CHAPTER 7

Exotic Optical Components

> To know a thing well, know its limits. Only when pushed beyond its tolerance will its true nature be seen.
>
> —Frank Herbert, *Dune*

7.1 INTRODUCTION

The mundane optical tasks of directing and transforming beams are mostly done with lenses and mirrors, with the occasional polarizer or wave plate thrown in, and this is enough for cameras and microscopes. They're the workhorses: steady, dependable, and reasonably docile, but not very fast. Building advanced instruments is more like racing. The more exotic components like fibers, gratings, scanners, and modulators are thoroughbreds—they're good at what they're good at, and aren't much use otherwise. (Watch your step in this part of the paddock, by the way.) Fibers are left until Chapter 8, but there are lots of connections between there and here.

7.2 GRATINGS

A diffraction grating is an optical surface with grooves in it, shaped and spaced so as to disperse incident polychromatic light into a sharply defined spectrum. There are lots of variations, but they're all basically holograms—the grooves reproduce the interference pattern between an incident monochromatic beam and the desired diffracted beam, and so form an optical element that transforms the one into the other. Some gratings are made holographically, and some are ruled mechanically, but the operating principle is the same.

The most common type is the *classical plane grating*, a flat rectangular surface with equally spaced grooves running parallel to one edge. Phase matching at the surface governs their operation; as with a planar dielectric interface, this condition can be satisfied over a broad continuous range of incident **k** vectors.

There are also Bragg gratings, where the grating structure runs throughout some volume, a more complex structure that is important in fiber devices, holograms, acousto-optic cells, and some diode lasers, as we'll see later in this chapter.

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

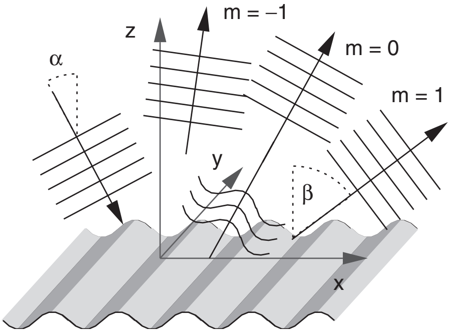


Figure 7.1. Plane diffraction grating.

7.2.1 Diffraction Orders

We can make this a bit easier to understand by looking at the special case of Figure 7.1. A metallic plane grating lies in the **xy** plane, with G sinusoidal grooves per unit length,

$$h(x, y) = a \sin(2\pi Gx), \tag{7.1}$$

where $a \ll \lambda$ and G, k_x, and k_y are small compared to k (i.e., a weak grating of low spatial frequency). A plane wave $\exp(i\mathbf{k}_{\text{inc}} \cdot \mathbf{x})$ hits the surface and gets its phase modulated by the changing height of the surface. If the medium is isotropic, linear, and time invariant, the modulus of k can't change,[†] so we've got a pure spatial phase modulation, as in Section 13.3.7. Thus the phase modulation produces mixing products (see Chapter 13) with wave vectors $\mathbf{k}_{Dm}$,

$$\mathbf{k}_{Dm} = \mathbf{k}_i + m\mathbf{k}_G, \tag{7.2}$$

where $k_G = 2\pi G\hat{\mathbf{x}}$ and $m = \ldots, -1, 0, 1, 2, \ldots$. Equation (7.2) can also be derived immediately from the phase matching condition: because the surface is periodic, the fields have to be invariant (apart from an overall phase) if we shift it by an integer number of cycles. Only a finite range of m produces propagating waves at the output—only those whose $|k_{\text{xm}}| < k$; that is,

$$\frac{-\sqrt{k^2 - k_y^2} - k_x}{k_G} < m < \frac{\sqrt{k^2 - k_y^2} - k_x}{k_G}. \tag{7.3}$$

Although we've kept k_y in this formula, gratings are nearly always oriented to make k_y as small as possible, so that it enters only quadratically in θ_d. Since G is wavelength independent, we can solve (7.2) (in the special case $k_y = 0$) for λ, yielding the *grating equation*[‡]

$$\lambda_m = \frac{\sin\beta - \sin\alpha}{mG}. \tag{7.4}$$

[†] When we get to acousto-optic modulators, this won't be true anymore, and frequency shifts will occur.

[‡] You sometimes see it used with the other sign convention, so that there is a plus sign in $\Delta u = \sin\theta_d - \sin\theta_i$; in any event, specular reflection ($m = 0$) has $\Delta u = 0$.

The illuminated patch on the grating is the same width for both beams, but because of obliquity, the diffracted beam undergoes an anamorphic magnification in x of

$$M = \frac{\cos\beta}{\cos\alpha}. \tag{7.5}$$

In general, the spectrum gets spread out in a cone, but in the $k_y = 0$ case, it gets spread out in a line, with only a slight curvature due to the inescapable finite range of k_y in real apparatus. If we send broadband light in at θ_i, we can select a narrow wavelength band centered on λ by spatial filtering.

The nonlinear relation of θ_d to θ_i for a given wavelength means that classical plane gratings cause aberrations if the light hitting them is not collimated in the **x** direction. These aberrations reduce the resolution of the spectrum and cause spectral artifacts, so we normally allow only a single k_x and a limited range of k_y.

Example 7.1: Czerny–Turner Monochromator. A monochromator is a narrowband tunable optical filter, based on the Fourier optics of gratings and slits. The classical design is the Czerny–Turner, shown in Figure 7.3. Polychromatic light from the entrance slit is (spatially) Fourier transformed by spherical mirror M1, so that every point in the slit produces its own plane wave at each wavelength. These are then dispersed by the (rotatable) grating and transformed back by M2 to produce a set of images of the entrance slit on the plane of the exit slit, of which only one is allowed to escape. Rotating the grating moves the dispersed spectrum across the exit slit, so that a different λ emerges.

The spherical mirrors are used well off-axis, so there is a significant amount of coma and astigmatism as well as spherical aberration, which must be accounted for in the design. Note the anamorphic magnification and pupil shift in the figure; normally we use mirrors that are somewhat larger than the grating to combat this.

In designing monochromators and spectrometers, we have to remember that most of the light doesn't make it out the exit slit, but bounces around inside the box, so we need good baffles. Otherwise this stray light would bounce all over the place inside, and some of it would eventually escape through the exit slit and contaminate the output spectrum. There's no place in Figure 7.2 to put a baffle that won't obscure the optical path, so real Czerny–Turners don't have planar layouts. The mirrors are canted down a bit (into the

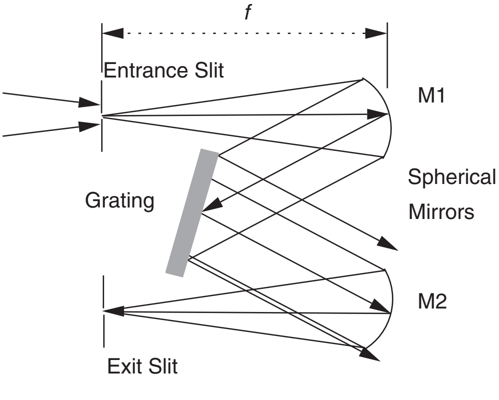


Figure 7.2. Czerny–Turner monochromator.

page), which depresses the grating position enough to fit a good baffle in over top, which helps a lot.

Another thing to watch out for in grating instruments, especially adjustable ones, is temperature-compensating the slit opening. Carelessness here, such as using brass slits on steel screws and a long mechanical path, typically leads to throughput changes of *several percent* per °C.

The other main problem with spectrometers is that they're all polarization sensitive. The p-to-s diffraction efficiency ratio of a transmission grating is massively wavelength dependent, and mirrors used off-axis can easily contribute several percent polarization (see Example 5.2).

7.3 GRATING PATHOLOGIES

So far, a grating is a reasonable spatial analogy to a heterodyne mixer (see Section 13.7.1). The analogy can be pressed further, because the grating also has analogues of LO phase noise (scattered light), LO spurs (periodic errors in the grating phase, giving rise to *ghosts*), and spurs due to LO harmonics (multiple orders, leading to overlap). It starts to break down when the 3D character of the intermodulation starts entering in; spatial frequency differences can arise from shifts in ω or θ_i, but the resulting fields are quite different.

7.3.1 Order Overlap

For any grating and any θ_i, the function $\lambda(\theta_d)$ is multivalued, so that more than one wavelength will make it through a monochromator at any given setting. Simple grating spectrometers are limited to a 1-octave range in the first order, as shown in Figure 7.3, and the limitation gets tighter at higher orders. The best way to reject light that would be aliased is to use *cross-dispersion*: just put a second grating or prism at 90° to separate out

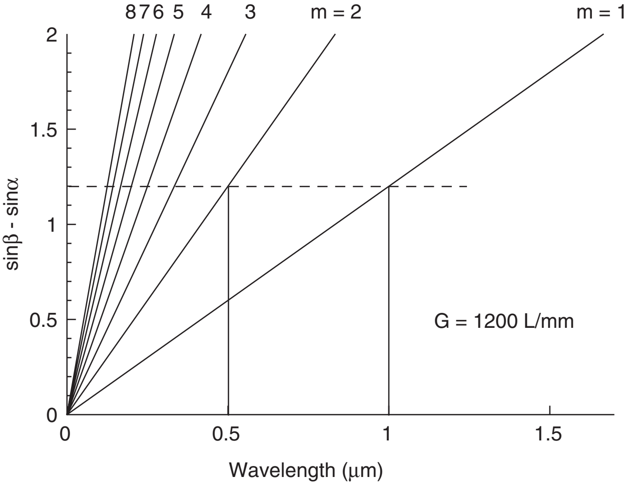


Figure 7.3. *M*th order wavelength as a function of $\Delta \sin\theta$.

the orders; the second grating's dispersion limits the allowable slit length. Cross-dispersed gratings are a good match to 2D detector arrays such as CCDs.

7.3.2 Ghosts and Stray Light

If we consider a grating as a frequency mixer, it isn't surprising that irregularities in the fringe spacing get transformed into artifacts in the diffracted spectrum. Small-scale irregularities give rise to a smoothly varying diffuse background of scattered light, which sets the maximum rejection ratio of a spectrometer. Low frequency variations in the grating pitch, caused, for example, by diurnal temperature variation in the ruling engine during a run, produce close-in sidebands on the diffraction spectrum just the way they would in an RF mixer; these close-in artifacts are called *Rowland ghosts*. Ghosts occurring further away are called *Lyman ghosts* and are even more objectionable, since it's much harder to connect them to their true source. As we saw, baffles help the stray light a lot, but ghosts are more of a problem, since they follow the normal light path through the exit slit. Both can be dramatically reduced by using another grating and a third slit, to make a *double monochromator*, and for some very delicate measurements such as Raman spectroscopy, people even use triple monochromators. It's amazing that any light makes it through all that, but it does if you do it properly, and you get double or triple the linear dispersion, too.

7.4 TYPES OF GRATINGS

Classical plane gratings are wonderful at dispersing different wavelengths, but bad at almost everything else—they cost a lot, aberrate converging beams, treat s and p polarizations differently enough to be a bother but not enough to be useful, require throwing away most of our light to get high spectral resolution, the list goes on and on. Lots of different kinds of gratings have been developed to try to deal with some of these difficulties.

Nearly all gratings sold are replicas, made by casting a thin polymer layer (e.g., epoxy) between the master grating and a glass blank (using a release agent to make sure it sticks to the glass and not the master). Reflection gratings (the usual kind) are then metallized.

Aside: Grating Specifications. Since everything about gratings depends on $\mathbf{k}_{||}$, their properties are usually specified for use in the *Littrow* configuration, where $\mathbf{k}_d = -\mathbf{k}_i$ (i.e., the diffracted light retraces its path). This implies that $k_y = 0$, and that there is no anamorphic magnification of the beam, which simplifies things a lot, but isn't necessarily representative of what you should expect far from Littrow.

7.4.1 Reflection and Transmission Gratings

The essential function of a grating is to apply a very precise spatial phase modulation to an incoming beam. This can be done by reflecting from a corrugated surface, or by transmission through different thicknesses of material. Transmission gratings have the advantages of lenses: compact multielement optical systems, lower sensitivity to flatness errors, and less tendency for the elements to get in one another's way. On the other hand, multiple reflections inside the transmission grating structure give rise to strong artifacts, making them unsuitable for high resolution measurements in general. With the exception of hologon scanners, you should use reflection gratings for nearly everything.

Reflection gratings are usually supplied with very delicate bare aluminum coatings. Touching a grating surface will round off the corners of the grooves and ruin the grating, and anyone who tries to clean one with lens paper will only do it once. Pellicles can be used to protect gratings from dust. Gold-coated gratings are useful, especially around 800 nm where the efficiency of aluminum is poor.

7.4.2 Ruled Gratings

Ruled gratings have nice triangular grooves, which allow high diffraction efficiency, but since the process of cutting the grooves produces minute hummocks and irregularities in the surface, they also have relatively high scatter. The increasing scatter limits ruled gratings to a practical maximum of 1800 lines/mm.

Ruled gratings can be *blazed* by tipping the cutting point so that the grooves are asymmetrical; by the array theorem of Fourier transforms,[†] a regular grating illuminated with a plane wave will have an angular spectrum equal to the product of the angular spectrum of a single grating line (the envelope) times the line spectrum from the Ш function. Blazing is a way of making the peak of the envelope coincide with the diffracted order, by tipping each grating line so that the specular reflection from that line is in the diffracted direction. The gratings of Figure 7.4(a) and (b) are blazed.

A grating blazed at λ_B works well from about λ_B to $1.5\lambda_B$, but falls off badly at shorter wavelengths. Like most other grating parameters, the blaze wavelength is quoted for Littrow incidence.

7.4.3 Holographic Gratings

Holographic gratings are frozen interference fringes, made by shining two really, really good quality laser beams on a photoresist-coated substrate. The grating period can be adjusted by changing the angle between the beams, or a transparent substrate can be suspended in an aerial interference pattern with fringe frequency $\Delta\mathbf{k}$, for example, from a laser bouncing off a mirror, and tipped to change the spatial frequency $\Delta\mathbf{k}_{||}$ at the surface. If the resist layer is thin and is developed after exposure (i.e., unexposed resist is washed away), a surface grating results. This grating is then transferred to a durable metal surface by plating it, attaching the plated surface to a stable support (e.g., a glass blank), and then chemically or mechanically stripping the resist, leaving its image in the metal. This metal submaster is then used to make replicas.

The grooves in a holographic grating are ideally sinusoidal in shape (although they can be blazed by ion milling or evaporation of metal from oblique incidence, or by special lithography techniques). The peak-to-peak phase modulation in the reflected wavefront is then roughly

$$\Delta\phi = 2k_Z d, \tag{7.6}$$

where d is the peak-to-valley groove height, and shadowing has been neglected.

Best overall efficiency with a sinusoidal modulation is obtained when the specular order goes to 0, which happens with $\Delta\phi = 4.8$ radians (in RF terms, the first Bessel

[†] We'll see it in Example 13.8 and Section 17.4.3 as convolution with a Ш function.

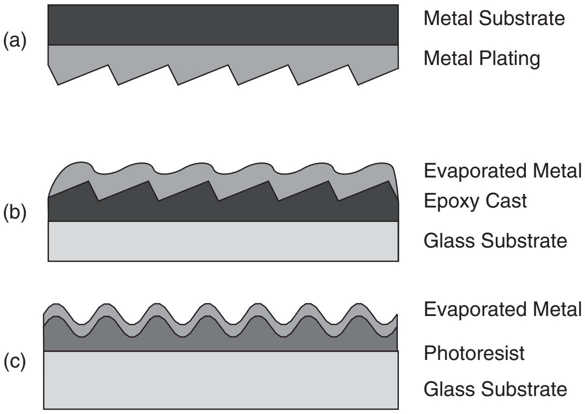


Figure 7.4. Diffraction gratings: (a) ruled, (b) replicated, and (c) holographic.

null at a modulation index of 2.405—see Section 13.3.7). Deep gratings are therefore used for long wavelengths, and shallower ones for short. Holographic gratings have less small-scale nonuniformity than ruled ones, so they exhibit less scatter and possibly fewer ghosts. The diffraction efficiency of holographic gratings is less strongly peaked than blazed ruled gratings, so they may be a better choice for weird uses.

7.4.4 Concave Gratings

From a diffraction point of view, there's nothing special about a flat surface; since the grooves embody the interference pattern between the incident and diffracted light, we can sample the pattern anywhere we like. Concave gratings combine the collimating, focusing, and diffracting functions in a single element. They are very expensive but are worth it in the deep UV, where mirror coatings are very poor ($R \approx 0.20$–0.35), so extra bounces cost a lot of photons.

The trade-off is bad aberrations; the focusing mirror is used off-axis, and the diffracted beam fails to obey the law of reflection on which mirror design is based. Concave mirror spectrographs are therefore usually highly astigmatic. This is not as bad as it sounds, since the image of a point in an astigmatic system is a line in the tangential plane (at the sagittal focus) or in the sagittal plane (at the tangential focus) (see Section 9.4.3). Providing the slits are at the tangential focus, the errors caused by astigmatism can be kept small, since the astigmatism's main effect is then to smear out the image along the slit. Of course, in a real situation the tangential image is not a perfectly straight line, and its curvature does degrade the spectral resolution somewhat.

There exist aberration-reduced holographic concave gratings, where the groove spacing is intentionally made nonuniform in such a way as to nearly cancel the leading aberrations over some relatively narrow range of θ_i and λ, which are great if you have the budget for a custom master, or a catalog item happens to fit your needs.

7.4.5 Echelles

An echelle grating (shown in Figure 7.5) is a coarse-ruled grating used in a high order, near grazing incidence; the scattering surface used is the short side of the groove instead

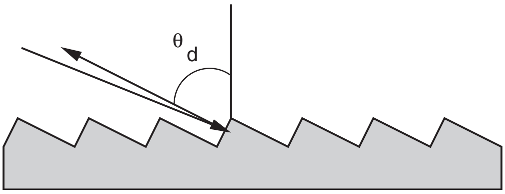


Figure 7.5. Echelle grating.

of the long side (like the risers of the stairs, rather than the treads). It is not unusual for an echelle to be used in the 100th order. Echelles usually have between 30 and 500 lines per millimeter and a really big one (400 mm) can achieve a resolving power of $2W/\lambda \approx 10^6$, a stunning performance. Because of the angular restriction, echelles are usually used near Littrow.

Problems with echelles include expense and severe grating order overlap. The expense is due to the difficulty in maintaining precision while ruling the coarse, deep grooves required, and of course to the low manufacturing volumes; the overlap requires cross-dispersion or a very well understood line spectrum.

7.5 RESOLUTION OF GRATING INSTRUMENTS

7.5.1 Spectral Selectivity and Slits

The usefulness of grating instruments lies in their selectivity—their ability to treat the different wavelengths differently. It's no use dispersing the light at one θ_i if it instantly remixes with other components at different θ_i, different ω, but the same θ_d (take a grating outside on a cloudy day, and try seeing colors). Just how we want to treat each wavelength varies; in a monochromator, we use a spherical mirror to image the angular spectrum on a slit, in order to select only one component, whereas in a pulse compressor, we just add a wavelength-dependent delay before recombining.

7.5.2 Angular Dispersion Factor

In computing the resolution of a grating instrument, we need to know the scale factor between λ and θ_d, the *angular dispersion D*:

$$D = \frac{\partial \beta}{\partial \lambda} = GM \sec \beta = \frac{(\sin \beta - \sin \alpha)}{\lambda \cos \beta}. \tag{7.7}$$

7.5.3 Diffraction Limit

The wavelength resolution of an ideal grating is limited by size, order, and operating wavelength. A uniformly illuminated square grating of side W has a sinc function lineshape as a function of $\mathbf{k}_\|$,[†]

$$E(u, v) = E_{0m} \frac{W}{\lambda} \operatorname{sinc}\left((u - u_i - mG)\frac{W}{\lambda}\right) \operatorname{sinc}\left((v - v_i)\frac{W}{\lambda}\right) \tag{7.8}$$

[†] The xy projection of $\mathbf{k}$, that is, (k_x, k_y).

(with $u = k_x/k$ and $v = k_y/k$ as usual, and assuming that $\mathbf{k}$ is along $\mathbf{x}$). By smearing out the angular spectrum, this effect sets a lower limit to the useful slit width in a spectrometer. As we'll see in great detail in Section 17.6, this sinc function is a nuisance, having slowly decaying sidelobes in its transform. (Don't worry about the sinc functions due to the slits—the exit slit is at an image of the entrance slit, whereas those sidelobes are in the pupil.)

In applications needing the cleanest peak shapes, it is sometimes worth apodizing the incoming light so that its diffraction pattern decays before hitting the edge of the grating. A carefully aligned linear fiber bundle can do a good job of this, provided the aperture of the spectrometer is a bit larger than the fibers'.

The equivalent width of this sinc function is $\Delta \sin\theta_d = \Delta u = \lambda/W$, or in angular terms,

$$\Delta\beta = \frac{\lambda}{W\cos\beta}. \tag{7.9}$$

The theoretical *resolving power* R of a grating is the reciprocal of the diffraction-limited fractional bandwidth:

$$R = \frac{k}{\Delta k} = \frac{\lambda}{\Delta\lambda} = mN = \frac{(\sin\beta - \sin\alpha)\,W}{\lambda}, \tag{7.10}$$

where N is the number of grating lines illuminated. This gives a useful upper limit on R,

$$R_{\text{max}} = \frac{2W}{\lambda}, \tag{7.11}$$

with the limit achieved when $\sin\theta_d = -\sin\theta_i = 1$ (i.e., coming in at grazing incidence and coming back along the incident path). The resolving power is a somewhat squishy number, of course, because the definition of resolution is arbitrary, you never use slits as narrow as the diffraction limit, the grating is never perfectly flat nor perfectly uniformly ruled, and the effects of coma are a limiting factor anyway. Typical values of R_{max} range from 10^4 to as high as 10^6 for very large UV gratings.

7.5.4 Slit-Limited Resolution

We normally don't try to achieve the diffraction-limited resolution, because it requires extremely narrow slits, which are very fiddly and let through very little light. Thus in most cases, it's sensible to use ray optics to discuss spectrometer resolution.

A slit of width w used with a mirror of focal length f produces a beam with an angular range of w/f radians. The grating and the second mirror will reimage the entrance slit on the exit slit, with an angular magnification of $1/M = \cos\theta_i/\cos\theta_d$. The spatial pattern that results is the product of the two slit patterns, so the angular smear is the convolution of that of the two slits, taking account of magnification,

$$A(\beta) = \text{rect}\left(\frac{f(\beta-\beta_0)}{w_{\text{exit}}}\right) \star M\ \text{rect}\left(\frac{Mf(\beta-\beta_0)}{w_{\text{ent}}}\right), \tag{7.12}$$

which has equivalent width

$$\Delta\theta_{d\text{slit}} = (w_{\text{exit}} + w_{\text{ent}}/M)/f. \tag{7.13}$$

This can be translated into spectral resolution by dividing by $\partial\theta_d/\partial\lambda$,

$$\left.\frac{\Delta\lambda}{\lambda}\right|_{\text{slit}} = \frac{\Delta\beta}{\lambda D} = \frac{(w_{\text{exit}}\cos\beta + w_{\text{ent}}\cos\alpha)}{\sin\beta - \sin\alpha}. \tag{7.14}$$

7.5.5 Étendue

Neglecting diffraction, the étendue $n^2 A\Omega'$ of a grating spectrometer is the product of the entrance slit area wL and the projected solid angle of the grating as seen from the entrance slit, which is approximately W^2/f^2 at normal incidence. The oblique projection of the grating area goes down by cos θ_i, and the anamorphic magnification $M = \cos\theta_d/\cos\theta_i$ changes the effective size of the exit slit, but those are both effects of order 1, so we'll sweep them under the rug and say

$$A\Omega' = \frac{wLW^2}{f^2}, \tag{7.15}$$

which is tiny; if $f = 250$ mm, a 25 mm grating with a 5 mm × 20 μm slit (4× the diffraction limit at $\lambda = 500$ nm, about typical), $n^2 A\Omega = 10^{-5}\text{cm}^2\cdot\text{sr}$. Unfortunately, the resolution goes as $1/w$, so we are faced with a 1:1 trade-off of resolution versus photon efficiency for a fixed grating size. The diffraction efficiency of the grating isn't that great, about 0.8 if we're lucky, and we lose another 5% or so at each mirror. Furthermore, it is only at the centre wavelength that the entrance slit is imaged at the exit slit; at other wavelengths it is offset more and more until the overlap goes to 0, so the total efficiency is reduced by almost half. A good ballpark figure is that the efficiency of a monochromator is around 30%.

7.6 FINE POINTS OF GRATINGS

7.6.1 Order Strengths

In general, accurately predicting the strengths of the various diffraction orders requires a vector-field calculation using the true boundary conditions, but there are some general features we can pull out by waving our arms.

Right away, we can see by conservation of energy that there are liable to be sharp changes in the grating efficiency whenever some m enters or leaves the range (7.3), whether by changing θ_i or λ. These are the so-called Wood's anomalies, which are sharp peaks and troughs in the diffraction efficiency curves. With some grating profiles, these anomalies are extremely large and sharp. Figure 7.6 shows the calculated diffraction efficiency of a ruled plane grating with a 9° blaze angle, used in Littrow, which displays strong polarization sensitivity and Wood's anomalies.

7.6.2 Polarization Dependence

The diffraction efficiency, of a grating is usually a strong function of polarization, being highest for light polarized across the grooves (i.e., p-polarized, when $k_y = 0$). This is intuitively reasonable when we consider the behavior of wire antennas—light polarized

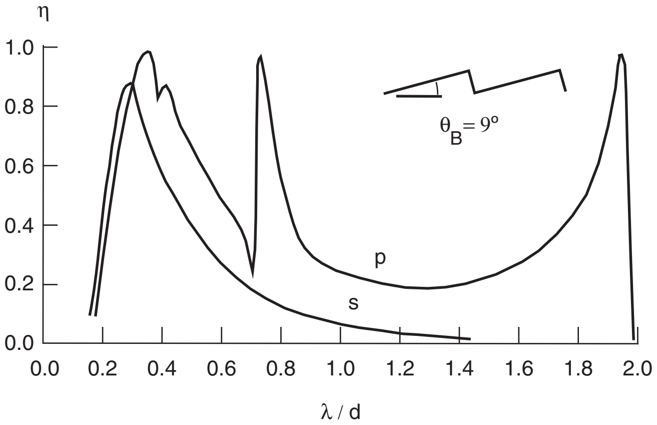


Figure 7.6. Theoretical diffraction efficiency in Littrow of a 9° blazed grating. Note the strong Wood's anomalies and polarization dependence. (Courtesy of The Richardson Grating Laboratory. Note that the Richardson book interchanges p and s.)

along the wires (s) causes strong currents to flow in the wires, which therefore reflect it (or absorb, if there's a load attached). The current is not interrupted by a change in the direction of the surface, because it's flowing along the grooves; thus there is little light scattered. On the other hand, light polarized across the grooves (p) causes currents that have to go over the top of the grooves, leading to strong scatter. By and large, this is how gratings behave, although there are lots of subtleties and exceptions.

7.6.3 Bragg Gratings

The detailed theory of free-space Bragg gratings is fairly involved because of multiple scattering, but weak ones with constant fringe spacing can be treated by coupled-mode theory, since there are only a few orders to worry about.

The main feature of Bragg gratings is that the phase matching condition has to apply throughout the volume, rather than just at one plane, leading to the Bragg condition, which for a sinusoidal grating is

$$\mathbf{k}_d - \mathbf{k}_i = \pm\mathbf{k}_G, \tag{7.16}$$

which is known as the *Bragg condition*.

It's a very stiff condition, since as Figure 7.7 shows, $\mathbf{k}_G$ is the base of the isosceles triangle made up of $\mathbf{k}_i$ and $\mathbf{k}_d$; for a given wavelength and an infinite grating, there's only a single choice of $\mathbf{k}_i$ that works. This $\mathbf{k}_i$ selectivity is smeared out by the finite depth of the grating, just as the angular spectrum is by the finite width of the grating. Nonetheless, a deep Bragg grating with many fringes is a high-Q filter in **k**-space.

Bragg gratings can have diffraction efficiencies approaching unity, and in fact since the diffracted wave meets the Bragg condition for being diffracted back into the incident wave, Bragg diffraction is a coupled-modes problem of the sort we'll encounter in Section 8.3.3.

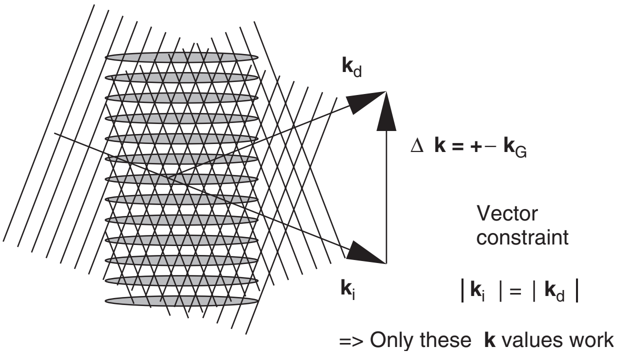

$\mathbf{k}_d$
Δ **k** = +− $\mathbf{k}_G$
Vector constraint
$\mathbf{k}_i$
$|\mathbf{k}_i| = |\mathbf{k}_d|$
=> Only these **k** values work

Figure 7.7. Bragg grating.

7.7 HOLOGRAPHIC OPTICAL ELEMENTS

We noted that all gratings are holograms, in the broad sense of embodying the interference pattern of the incident light with the diffracted light. The class of holograms is of course much more general than just diffraction gratings. Fresnel zone plates are basically holograms of lenses, for example, and with the development of computer-generated holograms we can conveniently make more general patterns, for example, holographic null correctors for testing aspheric optics with spherical reference surfaces, and even beamsplitters producing several diffracted beams in different directions with widely differing strengths.

It is somewhat ironic, but a consequence of their very generality, that holographic elements tend to be application specific. You probably won't find an off-the-shelf item that does what you want, so holograms are used mostly in high volume applications such as bar code scanners.

One exception is the holographic diffuser, whose range of scattering angles can range from 1° to 60°. These work poorly with lasers due to speckle but are just the ticket for situations where an ugly illumination function has to be spread out, for example, LEDs, fiber bundles, and liquid light pipes.

Holograms function differently depending on the number of distinct phase or amplitude levels available. Simple holograms have only two levels: zone plates made from annular holes in chrome masks or from a single layer of photoresist. This presents some problems. For one thing, a two-level zone plate functions as both a positive and negative lens, since with only two levels, $\exp(i\mathbf{k}\cdot\mathbf{x})$ is indistinguishable from $\exp(-i\mathbf{k}\cdot\mathbf{x})$. Just as adding bits helps a DAC to mimic an analog signal more accurately, so more levels of phase or amplitude improves holographic optics. The most popular way to implement this idea is *binary optics*: add more levels of lithography, with binary-weighted phase shifts, just as in a DAC.

7.7.1 Combining Dispersing Elements

It is often convenient to combine two dispersive elements in one system, such that their angular dispersions add or subtract. The usual rule is that if the diffraction direction tends to straighten out, the dispersion effects tend to subtract, whereas if the beam is being sent round and round in a circle, the effects add. This of course applies only for elements of

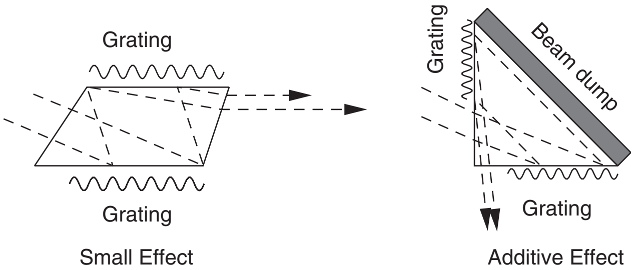


Figure 7.8. Cobbling gratings to add or subtract the diffractive effects, while possibly changing the beam shape anamorphically. Leaving the second grating near grazing multiplies the tuning sensitivity of both diffractions.

the same sort; gratings bend shorter wavelengths less, whereas prisms bend them more. A combination grating and prism could have additive effects even with a nearly straight optical path, like an Amici prism (Section 4.9.3).

Gratings and prisms can be used to change the beam diameter anamorphically (i.e., with magnification different in x and y). The gratings in Figure 7.8 first expand and then contract the beam. The width out of the page is of course unaltered. This anamorphic property is often useful with diode lasers, and Figure 7.8 shows how it can be accomplished with and without tuning sensitivity of the angular deflection.

7.8 RETROREFLECTIVE MATERIALS

We're all familiar with retroreflecting materials, used for bicycle reflectors, traffic signs, and safety clothing. They have lots of possibilities in instrument design too, and so everyone should know something about them. The basic idea is to use large numbers of small, poor quality retroreflectors, so as to return incident light generally back toward its source, with an angular spread of 0.5° to 5° in half-angle and (unlike larger corner cubes) no significant lateral shift.

There are two main kinds: glass beads and corner cube arrays embossed in plastic (see Section 4.9.8). A sphere in air makes a pretty good retroreflector if its index is chosen so that incident light is focused on the opposite surface of the sphere. A ray at height h, parallel to the axis, has an angle of incidence $\sin\theta_i = h/R$, and to focus on the back surface, $\sin\,\theta_d = h/(2R)$, so by Snell's law,

$$\frac{n_2}{n_1} = \frac{\sin\theta_i}{\sin\theta_r} = \frac{h/R}{h/(2R)} = 2, \tag{7.17}$$

which can just about be done in glass. The angular spread can be adjusted via defocus, by varying n. To prevent most of the light getting out the other side, aluminum-coated beads are used, with the aluminum etched off the top surface after the coating is applied. The embossed plastic stuff has a narrower acceptance angle than the spheres, because away from the symmetry axis of the corner, more and more of the light fails to make the three TIR bounces required for retroreflection. It's also a bit of a puzzle to mount, because you can't touch the TIR surface or you'll deactivate all the little cubes. The standard solution is to use a quilted foam backing that touches only a few percent of the area, which makes the retroreflection spatially nonuniform.

The figure of merit for a retroreflective material is its *specific brightness*, which is measured in inverse steradians, although it's quoted as lux per steradian per lux or other cute units that amount to the same thing: if you put in 1 W/m^2, how much flux is radiated per steradian at specified angular offsets from exact back-reflection? For a retroreflector with an RMS beam spread half-angle of $\Delta\theta$ and photon efficiency η, the specific brightness is

$$B \approx \frac{\eta}{\pi(\Delta\theta)^2}. \tag{7.18}$$

For a $\Delta\theta$ of 0.5°, this is around 4000—a factor of 13,000 over a Lambertian ($\Omega' = \pi$) reflector, assuming that $\eta = 1$. The best real materials are more like 900 or 1000, a factor of 3000 or so over Lambertian. This is an enormous effect—a strip of tape in the right place can get you a 70 dB (electrical) signal level increase, which is well worth the labor and the compromises it takes to use these materials (they're very speckly when used with lasers, for example).

The stuff made for use in signs and clothing isn't that great; the specific brightness is usually below 300, and 60 is not uncommon, but those numbers still represent big signal increases in systems where most of the illumination is not collected. There does exist material that gets into the 10^3 range, but it isn't easy to find. You can also buy just the beads,[†] made usually from barium titanate glass with an index of 1.9 or a bit higher. They're used for spraying onto traffic paint before it dries and may be helpful for special situations where it's inconvenient to use the made-up sheet material. Other considerations include rain—glass bead retroreflector relies on refraction at the air–glass boundary, and so doesn't work well when it's wet. Assuming the TIR surfaces remain dry, the corner cube stuff is almost unaffected by rain.

The best retroreflective materials in the 3M catalog for instrument use are Scotchlite 2000X Very High Gain Sheeting (corner cubes) and Scotchlite 7610 High Gain Reflective Sheeting (spheres). Both are available in tape rolls and can really make a difference to your SNR, especially in a fiber-coupled instrument. The other main manufacturer of this stuff, Reflexite, also has cube material tailored to produce a nearly fixed offset angle (not 180°).

TIR film will mess up the polarization of your beam, as in Section 4.9.8. Reflexite makes metallized corner cube material, which reduces this problem. Plastic retroreflector is naturally of no use in the UV, but the high brightness glass bead stuff is quite good, as it has no plastic overcoat on top of the spheres.

7.9 SCANNERS

Scanning systems are one of the thorniest areas of electro-optical instrument building. The whole point of scanning is to interrogate a huge $A\Omega$ volume sequentially with a low-$A\Omega$ system. None of the available methods is as good as we'd like, and the cost of commercial solutions is enough to make you gasp (how about \$4000 for a 25 mm, two-axis galvo scanner with analog driver cards and no power supply?).

[†] Suppliers include Potters Industries and Cataphote.

The difficult design issues, together with the very high cost of playing it safe, make scanner design worth a considerable amount of attention. The main points to worry about are scanner and encoder accuracy, rectilinearity, range, speed, jitter, aberration buildup, field flatness, temperature drift, and power consumption (other than that, it's easy). We'll start with mechanical scanners.

Before we begin, it is useful to have a bit of vocabulary to describe different scanner vices. Nonrepeatable motions, caused, for example, by out-of-round bearings, are called *jitter* if they're in the scan direction and *wobble* otherwise. Repeatable errors are *scan nonlinearity* along the scan direction, and *scan curvature* out of it. Temperature drift, which is sometimes very serious, comprises offset (zero) drift and gain drift. Scanning is often combined with signal averaging to reject noise and spurious signals; see Section 10.9.2.

7.9.1 Galvos

Galvanometer scanners are electric motors that don't turn very far ($\pm 45^\circ$, maximum) but can accelerate very fast and move very accurately. They usually run on ball bearings, but some use flexure bearings. Single- and double-axis galvos are available that provide 12 bit angular accuracy with settling times of several milliseconds. Small ones are much faster than large ones, because the moment of inertia I goes as mr^2, which tends to grow as r^5. The angular accuracy of the encoders isn't all that great over temperature, typically 10 arc seconds drift and 0.05% gain/°C, with some being much worse. If you really need that 12 bits, you have to compensate for those in some way. Jitter typically runs 10–20 arc seconds, and wobble more like 5–10. Those are really what limit the resolution. The torque from any motor tends to go as the volume—due to field and current limitations, the available force per unit area is approximately constant, so the torque goes as the surface area times the diameter. The moment of inertia grows faster than that, so big galvos tend to be slow.

Resonant galvos shrink the mirror and the motor, and add a big torsion spring to get their restoring force, which enormously reduces their moment of inertia. This makes them quite a bit faster (500 Hz), but being high-Q resonant systems, they cannot be controlled within a cycle; only the amplitude and phase of their sinusoidal oscillation can be changed, and many cycles are required for settling afterwards. Thus resonant galvos are good fast-scan devices, where they compete with rotating polygons and hologons; the trade-off is adjustable angular range and sinusoidal scan versus uniform scan speed through a fixed angular range.

7.9.2 Rotating Scanners

All reciprocating scanners have to slow down and turn around at the end of their travel, which makes them relatively slow. What's worse, their varying scan speed makes it relatively difficult to take data points equally spaced in angle—it requires a continuously varying clock frequency. This can be implemented with a lookup table in hardware, or done by resampling the data afterwards (see Section 17.8). Since the dwell time on each pixel is different, more lookup tables may be needed to take out the resulting gain error and offset. All these lookups, whose contents depend on unit-to-unit differences such as the details of loop damping and rotor imbalance, require an onerous calibration procedure for high accuracy applications.

Nonconstant scan speed is particularly obnoxious when you're using a continuous frame scan, since it leads to hooks at the ends of the scan line. Large amounts of overscan are necessary to combat it. A scanner with good acceleration can make sharper turns, so less overscan is needed with high torque motors, low moment of inertia (i.e., small mirrors), and slower scan rates. This is all manageable, but still a constant, fast-scan speed (constant in m/s or rad/s depending on the application) would be useful for raster applications.

One partial solution is a continuously rotating scanner, such as a polygon mirror or holographic scanner.

7.9.3 Polygon Scanners

A polygon scanner is a spinning wheel with flat mirror facets on its periphery (Figure 7.9). These may be oriented normal to the radius vector, so that the wheel is a polygonal cylinder, or angled like a cone. In order to get a straight-line scan with a cylindrical polygon, the beam has to come in normal to the rotation axis, although other configurations exist with tilted facets.

With the beam so aligned, rotating a mirror through $\theta/2$ deviates the reflected light by θ, so an n-facet polygon deflects light through an angular range $\Delta\theta$ of

$$\Delta\theta = \frac{4\pi}{n}, \tag{7.19}$$

although you can't use all that range, since at some point the edge of the facet has to cross your beam, leading to a dead time on each facet. A polygon rotating at constant speed naturally produces a constant angular velocity (CAV) scan, which is great for some things (e.g., lidar) but a problem for others (e.g., document scanning, where a constant linear velocity is much more convenient). Polygons can go very fast; speeds over 50,000 rpm can be achieved with a small solid beryllium scanner in a partial vacuum, though not easily or cheaply. The ultimate limit is set by deformation and then failure of the polygon itself. High end printers ($1M) use polygons with 10–12 facets running at up to 40,000 rpm on air bearings (with 10 beams, that's a 70 kHz line rate), but in more pedestrian applications, keep below 7000 rpm, and remember that below 3000

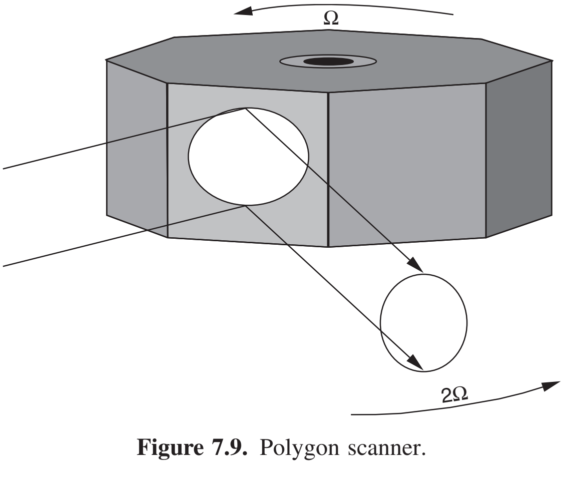


Figure 7.9. Polygon scanner.

things get much easier. Polygons cause no aberration of the scanned beam. They have a unidirectional scan pattern, with a retrace interval as the edge between two facets crosses the beam. Well-made polygons can have a facet-to-facet angular accuracy of a few arc seconds, which is good enough for most applications. Cheap polygons are closer to an arc minute, but cost only a few dollars.

7.9.4 Polygon Drawbacks

Polygons have a number of drawbacks. Their moment of inertia is high, so that the scan rate cannot be changed rapidly. We usually use them to get high speed, so that the kinetic energy is high, which compounds the speed problem and adds wind resistance, turbulence, and noise, ultimately forcing us to use a vacuum chamber.

The constant angular velocity of the beam means that it scans a flat surface at a nonuniform rate, unless we do something about it. A subtler difficulty is that since the rotation axis does not pass through the surface of the mirror, as it does with a galvanometer, the scanned beam translates as well as pivoting during a line. Thus a polygon-scanned system lacks a true pupil. You can get (very expensive) f-θ lenses, which have just the right amount of built-in barrel distortion to convert a constant angular velocity scanned beam into a constant linear velocity spot, and a flat field; they're big chunks of glass, used at very low aperture (e.g., a "250 mm $f/16$" lens whose front element is 90 mm across). It is somewhat galling to have to use a \$1200 lens with a \$25 polygon.

The remaining trouble is their very high sensitivity to shaft wobble. A polygon accurate to a few arc seconds is very much easier to get than a motor whose shaft is straight and stable to that accuracy, especially at the end of its life. Air bearings are a good choice for high speed polygons, but they don't usually work as well at low speed.

7.9.5 Butterfly Scanners

In Section 4.9.4, we encountered the pentaprism, which is a constant deviation 90° prism that works by having two reflections; tipping the prism increases one angle while decreasing the other, making the total constant. The same principle can be applied to scanning, resulting in the butterfly scanner of Figure 7.10, which is a nearly complete solution to the shaft-wobble problem; drawbacks are complexity, expense, probably worse fixed facet-to-facet errors, and much higher air resistance, noise, and turbulence.

7.9.6 Correcting Rasters

Once the shaft wobble has been corrected, the scan is still not perfect. To get really good rasters from any multisegment device, you really have to have software that knows which segment you're on, and dials in the appropriate offsets. While this requires calibration, it's not a difficult problem since it's only line-to-line variation and can be done in the instrument itself using a vertical bar test pattern. Furthermore, the dimensional stability of the hologon or polygon means that it can be done once and forgotten about.

In-plane errors cause only timing trouble, which can be eliminated completely. Out-of-plane errors are more obnoxious visually, causing obvious nonuniformity in raster line spacing, and are more difficult to handle, requiring an additional deflection element such as a Bragg cell.

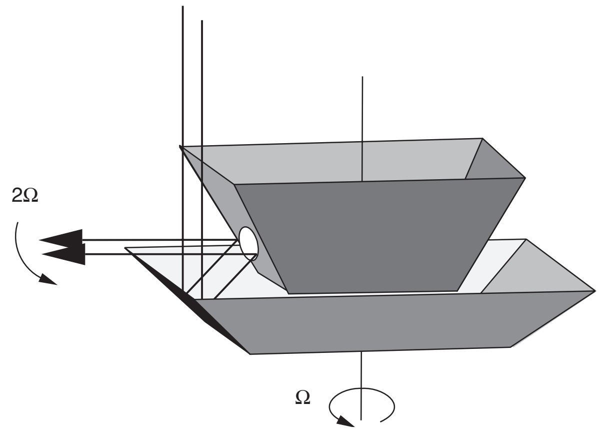


Figure 7.10. The butterfly scanner works on the pentaprism principle.

7.9.7 Descanning

In order to reduce the required $n^2 A\Omega'$ of the detection system (which makes it smaller, faster, cheaper, and dramatically more resistant to background light), we usually need to *descan* the received light. There's an easy and a hard way to do this.

The easy way is to work in backscatter, and use the same scanner on the transmit and receive sides of the instrument, for example, to interrogate a sample one point at a time. If the scanner is at the pupil of the objective lens, light backscattered from the sample will be recollimated, so that it comes back antiparallel to the transmit beam. The mirror won't have had time to move far during the time of flight, so by symmetry, both the scan angle and any angular jitter are removed, leaving only a bit of Doppler shift.

The hard way is to use a second scanner, synchronized to the first. You have to do this sometimes, for example, in a long path sensor where you're sweeping a focused beam back and forth without steering it, and can't work in backscatter for some reason. This really is a miserable way to make a living, so work hard to avoid it; a corner cube or some retroreflecting tape will often let you off this hook.

Aside: Preobjective and Postobjective Scanning. Before scanning, our beam is usually collimated, so its NA is very low. It seems a bit perverse to take something like that, which we can focus very well with inexpensive lenses, scan it through perhaps 45°, and then try to focus it with a very expensive f-θ lens. Couldn't we focus first and scan afterwards?

If the NA is low enough and the working distance long enough, this is a good possibility. The major drawback is that the field flatness and nonuniform scan speed (in m/s on the sample surface) are uncorrected unless you do something fancy yourself. This may not matter in your application, in which case this *postobjective* scan strategy will work well. Just don't try it with hologons (see Section 7.9.9). A hybrid scheme, where the line scan is preobjective and the frame scan is postobjective, is also widely used.

7.9.8 Constant Linear Scan Speed

It isn't that easy to achieve constant scan speed with a mechanical scanner. Rotating a mirror at a constant speed $\dot{\theta}/2$ produces a reflected beam whose angular speed $\dot{\theta}$ is constant; if the scanner is a distance h above a planar surface, the scan position x on the surface is

$$x = h \tan\theta, \tag{7.20}$$

which is stretched out at large angles, just like late-afternoon shadows. Reciprocating scanners such as galvanometers and voice coils slow down and stop at the ends of their angular travel, so they tend to scan more slowly at the edges; if the beam scans sinusoidally through $\pm\theta_{\text{pk}}$ at radian frequency ω, then the scan speed v is

$$v(\theta) = h\omega\sqrt{\theta_{\text{pk}}^2 - \theta^2}\,\sec^2\theta \quad (\dot{\theta} > 0). \tag{7.21}$$

The slowdown of $\dot{\theta}$ at large θ approximately compensates the stretching out of x, so that the resulting curves look like the Chebyshev filters of Section 15.8.3; choosing $\theta_{\text{pk}} = 40.6^\circ$ produces a maximally flat response. Table 7.1 shows optimal scan parameters for an equiripple error from $\pm 0.01\%$ to $\pm 5\%$: tolerance, linear range θ_L, peak range θ_{pk}, duty cycle (or scan efficiency), and the corresponding error with a CAV scan of $\pm\theta_L$. Note how the duty cycle improves as the error band is relaxed, and how much lower the maximum error is for the galvo versus the polygon, at least when used unaided. Usually we have to accept more scan speed variation and compensate for it with slightly nonsinusoidal motion (easiest), nonuniform pixel clock speed, resampling digitally afterwards, or (as an expensive last resort) an f-θ lens.

If we need to focus the beam on the surface, it's usually enough to put the galvo at the pupil of a flat field lens, with due attention to the lens's geometric distortion.

7.9.9 Hologons

A holographic scanner consists of a spinning surface containing holographic elements (Figure 7.11). The most common type is the *hologon*, short for holographic polygon. A hologon is a spinning glass disc containing radially segmented transmission gratings (like orange segments), with the grating lines oriented tangentially.

Hologon scanners are best operated near minimum deflection, that is, when the incoming and outgoing beam make equal angles with the surface normal. Small amounts of wobble in the shaft then cause only second-order angular deviations in the frame direction, which is an important advantage of holographic scanners over simple polygon mirrors, though butterfly scanners do even better. Beiser shows that for a scanner producing 90°

TABLE 7.1. Approximating a Constant Linear Velocity Scan with a Sinusoidal Galvo

Speed Tolerance (±%)	θ_L (±°)	θ_{pk} (±°)	Duty Cycle (%)	Constant AV Error (±%)
0.01	10.25	41.03	16.1	1.6
0.05	15.0	41.68	23.4	3.5
0.1	17.7	42.17	27.6	4.9
0.5	26.2	44.2	40.5	10.8
1	31.0	45.7	47.5	15
5	52.8	56.8	76	46

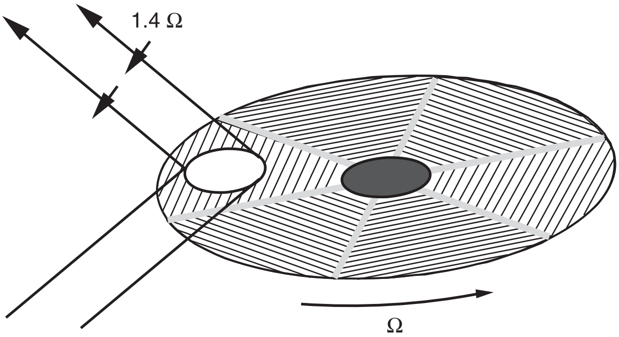


Figure 7.11. A hologon scanner is the diffractive analogue of a polygon. It isn't as efficient but has much lower jitter, weight, and wind resistance.

deviation ($\theta_i = \theta_o = 45^\circ$), a large shaft wobble of 0.1° (360 arc sec, or 1.75 mrad) produces a scan wobble in the frame direction of only 1.3 arc sec, an improvement of nearly 600:1 over a polygon, with even better performance for smaller wobbles.

The scan line produced by a hologon scanner is in general curved, because k_y can't always be 0 when we're rotating the grating in azimuth. By choosing $\theta_i = \theta_d = 45^\circ$, the scan can be made almost perfectly straight, which is how they are usually used. The deflection is easily found by applying phase matching; if the grating is rotated through an angle ϕ_{shaft} from the center of the facet, the change in k_x of the diffracted beam is equal to that of the grating, so in the middle of the scan line,

$$\frac{\partial\theta_{\text{az}}}{\partial\theta_{\text{shaft}}} \approx \frac{k_G}{k}, \tag{7.22}$$

which is equal to $\sqrt{2}$ for the 45°–45° scanner. The angular scan speed $\dot{\theta}$ is also mildly nonuniform, being slightly compressed at the ends of the travel (we saw that this is actually an advantage in scanning flat surfaces). The effect is nowhere near as strong as with galvos.

As a practical matter, hologons are restricted to collimated beams. A focused beam used with a collimated-beam hologon produces an astounding amount of astigmatism—dozens of waves over a 45° scan angle, even with an NA of only 0.01. Since they are holograms, it is possible to make a scanner that focuses as well as deflects the light. It might seem that the resultant saving of a lens would be very valuable, but in fact doing this throws away the major advantage of hologons, the insensitivity to shaft wobble. Resist the temptation to be too fancy here, unless your performance specs are modest (e.g., in hand-held bar code scanners). One possible exception would be adding a few waves of optical power to correct for aberrations in a simplified scan lens, because the wobble effect would then still be small. The angular accuracy of a hologon can be as good as 10 arc seconds facet to facet, although 30 is easier and hence cheaper.

If the facets are made with slightly different values of G they will deflect the beam at slightly different angles, so that an N-facet hologon by itself can perform an N line raster scan, which allows a useful trade-off between scan speed and alignment accuracy. (Doing this with a polygon would make it dynamically unbalanced.)

The diffraction efficiency of hologons is normally quite good—80–90%, but that isn't as good as a properly chosen mirror coating, so you'll pay a decibel or two in detected signal for using a hologon.

7.9.10 Fast and Cheap Scanners

If your scan range and accuracy requirements are modest, don't forget the obvious candidates, for example, mounting a laser diode on a piezoelectric translator or a bimorph, and letting the collimating lens do the work. Life doesn't get much easier than that.

7.9.11 Dispersive Scanning

It is also possible to scan over a small range by using a tunable source (e.g., a diode laser) with a dispersive element, such as the second compound grating device in Figure 7.8. This is a very good technique for some purposes, because it is extremely fast (∼20 resolvable spots in 3 ns), and although you do have to avoid mode jumps and cope with power variations, it presents few problems otherwise.

7.9.12 Raster Scanning

Raster scanning requires a 2D scanner or two 1D scanners in cascade. You can get two-axis voice coil type scanners, which tip a single mirror about two axes; they behave a bit like galvos but have only a few degrees' scan range and aren't as stable or repeatable, because the mirror mount usually relies on a single wire in the middle for its location, and the orthogonality of the tilts is not well controlled.

If we need to use two scanners, we must either accept a pupil that moves around a good deal (several centimeters with most galvos), or use a relay lens to image the center of one scanner on the center of the other. The usual approach is to use a small fast scanner first, to do the line scan, and a large, slow one afterwards for the frame scan, although the relay lens approach allows both scanners to be the same size. The moving pupil problem is worst with a pure preobjective scan approach, but if you can put the scan lens between the line and frame scanners, it gets a lot easier; in the ideal case of pure postobjective scanning, you can use a simple lens for focusing, with perhaps a weak toric field-flattening lens to correct for the different object distances at different scan positions.

7.9.13 Mechanical Scanning

Another approach is to keep the optical system very still and move the sample, as is done in some confocal microscopes and step-and-repeat photolithography tools. This is slow and prone to mechanical jitter, due to the requirement to start and stop the motion of a large mass quickly, and to the instabilities of translation stages. On the other hand, your point-spread function is really constant with position, and there is no limit on the number of resolvable spots. Another mechanical scanning method is to rotate or translate the entire optical system, as in a periscope or an astronomical telescope, which scans slowly to correct for the Earth's rotation.

7.10 MODULATORS

Diode lasers are unique among optical sources in being highly directional and easily modulated at high speed. Unfortunately, most sources of interest are not like that, so we need external modulators. Under this rubric lie a fairly motley collection of out-of-the-way physical effects, all of which have serious drawbacks, not all widely known. Modulators in general are troublesome devices if you need nice pure beams with uniform polarization, no etalon fringes, and smooth amplitude profiles.

Most modulators are based on bilinear interactions[†] between two fields in some material, for example, the electro-optic effect, where the applied electrostatic field causes the dielectric tensor to change, or the acousto-optic effect, where the optical wave diffracts from a sinusoidal phase grating produced by the traveling acoustic wave.

7.10.1 Pockels and Kerr Cells

Optical modulators based on the linear (Pockels) or quadratic (Kerr [‡]) electro-optic effects are shown in Figure 7.12. These are among the fastest modulators of all, but they are a genuine pain to use. Think of them as voltage-controlled wave plates.

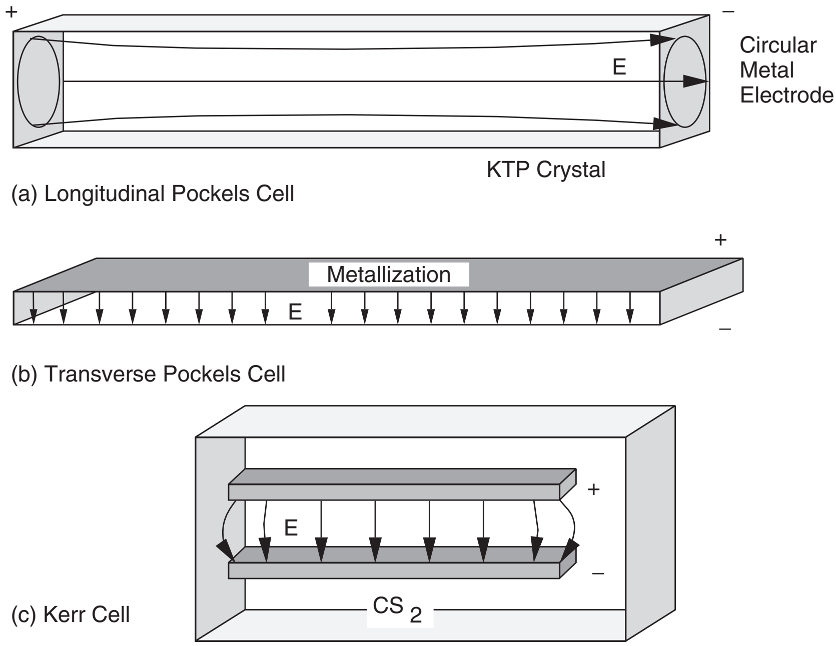


Figure 7.12. Pockels and Kerr cells.

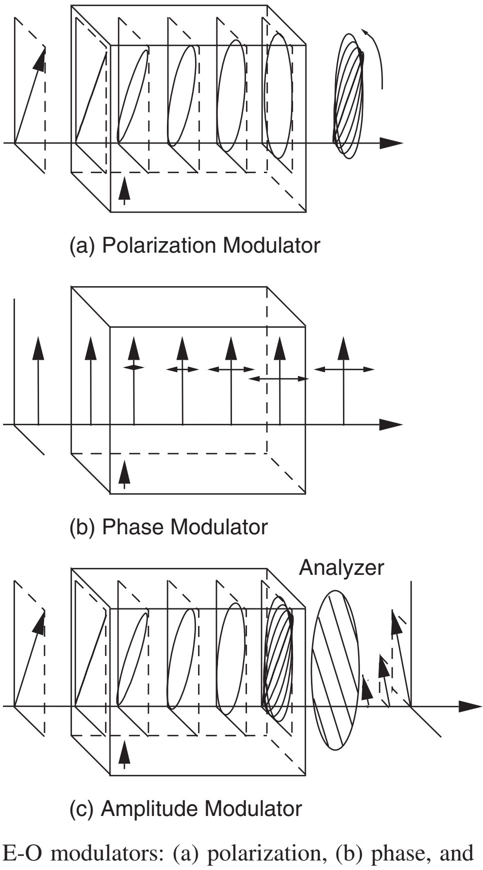
[†] A bilinear interaction is one that is linear in each of two independent variables; that is, it is expressible as $f(x, u) = g(x)h(u)$. An electronically variable attenuator is an example of a bilinear device if its gain is linear in its control voltage.

[‡] The electro-optic Kerr effect is not the same as the magneto-optic Kerr effect, which leads to polarization rotation in linearly polarized light reflected from a magnetized medium.

Kerr cells are based on a quadratic electro-optic effect in isotropic materials (usually nasty inflammable organic liquids such as carbon disulfide or nitrobenzene). Their quadratic characteristic makes them harder to use in applications needing linearity, of course, but since the static birefringence is 0, they are pretty predictable. Kerr cells are excited transversely by dunking capacitor plates into the liquid cell. They are normally used singly, with bias voltages around 10 kV. The organic liquids are strongly absorbing in the UV, so Kerr cells are generally limited to the visible and near IR. Getting decent uniformity requires limiting the fringing fields, which (as we'll see in Section 16.2.5) means making the plate dimensions several times their separation.

The variable retardation of electro-optic cells can be used to modulate the polarization and phase of a beam, as shown in Figure 7.13. Pockels cells are built from crystals such as potassium dihydrogen phosphate (KDP) or lithium niobate, nasty birefringent things whose dielectric tensor $\underline{\epsilon}$ depends on the applied **E**. The dependence of $\underline{\epsilon}$ is generally complicated; a given applied **E** can change all the coefficients of $\underline{\epsilon}$. Since the material is already anisotropic, the leading-order change is linear in applied field.

Pockels and Kerr cells are usually used as amplitude modulators, by following the polarization modulator with an analyzer. They can also be used as phase modulators, by aligning the polarization of the beam with one of the crystal axes, so that the polarization remains constant but n varies. It's hard to get this really right in a multielement cell,


(a) Polarization Modulator

(b) Phase Modulator

(c) Amplitude Modulator

Figure 7.13. E-O modulators: (a) polarization, (b) phase, and (c) amplitude.

because the optic axes of the elements may not be aligned well enough. Fancier things can be done, for example, frequency shifters made by putting a rotating **E** field on a crystal between crossed circular polarizers, but they tend to be rare compared with phase, amplitude, and polarization modulation applications.

There are two main designs for Pockels cells. Because the applied **E** changes more than one element of the dielectric tensor, the field can be applied longitudinally (parallel to **k**) or transversely. Longitudinal cells usually have a ring electrode around the outside of the face, which produces some polarization nonuniformity that limits their ultimate extinction ratios to a few hundred, even with long thin cells. Transparent electrodes such as indium–tin oxide are also used, but they're pretty lossy, and not conductive enough for the highest speed applications; for really fast stuff, the champion electrode is a low pressure neon plasma, which improves the extinction to a few thousand, even while improving the étendue.[†]

Transverse cells are simply metallized. The trade-off between the two styles is in étendue and capacitance versus voltage; a longitudinal Pockels cell has a reasonably large étendue (especially the ITO and neon plasma types) but requires high voltage, whereas a transverse one has high capacitance and (because it is long and narrow) very low étendue.

Since the effect is linear, the number of waves of birefringence in a longitudinal cell is proportional to the line integral of $\mathbf{E}\cdot d\mathbf{s}$ along the ray path, that is, to the voltage drop across the crystal. Since going from off to fully on requires a change in retardation of one-half wave (why?), the figure of merit for a given electro-optic material is the *half-wave voltage* V_π, which—most inconveniently—is usually several thousand volts.[‡] Because both polarizations get phase shifted in the same direction, the retardation is less than the phase modulation, so phase modulators can work at somewhat lower voltages.

The Pockels effect is fast; a properly designed transverse cell such as a 40 Gb/s telecom modulator can switch in 10 ps, and the intrinsic speed is even higher. The problem in bulk optics is that with a longitudinal cell whose half-wave voltage is (say) 4 kV, to get a 1 ns edge we have to make the bias voltage change at a rate of 4,000,000 V/μs, which is an interesting challenge—if the 50 Ω connecting cable is 10 cm long, it'll take 80 A for 1 ns just to charge the cable. You can do that with a spark gap or a series string of avalanche transistors, but only just; the usual method is a big thyratron producing a 10 ns edge, running into a ferrite-loaded coaxial pulse forming network.[§] All these are limited to shuttering applications since you can't stop an avalanche once it starts. (These techniques are discussed in Section 15.14.1.) Accordingly, people have worked out ways to ease the voltage requirement. The main way is to take many thin plates of electro-optic material dunked in index oil and drive them in parallel as in an interdigitated capacitor, as shown in Figure 7.12. You can get these down into sub-400 V territory, which is a lot easier although still not trivial.

The optical path in a Pockels cell contains a great big chunk of birefringent material, so it has a huge static retardation (many waves), and thus has a few percent nonuniformity

[†] Pockels cell people enjoy suffering almost as much as femtosecond dye laser people used to.

[‡] This is an example of the extreme linearity of optical materials—even if we get to pick an arbitrarily horrible material, and orient it for maximum effect, we still need thousands of volts to make an order-unity difference in the transmitted field.

[§] Saturation in the ferrite causes the back end of the pulse to move faster than the front end, which leads to shock formation, like the breaking of an ocean wave on a beach. You can get below 100 ps rise time for a 20 kV pulse in 50 Ω, but you have to really want to, and the EMI problems are, *ahem*, interesting.

of retardation across its face, fairly poor wavefront fidelity, and a significant amount of temperature drift. Many of the crystals used for Pockels cells are piezoelectric, and so they may exhibit low frequency resonances; those made with biaxial crystals have particularly nasty temperature dependence, since (unlike uniaxial crystals) the optic axes can move around with temperature. For a device with as much retardation as a Pockels cell, this can be a serious drawback. Low voltage devices have lots of etalon fringes too. For high accuracy applications, longitudinal Pockels cells need a lot of babysitting, and that confines them pretty much to lab applications.

7.10.2 House-Trained Pockels Cells: Resonant and Longitudinal

Many applications of modulators are relatively narrowband, so that we can stick the cell in an electrical resonator to reduce the required voltage by a factor of Q. Cells like that are available from 100 kHz up past 30 GHz, with operating voltages of 6–30 V.

Transverse cells tend to be long and thin because we win by a factor of L/d, which allows operating voltages down to the tens-of-volts range, a much better match to ordinary circuitry. This leads to higher capacitance, but since the electrical power goes as CV^2, we win anyway, and traveling-wave structures can be used to make the device look resistive when that becomes a problem.[†] The most serious limitation is their very small étendue. Even for light going right down the axis, the beam often has to be so small in diameter that diffraction limits the length of the cell. This is an especially serious limitation in the infrared, where diffraction is worse and more retardation is required to get a half-wave shift. Transverse modulators are most commonly found in integrated optics and fiber-coupled applications, where they are entirely practical; a single-mode waveguide made of electro-optic material needs very little étendue, the field confinement of the waveguide avoids any length limitation due to diffraction, and nonuniformity causes fewer problems since only one mode is involved. The really fast traveling-wave integrated-optic Pockels cells used for telecom usually need about 100–200 mW of RF power and have rise times as short as 12 ps or so. Telecom modulators are usually *zero chirp*, that is, they produce little or no phase modulation, which otherwise shows up as spurious FM sidebands. Chirp is one of the main limitations of directly modulated diode lasers, so this matters. If you really need fast beam modulation, consider using one of these and expanding the beam later.

7.10.3 Liquid Crystal

Another class of electro-optic devices is based on liquid crystals (LCs). These odd materials are liquids made of large molecules, which show some large scale orientation effects even in the liquid state. The physics of liquid crystals is very rich (read complicated). A very slightly grooved surface (e.g., glass that has been wiped in one direction with a cloth pad) can determine the orientation for fixed applications such as wave plates; an applied voltage can change their alignment, which changes the resulting birefringence. Because they rely on the physical motion of molecules, rather than electrons, all liquid crystal modulators are slow (1 μs to 1 ms). You use them like big, slow, low voltage Pockels cells, to modulate polarization, phase, or amplitude.

[†] Think of coaxial cable, which is 100 pF/m, but can handle gigahertz signals over many meters because of its traveling-wave character.

They come in two basic types: the extremely slow, continuously variable *nematic* ones, and the somewhat-faster, binary *ferroelectric* ones. One of the best things about LC devices is their huge étendue; you can get 100 mm diameters with $\Omega \approx 0.5$ sr. They are also nearly indestructible—their damage thresholds are so high they're not easy to measure.[†] Being liquids, they make intimate contact with the electrodes; because their birefringence is so high, they can be very thin. This makes it easy to build spatially addressable LC *spatial light modulators* (SLMs). Besides the familiar LCD displays, SLMs are used to make shutters, masks, and low resolution computer-generated holograms, among other things.

Example 7.2: Phase Flopping Interferometers. One especially good use of LC modulators is in making zero-background imaging measurements by inverting the phase of the signal but not the background, and frame subtracting. For example, many years ago a colleague of the author's, T. G. Van Kessel, built a very successful Nomarski interference system for measuring the latent image in photoresist. (The image is latent between exposure and development.) It was a normal Nomarski-type metallurgical microscope (see Example 10.2) with the addition of an early liquid crystal variable wave plate before the analyzer, oriented parallel to the Nomarski axis (45° to the analyzer axis). Changing the retardation from 0 to $\lambda/2$ on alternate video frames caused a π relative phase shift in the combined beams; this inverted the Nomarski contrast but preserved the background amplitude. Under frame subtraction, the weak phase contrast signals added and the strong background canceled out, making an electronically zero background measurement (see Section 10.8).

7.10.4 Acousto-optic Cells

The most common Bragg grating in bulk optics is the acousto-optic Bragg cell. We encounter the piezo-optic effect in Section 8.5.6, where it causes stress birefringence in lenses and prisms. Launching a strong acoustic plane wave in a material with a big piezo-optic coefficient makes a moving Bragg grating. Typical frequencies are 40 MHz to 2 GHz, which produce acoustic wavelengths of 2–100 μm. If the interaction zone is too skinny, phase matching perpendicular to $\mathbf{k}_A$ is no longer a constraint, so we get many weak diffraction orders, spaced at multiples of $\mathbf{k}_A$. This is the *Raman–Nath* regime, shown in Figure 7.14.

That grating has some unique properties: the diffracted light gets frequency-shifted by $\pm f_A$ depending on which direction it was diffracted. Also, the diffraction efficiency can be smoothly varied from 0% to 80% or so merely by changing the RF power from 0 to a watt or two (and more than that for some materials, e.g., glass).

The phase matching condition can be modified (and sometimes considerably relaxed) by using a birefringent material. By a suitable choice of cut, the change in k_{diff} with incidence angle or grating period can be compensated by that due to the change of n. This trick is used all the time in acousto-optic devices.

Acoustic waves in solids are tensor waves, which include scalar (longitudinal) and vector (transverse) waves, but more general shear waves can propagate too. Predicting the effects of a given order and type of diffraction can be done by classical field theory, but it is far easier and less blunder-prone to take a simplistic quantum view. We know

[†] That doesn't apply to the film polarizers on LC shutters, however.

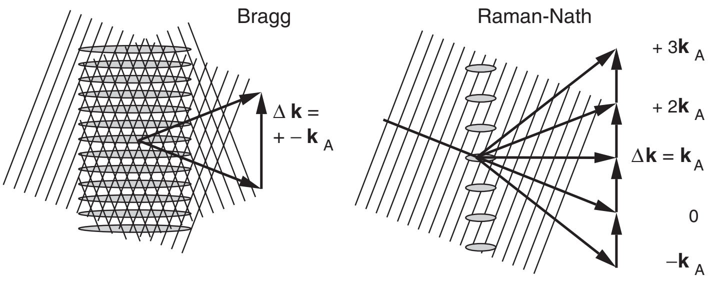


Figure 7.14. Acousto-optic cells: Raman–Nath and Bragg regimes.

that a photon has energy, momentum, and angular momentum; well, so do phonons, and they are all conserved during the interaction, on a quantum-by-quantum basis. A photon that absorbs a phonon (the + or *anti-Stokes* branch) gets its frequency upshifted ($E = \hbar\omega$), and is bent along the acoustic propagation direction ($\mathbf{p} = \hbar\mathbf{k}$)—the energies and momenta add. If instead it emits one (by stimulated emission, the − or *Stokes* branch), it's downshifted and bent away from $\mathbf{k}_{\text{acoustic}}$. Similarly, conservation of angular momentum means that a linearly polarized photon that absorbs or emits a shear phonon has its polarization shifted—it goes from s to p or p to s. A second-order diffraction gets twice the shift in each, because it involves emitting or absorbing two phonons, and so on.

Acousto-optic cells are usually used with laser beams, because their aperture is so small; the major use is as medium-speed amplitude modulators (DC to 10 MHz easily, DC to 100 MHz if you really work at it—focused beams, fast materials, small crystals). One exception is the acousto-optic tunable filter (AOTF), which achieves a wide field angle (and hence a decent étendue) by *noncritical phase matching*, where the curve of λ versus the phase-matched θ_i has a maximum, so the phase matching error is quadratic in $\Delta\theta_i$. These are available in both collinear and noncollinear designs. Narrowness of the passband requires more waves of interaction zone, as we saw, so the angular acceptance goes down as the selectivity goes up; a typical cell with $\Delta\nu/\nu = 0.2\%$ works over $\pm 5^\circ$ ($\Omega = 0.023$ sr), a much wider range than a grating instrument with the same resolution.

You can image through AOTFs, but it doesn't work very well unless you're careful. The images are corrupted by ghosts and prism aberrations, and the ultimate spectral selectivity is limited by a high background light level. Both of these problems are caused by sinc function sidelobes due to a finite interaction length and the random phase matching of all the light and acoustic energy bouncing around inside the crystal. Putting two in a row is an effective but pricey solution to the ghost problem, and putting the AOTF at an image instead of a pupil pretty well solves the aberration problem too.[†]

7.10.5 AO Deflectors

The same crystal-cutting tricks allow a reasonable range of output angles (± 2–3°) for a 1-octave frequency shift, making a narrow-range but very fast scanner with high

[†] Dennis R. Suhre et al., Telecentric confocal optics for aberration correction of acousto-optical tunable filters. *Appl. Optics* **43**(6), 1255–1260 (February 20, 2004).

diffraction efficiency, the acousto-optic deflector (AOD). In this case, we want θ_i to be constant over a wide range of acoustic frequency, a condition called *tangential phase matching* that can be met in TeO_2.

There are interesting tricks you can do with AODs. If you gently focus a beam at the center of an AOD, it will pivot about its focus with f. Recollimating produces a beam that moves from side to side without steering, which is very useful for long path sensors such as the extinction system of Section 10.8.3. Galvos work for this too, of course, but since an AOD has no moving parts, you can get a much higher scan frequency, with zero jitter and wobble.

Getting high resolution out of an AOD requires lots of fringes, just like any other Bragg grating; the number of resolvable spots is equal to the transit time–bandwidth product. Like Rayleigh and Sparrow resolution criteria, there's a factor of order 1 in front that we can argue about, depending on your beam profile and how much overlap you allow between distinct spots.

Aside: Acoustic Phase Delay. Bragg cells are often used in heterodyne interferometers, and it is sometimes important to remember that the acoustic propagation delay translates into a huge phase shift. This acoustic phase shift gets impressed on the optical phase, and so appears in the phase data. It is often far from negligible; if you're using two passes through an 80 MHz cell that 200λ delay has a phase sensitivity of 31 rad/MHz. This is a nuisance in wide-range AOD measurements, or where it makes the oscillator spurs show up in the data, but in fixed-frequency applications it can be useful—you can stabilize the operating point of your system by using feedback to adjust the acoustic frequency. Shear-wave TeO_2 devices are the best overall in the visible. Optics people used to birefringent materials with $(\delta n)/n$ of a percent or less are usually surprised that the slow shear wave in TeO_2 goes at 600 m/s while the longitudinal wave goes at 4200. A really big TeO_2 cell can have 1000 resolvable spots in a single pass, though several hundred is more typical.

While they're the best of the fast bulk-optics modulators, AO cells have some major drawbacks. Cells with small Bragg angles (longitudinal devices at lowish frequency) have severe etalon fringes. AODs are less prone to these, because of the high angles, high diffraction efficiency, and polarization shift. There is also usually some beam apodization due to acoustic nonuniformity, and ripples in the diffraction efficiency in space and frequency due to acoustic standing waves. The standing wave effect is eliminated in shear wave devices by cutting the bottom of the cell at 5°; because of the huge $\Delta\mathbf{v}$, this totally destroys the phase matching between the reflected acoustic wave and the light.

Some people say that AO cells have poor wavefront fidelity, but the author has never had a moment's problem with it. Scanning rapidly does introduce aberrations however. It takes some time for the wave to cross the beam, so a frequency ramp produces astigmatism by sending rays at different x in different directions; a frequency staircase with time allowed for settling avoids this problem. The polarization eigenstates of a shear wave cell are also very slightly elliptical, which one occasionally needs to remember. Overall, a slow shear wave AOD is a pretty trouble-free device.

7.10.6 Photoelastic Modulators

Besides a change in refractive index, the acousto-optic effect also induces stress birefringence in the crystal. Normally we don't worry too much about this, especially with TeO_2

devices, where the incoming beam has to be polarized in the right direction to get good performance. Photoelastic modulators are acousto-optic devices that exploit this effect. The basic idea is to run an AO cell at a very low frequency, 20–100 kHz, so that the acoustic transducer basically just shakes the entire crystal back and forth, and tune the frequency to the lowest order longitudinal vibration mode of the crystal—just like an organ pipe. The acousto-optic effect leads to a more or less uniform phase modulation, but the stress birefringence (the photoelastic effect) basically turns the crystal into an oscillating wave plate, whose retardation can reach $\pm\frac{1}{2}$ wave. Photoelastic modulators thus act a bit like acoustic Pockels cells, only much slower. Their big advantage is greater uniformity of the birefringence across their field.

7.10.7 Acousto-optic Laser Isolators

The acousto-optic effect is symmetrical, so reflected light propagating backwards along the beam will be diffracted back into the laser. The returned first-order beam enters the cell at $f \pm f_A$ from its first pass, and winds up at $f \pm 2f_A$, because the sign of the optical **k** vector has changed. The frequency shift is so small that the difference in deflection angle is negligible, and the light goes straight back into the laser.

The laser is a Fabry–Perot resonator, though, and provided f_{acoustic} has been properly chosen, and the cavity finesse is high (as in gas lasers), virtually none of that light will make it back into the laser cavity to cause feedback problems. Even with diode lasers, where the finesse is low, and a lot of the light does make it into the cavity, the beat frequency $2f_A$ is so much higher than the $\sim$100 kHz of mode hops that its effect is much reduced. (Why doesn't this work with the zero-order beam?)