runs together. (In a DFT power spectrum with only one run, the standard deviation is equal to the mean, so to get good results you have to average quite a few spectra or take the RMS sum over many adjacent frequency bins—see Section 17.5).

Alternatively, you can get a filter whose bandwidth coincides with your measurement bandwidth, and just measure the total power that gets through it, using a sufficiently fast RMS-DC converter (the author often uses an old HP3400A RMS voltmeter, which has a 10 MHz bandwidth). The noise power is the time-averaged total power of the signal, minus the DC power. Since the noise and the DC are uncorrelated, it doesn't matter whether you subtract the DC before or after taking the RMS. You can't do a good job with just a scope and no filter, because you won't know the bandwidth accurately enough.

Really work at relating the laser noise to the shot noise, because if you don't, you won't know how well or badly you're doing.

2.6 MORE SOURCES

2.6.1 LEDs

Light-emitting diodes (LEDs) are fairly narrowband continuum sources, which are nowadays ubiquitous because of their low cost and good properties. Their lifetimes are so long (hundreds of thousands of hours, without abuse) that you'll probably never wear one out. Their other properties are intermediate between those of tungsten bulbs and lasers; their spectral width is about 5–20% of their center frequency (say, 50–100 nm full width for a 600 nm LED) and their emitting areas are smallish (100 μm diameter or so). Their moderately wide bandwidth means that their temporal coherence length is extremely small, about 3–10 microns, which is often very useful where some spectral information is needed but etalon fringes must be avoided.

LEDs are available in a wide range of colors, from the near IR (900–1100 nm), where they are used for remote control, infrared LANs, and multimode fiber connections, to the ultraviolet (350 nm). Some LEDs have more than one spectral peak, so if this matters, make sure you measure yours.

Red LEDs are the most efficient, but the other wavelengths are catching up fast. Already the best LEDs are more efficient than tungsten bulbs, and their efficiencies continue to climb. White LEDs are really a species of fluorescent bulb—a UV LED excites a fluor in the package, and it's the fluor that limits the coherence and lifetime.

There are a few good reasons for the historical inefficiency of LEDs. The efficiency of generating photons from carriers is reasonably good, approaching 100% in direct bandgap devices. Unfortunately, these photons are generated deep inside a highly absorbing semiconductor material with a refractive index of about 3.3. This hurts in two ways: the semiconductor absorbs the light, and the critical angle for light exiting the semiconductor into air is arcsin(1/3.3), or about 18°, so that the solid angle through which photons can exit is only 0.3 steradians, or about 2.3% of a sphere. Fresnel losses on reflection limit the actual extraction efficiency of uniformly distributed light to below 2%. If the substrate is absorbing, the totally internally reflected photons will mostly be absorbed before getting another chance. The absorbing substrate problem is addressed by building heterojunction LED with transparent top layers (and newer ones have transparent substrates as well, with only the junction region being absorbing). The total internal reflection problem is helped by putting the LEDs in a plastic package ($n \approx 1.5$), which more than doubles the extraction efficiency, and by roughening the surfaces, which gives the light lots of chances to escape.

An LED's intermediate source size means that its spatial coherence is not very good compared with a single-mode laser, but is still better than you can do with a tungsten bulb. Again, this is usually a virtue except in interferometry.

Unlike nearly any other source, LEDs can have their intensities varied continuously over a huge range (from 0 to their maximum rated power) with little or no spectral shift, by merely changing their drive current. Spectral shifts remain small until maximum power is approached, at which point most LEDs shift slightly to the red,[†] due to high die temperatures (LED wavelengths tend to drift at about +100 ppm/K) and to high level injection effects, which narrow the effective bandgap. This throttling capability is of great value in extending the dynamic range of really low cost sensors.

Ordinary LEDs are fairly slow compared with laser diodes, being limited by carrier lifetime rather than stimulated emission. You can modulate an LED at a few megahertz, but garden variety ones don't go faster than that. IR LEDs intended for communications can go a bit faster, up to about 100 MHz.

[†] That is, toward longer wavelengths, not necessarily closer to 630 nm. This is really visible with yellow LEDs, which turn orange at very high drive currents.

The main problem with LEDs is their packaging. For efficiency reasons, nearly all LEDs come in molded plastic packages with cylindrical sides and very steep lenses on the top surface. The optical quality is poor and the die centration erratic. Furthermore, the wide angle of illumination of the sides leads to lots of internal reflections in the package. Together, these defects produce far-field patterns that look like a cheap flashlight's. The plastic package also limits the LED's power dissipation, because its thermal resistance is very high, so that the heat has to travel down the leads. For applications where lower spatial coherence is desirable, an integrating sphere will produce a nice Lambertian distribution, and even a ground glass, 3M Magic Invisible Tape, or holographic diffuser will help considerably. Some LEDs have glass powder mixed with the plastic, producing a frosted effect with better uniformity but lower efficiency and wider angular spread.

If higher spatial coherence is needed, consider using a surface-mount LED with a flat surface. Another way is to use a laser diode below threshold as a spatially coherent LED. The spatial characteristics are lithographically defined and the windows are good, so their beam qualities are excellent. Their coherence length is short and depends on how hard you drive them (it gets longer as threshold is approached), but is generally shorter than in multimode lasing operation, and of course very much shorter than in single-mode operation. Laser diodes are more expensive and (below threshold) won't produce as much total power as a bright LED. Compared to the work of removing the encapsulant from an LED or using LED chips and wire bonds, a diode laser below threshold is a bargain, unless you need more than a few hundred microwatts of power.

2.6.2 Superluminescent Diodes

The laser-diode-below-threshold idea can be extended to higher gains by removing the regeneration (e.g., by AR-coating one or both facets) or by gouging the back facet with a scriber. The resulting device has stimulated emission but little or no regeneration, and is called a superluminescent diode (SLD). Its coherence length is short, although not as short as a LED's, and it is a bit less sensitive to back-reflections than a diode laser. SLD linewidths are a few percent of the center wavelength.

SLDs are far from immune to back-reflection, however; residual resonator effects usually leave ugly peaks and valleys in their spectra, so the fringe visibility doesn't usually wash out to 0 as smoothly or as quickly as one would like. If you're using SLDs with fiber pigtails, make sure you measure the spectrum *after* the pigtail is attached, and generally be suspicious.

Commercial SLDs (sometimes confusingly called SLEDs) are available with output powers up to 10 mW or so. They're made with AR coatings and have to be run at very high current density due to the lack of regeneration. This makes them even more sensitive to optical feedback than the homemade ones. A perpendicularly cleaved fiber end is enough to make them lase, which (due to the high level of pumping) will usually blow them up. Their modulation bandwidths are more like LEDs' than laser diodes' ($\sim$10 MHz).

2.6.3 Amplified Spontaneous Emission (ASE) Devices

Optical amplifiers all have spontaneous emission. Leaving out the input signal and cranking up the pump power causes the spontaneous emission to be amplified like any other signal, producing an ASE source. Commercial ones produce tens of milliwatts from ytterbium (1000–1100 nm), neodymium (1060 nm), or erbium (1560 ± 40 nm), but are expensive. Their center wavelengths can vary a lot depending on the fiber material and the emission bands selected, and their linewidths are a few percent, becoming narrower at higher gains. A better behaved ASE device is the *frequency-shifted feedback* laser, with an acousto-optic cell in the cavity. These permit multipass gain while preventing normal lasing, because the Nth pass gets shifted by $2N$ times the acoustic frequency.

2.6.4 High Pressure arc Lamps

Thermal sources cannot exceed the brightness of a black body at the same temperature. The evaporation rate of the most refractory metals limit filament temperatures to below 3500 K, so plasmas are the only way to get brighter thermal sources. There are several types of plasma sources, divided into high and low pressure plasmas, either continuous (arc lamps) or pulsed (flashlamps).

High pressure devices are much brighter than low pressure ones. Although they are thermal sources, they are far from thermodynamic equilibrium, so their spectra are usually dominated by strong emission lines corresponding to electronic transitions in the component gases. These lines are very broad (tens of nanometers typically) due to collisional broadening from the high pressure of hot gas. There is also a weaker black body continuum from the plasma, and still weaker, lower temperature thermal radiation from the hot electrodes.

Arc lamps are available with a variety of fill gases, ranging from hydrogen to xenon, including especially mercury/argon and sodium/argon. Mercury arcs are the most efficient, but produce most of their output in the UV (especially near 254 nm), and their visible output is very green due to a strong line at 546 nm. Sodium vapor is nearly as efficient as mercury, but is very orange (590 nm). Xenon arcs appear white to the eye, which makes them popular for viewing applications. Hydrogen and deuterium lamps produce a strong continuum in the UV, with D_2 units going far into the vacuum UV, limited mainly by the windows (for VUV operation, they are used windowless, with differential pumping).

The emission lines get broader and weaker, and the continuum correspondingly stronger, as the gas pressure is increased; extremely high pressure arcs (200 atmospheres) can make even mercury vapor look almost like a 6000 K light bulb.

The plasmas are inherently unstable, so the spectrum and power output fluctuate with time in a $1/f$ fashion, with ±10% power variations being common. There are variations caused by strong temperature gradients in the tube, causing big changes with time and warmup. The pressure increases with temperature, which changes the broadening. The position of the arc is unstable with time as well, at the level of 0.1 to a few millimeters, which will cause the light from your condenser to wander around. In some bulb types, the spot moves erratically at a speed of 50 m/s or faster, leading to lots of noise in the kilohertz range. They are a great deal noisier than tungsten or LEDs, and their noise has a spatially dependent character that makes it hard to correct for accurately.

Commercial arc lamps are divided into short arc and long arc styles. The length of the arc partly governs the source size, and thus the spatial coherence of the light: you can do a better job of collimating the light from a short arc lamp. They come in two main package types: squat ceramic and metal cylinders with quartz windows and integral reflectors, and long quartz tubes with a bulge in the middle. The ceramic type is much easier to use but has a few idiosyncrasies. Most such lamps have an arc formed perpendicular to the window, so that the cathode gets in the way of the beam. The package contains a parabolic reflector, which roughly collimates the light reflected from it (typical angular spreads are $\pm 1^\circ$ to $\pm 3^\circ$, with the spread increasing as the bulb ages). Unreflected light floods out into a wide angle; the illumination function is thus rather strange looking, with a strong doughnut-shaped central lobe and a weaker diffuse component. On the other hand, the arc and reflector come prealigned, which substantially simplifies changing the bulb.

EG&G sells ceramic arc and flashlamps that circumvent this problem, by arranging the electrodes transversely (so that the current flow in the arc is parallel to the window and the electrodes are not in the way), and using a spherical reflector with the arc near its center of curvature. All the light therefore emanates from the arc, and both the strange pupil functions are avoided. You still have to collimate the light yourself, however. One choice will probably be better than the others in your application.

2.6.5 Flashlamps

A flashlamp or flashtube is a pulsed arc lamp, usually filled with xenon or krypton, and is used where very bright, hot, short-duration illumination is needed. Flashlamps can reach source temperatures of over 10^4 K, making them instantaneously about 100 times brighter than a tungsten source of the same size. When used at high power, their plasmas are *optically thick* (i.e., highly absorbing[†]), so that their radiation is dominated by the continuum component. At lower pulse energies, the plasma is cooler (closer to 5000 K than 10,000 K), and the line spectrum is more pronounced.

Flashlamps are powered by capacitor discharge; you charge up a special capacitor[‡] that is connected to the lamp through a very small inductor (which may be just the leads). When the capacitor is fully charged, you trigger the arc, either with a tickler coil, wound round the cathode end of the tube and driven with a 10 kV pulse, or by momentarily increasing the voltage across the tube with a pulse transformer.

Once the arc gets going, it discharges the capacitor in a time controlled by the impedance of the arc, wiring, and capacitor parasitic inductance and resistance (including dielectric losses), until the voltage drops below that needed to keep the arc going. This variable excitation means that the light intensity is strongly peaked with time; unless you do something special, there's no flat top to the light pulse, regardless of its length. In fact, the current can easily oscillate due to LC resonance in the capacitor and leads. Flashlamp manufacturers publish data allowing you to design a critically damped network, so that the total pulse width is minimized for a given tube, capacitor, and drive energy.

The total duration is usually between 100 μs and 10 ms, although active controllers (using big power MOSFETs) can make it as short as 1 μs. Pulse energy is limited by explosion of the bulb and ablation of the electrodes. Getting long life (10^5–10^7 flashes) requires operating the bulb a factor of 6–10 below the explosion energy. Tubes with explosion energies between 5 J and 10 kJ are readily available.

Peak power is limited by how fast you can push that much energy into the bulb, and average power is limited by cooling. Flashlamps can work at kilohertz repetition rates if the pulses are weak enough. There is considerable variation in the pulses; a well-designed flashlamp supply with a new bulb may achieve 0.1% rms pulse-to-pulse variation, but

[†]It is perfectly correct from a thermodynamic point of view to call the central plasma core of a 1 kJ flashlamp "black." Try it some time.

[‡]Capacitors intended for rapid discharge have very low internal inductance and resistance, to improve their discharge time constants and avoid dumping lots of heat into the capacitor itself.

typical ones are more like 1%. They have the other noise characteristics of regular arc lamps, for example, strong spatial variations, wandering arcs, and lots of $1/f$ noise. Their wall plug efficiency can be over 50% when integrated over all ν, but is more typically 10%.

Flash initiation depends on the generation of the first few ions in the gas, which is inherently jittery compared with laser pulses, for example. Typical jitters are in the hundreds of nanoseconds and are a strong function of the bulb design. Jitter can be reduced with a "simmer" circuit, which maintains a low current arc between flashes, or a pretrigger that initiates a weak arc slightly before the main pulse.

Flashlamps and arc lamps with concentric reflectors are intrinsically more resistant to source noise from arc wander, because the reflector projects an inverted image back onto the arc. If the arc moves to the left, the image moves to the right. For an optically thick plasma, the effect can even help stability a little; if the left side gets hotter, its image dumps heat into the right side.

2.6.6 Spark and Avalanche Sources

You can get 400 ps rise time from a spark in a mercury switch capsule.[†] If subnanosecond speed is enough, and you don't need narrow linewidth or high brightness, this can be a convenient and cheap alternative to a pulsed laser. The sharp current pulse will couple into everything in the neighborhood, so careful electrical shielding will be needed. Spark sources have all the jitter and pulse-to-pulse variation problems of flashlamps, and since the rapid breakdown is what we're after, they can't easily be fixed with simmer circuits and so on.

2.7 INCOHERENT LINE SOURCES

2.7.1 Low Pressure Discharges

We are all familiar with low pressure gas discharge tubes, such as Geissler tubes (beloved of undergraduate optics labs) and ordinary domestic fluorescent bulbs. Low pressure discharges produce a line spectrum by electrically exciting gas molecules and letting their excited states decay by radiating photons. The positions of the spectral lines are characteristic of the molecular species doing the radiating, while the line strengths and linewidths are more strongly affected by experimental conditions such as pumping strategy, temperature, pressure, and other gases present.

The gas fill is often a mixture of gases to improve arc characteristics and aid energy transfer to the desired excited state of the radiating molecule, for example, argon in mercury lamps and helium in neon bulbs. Argon is a particularly popular choice; besides being cheap and inert, it has a low first ionization potential, so that arcs are easily struck. The striking voltage depends on pressure, with a minimum at about 1 torr, rising at high vacuum and at high pressure.[‡]

Their fairly narrow linewidths (moderate temporal coherence) and low spatial coherence make low pressure lamps useful for qualitative imaging interferometers. The most

[†]Q. A. Kerns, F. A. Kirsten, and G. C. Cox, Generator of nanosecond light pulse for phototube testing. *Rev. Sci. Instrum*. **30**, 31–36 (January 1959).

[‡]M. J. Druyvesteyn and F. M. Penning, *Rev. Mod. Phys*. **12**, 87 (1940).

common example is preliminary testing of optical surfaces via Newton's rings and Fizeau fringes (see Section 12.6.2).

The exceptions to the line spectrum rule are low pressure hydrogen and deuterium lamps, which emit a bright continuum extending into the VUV. Deuterium lamps emit almost no visible light at all, which is helpful in UV experiments because the filters and gratings that would otherwise be needed are often highly absorbing. These arcs are often run at extremely low pressure, especially deuterium, where the interior of the bulb connects immediately with the vacuum system, and deuterium leakage is controlled only by differential pumping.

2.8 USING LOW COHERENCE SOURCES: CONDENSERS

2.8.1 Radiometry of Imaging Systems

We saw in Section 2.4.1 that the product of area and projected solid angle is invariant under magnification and Fourier transforming, so that in the absence of vignetting or loss, the source and image radiance are the same, and the average radiance is the same in the pupil as well. Unfortunately, it is not easy to avoid vignetting when the source radiates into 4π steradians.

It is a common error to assume that a higher power bulb will get you more photons for your measurement. Thermodynamics dictates that no point in the optical system can have a higher radiance than that of the source (i.e., the arc or filament). Since filament and arc temperatures are lifetime-limited, this means that there is an absolute upper limit to how bright you can make your illumination with thermal sources. A higher wattage bulb is merely larger, not brighter. Accordingly, upgrading the bulb will produce a larger illuminated area or aperture, but the same radiance. If that's what you need, a higher wattage bulb can help. If not, you'll have to improve your condenser and collection optics instead, or switch to laser light.

2.8.2 The Condenser Problem

The challenge in illumination systems is to achieve the characteristics needed for good measurements: constant angular distribution, stability in intensity and polarization, and uniformity of brightness and spectrum across the field. Ideally it should also use the available source photons reasonably efficiently. Optical systems using tungsten illumination are almost always extremely inefficient with photons, because being efficient is tough. It isn't usually worth the trouble, either, because the equipment is mains powered and tungsten bulb photons are very cheap,† but other illuminators don't have this luxury.

As shown in Figure 2.2, the simplest kind of condenser is an imaging system, basically a lens corralling the source photons and pointing them toward the sample. Trying to get more photons will involve increasing the solid angle of the light collected (i.e., the NA of the condenser on the bulb side).

The envelopes of some kinds of bulbs (especially tungsten halogen) are serious impediments to illuminator design. If the optical quality of the bulb envelope is good enough, it may be possible to use the reflector to put the image of the filament right next to the real filament, which will improve the illuminated area and hence the efficiency.

†About $\$10^{-23}$ each.

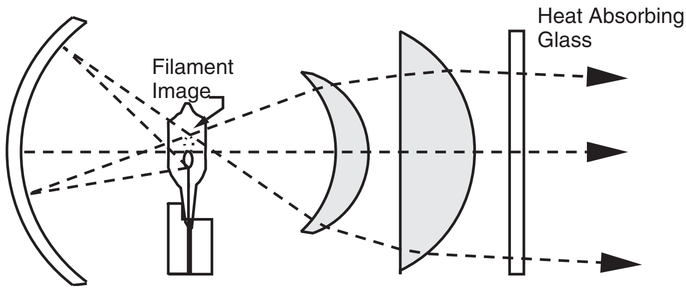


Figure 2.2. A typical condenser is a low quality imaging system with care taken in controlling thermal loads.

On the other hand, an irregularly shaped bulb envelope, or one with a poorly designed internal reflector, makes it nearly impossible to approach the thermodynamic limit for light coupling. Most fall into this category.

Example 2.1: Fiber Illuminators. Getting the highest radiance from a tungsten bulb may not always involve a condenser. For fiber bundle coupling, just image the filament on the bundle, using an NA (on the bundle side of the lens) high enough to overfill the fiber NA. The rest of the housing can just toss light back into the filament to reduce the electrical power required. If the bundle has a very high NA (some are 0.5), it may not be easy to get lenses like that. Figure 2.3 shows ways of dealing with this problem, using an array of lenses, a Fresnel lens, or a fast spherical mirror. With the bulb just off to one side of the center of curvature and the bundle at the other side, this does a good job of filling the bundle NA and diameter. The mirror doesn't have to be too good; there's no reason the image quality can't be reasonable, since both source and bundle are near the center of curvature. The mirror can have a cold mirror coating (which passes IR and reflects visible light) to reduce the heat loading of the fibers. (Pretty soon we'll just stick a white LED at the business end and get rid of the bundle entirely.)

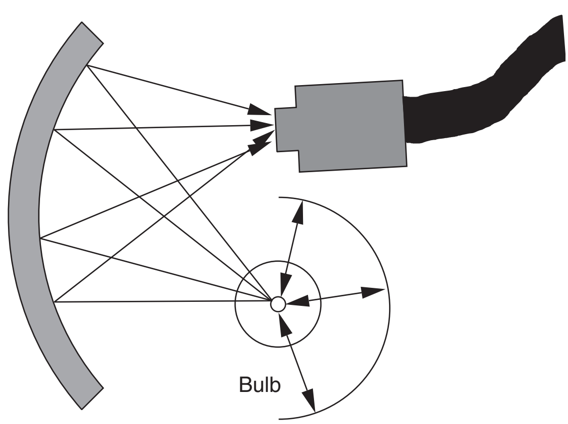


Figure 2.3. Condensers for fiber bundles: a simple lens or a spherical cold mirror.

Most people are better off buying condensers when they need them, since condenser design is a somewhat specialized art, and combination lamp housings and condensers are widely available commercially. If you design your own, take apart a couple of different commercial ones to see the tricks they use. Be sure to remember the heat absorbing glass, which is a tempered filter that dissipates the heat that would otherwise toast your optical system or your sample.

2.9 LASERS

Lasers have an immense literature, and come in a bewildering variety. We'll concentrate on the properties of the ones most commonly seen,† especially diode lasers. See Siegman‡ for much more theoretical and pedagogical detail.

Lasers rely on an external energy source, the *pump*, to provide a population inversion and hence laser gain. Oscillation requires a minimum pump power, known as the *threshold*. How many times above threshold the pumping is determines how rapidly the oscillation builds up, among many other parameters. For instruments, the distinctions between laser types center on wavelength, tunability, power level, pumping source, pulse width, and linewidth.

2.9.1 Mode Structure

A laser is a resonant cavity with a gain medium inside, and some way of coupling the light in the cavity to the outside, usually a partially transparent cavity mirror, the output coupler. The resonant condition requires that the cavity be carefully aligned, which is a fiddly process if it's user adjustable.

Cavities always have more than one mode, so that the laser's spatial and temporal behavior will be a superposition of cavity modes. Real lasers have complicated mode properties, but we can do a good job by decomposing the fields in terms of the modes of an unperturbed, empty resonator. These break down into the transverse modes, which are well described by Gaussian beams, and longitudinal modes, which are like the resonances of an organ pipe. All will have slightly different frequencies in general, constrained by the limits of the gain curve of the laser medium. For our purposes, transverse modes higher than TEM_{00} are unwanted, but fortunately they are usually easily suppressed, so that most commercial lasers have a single transverse mode.

The longitudinal modes are defined by the requirement that the round-trip phase in the laser resonator should be a multiple of 2π radians, so that multiple bounces will give contributions that add in phase and so produce a large total field. Accordingly, longitudinal modes are spaced by very nearly $\Delta\nu = c/(2n\ell)$, where 2ℓ is the round-trip distance (twice the cavity length) and n is as usual the refractive index of the material filling the cavity. For a large frame gas laser, $\Delta\nu$ is on the order of 200 MHz, but it ranges up to ~150 GHz for a cleaved cavity diode laser and 1 THz for a VCSEL. Diffraction effects and the pulling of the resonant frequency by the gain slope of the laser medium and coatings often prevent the longitudinal modes from being exact harmonics of each other,

†The most important neglected types are dye lasers. These were once very important for lab systems but have been eclipsed lately, except for those few applications requiring wide tunability in the visible.

‡Anthony E. Siegman, *Lasers*. University Science Books, Mill Valley, CA 1986.

which produces some noise and spurious signals in the detected photocurrent. Coupling between the modes will make them lock to each other if these effects are small enough.

Simplified laser theory suggests that a single longitudinal mode should always dominate, with a linewidth that narrows enormously for pumping rates well above threshold. In practice, multiple longitudinal modes are more commonly found, and pumping harder makes it worse. For interferometric applications, multiple modes are undesirable, because they cause the coherence length to be much shorter than it needs to be and produce noise and sensitivity variations with a periodicity of twice the cavity length. These applications often need single longitudinal mode (single-frequency) lasers.

A laser may even oscillate in more than one spectral line at once; this is suppressed by using a dispersing prism or grating as part of the cavity, so the cavity can only be aligned for one line at a time.

The equal spacing of these modes means that their configuration should be periodic with time, like an ideal piano string.

2.9.2 Relaxation Oscillation

Lasers work by running electrons through a loop consisting of three or four quantum transitions. Each of these transitions has its own time constant, which as any circuits person knows, can lead to instability due to accumulated phase shifts. Good lasers have one dominant time constant, namely, the long lifetime of the upper state of the laser transition. The other time constants lead to excess phase shift, which generally causes ringing in the laser output and level populations whenever a sharp change in pumping occurs, just like a feedback amplifier with too small a phase margin. Due to the historical fact that lasers weren't developed by circuits people, this ringing is miscalled *relaxation oscillation*. It isn't really oscillation, but it does limit the modulation response of the laser, and it causes excess noise near the peak—generally about 1 GHz for cleaved cavity diode lasers and more like 10 GHz for VCSELs. Diodes and solid state lasers show much stronger relaxation oscillations than gas lasers in general.

2.10 GAS LASERS

Gas lasers cover the entire visible, UV, and infrared range in isolated lines with very narrow tuning ranges. Typical examples roughly in decreasing order of usability are helium–neon (HeNe, 632.8 nm), argon ion (488 and 514.5 nm), helium–cadmium (HeCd, 442 and 325 nm), carbon dioxide (CW and pulsed IR at 10.6 μm), nitrogen (pulsed, 337 nm), and excimer (pulsed deep UV). Except for HeNe and CO_2, these are really suitable only for lab use or specialized applications such as laser light shows and medicine. Gas lasers are big, bulky, and fairly noisy (0.5–2% intensity noise).

Diode lasers are very popular just now because they're cheap, small, and mechanically rugged. Don't despise HeNe's, however—they are much less sensitive to optical feedback, are very frequency stable, have long coherence lengths (300 m for single-frequency units), are really well collimated, and work as soon as you plug them in. A HeNe is just the right medicine for a lot of ills. Besides, they come with their own power supplies and are totally invulnerable to electrostatic discharge (ESD) damage—if you don't smash it, it'll always just work. Low gain gas lasers such as HeNe's have no spatial side modes to worry about, so there aren't too many really weird sources of noise, apart from the occasional baseband mode beat (see Section 2.13.7).

HeCd lasers are very noisy, and N_2 lasers cannot be run CW due to bottlenecks in their level transition rates (you get 10 ns pulses no matter what you do—an example of relaxation oscillations that really oscillate). Excimer lasers are used for laser ablation of materials and are also widely used for semiconductor lithography. Their spatial coherence is very low; although their output may be below 200 nm, you can't focus their output more tightly than a millimeter or so.

Some gas lasers can be made single-frequency, either by reducing the gain and carefully tuning the cavity length (HeNe, $P < 1$ mW), or by putting a Fabry–Perot etalon (see Section 1.6.2) in the cavity (ion lasers). This is effective but painful, requiring temperature stabilization of one or both resonators to keep the tuning sufficiently stable.

Gas lasers are usually pumped by running an electric arc through the gas fill in a sealed tube with mirrors or Brewster windows fused to its ends. They are astonishingly inefficient, due to short upper state lifetimes and lots of deexcitation paths. For example, an Ar-ion laser is doing well to have a wall plug efficiency (laser emission/electrical input) of 0.02%, which leads to big AC supplies and water or fan cooling. The main exception is CO_2 lasers, which can hit 25% wall plug efficiency. All those Teslaesque features cause severe low frequency noise in most gas lasers, with the high power ones being worse.

Grating tuned gas lasers such as ion and metal vapour units usually have a diffraction grating or Littrow prism inside their cavities to select a single line, although multi-line units do exist. Getting a laser aligned is sometimes a painful process; it's discussed in Section 12.9.8. Apart from HeNe's and sealed CO_2s, gas lasers need significant amounts of tender loving care; if you're going to use one, learn how to clean and maintain the mirrors and windows properly, and if there's someone available who can show you how to align it, ask.

2.11 SOLID STATE LASERS

Solid state lasers are based on electronic transitions in impurity sites in solid crystals or glasses. Typical examples are Nd:YAG (1.06 μm), ruby (694 nm, pulsed), Alexandrite (700–820 nm), Ti:sapphire (0.66–1.2 μm, femtosecond pulses possible), and color center (0.8–4 μm, widely tunable).

Solid state lasers have better characteristics than gas lasers in general, and are much more flexible. Their efficiencies are limited by the pumping strategy employed; diode laser pumped Nd:YAGs have very respectable wall plug efficiencies (in the percents). The host crystals change both the center wavelengths and the strengths of the emission bands, so that the laser wavelength of a given ion will move by tens of nanometers depending on the host. Ruby (Cr^{3+}:Al_2O_3) was the first laser of all, and ruby units are still sold, though you'll probably never use one. Neodymium YAG (Nd^{3+} ion in yttrium aluminum garnet) lasers are much more common, because they have really good performance over a wide range of powers and pulse widths, from CW to the tens of picoseconds. The very long upper state lifetime of a YAG laser (250 μs) makes it efficient, and also allows it to store lots of energy. Powerful pulses can be achieved using Q-switching, where the cavity resonance is spoilt while the upper state population builds up, and is then rapidly restored (usually with an acousto-optic or electro-optic cell, but sometimes with a bleachable dye in low energy units). The cavity finds itself way above threshold and very rapidly emits a huge pulse, typically 1–10 ns wide.

A somewhat less well-behaved pulse can be obtained by letting the cavity build up a big circulating power, and then using the AO or EO cell to allow this energy to exit the cavity in one round-trip time, a technique called cavity dumping.

Pulsed YAGs are usually pumped with flashlamps, whose pulse width is a good match to the upper state lifetime. Efficiency suffers because the pump bands (the spectral regions in which pump light can be converted to laser light) are not well matched to the flashlamp spectrum. CW YAGs are nowadays usually pumped with 808 nm diode lasers. Single-longitudinal-mode diode-pumped YAGs are highly efficient, can be pretty quiet, too—the best ones are 10–30 dB above the shot noise at moderate modulation frequencies, and even closer above 20 MHz or so.

YAG lasers are often frequency doubled, yielding green light at 532 nm. These are familiar as low quality green laser pointers, but with a bit more work, A diode-pumped, frequency doubled, single-frequency YAG laser is a very attractive source: intense, efficient, quiet, and visible. High peak power units can be tripled or quadrupled into the UV. Doubling has to be done inside the cavity to get decent efficiency, and there are a number of sources of potential instability in the process.

Other neodymium-doped crystals are Nd:YV_{03} (yttrium vanadate) and Nd:YLF (yttrium lithium fluoride). Both have better thermal behavior than YAG—they're birefringent, which reduces losses due to thermal stress-induced birefringence (just as in PM fiber), and both have low dn/dT, which reduces thermal lensing and so improves beam quality. Vanadate lasers are also easier to run on the weaker 914 and 1340 nm lines. Neodymium-glass lasers can produce higher pulse energy than YAG, but the very low thermal conductivity of glass compared with YAG or vanadate severely limits the repetition rate, so that the average power is lower for the same size device. The spectrum is shifted slightly and shows a lot of inhomogeneous broadening. The upper state lifetime is much shorter and hence efficiency is lower.

Titanium–sapphire lasers are now the predominant choice for femtosecond pulses, having taken over the honor from colliding-pulse modelocked (CPM) dye lasers.

Diode-pumped YAGs, usually called diode-pumped solid state (DPSS) lasers, are potentially extremely quiet; their long upper state lifetime has the effect of filtering out the wideband AM noise of the pump, and the pump laser can be power-controlled over a wide bandwidth, so in the absence of mode jumps, mode-partition noise, thermal lensing, and photorefractive instability, a single-frequency DPSS laser is a very well-behaved device.

Decent solid state lasers are not simple or cheap devices and must be pumped with another light source, usually a flashlamp or diode laser. They typically run \$5000 to \$50,000, and lots more for the really fancy ones. Sometimes there is no substitute, but you'd better have a decent grant or be selling your system for a mint of money.

2.11.1 Modelocked Lasers, Optical Parametric Oscillators, and Other Exotica

Like vacuum systems, laser design is an arcane and lore-rich subject on its own—see Koechner's *Solid State Laser Engineering* and Siegman's *Lasers* if you need to know more details. All there's space for here is that the longitudinal modes of an ideal laser form a complete basis set for representing any pulse shape you like—including a δ-function. By forcing all the modes to oscillate in phase at time t_0, $t_0 + \Delta$, $t_0 + 2\Delta \ldots$, many cavity modes combine to produce a train of narrow pulses, with the period of cavity round-trip time. The phase and amplitude shaping is done by a modulated attenuator or by a saturable absorber, which attenuates bright peaks less than their dimmer wings, thus narrowing the pulse on every bounce. Until the limit set by loss and dispersion is reached.

Pulsed lasers are much harder to maintain than CW units in general, with the high pulse power, low rep rate, lamp-pumped modelocked units being generally the worst, on account of the thermal transients, vibration, coating damage, and general beating up that the crystals take when they're clobbered that hard. If you're exclusively a CW person, count your blessings—you've led a sheltered life.

Aside: Beam Quality. A complicated system such as a picosecond optical parametric generator pumped by the third harmonic of a lamp-pumped YAG laser is not going to have the beam quality of a TEM_{00} HeNe, no matter what. The good ones have decent spots in the near field that turn into mildly swirly messes in the far field; bad ones produce beams that resemble speckle patterns regardless of where you look. Small changes in conditions can produce severe beam degradation, so make sure that whatever source you have, you can measure its beam profile accurately and easily. Especially in the IR, it is amazing how many people trying to make complicated measurements (e.g., sum-frequency generation spectroscopy) don't have any idea of their beam quality. Many kinds of nonlinear sources, especially OPOs, have beam profiles that are pulse-height dependent, so you can't normalize them with an energy meter no matter how hard you try.

2.12 DIODE LASERS

The champs in efficiency and cost effectiveness are diode lasers. Diode laser photons are more expensive than tungsten photons, but cheap for lasers, and their output quality is good. Linewidths of 10–100 MHz or so are typical for single frequency units. Wall plug efficiencies can reach 40%, with slope efficiencies ($\partial P_{out}/\partial P_{\text{in}}$) of 80% or even higher.

Most diode lasers are of the cleaved cavity (Fabry–Perot) type: the die is cleaved to separate the lasers, and the cleavage planes form the cavity mirrors. They are usually coated at the rear facet with a moderately high reflector. The front facet can be left uncoated or may have a passivation layer. The Fresnel reflection on leaving the die is sufficiently strong (40% or so) to sustain oscillation with the extremely high gains available in the active region.

Current is confined to a very small active region by giving the surrounding material a slightly higher bandgap, so that the forward voltage of the diode in the active region is smaller, and it hogs all the current. The resulting spatial restriction of laser gain helps guide the beam, and lasers relying on this are said to be *gain guided*. A better method is *index guiding*, where the refractive index profile makes the active region a stable waveguide as well as confining the current.

Due to this monolithic construction, diode lasers are mechanically extremely rugged, although their associated collimators generally are not. Their packages are similar to old-style metal transistor cans, for example, TO-3, 9 mm (TO-8), and 5.6 mm (TO-72) for the smallest ones. Most have a monitor photodiode in the package, to sense the light emitted from the rear facet of the diode for intensity control. The monitor is big enough that it is not completely covered by the laser die, so that it will pick up light scattered back into the diode package, which makes it somewhat unreliable in practice.

Diode lasers are available in narrow wavelength ranges centered on 380, 405, 635, 650–690, 750–790, 808, 830, 850, 915, 940, 980, 1310, 1480, and 1550 nm (plus a few at oddball wavelengths like 1.06 and 1.95 μm), which are defined by their intended use. The shortest-wavelength diodes widely available are 405–410 nm high power multimode units (from 20 mW up to 100–200 mW, linewidth 1–2 THz) for Blu-Ray discs. At this writing (late 2008) the cheapest way to get these is to cannibalize them from the drives, because they cost $1000 apiece otherwise. There's a big hole between violet (405 nm) and red (633 nm), where there are currently no good laser diodes, so diode-pumped solid state lasers are the usual choice. Lasers used in optical drives have lots of power—roughly 150–200 mW for 658 nm DVD burner diodes. Longer wavelengths, 2–3 μm, can be obtained with quantum cascade lasers based on superlattices.

The market for diode lasers is dominated by optical storage and telecommunications, so diode types come and go, and they're expensive in ones and twos. CD/DVD player lasers are the cheapest: 650 nm, 5–7 mW or less. These cost around $10 unless you're a DVD player manufacturer and get them for $0.50 each in 100,000 piece quantities. Fancier ones, such as the SDL-5400 series of Figure 2.4, can give you >100 mW of single frequency power. Multiple longitudinal mode diodes reach about 1 W, and multiple transverse mode units (based usually on many apertures side by side) are approaching

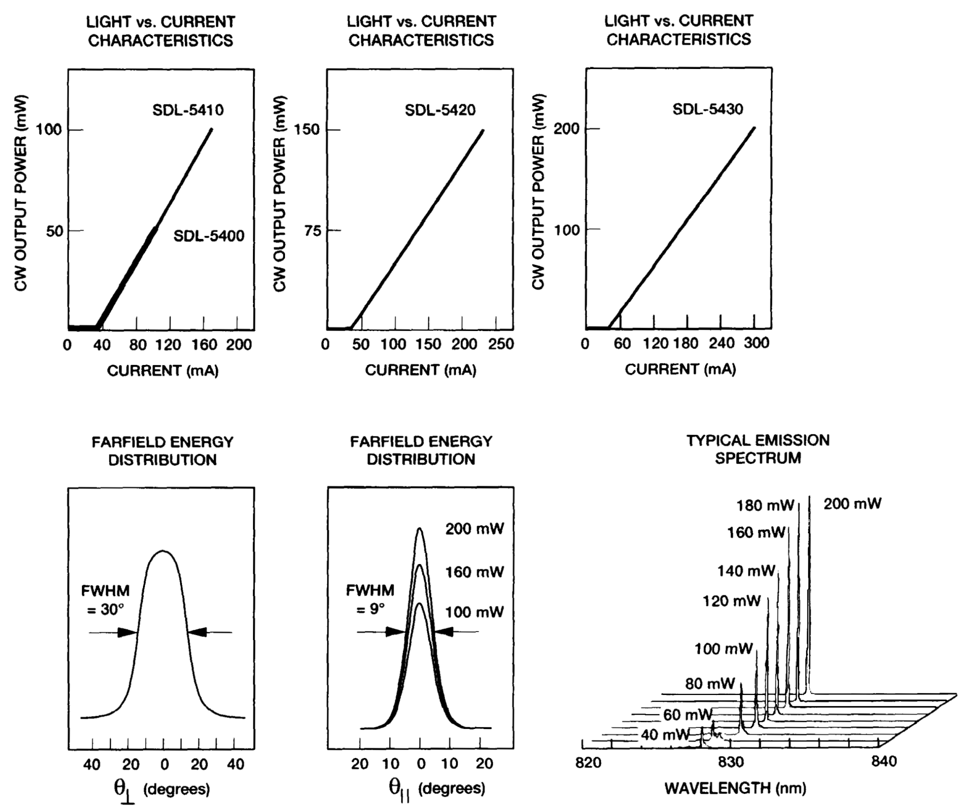


Figure 2.4. Beam parameters of Spectra Diode Labs (now JDSU) 5420 diode laser (From SDL, Inc., and reprinted with permission © 1994, 1997 SDL, Inc.)

1 kW CW. These bar-type diodes are best used as light bulbs, because their coherence properties are poor.

The bad things about diode lasers are that they are extremely delicate electrically, that they are inconvenient to drive and to collimate, that their tuning depends on everything under the sun, and that their mode hops and instabilities will drive you nuts if you're unprepared or need better-than-ordinary performance.

The other major drawback with all these low cost devices is that they are tightly tied to consumer applications in printers, DVD players, and so on. When the consumer technology moves on, the lasers go away; if you're using them in your instrument, and you don't have a big stock of them, you're out of luck.

2.12.1 Visible Laser Diodes

Working in the infrared is more difficult than in the visible, and the available spatial resolution is less for a given NA. Commercial visible laser diodes (VLDs) work near 670, 650, 630, and 405 nm, which is convenient for human eyes. High power VLDs are used in CDR and DVDR drives, so they've become quite common. Unfortunately, VLDs behave significantly worse than their IR brethren, with poorer tunability and frequency stability and more mode hopping. If you need good noise performance from your diode laser, look elsewhere.

Aside: Wavelength-Division Multiplexing (WDM) and Chirp. The intrinsic bandwidth of optical fiber is extremely wide—much wider than any foreseeable electronic switching frequency. Thus the most convenient way to get more capacity from a fiber-optic data link is by sending many independent signals down the same fiber, using a grating device or a sequence of narrowband couplers to combine the signals at one end and separate them at the far end. The wavelengths of the optical carriers conform to a standard spacing, the so-called *ITU grid*. The grid frequencies are multiples of 100.0 GHz, from 184.8 to 201.1 THz, with 50 GHz offsets also available. Achieving low crosstalk between channels requires that the laser tuning be very stable, both with time and temperature and (which is more of a challenge) with fast modulation. Normally the frequency of a semiconductor laser changes quite a bit—as much as a couple of nanometers ($\sim$250 GHz at 1.5 μm) [†] during a fast turn-on, which would scribble all over the grid if it weren't controlled. If the chirp is linear in time, it can be used with a grating to compress the pulse, but most of the time chirp is just a nuisance. VCSELs, even multimode ones, have much wider spectra but only a few modes; they have much lower chirp than FP lasers, which have many modes for energy to slosh about in (see Section 2.12.7).

2.12.2 Distributed Feedback and Distributed Bragg Reflector

Diode lasers need not use the cleaved ends of the die as mirrors. Distributed feedback (DFB) lasers use an active waveguide with a grating superimposed on it, resulting in very high selectivity. The similar distributed Bragg reflector (DBR) scheme uses separate gratings in passive waveguide regions, which can be tuned separately with a second bias

[†]Paul Melman and W. John Carlsen, Interferometric measurement of time-varying longitudinal cavity modes in GaAs diode lasers. *Appl. Opt*. **20**(15), 2694–2697 (1981).

current (some have more than one tunable grating segment). At one time, DFB lasers had better tunability since the same mechanisms that tune the frequency also tune the gratings to match, resulting in wide temperature and current tuning ranges with few or no mode hops. DFB lasers are expensive and specialized, so they're only available in the telecom bands. Chirp is very damaging there, as we saw, so lots of work has gone into reducing it; although you can tune modern DFB lasers with temperature, they hardly current-tune at all. For wider current tuning, DBR lasers are superior, if a bit less convenient to drive.

2.12.3 Tuning Properties

Everything in the world tunes diode lasers: temperature, current, pressure, and any sort of optical feedback (see Section 2.13.6). All of these effects can be controlled, but you have to be aware of the need. Highly stable diode laser instruments need clever control systems for temperature, current, and mode hop control, which we discuss in Chapter 20 (http://electrooptical.net/www/beos2e/thermal2.pdf) and Section 15.9.1.

Among Fabry–Perot lasers, the 800 nm ones are the best behaved. They tune at rates of around -0.015 to -0.08 cm^{-1}/mA and -0.1 cm^{-1}/K in a single mode (for a 5 mW unit), and around -4 cm^{-1}/K on average due to mode jumps. They can be tuned through 1–2 cm^{-1} by current in between mode jumps, and much further via temperature.

Tuning red VLDs is a different story. A small bandgap difference between the active area and its surroundings hurts the current confinement and makes VLDs very temperature sensitive and generally unstable. It is not trivial to find a nice single-mode operating region with a VLD, although it can be done in the lab if you don't need more than 0.5 cm^{-1} of current tuning range and are prepared to hunt. Violet and UV lasers are generally multimode. Stick with the 750–850 nm ones if you need to do anything fancy.

2.12.4 Mode Jumps

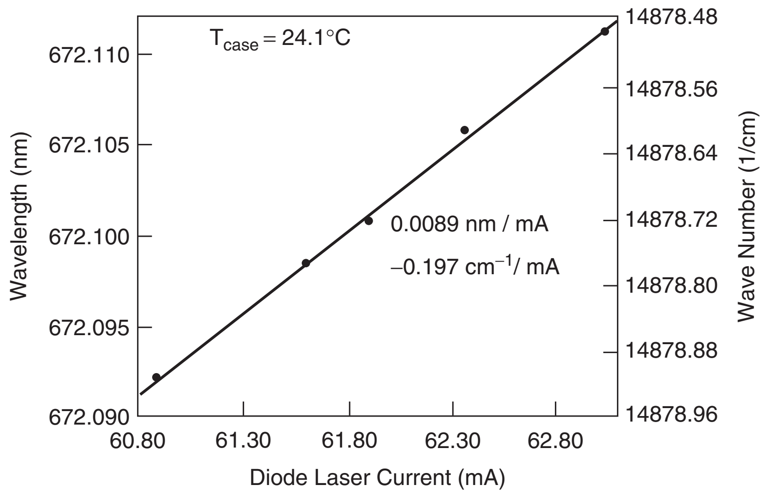
The tuning of single-mode diode lasers is discontinuous in both temperature and current, as shown in Figure 2.5, and the discontinuities unfortunately move around slowly with laser aging; even with perfect temperature and current control, your continuous tuning range won't stay the same. Another odd thing is that the tuning curve is multivalued: in some regions, the same temperature and current can support oscillation at two different frequencies. These two modes typically do not oscillate at the same time; it's one or the other, depending on history, a phenomenon called *hysteresis*.

The mode jumps move slowly downwards through the operating current range as you increase the temperature; they move much more slowly than the tuning, unfortunately, so you can't always get to your desired wavelength without external cavity stabilization.

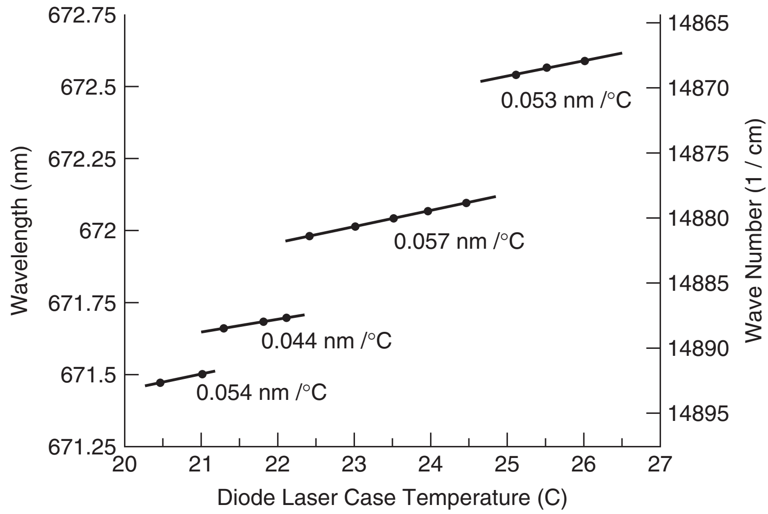
If you're building instruments relying on tuning F-P diodes, you normally have to ensure that you're working at a point where ν is a single valued function of T and I, and that takes some work. On the other hand, you only get DFB lasers at 1.3 and 1.55 μm, and they are two orders of magnitude more expensive, so fixing up F-P diodes to tune reliably is a worthwhile thing to do.

2.12.5 Regions of Stability

Besides the intrinsic Fabry–Perot mode structure, diode lasers in real optical systems always have some feedback, which is liable to lead to instability, and certainly modifies



(a)



(b)

Figure 2.5. Tuning properties of an FP visible diode laser (Toshiba TOLD9211): (a) versus current and (b) versus temperature.

the stable regions. If you vary the temperature and bias current to a diode laser, and look at its noise, you find two-dimensional islands of stability surrounded by a sea of mode hopping. There is some bistability at the edges, since the edges of the islands sometimes overlap, as we saw (they remain separate if you consider wavelength as a third dimension).

These islands can be mapped out fairly easily, and if you are using diode lasers in instruments it is a good idea to do that. Use some simple software for the job, because otherwise you'll be at it quite awhile; since the noise goes up so much during mode hopping, it is straightforward to design self-testing software for the instrument. The islands change slowly with time. VLDs have much smaller islands than 800 nm diodes.

If you want the islands to be as stable as possible, temperature control the collimating lens and the lens-to-laser spacing too. A collimator design that makes this easy is shown in Example 20.7 at http://electrooptical.net/www/beos2e/thermal2.pdf.

2.12.6 Checking the Mode Structure

Use a fine pitch grating (e.g., 2400 lines/mm at 670 nm), aligned so that the first-order beam exits near grazing incidence. If the laser is well collimated, the mode structure will show up clearly. You get a bunch of dim dots at the far-off Fabry–Perot transmission maxima, which are just filtered spontaneous emission, and a few bright ones that are actually lasing. You can spot mode jumps right away with this trick, and sticking a photodiode in just one of the bright spots will give you a visceral feel for just how bad mode partition noise can be.

2.12.7 Vertical Cavity Surface-Emitting Lasers

Almost all lasers have cavities that are long and skinny. This means that they have a great many longitudinal modes to jump around in, which causes noise problems. In addition, they emit out the ends, which is okay for built-up lasers but is inconvenient for diodes. Diode lasers are hard to test before dicing and can't easily be made in two-dimensional arrays. They also have that nasty beam asymmetry and astigmatism to deal with.

A partial solution to some of these problems is the *vertical cavity surface-emitting laser* (VCSEL). A VCSEL is rather like a DFB laser built on end. The cavity mirrors and active region are made by multilayer deposition, and the bias current flows vertically right through the whole thing. The high reflectance of the mirrors allows the active region to be very thin, so that the longitudinal mode spacing is very wide (1 THz). That and the limited mirror bandwidth mean that a typical VCSEL has only one longitudinal mode. The NA of a typical VCSEL is 0.3–0.4, and due to its approximate rotational symmetry, its beam is roughly circular. That same symmetry means that its polarization usually wanders all over the place.

VCSELs also have lots of transverse modes, and they're nearly all multimode. It is quite common for a VCSEL to jump between two or three spatial modes as the current is increased, and wind up in some $N = 6$ mode that looks like a chrysanthemum (See Figure 2.6.) They can be really quick (20 GHz modulation bandwidth), so they work well for things like high speed communication via multimode fiber. There has been some progress made in improving the polarization purity and mode structure, usually by breaking the circular symmetry somehow, but VCSELs are far from a panacea.

Aside: VCSEL Pathologies. Even in datacom sorts of jobs, some VCSELs exhibit weird turn-on artifacts at low repetition rates. The pump current has to flow through all those layers, so there are lots of opportunities for trap states to form. You have to fill them all before the laser will turn on, and at low rep rates, they'll all have recombined before the next pulse arrives. Repopulating them each time slows down the leading edge badly.

VCSELs also tend to have really horrible $1/f$ mode partition noise, which is worse than in ordinary multimode lasers since the modes are randomly polarized. Here's another instance where a good quality polarizer at the laser can help a lot in a differential measurement.

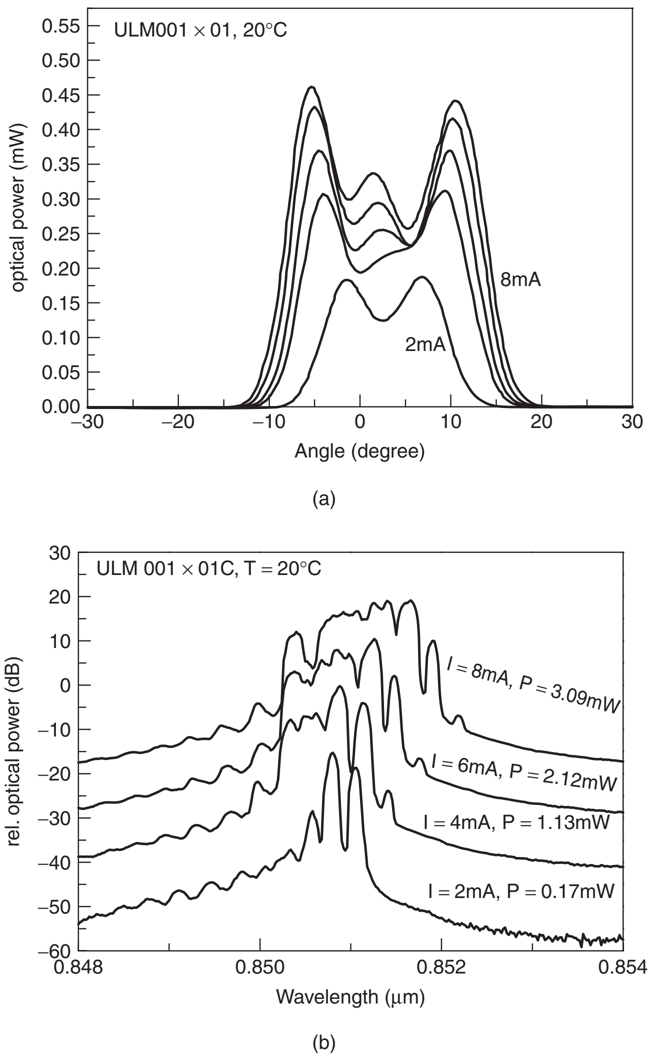


Figure 2.6. Angular beam width (a) and spectrum (b) of an 850 nm multimode VCSEL versus bias current. (Reproduced by courtesy of U-L-M Photonics GmbH.)

2.12.8 Modulation Behavior

You can modulate a laser diode fast enough for almost anything. Generic diodes can be modulated strongly up to 1 GHz, and special ones such as 850 nm communications VCSELs can go much faster, up to 20 GHz. The limiting factors are parasitic inductance and capacitance, and the intrinsic speed of the laser mechanism due to transition rates and the relaxation oscillation peak.

The modulation sensitivity increases toward the relaxation peak and drops off very rapidly beyond it, but you don't want to be working there anyway. The relaxation frequency (and hence the maximum modulation frequency) generally increases as the operating current goes up, up to somewhere near maximum output power, and then starts to slow down slightly.

Like an RF power transistor, the modulating impedance of the diode is normally low, which leads to vulnerability to parasitic inductance; in fact, wiring an RF power transistor right across the diode is a good way to modulate a high power device, but watch out for oscillations at 300 MHz to 3 GHz when you do this—it's possible to have too much of a good thing, and an anti-snivet resistor can come in very handy (see Section 19.7.4).

The main problem in modulating diode lasers is that the intensity and frequency modulate together, so that it isn't easy to get pure AM or pure FM from current tuning. If you're doing a spectroscopic measurement with a diode laser, you need some way of suppressing the giant intensity slope superimposed on your data. Typical methods are FM spectroscopy (Section 10.6) and noise canceler spectroscopy (see Section 10.8.6).

2.12.9 ESD Sensitivity

Diode lasers are so extremely sensitive to electrostatic discharge that ESD is the major cause of reliability problems. ESD in reverse bias causes hot carrier damage to the junction region, increasing the threshold, but forward bias ESD is much more spectacular: because the laser radiation grows so fast, a carpet shock with a 1 ns rise time can generate such a high peak laser power that the output facet of the laser gets blown right off. This is usually easy to spot; it puts stripes in the laser output pattern and dramatically reduces the output power.

Always use ground straps, and design protection into your diode laser mounts: use a DIP reed relay to keep the anode and cathode shorted together when the laser is not in use. A big bypass capacitor (1 μF) will absorb carpet shocks without letting the laser blow up, although this severely constrains the AC modulation possibilities. One way round this is to use a transformer with its secondary wired in series with the the diode to get fast modulation without ESD risk.

2.12.10 Difficulty in Collimating

Edge-emitting diode lasers emit radiation from a stripe-shaped aperture about 1 μm wide by a few microns long. The aperture is small, so that light comes out into a large solid angle, which is very asymmetrical: typical NAs are 0.09–0.15 in one plane and 0.3–0.5 in the other, roughly an elliptical Gaussian beam. This large NA means that the laser–lens distance is very critical; the depth of focus is only a few microns, so that extreme stability is required. Very thin layers of UV epoxy are a good way of affixing the lens and laser together.† Stability against mode hopping requires significantly tighter control than focus stability does.

The light is strongly polarized with **E** along the minor axis of the ellipse. The beam from a gain guided laser is a complete mess. It has such large amounts of astigmatism that getting decent collimation is extremely difficult (they're getting rare now anyway). Index guided lasers have some astigmatism, usually about 0.6 wave. This amount is big enough that it needs correction, which is a pain, but not so large that correcting it is unduly difficult if you have a measuring interferometer.

Astigmatism can be reduced with a weak cylindrical lens, but the commercially available ones are normally a poor match to your laser, unless you're very lucky. Fortunately, good collimation can be reliably obtained without the use of astigmatism correction optics

†See Example 20.7 at http://electroptical.net/www/beos2e/thermal2.pdf.

(0.95 Strehl ratio is often attainable).[†] The trick is to compensate for the astigmatism with a small amount of defocus, so that instead of the wavefront being a cylinder of 0.6λ p-p, it is a saddle shape with a p-p error of 0.3λ. You can also use the collimator slightly off axis, to get a bit of coma and astigmatism to knock the laser's astigmatism flatter still. This approach is unfortunately a bit fiddly; in a production system, it may be better to use a more powerful laser and just chop off all but the center 20% with a circular aperture.

The elliptical beam can be circularized with one or two prisms, where the beam enters near grazing (producing a patch elongated into a circle) and leaves near normal, so that the circular patch defines the refracted beam profile. Gratings can be used in the same sort of way, with lower efficiency. If you're using gratings, make sure you use the deviation-canceling configuration in Figure 7.8, or your beam will wander around with tuning.

An anamorphic lens system,[‡] in this case a telescope made of cylindrical lenses, can circularize the beam as well as cancel the astigmatism, but these are limited by the difficulty of fabricating good quality cylindrical lenses at low cost.

Blue Sky produces lasers with diffraction-limited circular output beams by using a cylindrical microlens right near the laser die, inside the package, which in principle makes a lot of things easier. These are not particularly cheap, but they're good for lab use. (You still have to collimate them yourself.)

2.12.11 Other Diode Laser Foibles

Diode lasers have a number of other funny properties. One is that dust and crud are photophoretically attracted to regions of high light intensity, which unfortunately means the diode laser facet or the center of the laser window. In a high power system, the beam path will get dirty long before the rest of the system, so make sure to control outgassing and dust inside your diode laser-based instruments. Interestingly, dry air turns out to be much better than inert gas—having oxygen around gives the organic crud a chance to oxidize before it builds up enough to cause damage.[§]

2.13 LASER NOISE

Lasers exhibit noise in both intensity and frequency. Apart from ultrahigh resolution spectroscopy, most measurements suffer more from intensity noise than frequency noise. As we'll see in Chapter 10, a lot of ingenuity has been expended on getting rid of the effects of laser noise.

2.13.1 Intensity Noise

Intensity noise pollutes laser-based measurements in two ways. Most measurements have a nonzero baseline, so that a fluctuation in the laser power causes a fluctuation in the

[†]The Strehl ratio is the best single measure of beam quality for instrument purposes, where the beams aren't too ugly. It cannot be greater than 1, and 0.8 corresponds roughly to Rayleigh's λ/4 criterion for diffraction-limited image quality. See Section 9.5.4 for more discussion of the aberration terms used here.

[‡]That is, one with different magnifications in x and y.

[§]R. Jollay et al., *Proc. SPIE* **2714**, 679–682 (1996).

background signal, which shows up as additive noise in the measurement (additive means that this noise doesn't depend on the signal strength). This additive noise reduces the sensitivity of the measurement and, since laser noise tends to be non-Gaussian in character, may not average out well.

The other effect of laser intensity noise is to cause the signal itself to fluctuate in strength. In nearly all laser-based measurements, the detected signal is proportional to the laser power, so an intensity fluctuation produces a signal strength fluctuation. This is called *multiplicative* noise, or noise intermodulation. For example, consider a single-beam tunable diode laser spectrometer looking at a gas cell. The spectrum consists of absorption lines on a smooth baseline. Intensity noise causes the baseline to fluctuate (additive noise), but also causes the absorption peaks to fluctuate in magnitude (noise intermodulation). (See Section 13.6.11.)

There are a variety of differential and ratiometric detection schemes to help with this problem, of which laser noise cancellation is by far the most effective; it can provide as much as 70 dB reduction in both additive laser noise and noise intermodulation, and gets you down to the shot noise reliably, even with noisy lasers. It needs an extra beam plus about $10 worth of simple electronics. If your measurement suffers from laser intensity noise, have a look in Section 10.8.6.

2.13.2 Frequency Noise

An oscillator is composed of an amplifier plus a frequency determining device, usually a resonator. The resonator attenuates signals outside its narrow passband but, more to the point, exhibits a phase shift that varies rapidly with frequency. Since oscillation requires that the round-trip phase be an exact multiple of 2π, changes of the resonator length force the frequency to move. The frequency noise of the oscillator is determined by the combination of the amplifier's phase fluctuations and the phase slope of the resonator. Lasers are a bit more complicated in that the resonator may exhibit phase fluctuations too, as when fan vibrations or cooling water turbulence jiggles the mirrors of an argon ion laser. Resonator noise is confined to low frequencies, so it can be dealt with separately by mechanical means.

This picture of frequency noise suggests that lasers can be stabilized by using longer cavities with higher Q, which is true; external cavity diode lasers have linewidths a factor of 10^3–10^4 narrower than when the diode's F-P resonator is used. There are also various active locking techniques such as Pound–Drever stabilization, which are beyond our scope but are covered well in Ohtsu.

2.13.3 Mode Hopping

Most lasers have a large number of possible oscillation modes within the gain curve of the medium. Ideally, one of these should dominate all others, but this is often not the case, due to spatial hole burning.† Even when it is, the difference between adjacent modes is often small enough that it takes very little perturbation to switch the laser from one mode to another. This will happen during warmup, for example. Sometimes, though,

†Hole burning is the picturesque name given to the local reduction of laser gain near peaks of the standing wave pattern in the gain medium. This reduces the gain of the strongest mode, without reducing that of the others equally. See Siegman for more details.

there is no one stable mode of oscillation (usually due to spurious feedback of one kind or another coupling the modes), leading to mode hopping.

Diode lasers are especially sensitive to mode hopping, because their cavities are strongly coupled to the outside (reflectivities of about 40%, vs. 99.5% for a HeNe). A spurious reflection on the order of 1 part in 10^6 can set a diode laser into mode hopping; this leads to a lot of flaky failures that come and go, making testing difficult. The actual mechanism of mode hopping is a complicated interaction between the thermal, plasma-optical, and current confinement behaviors of the diode; a change in the laser tuning changes the power output, which changes the dissipation in the channel, which changes the temperature, which changes the cavity length and index, which changes the tuning, and so forth. Since the active region can be cooled very rapidly by the laser output itself, there is a strong coupling between die temperature, tuning, and power output that leads to instability. Visible diode lasers are the worst for this. VCSELs are designed with cavities so short that their free spectral range is wider than the natural linewidth, so they don't hop between longitudinal modes, but frequently they do between transverse modes.

Mode hopping is most obviously a frequency noise phenomenon, but it results in strong (0.1–1%) intensity noise as well, because the gains of the laser system in adjacent modes are not the same. Mode hopping causes irregular jumps and spikes in the laser power, at rates of 100 kHz or so. Even in a pure intensity measurement, diode laser mode hopping due to incidental feedback is very obnoxious and can render your measurement extremely difficult. Mode hopping makes all your etalon fringes dance around, so that the frequency noise gets converted to intensity noise as well.

2.13.4 Mode-Partition Noise

The total power output of a laser is limited by the pump power among other things, and various saturation effects couple the intensities of the modes, so that the instantaneous power of the laser varies less than that of the individual modes. The switching back and forth of the laser power is called *mode-partition noise*. It is insidious, because it doesn't show up on a power meter but is silently at work screwing up your measurement and producing seemingly irrational results; with a gas laser, it is easily possible for a spatial filter, a knife edge, or even an iris diaphragm to cause the intensity noise to go up by 20 dB; a stray etalon fringe can do the same, since the different modes will see different phase shifts and hence will be demodulated differently. It's pretty puzzling if you don't know the secret.

The quietest lasers are single longitudinal mode units, followed by highly multimode ones, and the worst usually have only a few (2–10) modes. The same is true of optical fiber devices (see Section 8.4.3). It's worth trying a quick experiment with a few-mode diode laser and a grating—if you catch one mode with a photodiode, you'll usually find it much noisier than the entire beam, even in absolute terms.

2.13.5 Gotcha: Surface Near a Focus

Unless you use really effective—60 dB or more, optical—Faraday isolators to protect the diode laser from feedback, make sure that there is no surface in your optical system that coincides or even nearly coincides with a focus. Even if the specular reflection goes off at a steep angle, and so misses the laser, there will be enough scatter to make the laser mode hop. If you have a diode laser system that mode hops, and you're sure you've eliminated all the near-normal surfaces, look for this one. If it isn't a surface near a focus, it's probably feedback right inside the collimator.

An insidious possibility you may need to watch for is that the focus in question may not be a focus of the main beam, but of a converging stray reflection; a concave coated surface will give rise to a converging beam 1% as strong as the main beam, and even after two more bounces to get back into the laser, it can easily get to the 10^{-6} mode hopping danger level. The ISICL sensor of Example 1.12 has this difficulty if it isn't mounted properly.

2.13.6 Pulling

The oscillation frequency of a laser has to be at a frequency where the round-trip phase delay is an integral multiple of 2π radians. Normally, as the cavity length ℓ changes slightly, or small amounts of contamination collect on the mirrors, the oscillation frequency changes so as to preserve the round-trip phase. However, if one of the mirrors suddenly were to develop a time-dependent phase shift of its own, the oscillation frequency would have to respond to it by shifting, even if the cavity length were perfectly stable.

This is more or less what happens with pulling; some external influence, for example a Fabry–Perot resonator such as an optical spectrum analyzer or merely an incidental reflection, sends delayed light back to the output coupler of the laser. Interference between the light inside the laser cavity and this spurious reflection causes the phase of the light returned from the output coupler to change.

The resulting frequency shift is

$$\Delta\nu \approx \frac{-\Delta\phi}{\partial\phi/\partial\nu + 2\pi\ell/c}, \tag{2.12}$$

where $\partial\phi/\partial\nu$ is the phase slope of the total reflection. Note that even if the spurious reflection is weak, so that the total phase excursion is small (see Section 13.6.9), $\partial\phi/\partial\nu$ can be made very large by using a sufficiently long delay. As a result, the stray reflection can in principle take over the frequency determining role from the cavity. This happens especially in diode lasers, where the cavity is short and the mirrors leaky; sometimes their tuning behavior is determined more by the back-reflection from their collimator than by the cavity mirror.

2.13.7 Mode Beats

If you shine a HeNe laser onto a photodiode and examine the results on a spectrum analyzer set to DC–10 MHz, you'll probably see the occasional strong, narrow spur [†] appear, sweep toward 0, and disappear when it gets to 1 MHz or so.[‡] It may be as strong as 0.1% of the total photocurrent. These odd objects are baseband mode beats. Lasers, like all oscillators without automatic level control, operate in a strongly nonlinear region, which means that all their (supposedly orthogonal) modes are in fact coupled together

[†]That is, spurious signal, see Section 13.5.

[‡]If it doesn't do it right away, put your hand on one end of the housing to cool it down a bit, and then look. A bit of bending caused by thermal gradients will unlock the modes.

with each other. Those mode beats are caused by one mode mixing with a third-order intermodulation product of two others (Section 13.5.3 for more details). Since optics people always have fancier names for things, this intermodulation is called *four-wave mixing*. Small anharmonicities in the laser medium or cavity cause the two products to differ in frequency by about 1 part in 10^9, causing the few-hundred-kilohertz mode beats we see. The disappearance at low frequency is caused by the modes jumping into lock with each other.

2.13.8 Power Supply Ripple and Pump Noise

Like other oscillators, lasers are sensitive to power supply ripple. The laser output depends on how much gain is available, which depends on how hard it's pumped, which depends on the supply voltage. This is usually worst in big gas lasers, whose large power requirements make quiet supplies harder to build. The advent of good quality switching supplies, whose high operating frequencies make filtering much easier, have improved this problem, but it still persists. You'll definitely see your power supply's operating frequency come through loud and clear. If it sits still, you can usually avoid it, but some supplies change their frequency with load, and that makes it much harder to avoid. Make sure you know this about your particular laser.

Diode lasers have very high electrical to optical conversion efficiencies, so the current noise on the supply translates more or less directly into photon noise. Most diode laser supplies use lousy voltage references, inadequately filtered, to define their output current, which makes the lasers themselves noisier. For bright-field measurements, it is really worthwhile to make sure that your diode laser's bias supply is quieter than the shot noise. This isn't too hard to do—see Section 14.6.7.

Lasers that are pumped optically will also be modulated by the noise of the pump source. Flashlamps are usually the prime contributors, but ion laser pumped dye lasers also suffer from instability due to mode noise in the pump laser.[†] As noted in Section 2.11, single longitudinal mode DPY lasers are very quiet.

2.13.9 Microphonics

The narrow linewidth of a laser comes not from the narrowness of the spectral line doing the lasing, but rather from its being put in a Fabry–Perot interferometer a meter long, and then subjected to regeneration, which narrows it further. Since the narrowness comes from the cavity selectivity, anything that causes the cavity length to fluctuate will produce frequency noise in the laser. This includes vibrations from ambient sound (microphonics), cooling water turbulence, fans, and conducted vibrations from the table. Lasers whose mirrors are firmly positioned (e.g., sealed HeNe units and especially diode lasers) are much less prone to this problem.

Some types of lasers (e.g., medium power diodes and diode pumped YAGs) come in styles with and without fans in the laser head. Avoid fans wherever possible.

[†]Martin C. Nuss, Ursula H. Keller, George T. Harvey, Michael S. Heutmaker, and Peter R. Smith, Amplitude noise reduction of 50 dB in colliding-pulse mode-locking dye lasers. *Opt. Lett*. **15**(18), 1026–1028 (1990).

2.13.10 Frequency Noise

Frequency noise in an oscillator comes from the noise of the active element, restrained by the selectivity of the resonator. The low frequency noise of the amplifier (and any noise or instability in the resonator) gets translated up to become low modulation frequency sidebands on the laser oscillation (see Section 15.9.4), and high frequency noise becomes a more or less white phase noise background. A resonator run at its 10^6th overtone, such as a laser cavity, makes this a bit more complicated by introducing competition between modes, and the complexities of the laser gain mechanism contribute intrinsic noise and instability, but this picture is still basically right.

Laser frequency noise can be reduced by reducing the noise forcing: quieting down the pumping, using a stable gain medium (e.g., Nd:YAG rather than N_2) when possible, or using a highly mechanically stable cavity. It can also be improved by using a more selective resonator (a longer or lower loss one). An external cavity stabilized diode laser uses both approaches.

There is a region of optical feedback in which temporal coherence collapses completely, and the laser output becomes chaotic. You'll recognize it if it happens to you.

2.13.11 Spatial and Polarization Dependence of Noise, Wiggle Noise

Laser noise is not merely a well-behaved wandering around of the intensity or frequency of the entire beam at once. There are important variations in the noise with position and polarization; for example, vignetting the beam of a single frequency argon ion laser has been known to increase its residual intensity noise (RIN) by an order of magnitude; a weak side mode that was previously orthogonal to the main beam was made to interfere by the vignetting, producing a huge noise increase. Diode lasers are especially prone to this pathology for some reason, but all lasers have spatially dependent noise.

Interference with small amounts of laser light and spontaneous emission whose spatial pattern is different causes the laser beam to wiggle back and forth ever so slightly, a phenomenon called wiggle noise.[†] A gas laser with a second spatial mode that is close to oscillating, or a "single-mode" fiber run at too short a wavelength (so there are really two or three or five modes) are especially bad for this; you don't know what pointing instability is like until you've used a system like that. On the other hand, a truly single-mode fiber with all the extraneous light stripped out is unsurpassed for pointing stability (see Section 8.2.2). The angular size of the wiggle goes as the ratio of the amplitudes of the laser mode and the spurious signal, that is, the square root of their intensity ratio, so this isn't as small an effect as you might think.

Polarization dependence is easier to understand; for example, lasers produce spontaneous emission, which is more or less unpolarized. Because it has a different dependence on pump power, the modulation of the spontaneous emission by pump noise will be different. Because the laser radiation is usually highly polarized, the detected noise will differ in character depending on the rotation of an analyzer.[‡]

[†] M. D. Levenson, W. H. Richardson, and S. H. Perlmutter, Stochastic noise in TEM_{00} laser beam position. *Opt. Lett*. **14**(15), 779–781 (1989).

[‡] An analyzer is just the same as a polarizer, but the name specifies that it is being used to select one polarization for detection, rather than to prepare a single polarization state for subsequent use.

2.14 DIODE LASER COHERENCE CONTROL

2.14.1 External Cavity Diode Lasers

The tuning and frequency stability of a Fabry–Perot diode laser are limited mainly by the poor selectivity of its cavity. It is possible to use a diode as the gain medium in a conventional laser cavity; this is done by antireflection coating the output facet to get rid of etalon resonances in the diode, and using an external reflector and a grating for selectivity. In these external cavity diode lasers (ECDLs), the external cavity takes over the frequency determining function; because of its length and the good optical properties of air, the linewidth is narrower and the tuning much more stable. Because the diode has such high gain, the cavity needn't be all that efficient, and typically fixed-tuned ECDLs use a grating in Littrow as the cavity mirror, with the specular reflection being the laser output. This makes the output beam alignment tuning-sensitive, so it is usually restricted to fixed-tuned applications. Tunable ECDLs use two bounces off a fixed grating, with a rotatable mirror as the tuning element; that makes the output beam pointing very stable. Very fast-tuning ECDLs use polygon mirrors but these have amplitude spurs due to Doppler shifts. ECDLs can also be made from uncoated diodes, but the two competing cavities make it much harder to find a stable operating region, and the tuning is no longer continuous as it is with the grating-tuned, AR-coated version. You'd think that the grating feedback would help to prevent mode hops by sharply distinguishing the allowed modes, but in practice it doesn't really. Even fancy commercial ECDLs are highly sensitive to feedback, so budget for a two-stage free-space Faraday isolator in addition to the laser itself.

2.14.2 Injection Locking and MOPA

Small diode lasers have better mode characteristics than large ones, just as small signal transistors make better oscillators than power units. Larger lasers can have their facets AR coated to eliminate regeneration, turning them into amplifiers instead. They can be used as power amplifiers for the radiation from smaller ones, the so-called master oscillator–power amplifier (MOPA) approach. The main drawback of MOPA, besides its cost and complexity, is the large amount of spontaneous emission from the amplifier.

Alternatively, the power stage can be left to oscillate, but seeded by shining the master oscillator's output into it. Under the right conditions, the power stage will *injection lock* to the seed. (This is an old microwave trick and also helps a lot with OPOs.) Injection locking requires less pump power than MOPA, and the Fabry–Perot resonance of the power stage filters out most of the spontaneous emission, but it's just flakier. Its bad behavior arises from the coupling between the two resonators, which are sensitive to temperature, current, and the phases of the moon (microwave versions have amplifiers with good input–output isolation, a luxury we'll have more than one occasion to envy before this book is finished). MOPA seems to be a much more reliable approach.

2.14.3 Strong UHF Modulation

Rather than increase the temporal coherence of a laser, sometimes it's more helpful to destroy it. Mode hops can be eliminated by the use of *quenching*, by analogy with the superregenerative detectors of early radio. When gain was very expensive, positive (regenerative) feedback offered an appealing if unstable method of getting more gain for less money. Superregens work by coupling the input into an oscillator circuit that is turned on and off at ultrasonic frequency by a second (*quench*) oscillator. The exponential buildup of the oscillations produces a waveform proportional to the size of the input, but many times larger, with the amplification and linearity controlled by the quench frequency (see Terman, it's beautiful). Lower quench rates give a logarithmic response.

Laser quenching isn't as pretty, but it's still useful. In a mode hopping laser, the situation is a bit more complicated, since the laser oscillation itself builds up very rapidly (1 ns or faster) and it's the mode hops we want to quench. Using large-signal UHF modulation to essentially turn the laser on and off at 300–500 MHz suppresses mode hopping completely, at the cost of enormously increased linewidth. Commercial ICs are available, or you can use an RF transistor in parallel with your diode laser. (Gallium nitride RF FETs are especially good for this.) This trick was widely used in magneto-optical storage applications and DVD players—there were even self-pulsating diode lasers that turned themselves on and off at UHF rates. (Note that the linewidth is still fairly small compared with a light bulb, so that interferometers built with these will have phase noise problems unless their path differences are extremely small.)