On the problem of single-surface amplified total internal reflection

Tobias S. Mansuripur

Department of Physics, Harvard University, Cambridge, MA 02139[*]

Masud Mansuripur

College of Optical Sciences, The University of Arizona, Tucson, AZ 85721

Abstract

We theoretically investigate the claim that the total internal reflection (TIR) of light from a transparent medium onto a lower index, semi-infinite gain medium has a reflection amplitude greater than unity. By analyzing the cases of a finite thickness gain medium, we demonstrate the importance of the 'round-trip coefficient'–the factor by which the amplitude of a plane wave is multiplied upon traveling one round-trip in the gainy slab–on the qualitative behavior of reflection, both below and above the critical angle. Analytical simulations of a Gaussian beam show that the 'side-tail' of the beam enters the slab before the central part of the beam; when the magnitude of the round-trip coefficient is greater than one, this weak pre-excitation of the amplifying medium gains more during propagation than it loses to transmission at the back facet, and returns to interfere with the central part of the beam at the first interface, giving rise to an amplified reflection of the primary beam. The angle at which the round-trip coefficient exceeds one can be smaller than the critical angle, and appears to be more relevant than the critical angle in determining whether the primary reflection is amplified. Because the back facet is instrumental in generating the amplified reflection, we argue against the existence of amplified TIR from a semi-infinite medium.

[*] mansuripur@physics.harvard.edu

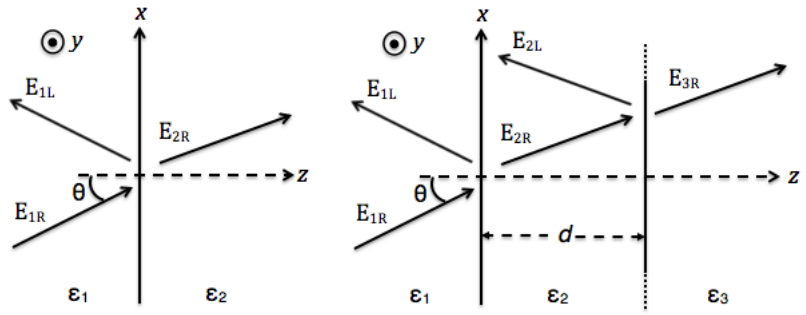
I. INTRODUCTION

Total internal reflection (TIR) is a well-known phenomenon in optics. Light incident from one transparent medium with refractive index n_1 onto a lower-index transparent medium with index n_2 at beyond the critical angle $\theta_c = \arcsin(n_2/n_1)$ will give rise to a reflected wave with the same amplitude as the incident wave. In the second medium there is no propagating wave, only an exponentially decaying 'evanescent' field which carries no energy into the second medium. When the second medium has loss or gain, the transmitted wave has both evanescent and propagating character, regardless of the angle of incidence, so there is always a non-zero time-averaged component of the Poynting vector in the second medium along the direction perpendicular to the interface. The field amplitudes can be calculated by assuming an incident, reflected, and transmitted (inhomogeneous) plane wave, and using the boundary conditions imposed by Maxwell's equations to equate on both sides of the interface the components of the E and H-fields parallel to the interface. However, things are not quite so simple, since Maxwell's equations admit two possible plane wave solutions in medium two with opposite signs for the z-component of the k-vector: for a gainy (lossy) medium, one solution whose amplitude grows (decays) while carrying energy away from the interface and another whose amplitude decays (grows) along z while carrying energy towards the interface. Only one of these must be chosen in order to unambiguously match the boundary conditions. In the case where the second medium is absorptive, everyone agrees that the amplitude of the transmitted wave must decay as it propagates away from the interface. This choice gives rise to a reflection coefficient whose magnitude is always below unity, which is well substantiated by experiment and not under dispute.

When the second medium has gain, however, it is not possible to rule out the exponentially growing wave by invoking energy conservation, as is done in the absorptive case. In fact, for angles of incidence below the critical angle, everyone agrees that the transmitted wave in the second medium must carry energy away from the interface, and the field amplitude grows exponentially ad infinitum. Above the critical angle, however, it has been proposed that the wave in the gain medium propagates toward the interface [1–3], carrying energy from medium two to medium one. Historically, this choice was motivated by an experiment which demonstrated that the light propagating in a fiber with a high-index passive core and low-index gainy cladding was amplified [1]. Since the only way to achieve a

Fresnel reflection coefficient with amplitude greater than unity is to choose the 'backwards' traveling wave in the second medium, this experiment appeared to justify the otherwise ad hoc decision to use one wave below the critical angle, and another wave above the critical angle, in deriving the Fresnel reflection coefficient from a single interface. This choice leads to so-called single-surface amplified TIR (ss-ATIR), a phenomenon which appears to nicely explain the experiments as well as serves as a counterpoint to attenuated-TIR from a lossy medium. However, in any experiment, the gainy cladding layer is always of finite thickness, surrounded by a third medium (typically air). When this third medium is taken into account, the reflection of the forward propagating wave in the gainy cladding off the back interface generates the backward traveling wave, and amplified reflection in medium one is possible under certain conditions (both below and above the critical angle). Thus, the debate over ss-ATIR is about one question: is the presence of the second interface necessary to generate the amplified reflection or not?

In this paper we begin in Sec. II by reviewing the case of two semi-infinite media and the arguments which have been made in favor of and against each choice for the wave in medium two. In Sec. III, we examine the case of a finite thickness gainy slab. By looking at analytical simulations of a Gaussian beam and pulse incident on the slab, we demonstrate novel phenomena which likely could not have been predicted by examining the individual plane wave solutions to the problem. To explain these results, we must first define a few quantities. The *round-trip coefficient* is the factor by which the amplitude of a plane wave is multiplied upon traveling one round-trip in the gainy slab; it is a well-defined quantity for both homogeneous and inhomogeneous plane waves. The round-trip coefficient is a function of the angle of incidence; when gain is present, the magnitude of the round-trip coefficient can exceed one at an angle–referred to as the *round-trip angle*–which is smaller than the critical angle. The *primary reflected beam* refers to the reflected beam in the incidence medium which originates from the point on the interface at which the incident beam strikes, as opposed to the beams which arise due to multiple reflections in the slab. The central result of our paper is this: the primary reflected beam can experience gain only when the incidence angle exceeds the round-trip angle. We show simulations which suggest a simple mechanism for this behavior: the Gaussian beam has a 'side-tail' that enters the slab before the main portion of the beam. When the round-trip coefficient exceeds one, the side-tail gains more energy during propagation than it loses to transmission at the back interface,



(a) (b)

FIG. 1. Geometry of the system under consideration, and the various k-vectors that need to be considered for the case of (a) two semi-infinite media and (b) a finite slab of thickness d. The values of ϵ_1 and ϵ_3 are real and positive, while $\epsilon_2 = \epsilon_2' + i\epsilon_2''$ has a positive real part and negative (gainy) or positive (lossy) imaginary part.

and therefore returns to the first interface where it contributes to the amplification of the primary reflected beam. For suitable parameters, this process occurs for incidence angles which are less than the critical angle. We argue that the round-trip angle is therefore more important than the critical angle in determining whether the primary reflected beam will be amplified. Our simulations show that this amplification mechanism relies on the presence of a second interface to reflect the side-tail back towards the first interface. Since no such mechanism is possible for the case of two semi-infinite media, we believe that single surface amplified total reflection is unlikely.

II. TWO SEMI-INFINITE MEDIA

We will describe the electromagnetic field using complex phasor notation, with the convention that the time dependence for a plane wave is given by $\exp(-i\omega t)$, where ω is the angular temporal frequency of the wave. The geometry is shown in Fig. 1(a). The transparent incidence medium fills the left half-space $z < 0$ and is described by the real-valued and positive relative permittivity ϵ_1. The second medium exists in the right half-space $z > 0$ and has relative permittivity $\epsilon_2 = \epsilon_2' + i\epsilon_2''$, where ϵ_2' is real-valued and positive, while ϵ_2'' is real-valued and can be positive or negative. It can be deduced from the standard Lorentz oscillator model that the material is lossy for $\epsilon_2'' > 0$ and gainy for $\epsilon_2'' < 0$. (Although not relevant to the present discussion, we note that the Lorentz oscillator model does not yield

the correct dispersion relation for gain media; a quantum-mechanical analysis of the gain medium is needed in order to arrive at a realistic frequency dependence for the medium's permittivity; namely, one that has a negative oscillator strength rather than a negative damping coefficient [5]. Also, modeling the gain by using a permittivity with negative imaginary part is valid only for small signals. For larger signals, gain saturation becomes a problem and must be treated in a different way. We assume that the small signal limit is always valid, which is easily achieved in theory by arbitrarily reducing the intensity of the incident wave.) For simplicity, we set the relative magnetic permeabilities μ for both materials equal to 1.0. We prefer to specify each material by its dielectric constant ϵ rather than its refractive index $n = \sqrt{\epsilon}$, because this square root requires a choice for the sign, and since the choice of sign is the primary source of confusion, we will do it only once when calculating the propagation vector below. The incident, reflected, and transmitted plane waves are given by

$$\vec{E}_{1R}(x,z) = E_{1R} e^{i(k_x x + k_{z1} z)} \hat{y} \tag{1}$$

$$\vec{E}_{1L}(x,z) = E_{1L} e^{i(k_x x - k_{z1} z)} \hat{y} \tag{2}$$

$$\vec{E}_{2R}(x,z) = E_{2R} e^{i(k_x x + k_{z2} z)} \hat{y}, \tag{3}$$

where the subscript $1R$, for example, denotes a right-traveling wave in medium one. Note that, in medium two, the time-averaged Poynting vector is

$$\langle \vec{S}(x,z) \rangle = \frac{|E_{2R}|^2}{2\omega\mu_0} \exp[-2\ \mathrm{Im}(k_{z2}) z] \left[k_x \hat{x} + \mathrm{Re}\,(k_{z2})\, \hat{z} \right]. \tag{4}$$

Therefore, the phase velocity along z is in the same direction as the energy flow along z. Hence, the description of a wave as 'right-traveling' or 'left-traveling' refers to both the phase velocity and the energy flow in this medium. (This nomenclature would need to be modified for negative-index media, for which the energy flow and phase velocity can point in different directions.) We will restrict our discussion to s-polarized light, although the arguments apply equally to p-polarized light. We also ignore any variation of the fields in the y-direction. The component k_x is determined by the angle of incidence and the index of medium one, $k_x = k_0 \sqrt{\epsilon_1} \sin\theta$, where $k_0 = \omega/c$ is the magnitude of the k-vector in vacuum. We have the freedom to choose k_x and k_{z1} to be positive, so that the incident beam is traveling up and to the right in our coordinate system. We have already made use of the fact that all three waves must have the same k_x in order to match the boundary conditions everywhere along the interface, and that the z-components of the k-vectors for the incident

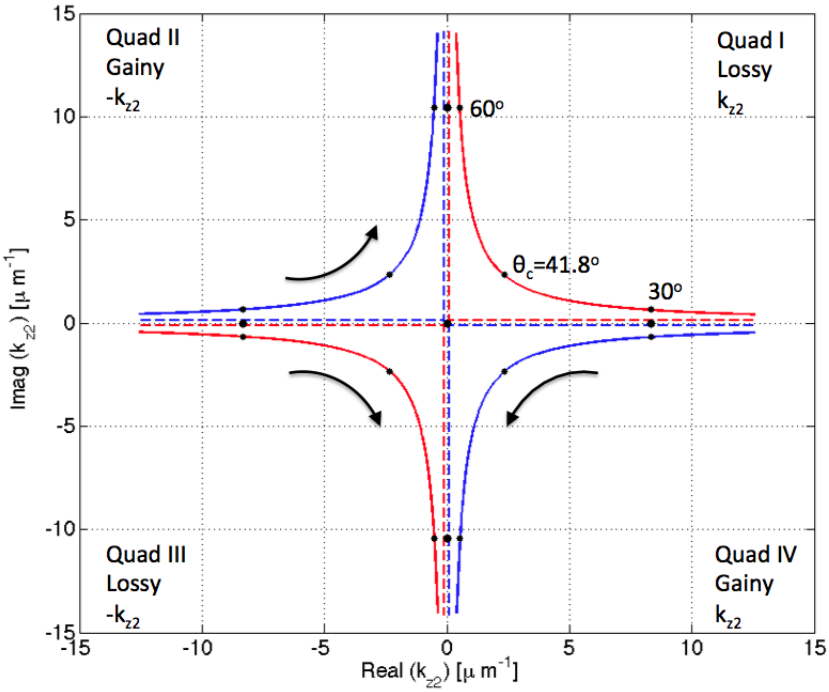


FIG. 2. $\lambda_0 = 600$ nm, $\epsilon_1 = 3.24$, $\epsilon_2' = 1.44$, $\theta_c = 41.8°$. Complex-plane plot of the two possible choices for k_{z2}, parameterized by θ, in both lossy (red, quadrants 1 and 3) and gainy (blue, quadrants 2 and 4) cases. The black dots on each curve indicate $\theta = 30°, 41.8°$, and $60°$, and the arrows point in the direction of increasing θ. The solid lines correspond to $\epsilon_2'' = 0.1$ (red) or $\epsilon_2'' = -0.1$ (blue). The dashed lines are the limiting cases for k_{z2} as ϵ_2'' approaches zero from above (lossy) and below (gainy).

and reflected waves differ only by a sign. From these fields and the associated H-fields, the Fresnel reflection coefficient is easily shown to be

$$\frac{E_{1L}}{E_{1R}} = \frac{k_{z1} - k_{z2}}{k_{z1} + k_{z2}}. \tag{5}$$

All that remains is to determine k_{z2}, which by the dispersion relation $k_x^2 + k_{z2}^2 = k_0^2 \epsilon_2$ can take only one of two values:

$$k_{z2} = \pm k_0 \sqrt{(\epsilon_2' - \epsilon_1 \sin^2\theta) + i\epsilon_2''}. \tag{6}$$

In Fig. 2, we examine k_{z2} of Eq. 6 in both lossy and gainy cases as we vary θ. For a lossy medium, $\epsilon_2'' > 0$, and so below the critical angle θ_c, the quantity under the radical is in the first quadrant of the complex plane, and therefore k_{z2} lies either in the first or third quadrant. The first quadrant solution corresponds to a right-propagating wave whose

amplitude decreases to the right, while the third quadrant solution corresponds to a left-propagating wave whose amplitude decreases to the left. In both cases the wave amplitude decays along the propagation direction. It is clear that the proper choice is always the first quadrant solution, i.e. the one for which the real part of k_{z2} is positive, for two reasons: 1) we expect that energy in medium two should be flowing to the right, and 2) choosing the third quadrant solution leads to a reflection coefficient greater than unity in magnitude (see Eq. 5), which is clearly nonsense for reflection from a lossy medium. Above the critical angle, the quantity under the radical is in the second quadrant, but the values of k_{z2} remain in the first and third quadrants, and we choose the first quadrant solution again for the same reasons as above.

For a gainy medium, $\epsilon_2'' < 0$, and so the quantity under the radical is in the fourth quadrant of the complex-plane for $\theta < \theta_c$, and in the third quadrant for $\theta > \theta_c$. Either way, k_{z2} lies in the fourth or second quadrant. The fourth quadrant solution corresponds to a right-propagating wave whose amplitude increases to the right, while the second quadrant solution corresponds to a left-propagating wave whose amplitude increases to the left. In both cases the wave amplitude grows along the propagation direction. Below the critical angle, it is universally accepted that the fourth quadrant solution, which propagates to the right and exponentially grows with increasing z, is the correct one. This is the principle behind a single pass optical amplifier. Put another way, if we imagine terminating the gain medium at some finite z with a perfect absorber, we expect the absorber to receive increasing amounts of energy with increasing values of z.

The controversy over whether to choose the fourth or second quadrant solution arises for angles of incidence greater than θ_c. The quadrant four solution carries energy away from the interface and leads to a reflection coefficient with magnitude less than one, which does not allow for ss-ATIR. There has been a reluctance to accept the fourth quadrant solution, for reasons that have mainly to do with the limiting case as the gain or absorption approaches zero. The argument is as follows. Imagine the second medium has a fixed ϵ_2', but its gain or absorption parameter ϵ_2'' can be tuned with a knob. Below the critical angle, k_{z2} approaches the same value (a real, positive number) whether ϵ_2'' approaches zero from below or from above (as seen in quadrants 1 and 4 of Fig. 2. Above the critical angle, however,

$$\lim_{\epsilon'' \to 0^-} k_{z2} = - \lim_{\epsilon'' \to 0^+} k_{z2}. \tag{7}$$

In other words, a gainy material tuned towards transparency will have a negative, imaginary k_{z2}, which implies an exponentially growing field in the absence of any gain. This seems absurd. It should be noted, however, that above the critical angle, as ϵ_2'' approaches zero (from above or below), the real part of k_{z2} approaches zero as well, while k_x, being determined only by the incident plane wave, remains finite. Thus, the wave in medium two is propagating primarily in the x-direction, so exponential growth in the z-direction may be justified despite a very small population inversion because the path length in the x-direction is very long. (This argument is also mentioned in [9], in a slightly different context.) This statement is not a formal justification that rapid exponential growth in the limit of small gain is acceptable. Our intention is simply to point out that the mathematical discontinuity in the two limiting cases may have a simple physical explanation.

The quadrant two solution, on the other hand, carries energy from medium two towards the interface, which allows the reflection coefficient in the first medium to be greater than one, thus yielding ss-ATIR [1, 2]. In the limit as ϵ_2'' approaches zero, this choice converges to the solution in a lossy medium. Still, the quadrant two solution is not without its flaws. The argument that is typically first raised is that this choice appears to violate causality: energy is flowing to the left despite the absence of any electromagnetic sources at $z = +\infty$ [4]. It has been countered that causality is upheld because 'no information flows in from the infinities' [6] due to the exponential decay of the field in the gain medium. One justification for choosing the second quadrant solution is that 'evanescently decaying waves remain decaying' [6], although there is no physical reason why evanescent waves cannot grow, especially considering the blurry distinction between evanescent and propagating waves when gain is present. Finite difference time domain (FDTD) simulations have been used to justify ss-ATIR [3], but it is worth emphasizing that the perfectly matched layer used to terminate the gain medium is never a perfect absorber, and even small numerical errors can be amplified to substantial values when gain is present. It has been argued, through the analysis of poles in the global permittivity function [7], that the choice between the second and fourth quadrant solutions depends on the dispersion of the dielectric constant. This conclusion has the undesirable consequence that knowledge of the dielectric constant at one frequency is not sufficient to determine the behavior of the medium at that frequency. Typically, for a linear, time-invariant medium we expect the response at one frequency to be independent of the response at any other frequency. Such back and forth arguments have led some to declare

the problem of a semi-infinite gain medium 'significantly unrealistic' [8], and to simply stop worrying about it.

We would like to emphasize one aspect of the problem which we believe has not received due attention in the literature. In conventional total internal reflection from transparent media, there is an abrupt change at the critical angle: k_{z2} is real-valued below θ_c, zero at θ_c, and purely imaginary above θ_c (see Fig. 2). However, in the case of lossy and gainy media, such an abrupt transition does not exist. Here, k_{z2} is a complex number with non-zero real and imaginary parts; as one increases the angle of incidence, k_{z2} changes smoothly through the critical angle. There is no hard distinction between propagating waves below θ_c and evanescent waves above θ_c. It is true that the rate of change of $\mathrm{Im}(k_{z2})$ accelerates beyond the critical angle, but even this rate-of-change is a smooth and continuous function of the incidence angle. There is in fact no parameter of the system that undergoes a discontinuous change (or a discontinuous rate-of-change) as θ crosses θ_c. Thus, within the framework adopted for the analysis (i.e., linear, homogeneous gain medium specified by a constant permittivity whose imaginary part is negative), there is no compelling reason to expect an abrupt jump of k_{z2} from the fourth to the second quadrant (and the accompanying discontinuity in the Poynting vector and phase velocity), as θ rises from slightly below to slightly above θ_c. We will demonstrate in the next section, for a finite slab, the manner in which the backward-propagating quadrant two solution is generated. We argue that without a back interface, it seems likely that only the quadrant four solution can exist.

III. THREE MEDIA

To understand the amplification of reflected light, it is necessary to consider three media, as shown in Fig. 1(b). The incidence medium fills the half-space $z < 0$ and has a positive, real-valued permittivity ϵ_1; medium two fills the space $0 < z < d$ and has a complex permittivity $\epsilon_2 = \epsilon_2' + i\epsilon_2''$, and medium three fills the space $z > d$ and has a positive, real-valued permittivity ϵ_3. We take the incident wave to be given by the same expression as in Eq. 1, and introduce two new plane waves E_{2L} and E_{3R}. There is no controversy about how to solve for E_{1L}, E_{2R}, E_{2L}, and E_{3R} in terms of the incident amplitude E_{1R} in this case, since both the right-traveling wave E_{2R} and left-traveling wave E_{2L} must coexist in the finite slab, with opposite signs for k_{z2}. It is instructive to express the reflection coefficient

in medium one as

$$\frac{E_{1L}}{E_{1R}} = r_{12} + \frac{t_{21}t_{12}r_{23}\exp(2ik_{z2}d)}{1 - r_{21}r_{23}\exp(2ik_{z2}d)}, \tag{8}$$

where r_{12} denotes the single-surface reflection coefficient from medium one looking into medium two, and the subscripts on r_{21}, r_{23}, t_{21} and t_{12} work the same way for the other single-surface reflection and transmission coefficients. This equation is valid for any angle of incidence, provided the reflection and transmission coefficients as well as k_{z2} are calculated at the desired angle. (Note: by our choice, k_{z2} has a positive real part and is associated with the wave E_{2R}, while the wave E_{2L} has $-k_{z2}$ as the z-component of its k-vector.) The quantity $r_{21}r_{23}\exp(2ik_{z2}d)$ is the round-trip coefficient. Traditionally, this quantity is important because if it equals 1.0, the plane wave inside the slab regenerates itself on a round-trip, which results in lasing. The possibility of lasing will be discussed later; for now we emphasize that there is no feedback in the x-direction because the slab is infinite along the x-axis, and so a pulse of light spontaneously emitted in any direction other than z will not be able to reinforce itself as it zig-zags up (or down) the slab. Due to this lack of feedback for any incidence angle θ other than zero, there is no problem with the round-trip coefficient taking values greater than unity. We will demonstrate that amplification of the primary reflected beam is possible when the magnitude of the round-trip coefficient is greater than unity, and does not depend on the critical angle.

A. Below the critical angle

In this section, we fix the parameters $\epsilon_1 = 3.24$, $\epsilon_2' = 1.44$, $\epsilon_3 = 3.24$, $d = 20$ μm, and $\lambda_0 = 600$ nm, for which the critical angle $\theta_c = 41.8°$. We begin by discussing the plane-wave solutions for this geometry at different values of the gain ϵ_2'', and also simulate a finite-diameter beam of light (and a finite duration pulse) incident upon the structure. The beam, with an assumed Gaussian spatial profile, is simulated as a superposition of plane-waves. While the beam is only approximately Gaussian due to our inability to sum over a continuum of plane waves, the simulation remains an analytical solution to Maxwell's equations. (See the appendix for simulation details.) In all simulations, the incident beam has a peak amplitude and intensity of 1.0 in arbitrary units.

First, we will examine the behavior of a beam for which the full width at half maximum (FWHM) of the beam's cross-section in the xz-plane is taken to be $30\lambda_1 = 10$ μm, where

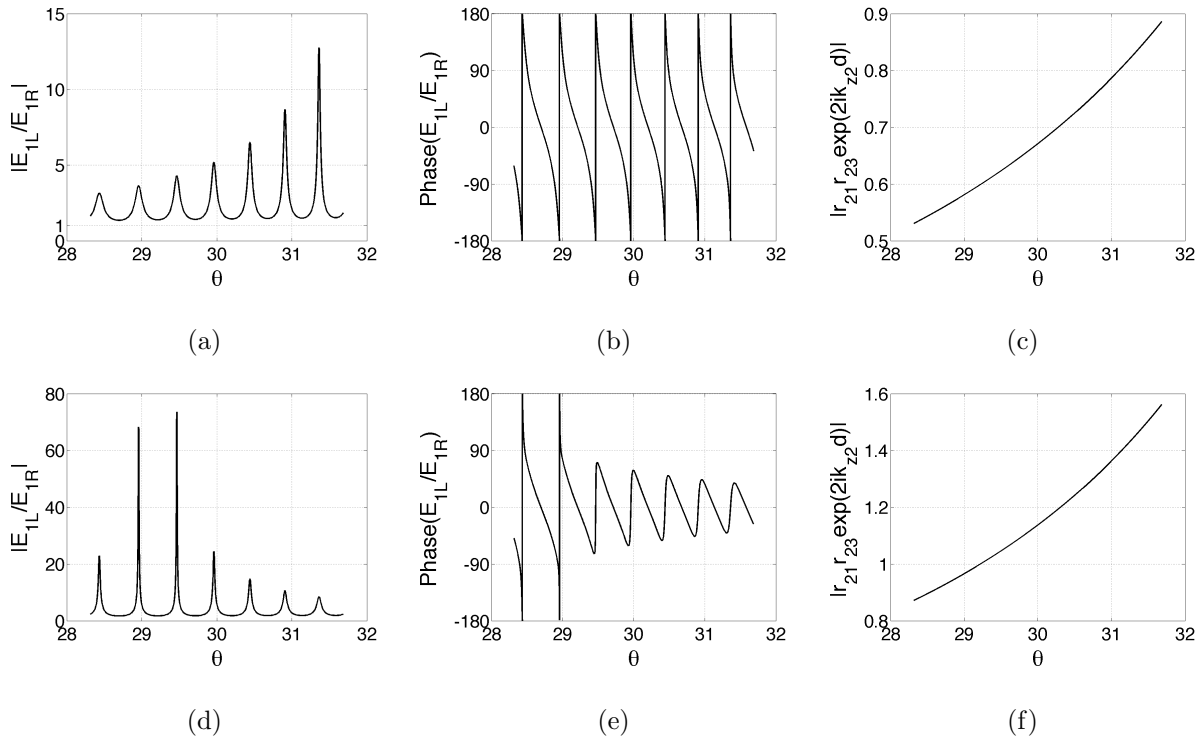


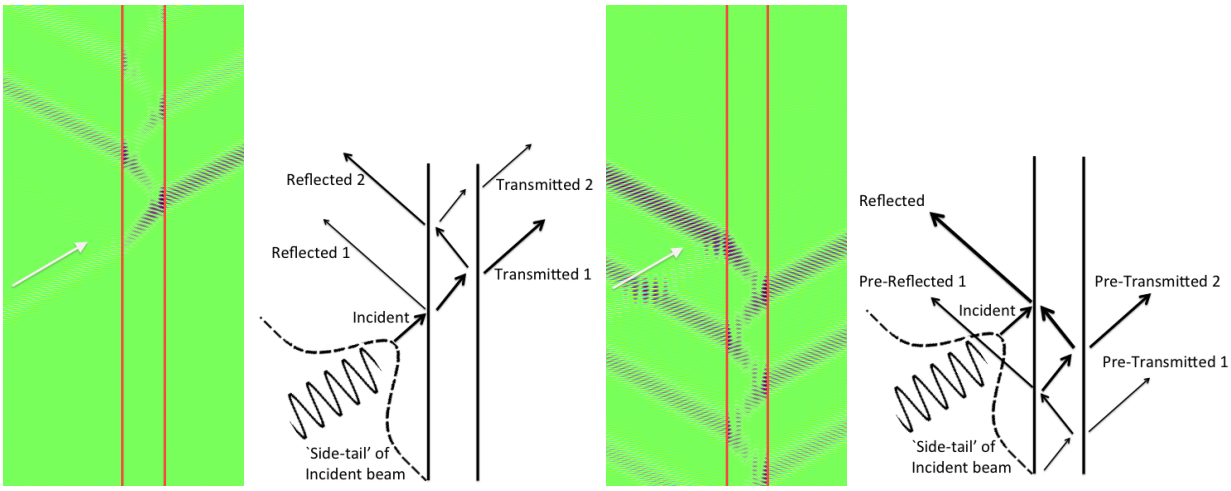
FIG. 3. For all plots, $\epsilon_1 = 3.24$, $\epsilon_2' = 1.44$, $\epsilon_3 = 3.24$, $d = 20$ μm, $\lambda_0 = 600$ nm, $\theta_c = 41.8°$. For $\epsilon_2'' = -0.007$: (a) magnitude of the reflection coefficient E_{1L}/E_{1R}, (b) phase of the reflection coefficient, and (c) magnitude of the roundtrip coefficient $r_{21}r_{23}\exp(2ik_{z2}d)$ versus angle of incidence θ. For $\epsilon_2'' = -0.009$, the same plots are shown in (d), (e), and (f). The angle of incidence is restricted to $28.3142° < \theta < 31.6858°$, which corresponds to the range of angles required to create a 10 μm FWHM Gaussian beam whose central k-vector is incident at $\theta = 30°$. In (a) and (d), the reflection coefficient peaks at certain resonances, and the behavior is qualitatively the same in both cases apart from the actual magnitudes of the reflectivity peaks. However, the behavior of the phase in (e) departs significantly from the phase shown in (b) at just above $\theta \approx 29°$. From (f), we see that this change corresponds with the angle above which the round-trip coefficient is greater than one, whereas the round-trip coefficient in (c) is always below one.

λ_1 is the wavelength inside the incidence medium. A very good approximation to this beam is achieved by superposing k-vectors whose angle of incidence lies in the range $28.3142° < \theta < 31.6858°$. In Fig. 3, we plot various quantities of interest for plane waves within this range of θ. Figures 3(a)-(c) correspond to $\epsilon_2'' = -0.007$, while Figs. 3(d)-(f) represent the slightly higher gain of $\epsilon_2'' = -0.009$. Figures 3(a) and 3(d) show the magnitude of the reflection coefficient in medium one, $|E_{1L}/E_{1R}|$. In both cases there exist resonant peaks,

which occur when the phase of the round-trip coefficient (not plotted) nears zero, and thus the denominator in Eq. 9 becomes small. Although the magnitudes of the reflectivity peaks are higher when the gain is higher, the behavior in both cases is qualitatively the same. Figures 3(b) and 3(e) show the phase of the reflection coefficient; just above $\theta \approx 29°$, the behavior of the phase in Fig. 3(e) departs significantly from the behavior in Fig. 3(b). To understand the physical reason for this sudden change, Figs. 3(c) and 3(f) show the magnitude of the round-trip coefficient. While this quantity is less than 1.0 for all angles of incidence in Fig. 3(c), in Fig. 3(f) the gain is a little larger, and the round-trip coefficient exceeds 1.0 when θ is greater than $\theta_{\mathrm{rt}} = 29.2°$, corresponding to the angle at which the phase behavior suddenly changes. We call θ_{rt} the round-trip angle and define it to be the incidence angle at which the magnitude of the round-trip coefficient equals 1.0.

This change has important consequences for the behavior of a Gaussian beam, which is determined by the superposition of such plane waves. Figure 4(a) shows the E-field amplitude of the beam in the case of $\epsilon_2'' = -0.007$. In fact, the behavior of this beam is very intuitive because it resembles what we would expect for a transparent slab: multiple reflections in the slab giving rise to multiple reflected beams in medium one and transmitted beams in medium three. The difference from the transparent case is that the gain allows this zig-zagging process to continue for many more round trips, and the sum of the energy flux of the reflected beams must be greater than the energy flux of the incident beam, simply because the magnitude of the reflection coefficient is greater than unity for all constituent plane-waves of the beam (Parseval's theorem), as can be seen in Fig. 3(a). This process is shown schematically in Fig. 4(b). Note, however, that the primary reflected beam (which is barely visible in Fig. 4(a) because it is weak) does not by itself carry more energy than the incident beam. The gain due to stimulated emission only appears in the subsequently reflected beams which arise after the incident beam has traversed the gain region.

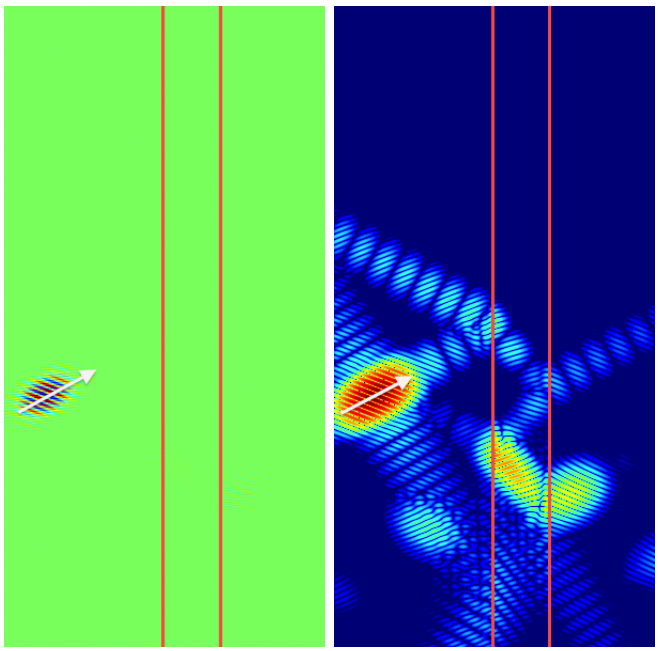
Figure 4(c) shows the E-field amplitude of the beam for $\epsilon_2'' = -0.009$. It appears that there is an excitation in the slab well before the beam arrives at the slab. To understand this, one needs to consider that the beam, though its FWHM is only 10 μm, has a Gaussian 'side-tail' which extends much further than 10 μm in both directions normal to the propagation direction of the beam, as depicted in Fig. 4(d). This side-tail decays exponentially and cannot be seen in Fig. 4(c), but it still enters the slab well before the central part of the beam. Because the round-trip coefficient is greater than unity, this excitation is amplified as



(a) (b) (c) (d)

FIG. 4. Simulations using the geometry of Fig. 1(b) with parameters $\epsilon_1 = 3.24$, $\epsilon_2' = 1.44$, $\epsilon_3 = 3.24$, $d = 20$ μm, $\lambda_0 = 600$ nm, $\theta_c = 41.8°$. (a) The E-field amplitude of a Gaussian beam (FWHM = 10 μm) whose central k-vector is incident at $\theta = 30°$. The white arrow is placed just above the incident beam. For this gain, the round-trip coefficient is less than unity for all constituent k-vectors, and the accompanying schematic (b) illustrates how the beam enters the slab and bounces back and forth, giving rise to many transmitted beams in medium three and reflected beams in medium one. (c) A simulation with all of the same parameters as in (a) except that $\epsilon_2'' = -0.009$, so that many of the constituent k-vectors have a round-trip coefficient greater than unity. In this case, the weak Gaussian side-tail of the beam, which enters medium two before the central part of the beam, can be amplified to observable values as it bounces back and forth in the slab, giving rise to many 'pre-reflected' and 'pre-transmitted' beams. This is shown schematically in (d).

it zig-zags to magnitudes which can be seen in the simulation. Along the way, it gives rise to multiple 'pre-reflected' and 'pre-transmitted' beams which emanate from points on the slab with $x < 0$, lower than the location at which the central portion of the beam arrives at the slab! Interestingly, the interaction between the amplified side-tail and the central part of the beam is such that when the two finally meet at the entrance facet of the slab, the result is a strongly amplified reflected beam in medium one, and a nearly extinguished beam in medium two. (The weak E-field in the region $x > 0$, other than the primary reflected beam, is due almost entirely to the k-vector constituents below θ_{rt}, for which the round-trip

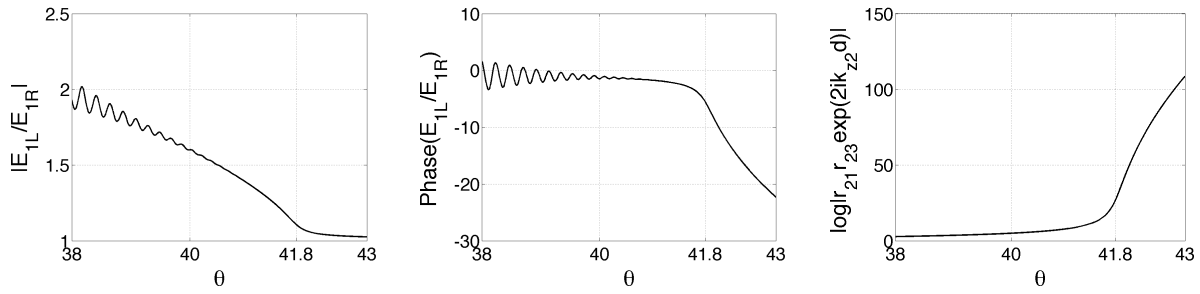


(a) (b)

FIG. 5. $\epsilon_1 = 3.24$, $\epsilon_2' = 1.44$, $\epsilon_3 = 3.24$, $d = 20$ μm, $\lambda_0 = 600$ nm, $\theta_c = 41.8°$. A Gaussian pulse (temporal FWHM=100 fs, spatial FWHM=10 μm, $\theta = 30°$), which travels in the direction of the white arrow, is shown at one instant of time before the central part of the pulse arrives at the slab. A larger value of gain $\epsilon_2'' = -0.018$, compared to $\epsilon_2'' = -0.009$ used in Fig. 4, is chosen for clarity because the larger gain leads to fewer pre-reflected pulses which would obscure the side-tail. (a) A linear plot of the E-field amplitude, which shows that the central part of the pulse has yet to arrive at the second medium. (b) A logarithmic plot of the E-field intensity. Red corresponds to the maximum intensity in the frame (= 1) and dark blue corresponds to intensity values of $\exp(-15)$ and smaller. Here, we see the interesting behavior of the side-tail, which enters the medium before the central part of the pulse arrives at the slab. At the instant of time shown in the simulation, one pre-reflected pulse has been generated, and a pre-transmitted pulse is about to leave the slab.

coefficient is less than unity.)

In order to visualize the side-tail more clearly, it is useful to simulate a finite-duration (as well as finite-diameter) pulse of light rather than a steady (cw) beam. Figure 5 illustrates such a pulse, Gaussian in space and time, whose FWHM spatial width is 10 μm and FWHM temporal width is 100 fs. Additionally, the gain is raised to $\epsilon_2'' = -0.018$, simply because larger gain leads to fewer pre-reflected pulses (discussed later) which would overlap and

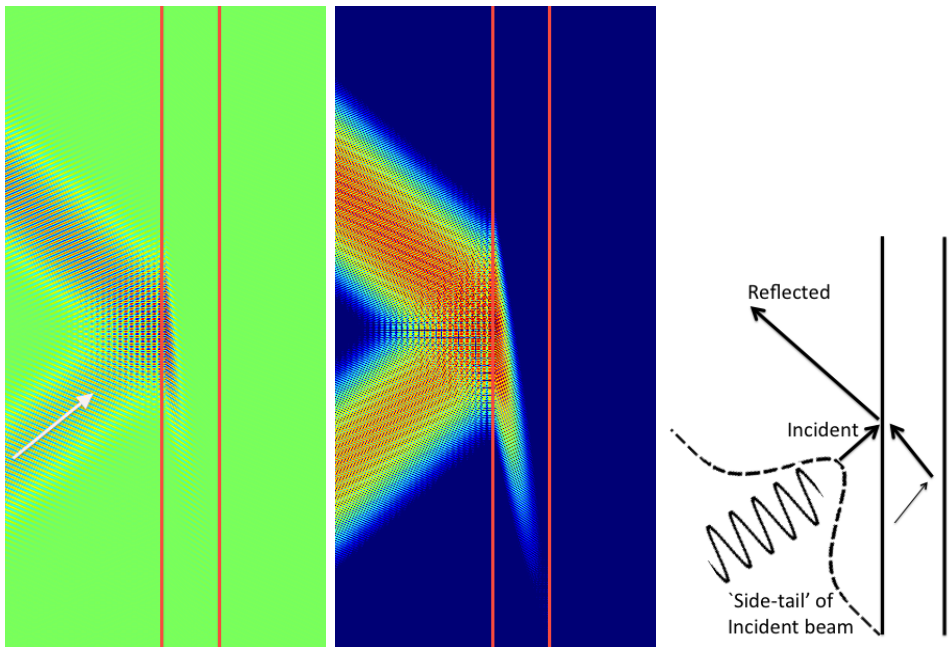


(a) (b) (c)

FIG. 6. For all plots, $\epsilon_1 = 3.24$, $\epsilon_2 = 1.44 - 0.009i$, $\epsilon_3 = 3.24$, $d = 20$ μm, $\lambda_0 = 600$ nm, $\theta_c = 41.8°$. (a) Magnitude of the reflection coefficient E_{1L}/E_{1R} versus the angle of incidence θ, (b) phase of the reflection coefficient versus θ, (c) natural logarithm of the magnitude of the round-trip coefficient $r_{21}r_{23}\exp(2ik_{z2}d)$ versus θ. As seen in (a) and (b), the magnitude and phase of the reflection coefficient change behavior from oscillatory to monotonically decreasing just above $\theta \approx 40°$, which is still below θ_c. Above 40°, the round-trip coefficient is so large that E_{2L} is vastly more significant than E_{2R}, which prevents the oscillatory behavior in the magnitude and phase of the reflection coefficient as a function of θ. As θ crosses θ_c, the decrease in the phase and increase in the round-trip coefficient both become steeper, but we emphasize that qualitatively, there is nothing different between this situation and the situation below θ_c.

obscure the side-tail. The peak intensity of the incident pulse is normalized to unity. A linear plot of the E-field amplitude is shown in Fig. 5(a), which indicates that the pulse has yet to arrive at the slab, and suggests that the reflection process has yet to begin. In Fig. 5(b), however, a logarithmic plot of the E-field intensity is shown which allows much smaller intensity values to be visualized; red corresponds to the maximum intensity in the frame (≈ 1) and dark blue corresponds to intensity values of $\exp(-15)$ and smaller. We see that the side-tail of the pulse has already excited the gain medium: one pre-reflected pulse is traveling to the left in medium one, a pre-transmitted pulse is about to leave the slab and enter medium three, and one can infer that a short time later, a second pre-reflected pulse will be created. (Some of these features can be seen in the linear plot of Fig. 5(a) if one looks closely.) It is clear that the process of reflection begins well before the central part of the pulse arrives at the interface with medium two.

Now that we have discussed the mechanism of pre-excitation in the slab for incidence angles just above $\theta_{\rm rt}$, we need to understand whether this mechanism changes as we approach,



(a) (b) (c)

FIG. 7. For all plots, $\epsilon_1 = 3.24$, $\epsilon_2 = 1.44 - 0.009i$, $\epsilon_3 = 3.24$, $d = 20$ μm, $\lambda_0 = 600$ nm, $\theta_c = 41.8°$. (a) A simulation of the E-field amplitude of a beam (FWHM $= 30$ μm) incident at $\theta = 41°$ (range of angles for the constituent k-vectors: $40.438° < \theta < 41.562°$) shows a similar behavior to that seen at 30° incidence, although there are no pre-reflected or pre-transmitted beams of any significant amplitude; instead, the side-tail enters the slab, travels to the right facet, then returns to amplify the reflection arising from the central portion of the beam. (b) The same simulation, plotted using a logarithmic scale, where red corresponds to the maximum intensity (5.09) and dark blue corresponds to intensity values of $\exp(-6)$ and smaller. (c) A schematic diagram of the behavior of the side-tail when the round-trip coefficient is much greater than one.

and then cross the critical angle. We return to the case of $\epsilon_2'' = -0.009$, for which we have already seen $\theta_{\rm rt} = 29.2°$ and $\theta_c = 41.8°$, and in Fig. 6 we analyze the behavior of plane waves in this system at angles near θ_c. In Fig. 6(a), the magnitude of the reflection coefficient changes behavior from oscillatory to monotonically decreasing just above $\theta \approx 40°$. The phase of the reflection coefficient shows the same change in behavior in Fig. 6(b). Figure 6(c) shows the natural logarithm of the magnitude of the round-trip coefficient; this quantity shoots up dramatically above θ_c because small increases in θ lead to large increases in the magnitude of the imaginary part of k_{z2}, as shown in Fig. 2. Still, it is important to notice that the round-trip coefficient is already very large before θ crosses θ_c: at $\theta = 40°$, the round-trip

coefficient has a magnitude of 144, and by $\theta = 41^\circ$ it has grown to 3505. As the round-trip coefficient becomes much larger than one, we can ignore the 1.0 in the denominator of Eq. 9. The exponential factor in the numerator and denominator then cancel, leaving us with the reflection coefficient

$$\frac{E_{1L}}{E_{1R}} = r_{12} - \frac{t_{21}t_{12}}{r_{21}}, \tag{9}$$

which is a smooth and featureless function of the incidence angle that approaches unity from above, as seen in Fig. 6(a) for $\theta > 40^\circ$. A more physical way to think about the behavior is that as the round-trip coefficient becomes larger, the behavior of E_{2L} dominates that of E_{2R}, since the former becomes much larger in magnitude than the latter. This leads to an end of the oscillatory behavior of the reflection coefficient's magnitude and phase, the reason being that the light that bounces around in the slab and interferes with itself gives rise to resonances only when the two interfering components have roughly equal magnitudes. The simulation of a beam having FWHM = 30 μm at the incidence angle of $\theta = 41^\circ$ (range of angles for the constituent k-vectors: $40.438^\circ < \theta < 41.562^\circ$) is shown in Fig. 7. Figure 7(a) is a linear plot of the E-field amplitude, while Fig. 7(b) is a logarithmic plot of the E-field intensity: all values are normalized to the peak intensity of the incident beam, dark blue corresponds to values of $\exp(-6)$ and below, and red corresponds to the maximum intensity in the frame (5.09). From the simulation, we see that the effect of the increasing round-trip coefficient (and corresponding approach of the reflection coefficient towards unity from above) is two-fold: i) the primary reflected beam resembles the incident beam, thus eliminating the multitude of pre-reflected and pre-transmitted beams seen at lower angles of incidence. There may still exist some pre-reflected and pre-transmitted beams, but they will be weak. (This is the reason the pulse simulation of Fig. 5 was calculated at the higher gain $\epsilon_2'' = -0.018$.) (ii) The continuity of the fields at the first interface implies that E_{2R} will be extremely small, though by no means will it be equal to zero. In Fig. 7, we still clearly see the amplified side-tail returning from the back interface to contribute energy to the primary reflected beam.

B. Above the critical angle

It is not possible to simulate a beam above the critical angle which shows the mechanism of pre-excitation as clearly as does Fig. 7(a) for the case of $\theta = 41^\circ$, for the reasons described

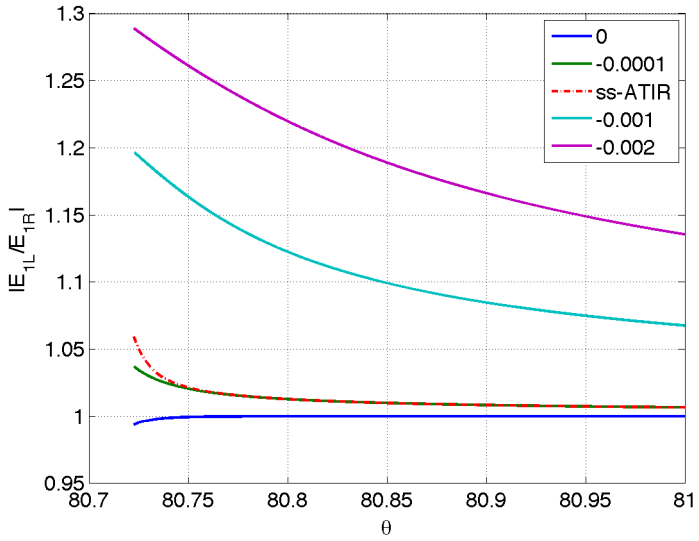


FIG. 8. The material parameters are those from Koester's fiber experiment [1]: $\epsilon_1 = 2.338$, $\epsilon_2' = 2.277$, $\epsilon_3 = 2.280$, $d = 26$ μm, $\lambda_0 = 1060$ nm. The plot shows the magnitude of the reflection coefficient, $|E_{1L}/E_{1R}|$, above the critical angle $\theta_{c12} = 80.72°$, for various values of $\epsilon_2'' = 0$ (blue), -0.0001 (green), -0.001 (turquoise), -0.002 (purple). The red dash-dot line is the single-surface amplified-TIR reflectivity for $\epsilon_2'' = -0.0001$, calculated using the single-surface formula (Eq. 5) with the quadrant IV choice for k_{z2}. Note that the ss-ATIR curve approaches the correct curve very quickly.

in the previous paragraph. The round-trip gain quickly becomes so large above the critical angle that E_{2R} is very small compared to E_{2L}. Also, the real part of k_{z2} quickly becomes so small above the critical angle that both the left-going and right-going waves in the slab propagate primarily in the x-direction, so the zig-zagging cannot be seen. This is just a concern for visualization, but not for the underlying physics. From the plots in Fig. 6, it is clear that, qualitatively, there is not much difference between the cases of slightly above and slightly below the critical angle. The right-going wave must reflect off the back interface in order to amplify the primary reflection.

We find it instructive to consider the same material and wavelength parameters as in Koester's gain-clad fiber experiment [1]: $\epsilon_1 = 2.338$ for the passive core, $\epsilon_2' = 2.277$ for the real part of the gainy cladding and $d = 26$ μm for the thickness, $\epsilon_3 = 2.280$ for the passive outer cladding, and incident light of vacuum wavelength $\lambda_0 = 2\pi c/\omega = 1060$ nm. Figure 8 shows the magnitude of the reflection coefficient as a function of θ above the critical

angle at various values of ϵ_2'': 0 (blue), -0.0001 (green), -0.001 (turquoise), -0.002 (purple). It should be noted that there are two critical angles, $\theta_{c12} = \arcsin(\sqrt{\epsilon_2'/\epsilon_1}) = 80.72°$ and $\theta_{c13} = \arcsin(\sqrt{\epsilon_3/\epsilon_1}) = 80.96°$. For a transparent medium two, the reflection coefficient will equal unity for $\theta \geq \theta_{c13}$, since no energy can penetrate medium three, and so any energy that enters medium two must eventually end up in the reflected beam E_{1L}. If we put just a little gain in the slab, $\epsilon_2'' = -0.0001$, the reflectivity will be about 1.04 just above θ_{c12}, but will quickly approach unity from above as θ increases. The same trend holds, although with higher reflectivities, for larger values of gain. It is interesting to compare these values with the ss-ATIR reflectivity, which is calculated using Eq. 5 with the quadrant IV solution for k_{z2} (and which completely ignores the effects of the back facet and medium three). This ss-ATIR curve is calculated for $\epsilon_2'' = -0.0001$ (red dash-dot), and is seen to quickly converge to the finite slab solution for $\theta > 80.75°$. For the higher gain curves, the ss-ATIR curves are not plotted, because they would overlap nearly perfectly with the corresponding finite slab solution.

The accuracy of the single-surface formula in determining the reflection coefficient is easily explained if we recall that when the round-trip coefficient is large E_{2L} is much larger than E_{2R} and dominates the behavior of the system. Although E_{2R} is small, it is not zero: the right-traveling wave is essential, as it generates the left-traveling wave when it reflects off the interface between media two and three. Note that as θ increases past θ_c, the round-trip coefficient becomes larger and yet the reflection coefficient in medium one quickly approaches unity from above. This means that as θ increases, E_{2R} quickly decreases; in other words, less light is able to penetrate the gainy slab, so that the round-trip gain, although it increases with θ, is acting on a smaller input.

IV. DISCUSSION

For a light beam incident on a gainy slab, we have demonstrated the importance of the round-trip angle–the angle at which the magnitude of the round-trip coefficient exceeds one–on the qualitative behavior of the reflection and transmission. We have shown simulations of particular cases for which the round-trip angle is less than the critical angle, and demonstrated that the round-trip angle is the more important parameter in determining whether the primary reflected beam is amplified or not. The simulations also elucidate the mecha-

nism for the amplification of the primary reflected beam: the incident Gaussian beam has a 'side-tail' that enters the slab before the main portion of the beam. Above the round-trip angle, the side-tail gains more energy during propagation than it loses to transmission at the back interface, and therefore returns to the first interface where it contributes to the amplification of the primary reflected beam. We have argued that the same mechanism should be at work above the critical angle, although for reasons we have explained it is difficult to visualize this effect in our simulations. In the problem of amplified TIR from a single-interface, the amplification mechanism that we have established cannot occur because it relies crucially on the presence of a second interface. Therefore, we conclude that ss-ATIR is unlikely. Instead, it is more plausible that the wave which penetrates the second medium in the semi-infinite case must under all circumstances carry energy into the second medium, whether the second medium is lossy or gainy, and regardless of the angle of incidence, in agreement with [4].

It is possible that the non-linear effects or time-dependent dynamics in a true amplifier give rise to more complicated mechanisms of reflection or amplification. For homogeneous media, however, Maxwell's equations allow only two possible choices for the k-vector in the second medium. We argue that one must always choose the wave which carries energy into the second medium, and that this choice upholds causality. We note that, following the advent of negative index materials, there was similar controversy in the literature regarding the sign choice for the k-vector of the transmitted wave. While some have argued that the Poynting vector in medium two can be directed towards medium one under certain conditions in negative index media [10], others have shown the consequences of that argument to be unphysical [11]. Interestingly, simulations similar to ours of a two-dimensional Gaussian beam incident on a finite thickness, non-magnetic ($\mu = 1$), gainy slab have demonstrated the pre-transmitted beam before [9, 12], although its peculiar location at $x < 0$ was ascribed to negative refraction within the slab. However, for a simulation of a single-frequency beam (which sees only the dielectric constant at that one frequency) in a non-magnetic medium, we find it difficult to see how negative index behavior could occur.

No real material is semi-infinite, of course, so some thought needs to be put into the design of an experimental test of the proposed behavior of the gain medium. With absorptive media, a finite thickness slab can be treated as effectively infinite. However, the nature of the side-tail demonstrated in this paper makes finding an effectively infinite gain medium very

difficult. One can try to increase the width of the gain medium in an effort to prevent the side-tail from having enough time to complete a round-trip and affect the primary reflection. However, the increased width leads to a larger round-trip coefficient, so earlier (and weaker) portions of the side-tail are still capable of completing a round-trip in the time it takes for the central portion of the pulse to arrive. The result is that the E_{2L} beam dominates the system, whereas in a true semi-infinite medium, we claim that only the E_{2R} beam exists. To get around this problem, a very sharp pulse (which does not have the long side-tail of a Gaussian pulse) must be used, and the gainy slab must be thick enough such that the weakest part of the incident pulse does not have time to travel to the back facet and return by the time the central part of the pulse arrives. Then, the amplitude of the reflected pulse will be less than the incident pulse, barring any non-linear effects. Since the intensity of the reflected pulse relative to the incident pulse will either be only slightly below one or slightly above one, the difference in the two theories could be difficult to measure. It has been suggested to look instead at the phase of the primary reflected beam with a suitable interferometry experiment [4], which differs greatly for the two theories.

V. APPENDIX: DESCRIPTION OF SIMULATION

The simulations of Gaussian beams and pulses incident on a finite thickness slab, described in Sec. III A, were performed using MATLAB. The simulations depict analytical solutions to Maxwell's equations in that the E-field at any pixel is determined by summing a large number of plane wave solutions. A single incident plane wave E_{1R} is characterized by the frequency ω and the z-component of its k-vector, k_{z1}. The components k_{z2} and k_{z3} can be calculated for the second and third medium using the dispersion relation and the appropriate dielectric constant, and by our convention both of these quantities are chosen to have a positive real part. The light is s-polarized. In the geometry of Fig. 1(b), E_{1R} is the amplitude of the incident plane wave at $(x = 0, z = 0)$, and the amplitudes of the four

other plane waves in the system are given by

$$E_{2R} = \frac{2k_{z1}(k_{z3}+k_{z2})E_{1R}}{(k_{z2}+k_{z1})(k_{z3}+k_{z2})+\exp(2ik_{z2}d)(k_{z3}-k_{z2})(k_{z2}-k_{z1})} \tag{10}$$

$$E_{2L} = \frac{-2k_{z1}(k_{z3}-k_{z2})E_{1R}}{(k_{z2}-k_{z1})(k_{z3}-k_{z2})+\exp(-2ik_{z2}d)(k_{z3}+k_{z2})(k_{z2}+k_{z1})} \tag{11}$$

$$E_{1L} = E_{2R} + E_{2L} - E_{1R} \tag{12}$$

$$E_{3R} = \frac{E_{2R}\exp(ik_{z2}d) + E_{2L}\exp(-ik_{z2}d)}{\exp(ik_{z3}d)}. \tag{13}$$

(These are the amplitudes in the $z = 0$ plane; to find the amplitude at another position, an appropriate phase factor must be included. The time dependence is $\exp(-i\omega t)$.) To construct the Gaussian beam from the plane wave solutions, we begin with the expression for the y-component of the E-field of an s-polarized beam traveling parallel to the z-axis, in the $z = 0$ plane

$$E_y(x, z = 0) = E_0 \exp\left(-\frac{x^2}{2\sigma_x^2}\right), \tag{14}$$

where E_0 is the peak amplitude and σ_x is directly proportional to the spatial FWHM

$$w_x = 2\sqrt{2\ln 2}\sigma_x. \tag{15}$$

By Fourier transforming and subsequently inverting the transform, the field can equivalently be written as an integral in k-space,

$$E_y(x, z) = \int_{-\infty}^{\infty} dk_x E_{1R}(k_x)\exp[i(k_x x + k_{z1} z)], \tag{16}$$

where

$$E_{1R}(k_x) = \frac{E_0\sigma_x}{\sqrt{2\pi}}\exp\left(\frac{-k_x^2}{2(1/\sigma_x)^2}\right), \tag{17}$$

and the FWHM is

$$w_k = 2\sqrt{2\ln 2}/\sigma_x. \tag{18}$$

Note that to each plane wave k_x in the integrand we have associated a k_{z1}, which allows us to propagate the field beyond the $z = 0$ cross-section. The component k_{z1} is a function of k_x and given specifically by the dispersion relation

$$k_{z1} = \sqrt{k_0^2\epsilon_1 - k_x^2}. \tag{19}$$

The dielectric constant ϵ_1 is used because the Gaussian beam is specified in medium one. At this point, we must discretize the calculation and sample the beam in the Fourier domain,

given by Eq. 17, by choosing a finite number of plane-wave samples N_s and a sampling width w_s, which we express as a multiple of the FWHM, $w_s = \alpha w_k$. It is typically sufficient to choose α equal to 2 or 3. The integral in Eq. 16 is approximated by the sum

$$E_y(x,z) = \sum_{k_x=-w_s/2}^{w_s/2} \Delta k_x E_{1R}(k_x) \exp[i(k_x x + k_{z1} z)], \tag{20}$$

where

$$\Delta k_x = \frac{w_s}{N_s - 1}. \tag{21}$$

At this point, it is helpful to think of E_{1R}, k_x, and k_{z1} as vectors containing N_s numerical elements each. To rotate the beam so that it travels at an angle θ to the z-axis, one performs the transformation

$$k_x \rightarrow \sin(\theta) k_z + \cos(\theta) k_x \tag{22}$$

$$k_z \rightarrow \cos(\theta) k_z - \sin(\theta) k_z \tag{23}$$

on each element of k_x and k_z. (The Fourier amplitude of each plane-wave $E_{1R}(k_x)$ is unaffected by the rotation in the case of s-polarized light.) Finally, to displace the waist of the Gaussian to some location (x_0, z_0) in the incidence medium, one must translate each Fourier amplitude by

$$E_{1R}(k_x) \rightarrow E_{1R}(k_x) \exp[-i(k_x x_0 + k_z z_0)]. \tag{24}$$

With these redefined values for E_{1R}, k_x, and k_z, the sum in Eq. 20 is a good approximation to a Gaussian beam traveling at an angle θ in a medium ϵ_1, whose waist w_x is centered at (x_0, y_0). The total field at any point in the system is given by

$$E_{\text{tot}}(x,z) = \begin{cases} \text{Real}\{\sum \Delta k_x \left(E_{1R}(k_x) \exp[i(k_x x + k_{z1} z)] + E_{1L}(k_x) \exp[i(k_x x - k_{z1} z)]\right)\}, & z < 0 \\ \text{Real}\{\sum \Delta k_x \left(E_{2R}(k_x) \exp[i(k_x x + k_{z2} z)] + E_{1L}(k_x) \exp[i(k_x x - k_{z2} z)]\right)\}, & 0 < z < d \\ \text{Real}\{\sum \Delta k_x E_{3R}(k_x) \exp[i(k_x x + k_{z3} z)]\}, & z > d \end{cases} \tag{25}$$

where E_{1L}, E_{2R}, E_{2L}, and E_{3R} are calculated element-wise from $E_{1R}(k_x)$ according to Eqs. 10-13. At each pixel in the simulated image, the value of the E-field is determined by the sum of plane-waves in Eq. 25, with the values of x and z indicating the location of the pixel. The resultant field is normalized to the maximum field value in the image, and displayed in color. The pulse simulation is calculated similarly, except that the field is Gaussian in

space and time, and so the field must be sampled in both spatial and temporal frequency. The calculation time is significantly longer for the pulse compared to the beam, and the simulations are only practical on a supercomputer.

The finite nature of the sampling has consequences which must be considered. The truncation of the Gaussian to the sampling width w_s in the Fourier domain leads to a convolution with a sinc function in the spatial domain; this effect is responsible for the ripples in the side-tail of the beam or pulse, as seen clearly in Fig. 5(b). Also, the finite number of samples N_s implies the spectrum of the signal is discrete, so the signal itself must be periodic. There are an infinite number of beams, spaced periodically along the x-axis by a distance $2\pi/\Delta k_x$, impinging on the slab. For example, in the beam simulations shown in Fig. 4, $\alpha = 2$ (which gives rise to the angular spread $28.3142^\circ < \theta < 31.6858^\circ$), and $N_s = 501$, which means the periodic beams are separated by $2\pi/\Delta k_x = 2830$ μm. If the sampling is increased to $N_s = 2001$, the period increases to 11330 μm, but the simulation images look identical to the ones with 501 samples. Therefore, 501 samples is sufficient in this case to ensure the beams do not interfere with each other, and the simulation is in fact a good representation of the field of a single beam. One might think that sufficient sampling could be impossible when the round-trip coefficient exceeds one, because the beam can zig-zag up the slab forever and eventually interfere with an adjacent beam, thus compromising the results. However, this is not what the simulations show. When the round-trip coefficient exceeds one, there are a limited number of pre-reflected and pre-transmitted beams, and very little energy propagates in the slab upwards of the point where the main portion of the incident beam hits the slab.

When simulating gain media, it is important to be wary of the possibility of lasing. If the gain is large enough to allow lasing, and there is a cavity for optical feedback, then the gain will clamp at 1.0, and in a real device a spontaneously emitted photon will be amplified, setting up a standing wave in the cavity. In the simulations, there is no spontaneous emission; nevertheless, if the material parameters are such that the round-trip coefficient equals 1.0 for a particular plane wave, then the reflection and transmission coefficients go to infinity, and the device radiates seemingly without being stimulated by the incident pulse. (The discrete nature of the Fourier transform used in the simulation means that the slab is in fact excited by a periodic train of pulses in time and space, and so the lasing is initiated in the simulation by one of these previous pulses, rather than by a spontaneous emission

event.) The lasing solution corresponds to a homogeneous solution of Maxwell's equations (response without a driving incident field), while the typical reflections and transmissions, which rely on the initial pulse, are the inhomogeneous solutions [13]. Still, the drawback of the simulations is that there is no gain clamping, and so one must be careful to keep ϵ_2'' below the lasing threshold. Otherwise, the effects we have discussed cannot be observed experimentally, since they will be obscured by the continuous lasing of the device. In our geometry, there is clearly an optical cavity for photons traveling in the $\pm z$ directions in medium two, with reflective facets at $z = 0$ and $z = d$. For the material parameters in Sec. III B, it can be shown that the threshold for lasing 'at normal incidence' for wavelengths near 600 nm is just above $\epsilon_2'' = -0.018$. All simulations in this paper were performed with ϵ_2'' below the threshold for lasing at normal incidence. It is also worth thinking about lasing 'at oblique incidence,' i.e. plane waves with nonzero k_x whose round-trip coefficient equals one. These exist, and give rise to a reflection coefficient of infinity. However, a photon is not a plane wave. Any photon (or pulse of finite width) with a non-zero k_x component in this geometry experiences no optical feedback, since the slab is infinite in the x-direction. (In practice, one could roughen or blacken the x-facets for a finite-size device.) Thus, it is acceptable and still physical for the round-trip gain at oblique angles to be greater than unity.

ACKNOWLEDGEMENTS

The pulse simulation in this paper was run on the Odyssey cluster supported by the Harvard FAS Research Computing Group. We thank Federico Capasso and Alexey Belyanin for helpful discussions. TSM is supported by an NSF Graduate Research Fellowship.

---

[1] C.J. Koester, "Laser action by enhanced total internal reflection," IEEE J. Quantum Electron. QE-2, 580-4 (1966)

[2] J. Fan, A. Dogariu, and L.J. Wang, "Amplified total internal reflection," Opt. Express, 11, 299-308 (2003).

[3] K.J. Willis, J.B. Schneider, and S.C. Hagness. "Amplified total internal reflection: theory, analysis, and demonstration of existence via FDTD," Opt. Express 16, 1903-14 (2008).

[4] A. Siegman, "Fresnel Reflection, and Evanescent Gain," OPN, 38–45 (Jan. 2010).

[5] G. Grynberg, A. Aspect, and C. Fabre. "Introduction to Quantum Optics." Cambridge University Press, 2010.

[6] S.A. Ramakrishna and O.J.F. Martin. "Resolving the wave vector in negative index media." Opt. Let., 30, 19, 2626-2628 (2005).

[7] J.O. Grepstad and J. Skaar. "Total internal reflection and evanescent gain." Opt. Express, 19, 21404 (2011).

[8] A. Lakhtakia, J.B. Geddes III, and T.G. Mackay. "When does the choice of the refractive index of a linear, homogeneous, isotropic, active, dielectric medium matter?" Opt. Express, 15, 17709 (2007).

[9] B. Nistad and J. Skaar. "Causality and elecromagnetic properties of active media." Phys. Rev E, 78, 036603 (2008).

[10] Y. Chen, P. Fischer, and F.W. Wise. "Negative Refraction at Optical Frequencies in Nonmagnetic Two-Component Molecular Media." Phys. Rev. Let., 95, (2005).

[11] S.A. Ramakrishna. "Comment on 'Negative Refraction at Optical Frequencies in Nonmagnetic Two-Component Molecular Media.'" Phys. Rev. Let., 98, (2007).

[12] J.B. Geddes III, T.G. Mackay, and A. Lakhtakia. "On the refractive index for a nonmagnetic two-component medium: Resolution of a controversy." Opt. Comm., 280, 120-125 (2007).

[13] H.E. Rowe. "Imperfections in Active Transmission Lines," Bell Syst. Tech. Journ., 43, 261, (1964).