



Fresnel Reflection, Lenserf Reflection, and Evanescent Gain

Anthony Siegman

Recent investigations into Fresnel reflection and total internal reflection have led to conflicting results when the evanescent medium is “gainy” rather than lossless or lossy. This controversy can be resolved by introducing the much less well known concept of Lenserf reflection—an idea put forward by the illustrious Dr. Tung Inn Cheek.

The total internal reflection (TIR) of light at certain angles of incidence on the interface between two lossless dielectric media is a well-known aspect of Fresnel reflection. It is also well understood that adding a small amount of loss or absorption in the evanescent region inside the reflecting medium will reduce the magnitude of this TIR reflection coefficient—an effect that provides the basis for so-called evanescent wave spectroscopy.

This situation becomes more controversial with laser gain. The question naturally arises: If the Fresnel reflection from an absorbing or lossy half space under TIR conditions has a magnitude less than unity, shouldn't Fresnel reflection from an

amplifying or gainy half space under the same TIR conditions have a magnitude greater than unity? Can one observe amplified total reflection, or “TIR with gain,” from a gainy reflecting medium?

Since the advent of the laser, at least a dozen refereed publications have claimed that amplified total reflection from a single gainy interface is real. This effect was patented in the 1970s, and two detailed calculations defending the phenomenon have been published recently in *Optics Express*. Thus, the primary objective of this article is to reexamine the basic physics of Fresnel reflection from an unbounded gainy medium, especially in the regime of TIR.



Augustin Fresnel observes his little known contemporary, Monsieur Nitsugua Lenserf (facing page).

A secondary goal is to enlighten the optics community about an important but often ignored family of alternative solutions to the Fresnel reflection equations that we might designate as “Lenserf reflection.” Understanding the distinctions between Fresnel and Lenserf reflection turns out to be helpful in understanding the properties (and even the reality) of amplified TIR and of so-called “evanescent gain” in optical waveguides.

Before we dive into our reexamination, however, readers might note that articles in OPN, although meant to present serious technical content, may on occasion present claims that are dubious, whimsical or even incorrect. Savvy OSA members will, I trust, be able to identify any such statements that follow.

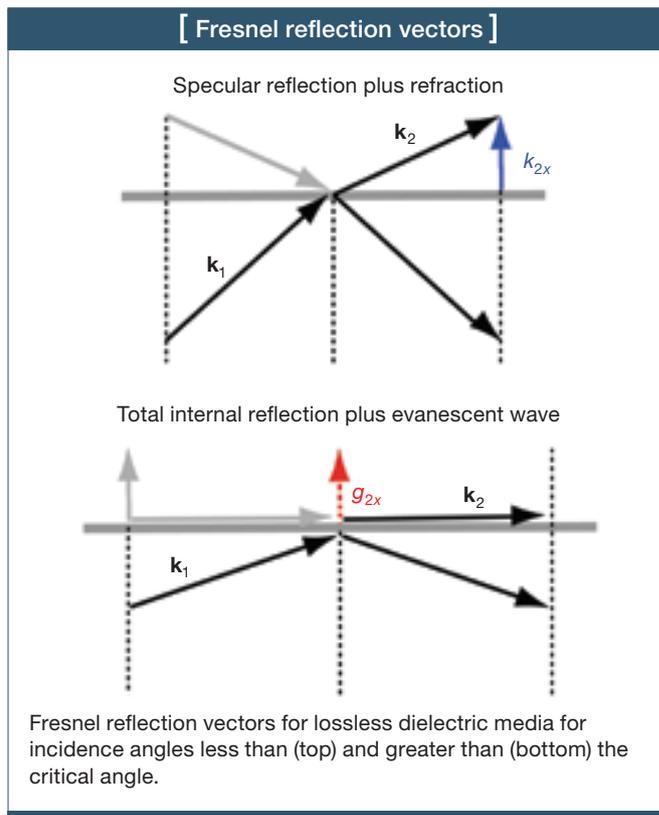
Lossless Fresnel reflection

To set the stage for this reexamination, let’s review the classic analysis of Fresnel reflection for an infinite plane wave incident from a higher-index medium onto a planar interface looking into an unbounded, lower-index reflecting medium, with both media assumed to be lossless. We take as our physical model a horizontal interface between two unbounded or infinite half spaces, as shown in the figure on the right, with the lower half space having the higher index of refraction. We use subscripts 1 and 2 to distinguish between the higher- and lower-index regions.

The solid black arrows indicate the real-valued phase vectors or “ \mathbf{k} ” vectors of the infinite plane waves that form the basis for the Fresnel analysis. These same black arrows may also, in real experiments, represent finite-width optical beams made up of superpositions of such infinite plane waves with their \mathbf{k} vectors distributed over a small angular spread. For simplicity, we limit our discussion to TE polarized waves only.

The upper sketch shows that a wave incident on the interface from below at an angle of incidence less than a certain critical angle will produce a specularly reflected wave that travels back down into the lower half space, plus a refracted wave that continues upward into the upper half space, with its \mathbf{k}_2 vector tilted at a larger angle from the normal than the incident \mathbf{k}_1 vector in accordance with Snell’s law. The vertical blue arrow above the interface indicates the x -directed component of the refracted \mathbf{k}_2 vector in this region.

As the angular tilt of the incident beam approaches the critical angle (which is taken to be 45 degrees for all illustrations in this article), the \mathbf{k} vector for the refracted wave tilts over more and more sharply, until the refracted \mathbf{k}_2 vector becomes exactly parallel to the interface when the incident \mathbf{k}_1 vector reaches the critical angle. At all angles of incidence beyond this, the \mathbf{k}_2 vector of the refracted wave remains exactly parallel to the interface, and the refracted wave is converted into a more unusual type of wave, commonly called an evanescent wave, which propagates exactly parallel to the interface rather than outward into the upper region; as a result, it carries no power further upward into the upper half space.



This regime is commonly called the “TIR regime,” since all the power or energy in the incident wave is now totally reflected back into the specularly reflected wave in the lower region. Without getting entangled in detailed mathematics, we can note that, in this TIR regime, the evanescent refracted wave becomes what is commonly called an inhomogeneous plane wave—that is, a wave having the general form $\epsilon_2(x_2, z_2) = \exp[-j\mathbf{k}_2 \cdot \mathbf{r}_2 + \mathbf{g}_2 \cdot \mathbf{r}_2]$, where \mathbf{r}_2 indicates the coordinates (x_2, z_2) perpendicular and parallel to the interface.

In fact, this evanescent wave has a phase vector \mathbf{k}_2 with vector components $\mathbf{k}_2 = (0, k_{2z})$, together with a “growth vector” \mathbf{g}_2 with vector components $\mathbf{g}_2 = (g_{2x}, 0)$. This vector has a negative value $g_{2x} < 0$, so that the amplitude of this evanescent wave decays exponentially (and generally quite rapidly) with distance in the direction normal to the interface. The red arrow in the lower sketch above is therefore shown as a dashed line to signify that it has a negative value with respect to the direction in which it points. The significance of the two additional shaded or grayed-out arrows that also appear in the upper left quadrants of these two sketches will emerge later.

Lossless Fresnel reflection coefficient

Satisfying the boundary conditions for the E and H fields at the interface then provides a recipe for finding the \mathbf{k}_2 and \mathbf{g}_2 vectors as a function of the incident \mathbf{k}_1 vector and its angle of incidence θ . Using this recipe, we can plot the values of $k_{2x}(\theta)$ and $g_{2x}(\theta)$ versus incidence angle θ for the lossless Fresnel case.

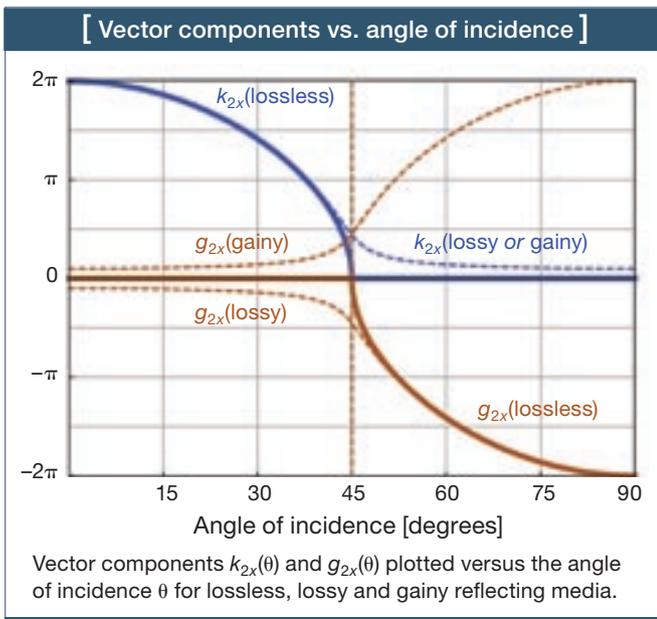
In the gainy case, the refracted traveling waves are amplified rather than absorbed, and thus they always increase in amplitude with distance away from the interface for both below-critical and above-critical angles of incidence.

For a lossless reflecting medium and for incidence angles $\theta < \theta_{\text{crit}}$, the “growth vector” g_{2x} is identically 0 and the phase vector \vec{k}_2 is purely real, corresponding to a refracted uniform plane wave that propagates upward and outward into the reflecting media, as described above.

For incident angles in the evanescent or TIR region $\theta > \theta_{\text{crit}}$, however, this behavior changes discontinuously to give $k_{2x} \equiv 0$, while the vertical component g_{2x} of the growth vector becomes real and negative. The evanescent fields in this regime thus propagate in the z direction but with an amplitude profile that decays exponentially perpendicular to the interface.

The final step in this classic derivation is to insert these results for $k_{2x}(\theta)$ and $g_{2x}(\theta)$ into a textbook formula for the Fresnel reflection coefficient $\hat{\rho}(\theta)$ — that is, the complex-valued ratio of reflected wave amplitude to incident wave amplitude just below the interface. A particularly instructive (but not particularly common) way of displaying this Fresnel reflection coefficient is to plot it as a contour in the complex plane, as a polar plot of Fresnel reflectivity $\hat{\rho}(\theta)$ for increasing values of θ . Such a plot is shown by the solid blue contour in the polar plot at the bottom right.

Displaying $\hat{\rho}(\theta)$ in this fashion makes clear that, for values of $\theta < \theta_{\text{crit}}$, the Fresnel reflection coefficient starts off at a small positive real value at normal incidence ($\theta = 0$) and increases toward unity magnitude, remaining positive and real as the angle of incidence approaches θ_{crit} . As the angle of incidence passes through the critical value, the reflection coefficient $\hat{\rho}$

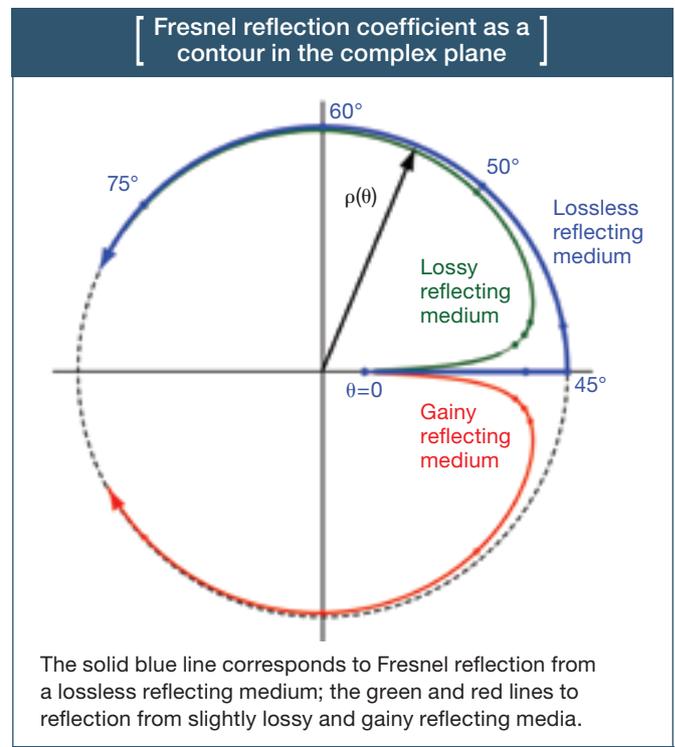


makes a discontinuous left turn and rotates around the upper half of the unit circle as the incidence angle increases further toward grazing incidence. The reflection coefficient therefore remains exactly unity, corresponding to TIR across this full range of incidence angles.

Fresnel reflection from lossy or gainy media

The textbook results for Fresnel reflection change in small but very significant ways if even small amounts of loss or gain are added in the reflecting half space. If one assumes that this gain or loss is uniformly distributed in the upper half space, one can evaluate these changes using the same physical model and repeating the same calculations, except that the index of refraction in the reflecting half space is now taken to have a complex value $\tilde{n}_2 \equiv n_{2r} + jn_{2i}$, with n_{2r} representing the pure real index in the lossless case and n_{2i} being a generally much smaller imaginary part that corresponds to a small absorption loss if n_{2i} is negative and to a correspondingly small linear gain coefficient if n_{2i} is positive.

If one carries through the same analysis as in the lossless case, taking care to keep proper track of the real and imaginary parts of all the now complex-valued physical variables and equations, one finds first of all that the $k_{2x}(\theta)$ and $g_{2x}(\theta)$ vector



components plotted earlier are modified as illustrated by the dashed curves. (The curves correspond to one specific small numerical value of gain or loss.) In particular, one sees that $k_{2x}(\theta)$ and $g_{2x}(\theta)$ now both become finite-valued and smoothly varying across all values of the incidence angle θ , with k_{2x} taking on the same positive value for equal values of either loss or gain, while g_{2x} takes on positive and negative signs for the gainy and lossy cases, respectively.

Based on the modified results for $k_{2x}(\theta)$ and $g_{2x}(\theta)$, the heavy blue contour in the polar plot of $\hat{\rho}(\theta)$ is then modified into the lighter green and red curves, corresponding to the lossy and gainy cases, respectively. A number of important results are immediately evident. For example, $\hat{\rho}(\theta)$ is no longer purely real for any angle of incidence. Instead, $\hat{\rho}(\theta)$ for θ close to zero shifts slightly in phase in equal and opposite directions for the lossy and gainy cases—exactly the result that one might expect for reflection at normal or near-normal incidence from a medium with a complex characteristic impedance.

As the incidence angle becomes larger, $\hat{\rho}(\theta)$ increases in magnitude in each case, while developing larger but still equal and opposite phase shifts for equal magnitudes of the loss and gain coefficients. As the incidence angle passes through what would have been the critical angle for the lossless case, the contours for the lossy and gainy reflection coefficients make rapid but smooth changes in direction, and then progress in opposite directions around the unit circle, except this time traveling slightly inside the unit circle. The primary result is therefore that $\hat{\rho}(\theta)$ becomes less than unity for both lossy and gainy cases at all angles of incidence, including angles within what was previously the TIR regime.

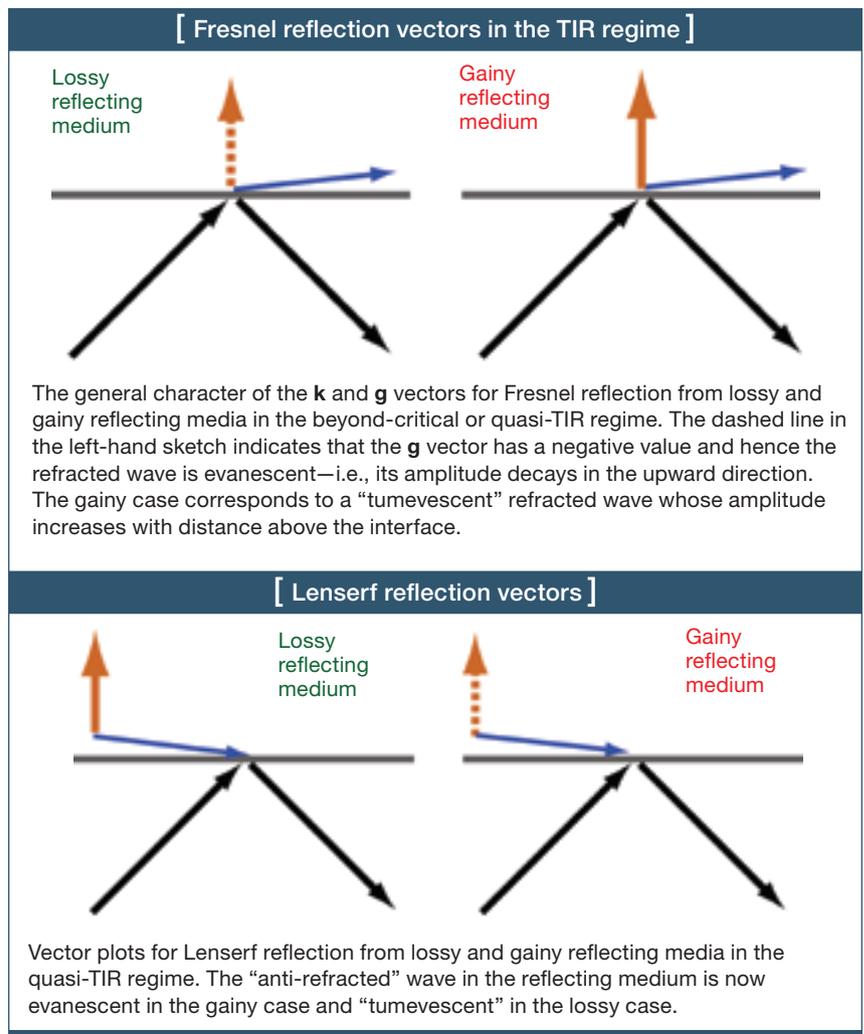
Notably, as soon as one adds any finite value of loss or gain in the reflecting half space, there is no longer any discrete critical angle, although the trajectory of $\hat{\rho}(\theta)$ does make a fairly high-speed turn as θ passes through what would have been the critical angle for the lossless case. In addition, the magnitude of the reflection coefficient always remains less than unity for any angle of incidence and for any value of loss or gain. Readers interested in the Goos-Hänchen shift will note that this shift now evidently takes on equal but opposite values for the lossy and gainy cases.

Further insight can be obtained by making plots of the \mathbf{k} and \mathbf{g} vectors for the incident, reflected and refracted waves with the modifications due to loss or gain taken into account. Two examples are shown in the top figure on the right, both corresponding to an incidence angle within the TIR regime but

assuming first a lossy and then a gainy reflecting medium. An essential feature of these sketches is that the \mathbf{k}_2 vectors of the “evanescent” waves even in the TIR regime now both remain tilted slightly upward at a small but finite angle away from the surface instead of becoming exactly parallel to the interface.

Note that the Poynting vector or power flow for a single inhomogeneous plane wave always points exactly along the \mathbf{k} vector of that plane wave, independent of the direction of the \mathbf{g} vector. Net power, therefore, evidently always flows upward, away from the interface and outward into the reflecting region, for both the lossy and gainy cases shown, although the tilt angles away from exact parallelism with the interface typically remain extremely small in the above-critical regime for any reasonable values of loss or gain.

Thus, in the lossy case, the amplitudes of these upward and outward refracted traveling waves always decrease with distance as they travel upward away from the interface, exactly as one might expect, in both the below-critical and the above-critical or quasi-TIR cases. The refracted fields for the lossy case can therefore be characterized as always decaying or evanescent in character for all angles of incidence—if one interprets the term



Even in the TIR region, the incident beam launches a quasi-refracted beam that grows indefinitely with distance as it travels at a finite (even if very small) angle away from the illuminated spot.

evanescent in the traditional fashion as referring to fields that decrease in amplitude away from the interface.

By contrast, in the gainy case, the refracted traveling waves are amplified rather than absorbed, and thus they always increase in amplitude with distance away from the interface for both below-critical and above-critical angles of incidence. These solutions are therefore the opposite of evanescent. Since there is no antonym to the term evanescent other than the awkward term “anti-evanescent,” we propose the term “tumevescent” to describe these upwardly growing fields in the gainy case. (The etymology of this term is left as an exercise for the reader.)

For both the lossy and gainy cases, the boundary condition at infinity (i.e., far above the interface) is not that the amplitude of the fields should approach zero at large distances. Rather, it is that the power flow should be outwardly directed everywhere above the interface. It is this condition—not a requirement that the field amplitudes approach zero at large distances, as stated in many optics texts—that makes the results presented thus far the physically valid solutions for Fresnel reflection from either lossy or gainy media.

We might further consider the behavior of each of these cases when the incident signal is not a single uniform and infinite plane wave but rather a beam that is well-collimated but finite in width. This finite-width beam can then be viewed as the coherent superposition of a continuous and fairly narrow distribution of infinite plane waves that are added together to make up this finite-width incident beam. This narrow distribution of incident plane waves will then create a similarly narrow distribution or coherent superposition of inhomogeneous plane waves in the region above the interface.

Suppose in particular that this incident beam—or more precisely the spread of \mathbf{k}_1 vectors that make up this incident beam—are predominantly limited to incident angles within the TIR or evanescent regions described earlier. In the lossless case, the resulting distribution of inhomogeneous plane waves will add together in such a way as to create a significant amount of evanescent energy only in a thin layer, having a thickness determined by the magnitude of the g_{2x} vector that is located immediately above the illuminated spot where the incident beam intersects the interface.

The same result will generally hold true in the lossy case: In the TIR regime, the incident beam will launch a small amount of energy into the distribution of quasi-evanescent (but lossy) waves propagating away at a very small angle above the interface. More detailed examination shows, however, that these waves are attenuated sufficiently rapidly as they travel away from the illuminated spot so that the evanescent fields remain

essentially confined to a very thin region immediately above the illuminated spot created by the incident beam on the interface.

This situation becomes very different for the gainy reflection case, however. In both the normally refracted region for incidence angles less than critical and the quasi-refracted or TIR regions, the inhomogeneous waves are amplified rather than attenuated as they travel from the illuminated spot. The net result is that, even in the TIR region, the incident beam launches a quasi-refracted beam that grows indefinitely with distance as it travels at a finite (even if very small) angle away from the illuminated spot. Understanding the effects of this outwardly growing beam is essential to gaining a full understanding of the Fresnel reflection behavior looking into a gainy medium.

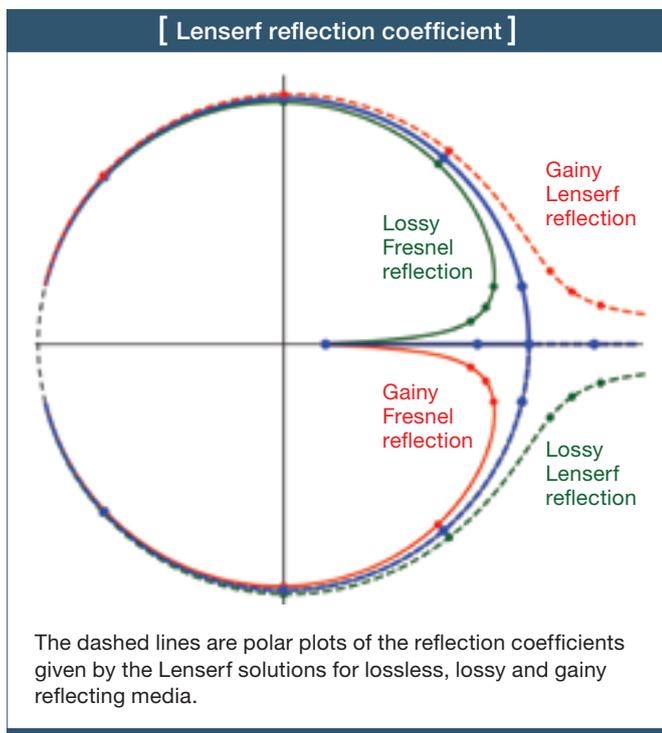
Lenserf reflection

Optics students who have trudged through the mathematical details of this Fresnel reflectivity derivation will recall that one of the crucial steps in the derivation involves taking the square root of a certain complex-valued expression, which contains the complex refractive indices of the incident and reflecting media together with the \mathbf{k}_1 vector (and potentially a \mathbf{g}_1 vector) of the incident beam, and then making a particular choice of sign for this square root to obtain all the results presented thus far.

This mathematical formulation also allows, however, a second solution to the same physical problem, simply by taking the opposite choice of sign for this square root; and this alternative would seem to be an equally valid solution, at least mathematically and perhaps also physically. This alternative in fact proves to be of genuine utility and physical significance—for example in understanding “layered” Fresnel reflectivity and in explaining and understanding the properties of dielectric waveguides having lossy or gainy cladding layers, especially when questions are raised as to the existence and interpretation of so-called evanescent gain.

A distinctive name for this alternative solution would be useful in discussing problems like these, and fortunately a very suitable one has recently emerged from historical studies carried out by my collaborator Dr. Tung Inn Cheek from the Kim Jong Il Institute for Scientific Truth in Pyongyang, North Korea. Dr. Cheek has proposed the name “Lenserf reflection” to honor a little known contemporary of Augustin Fresnel named Monsieur Nitsugua Lenserf, since it is surmised (albeit on very limited evidence) that Lenserf may have put forward this alternative solution to Fresnel’s basic equations.

Implementing the Lenserf solution in fact involves nothing more than reversing the signs of the $k_x(\theta)$ and $g_x(\theta)$ curves plotted earlier, and then examining the consequences. The



modified incident, reflected and refracted vector diagrams, and the modified polar plot of the Lensef reflection coefficient $\tilde{\rho}_L(\theta)$ that result from this sign change are illustrated in two additional sketches that are evidently in some ways very similar but in others quite different from the previous sketches for the Fresnel solution $\tilde{\rho}_F(\theta)$. I will not walk through the detailed characteristics of these alternative or Lensef results step by step, since their significance is fairly obvious from the illustrations. However, interested readers may want to work out the implications for themselves.

A source of concern may be that these Lensef solutions, although they satisfy all the physical and mathematical constraints of the Fresnel problem, seem to be nonphysical, at least in part. In particular they always seem to involve an “anti-refracted” wave that carries energy down from above within the reflecting medium toward the interface. These anti-refracted or Lensef waves for the lossless case directly correspond, in fact, to the grayed-out vectors in the opening illustration.

One response to this concern could be that the amplitude of the Lensef solution at least for the gainy case dies exponentially at large distances from the interface—a criterion that is commonly taken as a guarantee of physical reality in many textbook discussions of ordinary Fresnel reflection. A more dubious explanation could be that the carefully and coherently matched downward beams required by these solutions might be kindly generated by the ghosts of Lensef or Fresnel, or Cheek—or perhaps distinguished past OSA presidents. However this happens, one can discover that, just as the trapping of waves in lossless dielectric waveguides has a direct connection to Fresnel reflection, the propagation properties of optical

waveguides and fibers having gainy cladding regions can be described using Lensef reflection solutions.

Note that the Lensef solutions for any given case correspond exactly to the time-reversed solutions for a Fresnel solution, with the waves traveling in the reverse direction along the z axis and with gain changed to loss or loss changed to gain everywhere in the upper half space. Both solutions are thus at a minimum equally good solutions to the fundamental statement of the problem.

Conclusions and implications

Dr. Cheek and I find ourselves in full agreement on a few fundamental conclusions, the most basic of which is that single-surface “amplified TIR” really does not and cannot exist—at least, not from an unbounded and uniformly gainy half space. Whether one analyzes this problem using infinite plane waves or directs an actual finite-width beam at the interface looking into an unbounded, lower-index and uniformly gainy (or lossy) half space, the interface itself can only act (as it always does) as a lossless four-port beam splitter or unitary scattering element.

This interface, for any angle of incidence, simply splits the power in the incoming signal into a reflected beam or wave with a reflection coefficient of slightly less than unity, plus a refracted (or, if you like, quasi-refracted) beam or wave, typically of small initial amplitude, which is launched into the gainy half space and travels outward from the interface. This refracted signal then grows in amplitude as it travels outward into the gainy medium, whether it is traveling at a large angle in the below-critical-angle or Snell’s law regime, or at a small, near-grazing, but still finite angle within the TIR regime.

One must immediately add that Fresnel reflection from a layered or tapered gainy reflecting medium—that is, a reflecting medium with tapered or finite planar layers of gain at or just above the interface—is very different from the unbounded half-space case. The Fresnel and Lensef results for the unbounded case provide the tools to analyze the layered or tapered scenario. However, the reverse is not true, since the results obtained from detailed analysis of the layered case do not converge to the unbounded half space results described here, even in the limit of an infinitely thick gain layer. Fresnel reflection from a tapered or finite gain layer located just above an interface will in fact always have a reflection magnitude greater than unity throughout the entire TIR regime (except that this situation will not be simply Fresnel reflection, but rather a combination of Fresnel plus Lensef reflection).

An experimental test of our basic claim would obviously be desirable, and it is likely possible, if a bit tricky. A reasonably thick reflecting region that is switchable from lossy to gainy might be fabricated by placing a thick slab or prism of lower-index laser-doped and optically pumped laser glass or crystal in optical contact with a higher-index (and preferably low-loss) prism, or by placing an optically pumped gaseous gain medium in contact with a dielectric medium of any index value. As to measurement methods, the analysis in this article predicts

Dr. Cheek and I find ourselves in full agreement on a few fundamental conclusions, the most basic of which is that single-surface “amplified TIR” really does not and cannot exist—at least, not from an unbounded and uniformly gainy half space.

only a small change in reflection magnitude but a large change in reflection phase angle as the reflecting medium is switched from lossy to gainy. Rather than attempting to observe the small change in reflection magnitude caused by a lossy-gainy transition, one might instead use an appropriate interferometric setup to look for this phase change.

Given the subtlety of either of these effects, one would need to ensure that the switching process did not create any additional discrete index layers or nonlinear index changes at the interface or in contaminant or coating layers on the interface. One must also avoid layered reflectivity effects and address (or perhaps look for) the low-angle refracted beam that will be generated by the incident beam when it intersects a gainy medium. This small-angle refracted beam can then rattle around inside a finite gain region, being multiply reflected and repeatedly amplified. A picosecond mode-locked incident signal might provide an effective way to test for such effects.

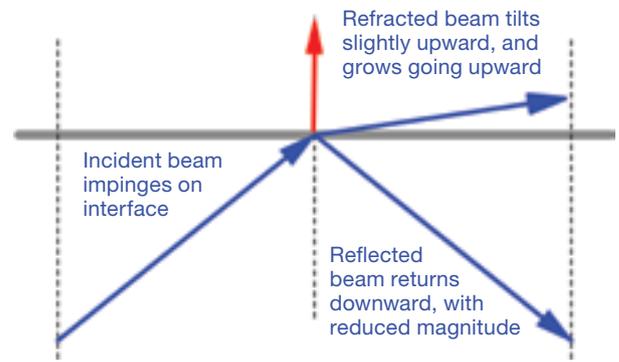
Finally, regarding evanescent waveguide gain, we can recall that TIR is the mechanism that traps radiation inside all standard types of dielectric waveguides and optical fibers. Laser atoms placed within the evanescent fields in the cladding regions of such waveguides will certainly produce stimulated emission, and so “evanescent gain” produced by gainy cladding material surrounding such waveguide structures might seem guaranteed to lead to amplifying (i.e., axially growing) modes.

Here again, however, the distinction between unbounded cladding layers and bounded or layered or tapered gainy regions is crucial. The conclusion that there is no amplified TIR from an unbounded but gainy half space would certainly seem to imply that there are no evanescently amplified modes for an optical waveguide immersed in an unbounded and gainy cladding; and, although some complexities remain to be resolved, this does in fact seem to be a valid conclusion.

Finite or tapered gainy cladding layers can, on the other hand, produce amplified Fresnel reflection, and hence can lead to amplifying eigenmodes in optical waveguides. The resulting behavior has a number of complexities in which both Fresnel and Lensef must become involved. One must also recognize that adding finite or tapered gainy layers to an existing and otherwise lossless waveguide will significantly modify the mode profiles in such a way that the fields in the gainy layers are no longer purely, or even primarily, evanescent. The resulting amplification should not really be characterized as evanescent gain but rather as a multi-layer waveguide with the finite gain layers viewed as really part of the waveguide core rather than its cladding.

More rigorous and general derivations of these assertions concerning waveguides with finite or tapered gainy cladding

[How amplified TIR really works]



An illustration showing why amplified total reflection looking into an unbounded uniformly gainy reflecting medium does not exist.

layers have yet to be obtained—but then, one might also point out that rigorous analytic solutions demonstrating stable and amplifying eigenmodes in any dielectric waveguides with unbounded gainy cladding regions are equally lacking. Further clarifications of these points—and, of course, the implementation of experimental tests—would thus be useful contributions for further research. ▲

The author would like to acknowledge the contributions made by Dr. Cheek, although Cheek himself has not consented to be added as a co-author. Given our geographical locations, I could not readily visit his institute nor be mine, and interaction via the Internet would have been difficult. During the many hours that I've lain awake worrying about the amplified TIR problem, however, Dr. Cheek was apparently at his desk in Pyongyang; we were thus able to establish a psychic communication channel. Our collaboration via this medium has greatly aided in the work reported here.



Anthony Siegman (siegman@stanford.edu) is the McMurtry Professor of Engineering Emeritus at Stanford University in Stanford, Calif., U.S.A.

[References and Resources]

- >> C.J. Koester. “Laser action by enhanced total internal reflection,” IEEE J. Quantum Electron. **QE-2**, 580-4 (1966).
- >> J. Fan et al. “Amplified total internal reflection,” Opt. Express **11**, 299-308 (2003).
- >> K.J. Willis et al. “Amplified total internal reflection: theory, analysis, and demonstration of existence via FDTD,” Opt. Express **16**, 1903-14 (2008).
- >> Additional information related to this article may be posted from time to time at www.stanford.edu/~siegman/.