THE MICROSCOPIC MECHANISM OF THIRD-ORDER OPTICAL NONLINEARITY IN InSb

D.A.B. MILLER, S.D. SMITH and B.S. WHERRY

Hertford College, Department of Physics, Reading, Birkbeck, UK

Received 23 June 1980
Revised manuscript received 17 July 1980

Microscopic models are proposed for the bandgap-resonant low intensity nonlinear refraction observed in InSb by Miller et al., predicting the self-defocusing deduced from beam patterns. The physical arguments differ from earlier treatments of third-order nonlinearity in semiconductors by explicitly including excitation and saturation effects, and can explain the exceptionally large experimental value of $\chi^{(3)} \sim 10^{-2}$ esu. The results are consistent with observations of optical bistability, two-beam differential gain and degenerate four wave mixing.

The discovery of a series of low power nonlinear optical effects in the semiconductor InSb by Miller et al. [1] has led to a macroscopic analysis, Weller et al. [2], in which it is shown that the intensity dependent refractive index $n_2$ has a value of $-6 \times 10^{-5}$ cm$^2$ W$^{-1}$ at 1886 cm$^{-1}$ near but below the bandgap of this material at 5 K. This value corresponds to a third-order susceptibility $\chi^{(3)} \sim 10^{-2}$ esu, which is several orders of magnitude larger than previously reported optical nonlinearities. This very large effect has already been exploited in devices by Miller et al. [3,4] including a nonlinear Fabry-Perot interferometer which acts as an optically bistable device, as well as a two-beam optical modulator (the "transphasor") which shows "optical transistor action" with signal gain of up to 10. In the original paper [1] we presented arguments to show that the mechanism was not thermal — viz. the thermal power requirements were one or two orders of magnitude greater than those observed and further considerations confirm this view [5]. Extensions to the experiments (with different samples and experimental conditions from ref. [2]) [4] (fig. 1) show that the effect has a strong resonance at the bandgap. The macroscopic treatment [2] of distortion of the propagating wavefront by nonlinear refraction indicates a significant difference in the near field beam patterns between the focusing the defocusing cases (in contrast to the far field); only the self-defocusing is consistent with experiment. We are therefore concerned to propose a mechanism that leads to a large, negative contribution to the nonlinear refractive index, resonating at the bandgap of a direct gap semiconductor. Four contending processes will be discussed, refraction associated with (i) the "dynamic Burstein-Moss effect" (BM), (ii) the induced free-carrier plasma (P), (iii) direct saturation of interband excitation (S), and (iv) saturation of exciton absorption. Of these (i) and (iii) are shown to

![Fig. 1. Comparison of theory for InSb (taking $E_G = 1899$ cm$^{-1}$) with experimental results [4] at 5 K.](image-url)
dominate in InSb and to give magnitudes for \( n_2 \) consistent with experiment.

The usual framework for calculation of nonlinear susceptibilities is the expansion of the electronic polarization, \( P \), in powers of the field strength, \( E(\omega) \). In the case of nonlinear refraction one has

\[
p^{(3)}(\omega) = \chi^{(3)}(\omega, \omega, -\omega, \omega) E(\omega) E(-\omega) E(\omega), \tag{1}
\]

where the monochromatic incident field is defined as

\[
E(t) = E(\omega)e^{-i\omega t} + E(-\omega)e^{i\omega t}.
\]

For sufficiently weak fields, perturbation theory [6] then gives the standard result:

\[
p^{(3)}(\omega) = \sum_{ijkl} \frac{1}{\hbar} \frac{\mu_i^p \mu_k^l \cdot E(\omega)}{\omega_{ij}^2 \omega_i \omega_j \omega_l \omega_k} \left( \frac{\omega_{ij} \omega_{kl}}{i \Gamma_{ij}^l} \right) \left( \frac{\omega_{ji} \omega_{lj}}{-i \Gamma_{ji}^l} \right) + \text{5 terms}
\]

\[
\text{corresponding to different time orderings}. \tag{2}
\]

The sum is over all electronic states \( i, j, k, l \) per unit volume; \( \mu_i^p \) is the momentum matrix element and the energy difference between states \( j \) and \( i \), \( \Gamma_{ij}^l \) is the inverse of the dephasing time of the excitation.

For frequencies well below the bandgap resonance, \( \omega < \omega_G/2 \) say, theoretical values for \( \chi^{(3)} \), calculated on the above basis, are in reasonable agreement with experiment and in the range \( 10^{-11} - 10^{-8} \text{ esu} \) for Si, GaAs, InSb, etc. [7]. The denominators in eq. (2) demonstrate resonant enhancement as \( \omega \) approaches \( \omega_G \). In order to obtain an estimate of the strength of enhancement it might be noted that for spin-flip Raman scattering in InSb (a similar form of nonlinear process that is described by \( \chi^{(3)}(\omega, \omega, -\omega, \omega) \)) the pump frequency is altered from \( \omega_G/2 \) to \( \omega_G \), \( \chi^{(3)} \) increases by a factor of order \( 10^3 \) [8,9]. Bearing the above numerical values in mind a further factor of order \( 10^3 \) is still needed in order to explain the present experimental observations.

Any proposed mechanism has also to explain a further result obtained from the nonlinear Fabry-Perot interferometer [3], namely that each successive order of the interferometer requires a successively larger change in increment of the intensity. The implication is that the nonlinear refraction is not represented merely by \( n = n_1 + n_2 I \) where \( n_1 \) and \( n_2 \) are constant, but that the nonlinearity is itself saturating (see fig. 2). This is readily explained by either of the two saturation mechanisms which we will discuss.

Let us suppose for the present that the nonlinearity is a consequence of the photo-excitation of free-carriers. Seiler and Hanes [10] have observed photoconductivity under the condition \( \omega < \omega_G \), indicating that optical excitation can occur for such frequencies. They suggest that transitions from impurity levels are responsible for the absorption. In our experiments however we observe the complete saturation of impurity effects at powers well below those of interest here [5] and therefore do not expect any intensity-dependent free-carrier concentration from this source. With impurity effects saturated, however, a substantially-linear absorption remains, increasing slowly as the laser frequency approaches the bandgap. If we ignore for the moment the discussion of the mechanism for absorption, we can calculate an upper limit to the photo-excited free-carrier population by assuming that all of the observed absorption excites electrons from the valence to conduction bands.

The bandgap for InSb at 5 K occurs at around 1900 cm\(^{-1}\). Our measured absorption coefficient, \( \alpha \), in the range 1850 cm\(^{-1}\) to 1886 cm\(^{-1}\) varies from 0.2 to 1.0 cm\(^{-1}\) for laser intensities of \( 10^2 - 10^3 \) W cm\(^2\). The corresponding rate of free-carrier generation is at most \( \alpha I/\hbar \omega \), that is between \( 6 \times 10^{20} \) to \( 3 \times 10^{22} \) carriers per cm\(^2\) per second. For a recombination time \( \tau_R \) of order a few hundred nanoseconds, as appropriate
to experimental conditions, this yields a steady state free-carrier density $N$ of between $\sim 1 \times 10^{14}$ and $\sim 6 \times 10^{15} \text{ cm}^3$ since

$$N = a \tau_R / (\hbar \omega).$$  \hspace{1cm} (3)

There are two direct contributions from such intensity-dependent free-carrier concentrations to the nonlinear refraction; (i) a dynamic Burstein-Moss shift and (ii) free carrier plasma refraction.

(i) The excitation of free-carriers has been invoked to explain the saturation of above-bandgap absorption of InSb [11,12]. In InSb the bandgap is at zero wavevector. The occupation of the lowest wavevector condition states blocks interband transitions close to bandgap. In the static case this is just the Burstein-Moss shift of the energy gap with appropriate donor or acceptor concentration. (Note that this “shift” is merely the filling of states at the “bottom” of the bands, thereby blocking the absorptive transitions between these states and shifting the wavelength of the onset of optical absorption; we do not refer to any many-body renormalization of the bandgap energy.) A simultaneous change in refractive index is also expected as the blocked transitions can no longer contribute to refraction. Thus if we use as a model conduction and heavy hole valence bands at 0 K, the change in refractive index due to blocking of states beneath a Fermi wavevector $k_F$ in the conduction band may be shown to be:

$$\Delta n = -\frac{2}{3\pi n h} \left( \frac{eP}{\hbar \omega} \right)^2 \left( \frac{2m}{\hbar} \right)^{3/2} \times (\omega_G - \omega)^{1/2} \left[ x_F - \tan^{-1} x_F \right].$$  \hspace{1cm} (4)

Here $P$ is defined conventionally as $i\hbar m^{-1} \langle S | P \rho | X \rangle$ and is the momentum matrix element between Bloch state function $S, X$ [13] and

$$x_F = \left( \frac{\hbar k_F^2}{2m} \right)^{1/2} \left( \frac{\omega_G}{\omega} \right)$$

is the Fermi wavevector in dimensionless units.

If we assume negligible initial carrier concentration, $\Delta n$ is dominated by a term proportional to the laser intensity in the case of photo-excitation because $k_F$ is proportional to $N^{1/3}$ and hence $I^{1/3}$, and $[x_F - \tan^{-1} x_F] \rightarrow x_F^3$ for small $x_F$. Hence

$$n_2(\text{B.M.}) = -\frac{2\pi}{3n} \left( \frac{eP}{\hbar \omega} \right)^2 \frac{\alpha \tau_R}{\hbar (\omega_G - \omega) \hbar \omega}.$$  \hspace{1cm} (5)

(ii) The refractive index of a free-carrier plasma is given by

$$n^2 = \left( \epsilon_{\infty} - 4\pi N e^2/m^* \omega^2 \right)$$

where $\epsilon_{\infty}$ is the optical dielectric constant and $m^*$ the carrier effective mass. Hence once again a photo-excited carrier concentration as given by equation (3) produces a nonlinear refraction, dominated by

$$n_2^\prime(P) = -\frac{2\pi}{3n} \frac{e^2}{m^* \hbar} \frac{\alpha \tau_R}{\hbar \omega}.$$  \hspace{1cm} (6)

Such a mechanism has been proposed by Jain and Klein [14] (although the excitation process is not discussed) in their interpretation of degenerate four-wave mixing in silicon. Comparing the plasma and Burstein-Moss contributions to $n_2$, and noting that $1/m^* \sim 4P^2/3\hbar^3 \omega_G$ for InSb [15] it can be seen that near resonance the Burstein-Moss term dominates, regardless of the absorption mechanism that creates the carriers:

$$n_2(\text{B.M.})/n_2^\prime(P) = \omega_G / 4(\omega_G - \omega).$$

For the carrier densities estimated above the Burstein-Moss refractive index contribution is of the order of the observed results.

(iii) We turn now to the excitation process itself. This will lead us to a slightly modified version of the Burstein-Moss index change, and to consider briefly a further source of nonlinearity. The major problem regarding excitation is that whilst the quoted work on saturation [11,12] and a plasma-induced refraction [15] has relied upon excitation at energies above the appropriate bandgap, in our case the excitation energy is less than the gap. Unfortunately, in this region just below the bandgap the mechanism for the absorption (with impurity effects already saturated) is not understood in InSb. It is, therefore, not clear that a real free-carrier population can be established, and it may not be correct to interpret eqs. (5), (6) too quantitatively.

We do, however, propose that interband transitions are responsible for the experimental results. In order to model excitation with $\omega < \omega_G$ we make a direct analogy to two-level systems, broadened by a dephasing time $T_2$. Nonlinear refraction is known to be associated with saturated absorption in a two-level system; it was proposed as early as 1966 by Javan and Kelley for the case of atomic vapours [16]. We describe the semiconductor interband excitation as a set of independent two-level systems, one for each allowed $k$-state, noting that radiative transitions couple only specific pairs of states (i.e. effectively between only states of the same $k$-value in different bands). (Complications arising from the mixed
spin nature of the light-hole valence band are ignored at present). Thus excitation into the $T_2$-broadened "band-tail" is assumed to be responsible for the nonlinear refraction.

The perturbation approach to $\chi^{(3)}$, described above, is inadequate to handle the saturation of the two-level system. It is appropriate to use density-matrix theory in order to account, in a consistent manner, for population redistribution. We have then [17]

$$P(\omega) = \sum_{i,k}^{\text{ie}} \frac{i}{m_\omega} (p_{ik} \rho_{ki}(\omega) + \rho_{ik}(\omega)p_{ki}),$$

(7)

where, employing the Schrödinger representation, $\rho_{ik}(\omega)$ indicates the $k,i$ density matrix element having frequency dependence $e^{-i\omega t}$. The diagonal elements of $\rho$ are the populations; saturation is accounted for by including in their rate equations non-radiative relaxation rates. It is just the introduction of non-radiative times $\tau_{ji}$ that distinguishes the present calculation from previous estimates of the nonlinear refractive index in solids [7].

Now for a set of uncoupled two-level systems the population redistribution can be accounted for exactly [18]. Thus, if we consider the total response to a monochromatic field,

$$P(\omega) = X(\omega,E) E(\omega),$$

(8)

where

$$X(\omega,E) = \frac{e^2}{\hbar m_\omega^2} \sum_i \left| \frac{p_{ji}^2}{(\omega_{ji} - \omega - i\Gamma_{ji})} \right|^2 \times \frac{(\omega_{ji} - \omega)^2 + \Gamma_{ji}^2}{((\omega_{ji} - \omega)^2 + \Gamma_{ji}^2) + 4 \Omega_{ji}^2 \Gamma_{ji}}$$

(9)

and $\Omega = (e/\hbar m_\omega)|p_{ji}^2| E(\omega)|$ is the Rabi flops frequency. Then, through $\Omega$, this expression for polarisation contains all odd orders of nonlinear terms. The analysis of the experimental results of present interest is in terms of the third-order susceptibility [4]. From equation (9):

$$\chi^{(3)}(\omega) = -i e^4 \sum_i \frac{p_{ji}^4}{\hbar^3} \frac{4\omega_{ji}\Gamma_{ji}}{(\omega_{ji} - \omega - i\Gamma_{ji})(\omega_{ji} - \omega)^2 + \Gamma_{ji}^2}.$$  

(10)

In particular the nonlinear refractive index, defined by $n = n_1 + n_2$, is

$$n_2(\omega) = (4\pi^2/n^2 c) \text{Re} \chi^{(3)}(\omega).$$

Following the perturbation theory technique we would have found in eq. (2) that the second energy denominator (the central one) vanishes if the states $i$ and $k$ are identical. This divergence of $P^{(3)}(\omega)$ manifests the fact that saturation has been ignored. Effectively the resonant denominator is replaced by $1/\tau_{ji}$ using density matrix theory. The presence of $\tau_{ji}$ rather than some large denominator — as seen in general frequency mixing expressions — is responsible for the high value of $\chi^{(3)}$ we ultimately calculate. In addition the right hand energy denominator will resonate for $\omega_{ji} \approx \omega$.

In the semiconductor situation of interest this occurs in particular for frequencies just below the band edge — a region in which absorption is minimal and yet dispersive effects are large. This is possible because we are dealing with a causal relation analogous to anomalous dispersion. Just as linear dispersion at a given frequency is effected by linear absorption at all frequencies, and is related to it through the Kramers-Kronig expressions, so the nonlinear refraction is associated (though not so simply) with the nonlinearity (saturation) in absorption at all frequencies.

For the present band model (where we assume no initial carrier concentration for simplicity) the sum over $i$ in eq. (9) can be reduced to an integration over wavevector $k$ and a sum over bands. For simplicity we consider dephasing times ($\Gamma_{ji}^{-1} = T_2$) and relaxation times ($\tau_{ji} = \tau$) that are independent of $k$.

We obtain

$$\chi^{(3)}(\omega) = -\hbar^{-3} \left( \frac{e^2}{\hbar \omega} \right)^3 \frac{15\pi^2}{8} \int_0^\infty dk \frac{k^2 \tau}{T_2}$$

$$\times \left| \omega_{cv}(k) - \omega - iT_2^2 \right|^2 (\omega_{cv}(k) - \omega - iT_2)^{-1}$$

(11)

$$\omega_{cv} = \omega_G + \hbar k^2/2m_e$$

is the heavy hole to conduction bandgap at wavevector $k$, and $m_e$ the reduced effective mass.

For frequencies below the bandgap, ($\omega_G - \omega \gg T_2^{-1}$), the expression simplifies to give a saturation-associated refractive index charge:

$$n_2(S) = -\frac{1}{\hbar^3} \left( \frac{e^2}{\hbar \omega} \right)^3 \frac{\tau}{T_2} \frac{2\pi}{15\pi^2 c} \left( \frac{2m_e}{\hbar^2} \right)^{3/2} (\omega_G - \omega)^{3/2}.$$  

(12)
The nonlinear refractive index so obtained is therefore negative, as required and resonates as \((\omega_G - \omega)^{-3/2}\).

In evaluating the magnitude of \(n_2\), the effective masses and momentum matrix element \(P\) are well known empirically using, for example, the linear absorption [19]. Only the exact position of the energy gap which may be influenced by many-body effects for example and the ratio of the energy relaxation time to the dephasing time are questionable parameters. Fig. 1 shows a comparison of theory and experiment [4]. Note that in order to simplify the theory we have assumed that \(T_2\), and more particularly the relaxation time \(\tau\), are independent of \(K\), and that different \(K\)-states are not coupled by the nonradiative processes. The basic resonance behaviour is obtained although experiment and theory differ on the detailed strength of the resonance. However, the order of magnitude of the nonlinearity is easily explained in these results with modest values of \(\tau/T_2\).

A similar \(T_2\) broadened level model has been used successfully in Ge in order to explain saturated absorption amongst the valence bands [20]. The direct saturation model developed above does of course also give an expression for the linear absorption, and it is useful to use this in order to compare eq. (12) with eqs. (5) and (6). This comparison emphasises the similarity of the direct saturation model with the dynamic Burstein-Moss effect (which is also a saturation model, although the saturation is now "indirect" because it is produced not directly by optical excitation but by scattering of excited carriers from other optically excited states). Under direct saturation each conduction state can be thought of as being partially occupied; there is no coupling between \(K\) states in the model. In the Burstein-Moss case an unspecified coupling such as intraband scattering (which does exist in practice) is assumed to take the excitation to the centre of the band, where a limited number of levels are considered to become fully populated (saturated), all others becoming empty. The same net occupancy of the conduction band must be present in the two cases if we assume equality of recombination \((\tau_R\) and relaxation \((\tau\) times. Eqs. (13) and (14) give the band-tail expression for the \(T_2\)-model and the resulting comparison of the three contributions to nonlinear refraction.

\[
\alpha = \frac{1}{3nc} \left( \frac{e P}{\hbar \omega} \right)^2 \left( \frac{2m_1}{\hbar} \right)^{3/2} \frac{\omega}{\hbar T_2} \left( \omega_G - \omega \right)^{-1/2},
\]

(13)

\[
n_2\text{ (B.M.)} = (5/3)n_2\text{ (S)} = (\omega_G/4(\omega_G - \omega))n_2\text{ (P)}.
\]

(14)

The Burstein-Moss effect is marginally larger than the direct saturation one because the occupied states are, on average, closer to resonance. The two descriptions are alternatives; the practical situation should lie between the two extremes. The plasma contribution on the other hand should be present for either model.

In the direct saturation model the two parameters \(T_2\) and \(\tau\) remain to be interpreted. \(T_2\) we expect to be related to the rate of intraband collisions which should randomise the quantum-mechanical wavefunctions (dephasing). In order to fit the experimental absorption coefficient at \(5\) K (\(\approx 1\) cm\(^{-1}\) at \(\sim 1886\) cm\(^{-1}\)) from eq. (13) a value of \(T_2 \sim 8 \times 10^{-11}\) s is required. This is longer than the estimated picosecond times between collisions in electrical mobility data but these electrical collisions may take place at different points in the bands from those relevant for absorption and may not be totally effective in dephasing at this low temperature. The largest possible value of \(\tau\) is the recombination time \(\tau_R\) and a value of \(\sim 500\) ns (using \(T_2 = 8 \times 10^{-11}\) s) in eq. (12) predicts the large experimental value [2] of \(n_2\) \((-6 \times 10^{-5}\) cm\(^{-1}\)). In this limit, the S and BM models are essentially the same. However, interband recombination is not the only mechanism by which the excitation of a particular "two-level" system can be effectively relaxed. After excitation, the electron (hole) can be scattered to another state inside the conduction (valence) band by the fast intraband processes thereby effectively relaxing the excitation of this particular two-level system (by transferring its population to other states). Intraband energy relaxation has been measured at about \(\sim 100\) ps [21] in InSb at \(5\) K albeit in a magnetic field. Note that this fast intraband contribution corresponds not to a relaxation of the total number of excited carriers but merely to a relaxation of the form of the distribution of the carriers within the bands, reflecting the fact that the change in refractive index is sensitive not only to the number but also the distribution of the carriers. However as it is faster it is also likely to be smaller, a fact reflected in (12) where a smaller \(\tau\) gives a smaller \(n_2\). To account more fully for this intraband contribution will require more detailed assumptions regarding scattering and absorption mechanisms.

The point of the direct saturation model is that, although it is crude, and the use of a fixed \(T_2\) value is known to give an absorption tail that is too strong at
frequencies well away from the band edge it does satisfy all the present requirements for \( n_2 \); namely (i) a resonant, negative \( n_2 \), (ii) a saturation of this nonlinearity at high intensity and (iii) a mechanism for absorption, which is essential even if the Burstein-Moss effect is to be invoked.

It is necessary to eliminate one other possible source of the nonlinearity, which has been used to explain a similar effect observed in GaAs [22,23]: saturation of excitonic absorption between discrete levels. (iv) Gibbs et al. consider the total number of excitons, essentially one per orbit volume, that can be generated in GaAs. Treating them as two-level systems of equal excitation energy one obtains saturation and hence nonlinear refraction of empirically exactly the same form as for atoms [19]. Whilst in the large gap material GaAs excitonic saturation might be expected to dominate band saturation, which we note should in principle still be present, the opposite is expected in InSb. Because of the small conduction band effective mass in InSb the excitonic Bohr radius is particularly large and the excitons are screened at low free-carrier densities, \( \sim 1.5 \times 10^{14} \) cm\(^{-3} \), compared to \( 2 \times 10^{16} \) cm\(^{-3} \) for GaAs. With an experimental value of the uncompensated impurity concentration \( N_D - N_A \sim 4 \times 10^{14} \) cm\(^{-3} \), and no known linear optical observation of excitons in InSb we are therefore justified in ignoring these excitations.

Despite some gross approximations, the simple theory we have proposed for the resonant nonlinear refractive index in InSb is in reasonable agreement with both the magnitude and resonance behaviour observed. By considering the consequences of saturation we have calculated values of third-order nonlinear susceptibility \( \chi^{(3)}(\omega; \omega, -\omega, \omega) \) to be more than five orders of magnitude greater than those previously predicted and of sign and magnitude consistent with the observed self-defocusing, optical bistability, two-beam differential gain and degenerate four-wave mixing [24].

The generality of the concept of interband saturation-induced nonlinear refraction introduced in this paper implies that similar effects should be observable in other semiconductor materials and at higher temperatures (indeed similar nonlinear refraction has already been observed in InSb at 77 K [11]). Given the current interest in nonlinear refraction for phase conjugation and optical bistability, the large effects predicted from this model may be of considerable practical importance.

References