Anomalous size dependence of optical nonlinearities due to excitonic coherence

2004 J. Phys.: Condens. Matter 16 R247

(http://iopscience.iop.org/0953-8984/16/8/R01)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **16** (2004) R247–R273 PII: S0953-8984(04)37618-6

TOPICAL REVIEW

Anomalous size dependence of optical nonlinearities due to excitonic coherence

H Ishihara

Department of Materials Engineering Science, Osaka University and CREST of the Japan Science and Technology Corporation (JST), Toyonaka, Osaka 560-8531, Japan

E-mail: ishi@mp.es.osaka-u.ac.jp

Received 4 December 2003
Published 13 February 2004
Online at stacks.iop.org/JPhysCM/16/R247 (DOI: 10.1088/0953-8984/16/8/R01)

Abstract

Theoretical studies of size dependent optical nonlinearities arising from excitonic coherence are reviewed in this article. In particular, we concentrate our attention on the phenomena in the size region beyond the long wavelength approximation (LWA) regime, where the peculiar size-resonant behaviour of the internal radiation field with nanoscale spatial structures causes anomalous size dependence of the nonlinear response. On the basis of microscopic nonlocal response theory, we demonstrate a size-resonant enhancement of the nonlinear signal for several types of nonlinear process, where the interplay between the spatial structures of the internal field and the wavefunction of confined excitons plays an essential role. The predicted effects have been observed in experiments, which shows that the mechanism of size dependence beyond the LWA regime could be a new guiding principle for developing highly efficient optical nonlinear materials and devices.

Contents

1. Introduction

One of the attractive aspects of the optical properties of nanostructures is their remarkable size dependence. Research in this area could provide a new degree of freedom for designing highly functional photomaterials with modern nanofabrication technologies. Recent studies of nanostructures have revealed a lot of peculiar optical effects not observed in bulk systems [1]. Many of them cannot be described with the conventional optical response theory based on the macroscopic treatment of the matter system. For a proper description, a microscopic treatment is essential: the quantum mechanical structures of the matter system must be properly incorporated into the theory. Coherence of the matter system plays a particularly important role in nanostructures because the electronic wavefunctions can be extended over the whole volume of the sample. This causes size quantization, and hence the size dependence of the discrete energy spectrum in dielectric functions, which makes the optical response of nanostructures quite different from the bulk one. The change of the spatial structure of the wavefunctions with the sample structure also introduces new elements in the optical response. The oscillator strength, which is a standard measure of the radiation–matter coupling strength in limited conditions, is generally determined through the spatial structures of the electronic wavefunctions, and it shows explicit sample structure dependence in the nanoscale size regime. One such effect, an enhancement of the oscillator strength in low dimensional systems, has been well studied. For two-dimensional excitons, for example, it has been considered that the excitonic binding energy becomes larger than the bulk one [2, 3], and so is the oscillator strength due to the shrinkage of the electron–hole relative wavefunction. A similar effect has been considered for quantum dots (zero-dimensional system), where the matrix element of the transition dipole moment becomes larger and the oscillator strength increases with decrease of the sample size [4, 5]. This size regime is called the strong confinement regime.

If the system size is much larger than the spatial extent of electron–hole relative wavefunctions (excitonic Bohr radius), the picture of the confinement of the excitonic centre-of-mass (c.m.) motion is appropriate for describing the optical processes [6]. This size regime is called the weak confinement regime. Since it is possible for the coherent length of the c.m. wavefunction to become very long, depending on the quality of the sample, peculiar size effects appear over a wide range of size. In the late 1980s, arguments on the size-linear enhancement of the third-order nonlinear susceptibility $\chi^{(3)}$ and the radiative decay rate [9, 10] attracted many researchers because they seem to be very favourable effects for developing highly effective optical devices. These effects are not due to the shrinkage of the e–h relative wavefunction, but due to the coherent extension of the c.m. wavefunction of excitons, whose basis lies in an early paper discussing the giant oscillator strength effect [7, 8]. That is, the oscillator strength is size-linearly enhanced as long as the c.m. wavefunction is coherent over the whole volume of the sample, and the same behaviour is expected for $\chi^{(3)}$ and the radiative decay rate. Although this picture seemed to be clear in a limited size region, there were ambiguous points such as how and at what size the enhancement is saturated, and how the values of $\chi^{(3)}$ and the radiative decay rate are connected to those of the macroscopic sample.

As mentioned above, the size dependences of the radiative decay rate and $\chi^{(3)}$ are usually discussed in terms of a common quantity, i.e., the oscillator strength. However, the saturation of their size-linear enhancements is based on different physical mechanisms. In the calculation

of $\chi^{(3)}$ by means of the perturbation expansion of the density matrix, there arise two types of term, i.e., the terms including only one-exciton states and those including two-exciton states. If we note the contributions of only the former type of terms, they show size-linear enhancement. When the size is so small that the level spacing between the quantized excitonic states is wider than the peak width, this enhancement is substantial. However, this enhancement should be saturated at a certain size. The saturation mechanism of this enhancement has been elucidated by a theoretical demonstration with a particular excitonic model (non-interacting Frenkel excitons on a one-dimensional chain with the size N) that enables us to derive an analytical expression for $\chi^{(3)}$ with a rigorous treatment [11, 13, 12, 14]. The calculated result shows that the cancellation between the contributions of the one-exciton states and those of the scattering two-exciton states begins to occur when the size becomes so large that the level spacing is narrower than the peak width, which leads to saturation of the size-linear enhancement. The maximum value of $\chi^{(3)}$ and the size where the enhancement is saturated (N_s) are determined by the inter-site transfer energy and the nonradiative damping constant. We should remark that these demonstrations have been carried out under the long wavelength approximation (LWA); that is, N_s is assumed to be much smaller than the scale of the spatial variation of the radiation field. If N_s is larger than the wavelength of the resonant light, the other mechanism should work to break the size-linear dependence before the enhancement is saturated with the cancellation mechanism, which will be mentioned below.

The argument of the size-linear enhancement of the radiative decay rate based on the same behaviour of the oscillator strength is also valid only in the LWA because the oscillator strength is defined only under the LWA. If the wavefunction of the resonant state is localized in a small volume where the spatial structure of the radiation field can be neglected, the LWA is a good approximation. However, this condition is not always satisfied because the coherent length of the c.m. wavefunction of excitons can be very long. As an effect of the spatial variation of the radiation field, a saturation of the size-linear enhancement of the radiative width has been theoretically demonstrated for one-dimensional chains consisting of two-level fine particles [15] and atoms [16]. In these demonstrations, it has been shown that an enhancement and its saturation of the higher excitonic levels follow the same behaviour as the lower level, one after another, as size increases. The picture of the oscillator strength can no longer be used to describe such behaviour of the higher excitonic levels because undoubtedly the interplay between the spatial structure of the radiation field and that of the extended excitonic wavefunctions plays an essential role. In such a case, the radiative width should be discussed as a generalized measure of the radiation–matter coupling [15]. As explained later (with the explicit expression for the interaction between induced polarizations, (8)), the light wavelength normalized by the refractive index (from the nonresonant contribution) should be compared with the excitonic coherent length for an examination of the validity of the LWA.

In the description of the linear and nonlinear polarizations, on the other hand, breakdown of the LWA is induced by the spatial variation of the self-consistent Maxwell field in the medium. In the first-principles treatment based on the quantum mechanical method, the induced polarization at a position $\mathbf{r}$ in the linear response, for example, is written in a nonlocal form, i.e.,

$$\boldsymbol{P}(\boldsymbol{r}) = \int \chi(\boldsymbol{r}, \boldsymbol{r}')\boldsymbol{E}(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}', \tag{1}$$

where χ is the linear susceptibility, $\boldsymbol{E}(\boldsymbol{r})$ is the electric field at position $\boldsymbol{r}$. If the scale of the spatial variation of $\boldsymbol{E}(\boldsymbol{r})$ is much longer than the coherent length of the resonance excited state, the LWA can be introduced by taking $\boldsymbol{E}(\boldsymbol{r})$ outside the integral sign. However, the validity of this condition is not self-evident because $\boldsymbol{E}(\boldsymbol{r})$ should be determined self-consistently with the induced polarization. Actually, it has been clarified that the LWA fails in an unexpected small

size region [17]. Induced polarizations generally have spatial structures similar to those of the wavefunctions of the resonant states owing to the nonlocal response and, hence, the response field also has nanoscale spatial structures reflecting those of the induced polarizations. In nanostructures confining excitons, a particular spatial pattern can be dominant according to the conditions of energy and size. In such a situation, the resonant behaviour of the internal field strongly affects the size dependence of the nonlinear response. Although a particular aspect of the resonant behaviour of the internal field has been studied as a problem of the interference of the exciton–polaritons in a thin film, little attention has been paid to its influence on the nonlinear response. Recently, we have theoretically proposed a new type of nonlinear response where the peculiar spectral and spatial structure of the internal field brings about its anomalous size dependence [17, 18], demonstrating nonlocal calculations for the pump–probe spectroscopy and the degenerate four-wave mixing (DFWM). A particularly noteworthy effect shown in these demonstrations is the size-resonant enhancement of nonlinear signals from the non-dipole-type excitonic states. This effect is the first manifestation of the nonlocality appearing in the nonlinear response. It is obvious that a picture based on a more general quantity than the oscillator strength is absolutely necessary for describing this phenomenon. As explained later, the components of the induced polarization associated with the respective quantized excitons and the radiative corrections of these excitons are fundamental quantities for understanding the nonlinear response beyond the LWA.

The above theoretical prediction has been examined in more recent experiments on DFWM with very high quality GaAs heterostructures [19, 20]. Systematic measurements of the thickness dependent DFWM signal have revealed a size-resonant enhancement of the DFWM from a non-dipole-type excitonic state for a particular thickness region. One of the interesting points of this nonlinear response is that the enhanced signal shows a very short radiative decay time: a couple of picoseconds [21, 22]. This phenomenon is also directly related to the breakdown of the LWA. In the LWA, the lowest excited state with the nodeless wavefunction always has the largest oscillator strength, whereas the second confined exciton with one node in its wavefunction exhibits the fastest radiative decay in the above experiment. The analysis of the exciton–radiation coupled modes with consideration of the spatial structures of both the radiation field and the excitonic c.m. wavefunction has clarified the peculiar size dependence of the radiative width of confined excitons in the size region beyond the LWA, which is not a size-linear dependence but size-resonant behaviour. In the case of GaAs thin layers, the thickness where the radiative width of the second excitonic level takes its maximum value is similar to the optimum thickness for the nonlinear response. This is very attractive from the applications point of view because it would provide a solution as regards how to develop efficient nonlinear devices combined with an ultrashort response time, going against the usual understanding of the trade-off between nonlinearity and fast response.

The purpose of this article is to review the recent studies of an anomalous size dependence of the nonlinear responses originating from long range excitonic coherence. In particular, we focus on the effects where the interplay between the spatial structure of the radiation field and that of the excitonic wavefunctions plays an essential role. Such a situation arises in the size region beyond the LWA and its study has been insufficient so far because the conventional optical response theory usually neglects the microscopic spatial structure of the radiation field. In a series of studies, we have made full use of the microscopic and nonlocal response theory to clarify the nonlinear processes peculiar to nanostructures.

This review is organized as follows. In section 2, the usual treatment of the LWA in analyses of optical properties of nanostructures is reviewed and the size dependence under the LWA is discussed. It is helpful for understanding what physics emerges from the generalized treatment beyond the LWA if we see how the approximated expressions are derived from the

general ones. The theoretical framework used to calculate the nonlinear response in a nonlocal way is explained in section 3. In section 4, we give several theoretical results showing the possibility of anomalous size dependence of the nonlinear response. Experimental results corresponding to the theoretical proposals are presented in sections 5 and 6. A summary and conclusion are given in section 7.

2. Size dependence of the optical response under the LWA

For a unified description of the optical response from the microscopic to the macroscopic system, we should start from the nonlocal expression for the induced polarization. That is, the resonant part of the linear polarization, for example, should be

$$\boldsymbol{P}(\boldsymbol{r},\omega) = \int \mathrm{d}\boldsymbol{r}'\, \chi(\boldsymbol{r},\boldsymbol{r}';\omega)\boldsymbol{E}(\boldsymbol{r}',\omega) \tag{2}$$

where

$$\chi(\boldsymbol{r},\boldsymbol{r}';\omega) = \sum_{\lambda} \bar{\chi}_{\lambda}^{(1)}(\omega)\boldsymbol{\rho}_{0\lambda}(\boldsymbol{r})\boldsymbol{\rho}_{\lambda 0}(\boldsymbol{r}') \tag{3}$$

and

$$\bar{\chi}_{\lambda}^{(1)} = 1/(E_{\lambda} - \hbar\omega - \mathrm{i}\gamma), \tag{4}$$

$$\boldsymbol{\rho}_{\lambda 0}(\boldsymbol{r}) = \langle\lambda|\hat{\boldsymbol{d}}(\boldsymbol{r})|0\rangle. \tag{5}$$

In these expressions, $\chi(\boldsymbol{r},\boldsymbol{r}';\omega)$ is the linear susceptibility, $\boldsymbol{E}(\boldsymbol{r},\omega)$ is the electric field at point $\boldsymbol{r}$ with the frequency ω, $\{E_{\lambda}\}$ are the eigenenergies of the eigenstates $\{|\lambda\rangle\}$ of the unperturbed system. In (4), γ is a positive infinitesimal value representing the adiabatic switching of the radiation–matter interaction. When we treat scattering mechanisms phenomenologically, γ is taken to be a positive finite value. In (5), $\hat{\boldsymbol{d}}(\boldsymbol{r})$ is the dipole density operator defined at the microscopic position $\boldsymbol{r}$. From the expressions (3)–(5), it is understood that if both the positions $\mathbf{r}$ and $\mathbf{r}'$ are within the coherent volume of a resonant state $|\lambda\rangle$, there arises a quantum mechanical correlation between these positions. If this coherent length l_{e} is much shorter than the light wavelength l, i.e., $l_{\mathrm{e}} \ll l$, we can take $\boldsymbol{E}(\boldsymbol{r})$ outside the integral sign. That is, we can rewrite equation (3), with some averaging process, as in the following form:

$$\boldsymbol{P}(\omega) = \chi(\omega)\boldsymbol{E}(\omega), \tag{6}$$

where

$$\chi(\omega) = \sum_{\lambda} \frac{C|\int \mathrm{d}\boldsymbol{r}\, \boldsymbol{\rho}_{0\lambda}(\boldsymbol{r})|^2}{E_{\lambda} - \hbar\omega - \mathrm{i}\gamma} \tag{7}$$

and C is a constant of proportionality. If the system size is macroscopic, P and E should have macroscopic position dependence. In the case where the resonant states are the excitons with the degree of freedom of the c.m. motion, $\boldsymbol{\rho}_{0\lambda}(\boldsymbol{r})$ is proportional to the c.m. wavefunction and the numerator of the rhs in equation (7) includes the well known form of the oscillator strength of excitons [8]. In this case, C includes the effect of the relative wavefunction and an electric dipole matrix element. For the sample size larger or smaller than the coherent length of the excitons, the integral volume in the numerator of the rhs in (7) should be small so that the internal field can be regarded as uniform to a good approximation. When the coherent length of excitonic c.m. motion is limited by the sample size, it is obvious that the oscillator strength appearing in (7) has size dependence. From the expression (7), it is understood that the oscillator strength and the uncoupled excitonic levels in the denominator determine the size dependent spectrum of the optical response. As for the third-order nonlinearity, the square of

the oscillator strength appears in the numerator of the nonlinear susceptibility and, hence, the susceptibility per volume shows explicit size dependence in the LWA, as shown later.

The oscillator strength also represents the radiative decay rate in the LWA. As we explain in the next section, the generalized measure of the radiation–matter coupling strength is the radiative width that is evaluated basically from the retarded interaction between the induced polarizations expressed as

$$A_{\lambda\lambda'} \equiv -4\pi q^2 \int\int \mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{r}'\,\boldsymbol{\rho}_{\lambda 0}(\boldsymbol{r}) \cdot \bar{\boldsymbol{G}}^{(\mathrm{T})}(\boldsymbol{r}, \boldsymbol{r}'; \omega) \cdot \boldsymbol{\rho}_{0\lambda'}(\boldsymbol{r}'), \tag{8}$$

where $\bar{\boldsymbol{G}}^{(\mathrm{T})}(\boldsymbol{r}, \boldsymbol{r}'; \omega)$ is the Green function that produces a transverse field. The explicit form of $\bar{\boldsymbol{G}}^{(\mathrm{T})}(\boldsymbol{r}, \boldsymbol{r}'; \omega)$ for free space is

$$\bar{\boldsymbol{G}}^{(\mathrm{T})}(\boldsymbol{r}, \boldsymbol{r}'; \omega) = G_q(\boldsymbol{r} - \boldsymbol{r}')\bar{\boldsymbol{I}} + \frac{1}{q^2 n}[G_q(\boldsymbol{r} - \boldsymbol{r}') - G_0(\boldsymbol{r} - \boldsymbol{r}')]\nabla'\nabla', \tag{9}$$

$$G_q(\boldsymbol{r}) = \frac{\mathrm{e}^{\mathrm{i}q\sqrt{n}|\boldsymbol{r}|}}{|\boldsymbol{r}|}, \tag{10}$$

where $\bar{\boldsymbol{I}}$ is the unit dyad, $q = \omega/c$ with c the light velocity, and n is the refractive index from the nonresonant contribution. If n has position dependence, the additional terms arising from the reflection at the boundaries between the regions with the different n values are added depending on the geometry of the sample. In the limited case where the off-diagonal elements of the retarded interaction can be neglected, the imaginary part of $A_{\lambda\lambda}$ corresponds to the radiative width of the λ state. As we can see in (8), this quantity includes information on the spatial structure of the electronic systems through $\boldsymbol{\rho}_{\lambda 0}(\boldsymbol{r})$ and information on the radiation field through $\bar{\boldsymbol{G}}^{(\mathrm{T})}(\boldsymbol{r}, \boldsymbol{r}'; \omega)$. If we neglect the spatial variation in $\bar{\boldsymbol{G}}^{(\mathrm{T})}$, this quantity also becomes proportional to $C|\int \mathrm{d}\boldsymbol{r}\,\boldsymbol{\rho}_{\lambda}(\boldsymbol{r})|^2$. In this way, the size dependences of the nonlinear response and the radiative decay rate are both discussed in terms of the oscillator strength in the LWA.

To see the size effect of the oscillator strength, let us consider the Frenkel excitons as a simple example of an excitonic model, whose eigenstate is

$$|K\rangle = \frac{1}{\sqrt{N}} \sum_n \mathrm{e}^{\mathrm{i}Kn} b_n^\dagger |0\rangle, \tag{11}$$

for

$$K = 2\bar{n}\pi/N \qquad (\bar{n} = 1, 2, \ldots, N), \tag{12}$$

where $b_n^\dagger$ is the creation operator for the exciton at site n, and N is the number of sites. If we take the dipole operator of the exciton as

$$\hat{d} = \sum_n (M b_n + M^* b_n^\dagger), \tag{13}$$

the transition dipole moment of the Frenkel exciton (11) in the LWA is

$$\langle 0|\hat{P}|K\rangle = \sqrt{N} M \delta_{K0}, \tag{14}$$

where M is the transition dipole moment of an atom (molecule). In this way, the dipole moment shows a macroscopic enhancement by the factor $\sqrt{N}$. This point holds also for the Wannier exciton, though there is a difference in the factor due to the additional freedom of internal motion. Because of this property of the exciton, the oscillator strength for the lowest state increases with the sample size as long as this state is coherent over the whole system. This effect is called the giant oscillator strength effect, which is based on basically the same idea as that for the bound excitons [7, 8]. According to this mechanism, it is possible to get a very large oscillator strength by controlling the system so that the oscillator strength is concentrated

in the lowest excitonic level. The size-linear enhancement of the radiative decay rate and the third-order nonlinear susceptibility $\chi^{(3)}$ were proposed by using this argument [9, 10]. As for the latter case, a size-linear enhancement of $\chi^{(3)}$ per unit volume was expected from the fact that each term of $\chi^{(3)}$ in the perturbation expansion is proportional to the square of the oscillator strength in the LWA.

These theoretical proposals seem attractive and some experimental results have shown a certain size enhancement of the nonlinear response [23, 24]. However, some ambiguous points exist in these arguments. Although the above size-linear enhancement was attributed to excitonic coherence, the size-linear factor appears also in a local two-level system in fact; that is, if we calculate the contribution to $\chi^{(3)}$ only from the one-atom excited states, it contains a factor N [13]. Because of the local nature of the system, the factor N cannot be related to the spatial extent of the excitonic wavefunctions. This difficulty can be eliminated by considering the cancellation between the terms including the ground state as the intermediate state and those including the two-exciton state. If double excitation at one site is allowed, all terms are cancelled out, which agrees with the fact that a pure bosonic system does not have nonlinearity. For the excitonic system, some degrees of freedom escape from the cancellation because of the Pauli exclusion principle prohibiting double excitation at one site. The remaining terms correspond to $\chi^{(3)}$ which does not have unphysical size dependence. In relation to this problem, Banyai *et al* [25] demonstrated the complete cancellation between the above two types of term when excitons are treated as pure bosons without energy dispersion. They discussed the size effect when the excitons deviate from the pure boson system, introducing biexciton states. After this review appeared, demonstrations of the cancellation effect of the excitonic system with energy dispersion were carried out by using a one-dimensional Frenkel excitonic system [11, 13], and the mechanism of the size-linear enhancement of $\chi^{(3)}$ and its saturation have been clarified in the LWA [12, 14]. That is, the remaining terms after cancellation contain the contributions of the one-exciton resonance and those of the two-exciton resonance, and they show size-linear enhancement at the one-exciton resonance energy when the system size is so small that each contribution is isolated energetically. However, as the size increases further, this enhancement is saturated because of the cancellation of the two contributions. The size region where $\chi^{(3)}$ shows the enhancement is determined by the transfer energy and nonradiative damping which are the elements related to the coherent nature of the excitons. Although this argument is very clear, we should keep it in mind that it is limited to the size regime where the LWA is valid. If the scale of the spatial variation of the internal field is smaller than N_s, a nonlocal theory is necessary to properly describe the size dependence of the nonlinear response.

In parenthesis, it is remarked that the cancellation behaviour is much influenced by the level scheme of the two-exciton states, and our recent studies have elucidated that the size dependence of $\chi^{(3)}$ at one-exciton resonance has a strong correlation with the exciton–exciton interaction. For example, if biexciton states are formed and split off from the band of unbound two-exciton states, the part escaping from the cancellation increases near the one-exciton resonance energy. This effect is more important for higher dimensions because a larger amount of escape from the cancellation occurs for higher dimensions. In order to avoid divergence from our topic, we refer the reader to our previous publications [13, 14, 26, 27] for detailed discussions of the size dependence of $\chi^{(3)}$ in the LWA and the cancellation problem, concentrating our attention on the size dependence beyond the LWA hereafter.

3. Nonlocal theory of nonlinear optical response

As we discussed in the previous section, in the conditions where the coherent length of excitons is longer than the wavelength of the internal field, the argument of $\chi^{(3)}$ alone is not sufficient to

explain the size dependence of the nonlinear response because the internal field also shows size-resonant behaviour. In such cases, nonlocal treatment of the optical response is indispensable.

In this section, the microscopic nonlocal response theory developed by Cho [28, 29] is explained and, then, we describe how one can extend it to the nonlinear response.

3.1. Microscopic nonlocal theory for linear response

The main task in the nonlocal response theory is to solve the Maxwell equation including the polarization expressed in the nonlocal form, namely

$$\nabla \times \nabla \times \boldsymbol{E}(\boldsymbol{r},\omega) - q^2\epsilon(\boldsymbol{r})\boldsymbol{E}(\boldsymbol{r},\omega) = 4\pi q^2 \boldsymbol{P}(\boldsymbol{r},\omega), \tag{15}$$

$$\boldsymbol{P}(\boldsymbol{r},\omega) = \int \mathrm{d}\boldsymbol{r}'\, \chi(\boldsymbol{r},\boldsymbol{r}';\omega)\boldsymbol{E}(\boldsymbol{r}',\omega), \tag{16}$$

where $\epsilon(\boldsymbol{r})$ is the background dielectric constant that generally depends on the position $\boldsymbol{r}$. The explicit form of $\chi(\boldsymbol{r},\boldsymbol{r}';\omega)$ is given in equations (3)–(5). By using a dyadic Green function, the solution of equation (15) can be written as

$$\boldsymbol{E}(\boldsymbol{r},\omega) = \boldsymbol{E}_{\mathrm{b}}(\boldsymbol{r},\omega) + 4\pi q^2 \int \mathrm{d}\boldsymbol{r}'\, \bar{\boldsymbol{G}}(\boldsymbol{r},\boldsymbol{r}',\omega)\cdot\boldsymbol{P}(\boldsymbol{r}',\omega), \tag{17}$$

where $\bar{\boldsymbol{G}}(\boldsymbol{r},\boldsymbol{r}',\omega)$ satisfies

$$\nabla \times \nabla \times \bar{\boldsymbol{G}}(\boldsymbol{r},\boldsymbol{r}',\omega) - q^2\epsilon(\boldsymbol{r})\bar{\boldsymbol{G}}(\boldsymbol{r},\boldsymbol{r}',\omega) = \bar{\boldsymbol{I}}\delta(\boldsymbol{r}-\boldsymbol{r}'), \tag{18}$$

and $\boldsymbol{E}_{\mathrm{b}}(\boldsymbol{r},\omega)$ is the field determined by just the background dielectric constant. Substituting equations (16) (and (3)) into (17), we obtain

$$\boldsymbol{E}(\boldsymbol{r},\omega) = \boldsymbol{E}_{\mathrm{b}}(\boldsymbol{r},\omega) + 4\pi q^2 \sum_{\lambda} \int\int \mathrm{d}\boldsymbol{r}'\,\boldsymbol{r}''\bar{\boldsymbol{G}}(\boldsymbol{r},\boldsymbol{r}';\omega)\cdot\frac{\boldsymbol{\rho}_{0\lambda}(\boldsymbol{r}')\boldsymbol{\rho}_{\lambda 0}(\boldsymbol{r}'')}{E_\lambda - \hbar\omega - \mathrm{i}\gamma}\cdot\boldsymbol{E}(\boldsymbol{r}'',\omega). \tag{19}$$

The key point of this theory is the separable form of the susceptibility with respect to the coordinates as shown in (3). This nature makes it possible to rewrite the equation (15) as a linear simultaneous equations system. That is, integrating both sides of equation (19) multiplied by $\boldsymbol{\rho}_{\lambda 0}(\boldsymbol{r})$, we obtain the linear equation for determining $\{X_{\lambda 0}\}$:

$$(E_\lambda - \hbar\omega - \mathrm{i}\gamma)X_{\lambda 0}(\omega) + \sum_{\lambda'} A_{\lambda\lambda'}(\omega)X_{\lambda' 0}(\omega) = X^{(0)}_{\lambda 0}(\omega), \tag{20}$$

where we define

$$A_{\lambda\lambda'}(\omega) \equiv -4\pi q^2 \int\int \mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{r}'\boldsymbol{\rho}_{\lambda 0}(\boldsymbol{r})\cdot\bar{\boldsymbol{G}}(\boldsymbol{r},\boldsymbol{r}';\omega)\cdot\boldsymbol{\rho}_{0\lambda'}(\boldsymbol{r}'), \tag{21}$$

$$X_{\lambda 0}(\omega) \equiv \frac{F_{\lambda 0}}{E_\lambda - \hbar\omega - \mathrm{i}\gamma}, \tag{22}$$

$$X^{(0)}_{\lambda 0}(\omega) \equiv \int \mathrm{d}\boldsymbol{r}\,\boldsymbol{\rho}_{\lambda 0}(\boldsymbol{r})\cdot\boldsymbol{E}_{\mathrm{b}}(\boldsymbol{r},\omega), \tag{23}$$

and

$$F_{\lambda 0}(\omega) \equiv \int \mathrm{d}\boldsymbol{r}\,\boldsymbol{\rho}_{\lambda 0}(\boldsymbol{r})\cdot\boldsymbol{E}(\boldsymbol{r},\omega). \tag{24}$$

In these definitions, $A_{\lambda\lambda'}$ is the interaction between the induced polarizations associated with the λ and λ' states, $X^{(0)}_{\lambda 0}$ is the interaction between the incident field and induced polarization, and $X_{0\lambda}$ is the amplitude of the λ component of the induced polarization. $F_{\lambda 0}$ is the λ component of the internal field when we expand it in the basis of the wavefunctions $\rho_{\lambda 0}(\boldsymbol{r})$. According to the definition of $\bar{\boldsymbol{G}}(\boldsymbol{r},\boldsymbol{r}';\omega)$, $A_{\lambda\lambda'}$ includes both the instantaneous Coulomb and retarded

interactions. Substituting the solutions of equation (20) into (19), the response field at an arbitrary place can be obtained. If we express equation (20) in a matrix form as

$$\mathbf{SX} = \mathbf{X}^{(0)}, \tag{25}$$

the roots of $\det \mathbf{S} = 0$ provide the complex eigenenergies (Ω_λ) of the radiation–matter coupled system whose real parts $\{\mathrm{Re}[\Omega_\lambda]\}$ include the radiative shift and whose imaginary parts $\{\mathrm{Im}[\Omega_\lambda]\}$ correspond to the radiative width. In the size region beyond the LWA, these quantities play a key role rather than the oscillator strength.

3.2. Microscopic nonlocal theory for the third-order nonlinear response

As for the third-order nonlinear response, $P^{(3)}$ can be calculated by means of the usual perturbation expansion of the density matrix [31], namely,

$$\boldsymbol{P}^{(3)}(\boldsymbol{r}, t) = (-\mathrm{i})^3 \int_{-\infty}^{t} \mathrm{d}t_1 \int_{-\infty}^{t_1} \mathrm{d}t_2 \int_{-\infty}^{t_2} \mathrm{d}t_3 \, \langle [[[\hat{\boldsymbol{d}}(\boldsymbol{r}, t), H'(t_1)], H'(t_2)], H'(t_3)] \rangle, \tag{26}$$

where the angular brackets mean a statistical average, $\hbar$ is taken to be unity, $\hat{\boldsymbol{d}}(\boldsymbol{r}, t)$ and $H'(t)$ are the interaction representations of the dipole density operator and the electron–radiation interaction, respectively:

$$\hat{\boldsymbol{d}}(\boldsymbol{r}, t) = \exp(\mathrm{i}H_0 t)\hat{\boldsymbol{d}}(\boldsymbol{r}) \exp(-\mathrm{i}H_0 t), \tag{27}$$

$$H'(t) = \exp(\mathrm{i}H_0 t)\left\{ -\sum_s \int \mathrm{d}\boldsymbol{r} \, \hat{\boldsymbol{d}}(\boldsymbol{r}) \cdot \boldsymbol{E}(\boldsymbol{r}, \omega_s) \exp(-\mathrm{i}\omega_s t + \gamma t) \right\} \exp(-H_0 t), \tag{28}$$

H_0 being the unperturbed Hamiltonian, $\gamma = 0^+$ the factor for adiabatic switching of the electron–radiation interaction. Decomposing the threefold commutator and carrying out the integration in (26), we get the expression

$$\begin{aligned} \boldsymbol{P}^{(3)}(\boldsymbol{r}, t) = \sum_p \sum_q \sum_s \int \mathrm{d}\boldsymbol{r}_1 \int \mathrm{d}\boldsymbol{r}_2 \int \mathrm{d}\boldsymbol{r}_3 \, \exp[-\mathrm{i}(\omega_p + \omega_q + \omega_s + 3\mathrm{i}\gamma)t] \\ \times \boldsymbol{E}(\boldsymbol{r}_1, \omega_p)\boldsymbol{E}(\boldsymbol{r}_2, \omega_q)\boldsymbol{E}(\boldsymbol{r}_3, \omega_s)\chi^{(3)}(\boldsymbol{r}, \boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3; \omega_p, \omega_q, \omega_s), \end{aligned} \tag{29}$$

where

$$\begin{aligned} \chi^{(3)}(\boldsymbol{r}, \boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3; \omega_p, \omega_q, \omega_s) = \sum_\lambda \sum_\mu \sum_\nu \Bigg(& \frac{\langle 0|\boldsymbol{d}(\boldsymbol{r})|\lambda\rangle\langle\lambda|\boldsymbol{d}(\boldsymbol{r}_1)|\mu\rangle\langle\mu|\boldsymbol{d}(\boldsymbol{r}_2)|\nu\rangle\langle\nu|\boldsymbol{d}(\boldsymbol{r}_3)|0\rangle}{(E_{\lambda 0} - \Omega_3')(E_{\mu 0} - \Omega_2')(E_{\nu 0} - \omega_s')} \\ & + \frac{\langle 0|\boldsymbol{d}(\boldsymbol{r}_3)|\lambda\rangle\langle\lambda|\boldsymbol{d}(\boldsymbol{r}_1)|\mu\rangle\langle\mu|\boldsymbol{d}(\boldsymbol{r})|\nu\rangle\langle\nu|\boldsymbol{d}(\boldsymbol{r}_2)|0\rangle}{(E_{\nu\mu} - \Omega_3')(E_{\nu\lambda} - \Omega_2')(E_{0\lambda} - \omega_s')} \\ & + \frac{\langle 0|\boldsymbol{d}(\boldsymbol{r}_3)|\lambda\rangle\langle\lambda|\boldsymbol{d}(\boldsymbol{r}_2)|\mu\rangle\langle\mu|\boldsymbol{d}(\boldsymbol{r})|\nu\rangle\langle\nu|\boldsymbol{d}(\boldsymbol{r}_1)|0\rangle}{(E_{\nu\mu} - \Omega_3')(E_{0\mu} - \Omega_2')(E_{0\lambda} - \omega_s')} \\ & + \frac{\langle 0|\boldsymbol{d}(\boldsymbol{r}_2)|\lambda\rangle\langle\lambda|\boldsymbol{d}(\boldsymbol{r}_1)|\mu\rangle\langle\mu|\boldsymbol{d}(\boldsymbol{r})|\nu\rangle\langle\nu|\boldsymbol{d}(\boldsymbol{r}_3)|0\rangle}{(E_{\nu\mu} - \Omega_3')(E_{\nu\lambda} - \Omega_2')(E_{\nu 0} - \omega_s')} \\ & + \frac{\langle 0|\boldsymbol{d}(\boldsymbol{r}_3)|\nu\rangle\langle\nu|\boldsymbol{d}(\boldsymbol{r}_2)|\mu\rangle\langle\mu|\boldsymbol{d}(\boldsymbol{r}_1)|\lambda\rangle\langle\lambda|\boldsymbol{d}(\boldsymbol{r})|0\rangle}{(E_{\lambda 0} + \Omega_3')(E_{\mu 0} + \Omega_2')(E_{\nu 0} + \omega_s')} \\ & + \frac{\langle 0|\boldsymbol{d}(\boldsymbol{r}_2)|\nu\rangle\langle\nu|\boldsymbol{d}(\boldsymbol{r})|\mu\rangle\langle\mu|\boldsymbol{d}(\boldsymbol{r}_1)|\lambda\rangle\langle\lambda|\boldsymbol{d}(\boldsymbol{r}_3)|0\rangle}{(E_{\nu\mu} + \Omega_3')(E_{\nu\lambda} + \Omega_2')(E_{0\lambda} + \omega_s')} \\ & + \frac{\langle 0|\boldsymbol{d}(\boldsymbol{r}_1)|\nu\rangle\langle\nu|\boldsymbol{d}(\boldsymbol{r})|\mu\rangle\langle\mu|\boldsymbol{d}(\boldsymbol{r}_2)|\lambda\rangle\langle\lambda|\boldsymbol{d}(\boldsymbol{r}_3)|0\rangle}{(E_{\nu\mu} + \Omega_3')(E_{0\mu} + \Omega_2')(E_{0\lambda} + \omega_s')} \\ & + \frac{\langle 0|\boldsymbol{d}(\boldsymbol{r}_3)|\nu\rangle\langle\nu|\boldsymbol{d}(\boldsymbol{r})|\mu\rangle\langle\mu|\boldsymbol{d}(\boldsymbol{r}_1)|\lambda\rangle\langle\lambda|\boldsymbol{d}(\boldsymbol{r}_2)|0\rangle}{(E_{\nu\mu} + \Omega_3')(E_{\nu\lambda} + \Omega_2')(E_{\nu 0} + \omega_s')} \Bigg). \end{aligned} \tag{30}$$

In these expressions, we assume $T = 0$ K, and

$$H_0|\xi\rangle = E_\xi|\xi\rangle \qquad (\xi = 0, \lambda, \mu, \nu), \tag{31}$$

$$E_{\xi\eta} = E_\xi - E_\eta, \tag{32}$$

$$\Omega_3' = \omega_p + \omega_q + \omega_s + 3\mathrm{i}\gamma, \tag{33}$$

$$\Omega_2' = \omega_q + \omega_s + 2\mathrm{i}\gamma, \tag{34}$$

$$\omega_s' = \omega_s + \mathrm{i}\gamma. \tag{35}$$

Here, we use the notation $A_1, A_2, \ldots, B_4$ in the order of appearance in (30). B_i $(i = 1, 2, 3, 4)$ shows the reversed order of operators compared with A_i and they are multiplied by -1.

In equation (30), $|\mu\rangle$ can be the ground state or doubly excited states. The former and latter cases are denoted as $A_i(0)$, $B_i(0)$ and $A_i(2)$, $B_i(2)$, respectively. As argued in the previous section, a cancellation between these two types of term occurs. Therefore no term should be omitted without proper examination of its contribution.

As in the case of linear response, the separable form of each term with respect to the coordinates is the key to solving the nonlinear Maxwell equation [30]. In $\chi^{(3)}$, there appear dipole density matrix elements of the form $\langle\sigma|\hat{\boldsymbol{d}}(\boldsymbol{r})|\tau\rangle$ where $\{|\sigma\rangle, |\tau\rangle\}$ is either a ground state or singly or doubly excited states. The matrix element of $\hat{\boldsymbol{d}}(\boldsymbol{r})$ connecting the one- and two-exciton states can be written as

$$\langle\lambda|\hat{\boldsymbol{d}}(\boldsymbol{r})|\mu\rangle = \boldsymbol{\rho}_{\lambda\mu}(\boldsymbol{r}). \tag{36}$$

Using this expression for the matrix element together with $\boldsymbol{\rho}_{0\lambda}(\boldsymbol{r})$ defined in (5), we can write $P^{(3)}$ with the parameters $\{F_{\sigma\tau}\}$. The solution of the Maxwell equation for a certain frequency from a particular combination of $(\omega_s, \omega_p, \omega_q)$ can be written as in (19) with the Green function $\bar{\boldsymbol{G}}(\boldsymbol{r}, \boldsymbol{r}'; \omega)$. Substituting the above solution in the definition of $F_{\sigma\tau}$, we can derive a simultaneous cubic equations system to determine a set of $\{F_{\sigma\tau}\}$.

In the nonlinear case, we have to consider a set of Maxwell equations for all the relevant frequency components of the field, and these components are coupled through the nonlinear terms. Let us suppose that there are altogether N_f frequency components of the field $(\omega_1, \omega_2, \ldots, \omega_{N_\mathrm{f}})$. In the Maxwell equation for frequency ω_{n_f}, the expansion coefficients $\{F_{\lambda 0}(\omega_{n_\mathrm{f}})\}$ appear from the first-order polarization $P^{(1)}(\boldsymbol{r})$. The contribution from $P^{(3)}(\boldsymbol{r})$ in the same equations contains the products of three expansion coefficients $\{F_{\sigma\tau}\}$. For example, the expression corresponding to $A_1(0)$ in $P^{(3)}(\boldsymbol{r})$ is

$$\boldsymbol{P}^{(3)}(\boldsymbol{r}, t)|_{A_1(0)} = -\sum_{p,q,s} \exp[-\mathrm{i}(\omega_p + \omega_q + \omega_s + 3\mathrm{i}\gamma)t] \times \sum_\lambda \sum_\nu \frac{\boldsymbol{\rho}_{0\lambda}(\boldsymbol{r}) F_{\lambda 0}(\omega_p) F_{0\nu}(\omega_q) F_{\nu 0}(\omega_s)}{(E_{\lambda 0} - \Omega_3')\Omega_2'(E_{\nu 0} - \omega_s')}, \tag{37}$$

where $\sum_{p,q,s}$ means the summation over all the combinations of $(\omega_p, \omega_q, \omega_s)$ which satisfy the condition $\omega_p + \omega_q + \omega_s = \omega_{n_\mathrm{f}}$, with ω_p, ω_q, and ω_s taken from the set $(\omega_1, \omega_2, \ldots, \omega_{N_\mathrm{f}})$.

The term corresponding to $A_1(2)$ is

$$\boldsymbol{P}^{(3)}(\boldsymbol{r}, t)|_{A_1(2)} = \sum_{p,q,s} \exp[-\mathrm{i}(\omega_p + \omega_q + \omega_s + 3\mathrm{i}\gamma)t] \times \sum_\lambda \sum_\mu \sum_\nu \frac{\boldsymbol{\rho}_{0\lambda}(\boldsymbol{r}) F_{\lambda\mu}(\omega_p) F_{\mu\nu}(\omega_q) F_{\nu 0}(\omega_s)}{(E_{\lambda 0} - \Omega_3')(E_{\mu 0} - \Omega_2')(E_{\nu 0} - \omega_s')}, \tag{38}$$

where

$$F_{\lambda\mu}(\omega_p) = \int \mathrm{d}\boldsymbol{r}\, \boldsymbol{\rho}_{\lambda\mu}(\boldsymbol{r}) \boldsymbol{E}(\boldsymbol{r}, \omega_p). \tag{39}$$

The term corresponding to $B_2(2)$ is

$$\boldsymbol{P}^{(3)}(\boldsymbol{r},t)|_{B_2(2)} = \sum_{p,q,s} \exp[-\mathrm{i}(\omega_p+\omega_q+\omega_s+3\mathrm{i}\gamma)t] \times \sum_{\lambda}\sum_{\mu}\sum_{\nu} \frac{\rho_{\nu\mu}(\boldsymbol{r})F_{0\nu}(\omega_q)F_{\mu\lambda}(\omega_p)F_{\lambda 0}(\omega_s)}{(E_{\nu\mu}+\Omega_3')(E_{\nu\lambda}+\Omega_2')(E_{0\lambda}+\omega_s')}. \tag{40}$$

Writing the remaining 13 terms in a similar way, we see that $\boldsymbol{P}^{(3)}(\boldsymbol{r})$ is written as a linear combination of the functions $\{\rho_{0\lambda}(\boldsymbol{r})\}$ ($\{\rho_{\lambda 0}(\boldsymbol{r})\}$) and $\{\rho_{\lambda\mu}(\boldsymbol{r})\}$ ($\{\rho_{\mu\lambda}(\boldsymbol{r})\}$).

In the coefficients of these functions, there appear products of three $F_{\sigma\tau}$s. If we regard $\{F_{0\lambda}(\omega_{n_\mathrm{f}})\}$ ($\{F_{\lambda 0}(\omega_{n_\mathrm{f}})\}$) and $\{F_{\lambda\mu}(\omega_{n_\mathrm{f}})\}$ ($\{F_{\mu\lambda}(\omega_{n_\mathrm{f}})\}$) as given constants, we can write the formal solutions of the Maxwell equation as in equation (19) with the Green function appropriate for the geometry considered. Substituting these solutions $\{E(\boldsymbol{r},\omega_{n_\mathrm{f}})\}$ in the definitions of the expansion coefficients (24), (39), we obtain the simultaneous cubic equations for $(N_\lambda + N_\lambda N_\mu)N_\mathrm{f}$ variables $\{F_{\sigma\tau}\}$, where N_λ and N_μ are the numbers of one- and two-exciton states. Since there are $(N_\lambda + N_\lambda N_\mu)N_\mathrm{f}$ equations, we can obtain unique solutions.

Though it may be difficult to solve these equations analytically, it is possible to solve them numerically in the case where not too large a number of states are required as the basis set. For the large system size, there are the cases where we omit the terms containing the higher levels that do not contribute in the relevant resonant energy region. As for the summation of the frequency, it can be restricted to the combinations which satisfy the resonant conditions and, actually, the calculations have been performed for particular cases of combinations of frequencies, and models of electronic systems. This method has been used for the analysis of pump–probe measurements [30, 17], four-wave mixing [32, 20], the input intensity dependence of a single incident beam [33], and so on. For the pump–probe measurements and the four-wave mixing, the Maxwell equations can be reduced to linear equations if the probe beam is assumed to be much weaker than the pump or control beam, which makes the problem very easy. On the other hand, for the analysis of the input intensity dependence for the case of a single incident beam, we have to solve the simultaneous cubic equations for unknown parameters $\{F_{\sigma\tau}\}$. The iterative method is useful for solving such cubic equations. In sweeping the input intensity, we start from an intensity near the linear response region, employing the solutions of $\{F_{\sigma\tau}\}$ for the linear response as the zeroth-order solutions, because the deviation of the convergent solutions from the linear one is not very large. After the convergent solutions are obtained, we move to a higher input intensity and perform the iterative calculation employing the previous solutions as zeroth-order solutions. Repeating such steps, we sweep in the input intensity domain. Interestingly, we can find another set of solutions by sweeping from the higher intensity to the lower, if bistable solutions are possible. It is difficult to obtain such solutions in a general way, because the simultaneous equations have many dimensions, and mathematically a large number of solutions exist. While detailed results of such calculations are not described in this review, the peculiar input–output characteristics due to nonlocality are presented in [33].

In the rest of this review, several examples of calculation for pump–probe measurements and degenerate four-wave mixing (DFWM) near excitonic resonance are shown, where the anomalous size dependence due to the excitonic coherence typically appears.

4. Anomalous size dependence of the nonlinear response beyond the LWA

In the size regime beyond the LWA, a remarkable spectral and spatial structure of the internal field strongly affects the size dependence of the nonlinear response and, thus, size-resonant behaviour appears in various types of nonlinear process, which is in contrast with the size-linear behaviour of the nonlinear susceptibility discussed in the LWA.

In the macroscopic theory of the optical response of solids, a microscopic spatial structure of the internal field is neglected, and hence it is considered that a dipole-type transition always provides the largest contribution in the elementary optical process. This, however, has been revealed to be untrue by analyses based on the microscopic response theory, especially for nanostructures [30, 32]. In the weak confinement regime, the spatial structure of the resonantly induced polarizations generally has similar patterns to those of the excitonic c.m. wavefunctions. On the other hand, the internal field is determined as a superposition of the incident light and radiated field by the induced polarizations. With such a self-consistent relationship, the spatial structure of the internal field strongly reflects those of the induced polarizations. As a result, a peculiar size dependence of optical responses is expected over a wide range of size regions. Although particular aspects of this effect have been studied as a problem of polariton interference in ultrathin films experimentally [34–36] and theoretically [37, 38, 6, 39], there has been no discussion on how this effect influences the nonlinear response. In this section, we demonstrate nonlocal calculations of the internal field within the linear response regime considering a specific model system and, then, a peculiar size dependence of the nonlinear response is discussed, especially for non-dipole-type excitonic states.

4.1. Resonant behaviour of the internal field with a nanoscale spatial structure

In order to make a smooth continuation to the model used in the nonlinear calculation, we handle the problem with a discrete lattice model, namely, a thin film consisting of N discrete layers where the c.m. motion is confined to along the surface normal direction and the relative motion is fixed. (The relative motion of excitons is reflected in the parameter of the exciton–radiation coupling per unit cell (e.g. LT splitting).) The excitation on each layer is specified by the layer number j and the 2D vector $\mathbf{k}$ in the lateral direction. The possible distortion of exciton wavefunctions near the film surfaces is neglected for simplicity (though more sophisticated treatment is possible, by considering the evanescent components of the wavefunction near the surfaces [40, 41]). Within the linear response, this model gives essentially the same result as the continuum model based on Wannier excitons with properly chosen parameter values for the transfer energy and the transition dipole moment per unit cell [42].

The unperturbed Hamiltonian of this model is

$$\mathcal{H}_0 = \sum_{j=0}^{N+1} \varepsilon_0 a_j^\dagger a_j - b \sum_{j=1}^{N+1} (a_{j-1}^\dagger a_j + a_j^\dagger a_{j-1}), \tag{41}$$

where $a_j^\dagger$ and a_j are the creation and annihilation operators of an exciton on the jth site, ε_0 is the excitation energy of each site, b is the transfer energy, and we introduce the virtual sites $j = 0$ and $N + 1$ on which the amplitude of the exciton is supposed to be zero. The lattice constant of the chain is taken to be the unit of length. The eigenvalues and eigenfunctions of one-exciton states are

$$E_1(n) = \varepsilon_0 - 2b\cos(K_n) \tag{42}$$

and

$$|K_n\rangle = \sqrt{\frac{2}{N+1}} \sum_j \sin(K_n j) a_j^\dagger |0\rangle, \tag{43}$$

respectively. The allowed values of K_n are

$$K_n = \frac{n\pi}{N+1} \qquad (n = 1, 2, \ldots, N). \tag{44}$$

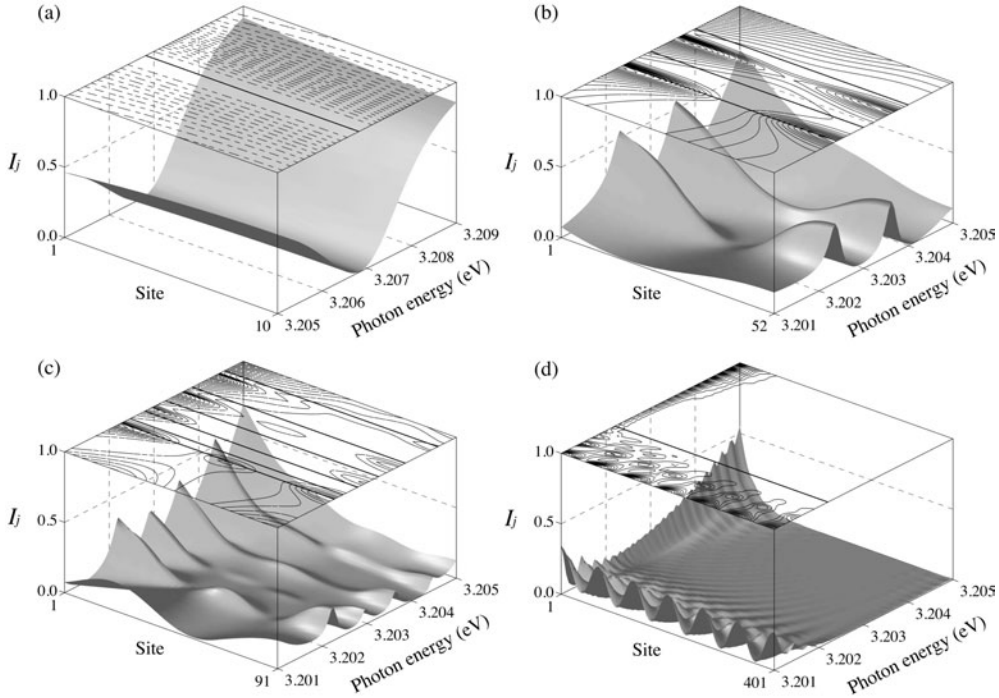

Figure 1. The intensity of the internal field I_j as a function of the discrete site j and the photon energy. (a) $N = 10$, (b) $N = 52$, (c) $N = 91$, (d) $N = 401$. $\Gamma = 0.06$ meV for every N. Contour lines of I_j and vertical lines indicating one-excitonic levels are shown at the top of each figure.

We apply this model to the calculation of the linear response by means of the microscopic nonlocal response theory explained in section 3.

As for an example of a single-component exciton with a small Bohr radius, we choose the material parameters of the Z_3 exciton of CuCl:

$$\begin{aligned} &\hbar\omega_{\mathrm{T}} = 3202.2 \text{ meV}, \qquad b = 57.0 \text{ meV}, \\ &\frac{4\pi|M|^2}{v_0} \ (\equiv \Delta_{\mathrm{LT}}) = 5.7 \text{ meV}, \qquad \varepsilon_{\mathrm{b}} = 5.6, \qquad a_0 = 5.4 \text{ Å}, \end{aligned} \tag{45}$$

where $\hbar\omega_{\mathrm{T}}$ is the transverse exciton energy, namely the energy of the bottom of the exciton band for $N \to \infty$, i.e., $\hbar\omega_{\mathrm{T}} = \varepsilon_0 - 2b$, v_0 is the unit cell volume, M is the transition dipole moment per cell, Δ_{LT} is the LT splitting energy, i.e., the separation between $\hbar\omega_{\mathrm{T}}$ and the longitudinal exciton energy, a_0 is the lattice constant.

The spatial distribution of the field intensity I_j $(=|\mathcal{E}_j|^2)$ in the film is shown in figure 1 for several values of the thickness, where $\mathcal{E}_j$ is the electric field at site j. We assume high quality of the film with the small nonradiative damping (Γ) of 0.06 meV [36, 38]. This choice means that we assume very high excitonic coherence in the sample.

In contrast with the case of $N = 10$, the spatial variation for $N = 52$ (91) near the lowest excitonic level is remarkable, reflecting the spatial pattern of induced polarization associated with the $n = 2$ ($n = 3$) excitonic state. An enhancement of the internal field at a particular size (a particular energy) can be regarded as an effect due to a nanoscale Fabry–Perot interference. For the thicker case ($N = 401$), the damping effect becomes dominant and the effect of interference between the transverse and longitudinal excitonic energies is not conspicuous.

In the case where the nonlocal effect is important, it is useful to express the nonlinear polarization using the quantities $F_{\sigma\tau}$ as in (37)–(40) and to observe the size-resonant behaviours of F_{0n} (or X_{0n}) rather than $\mathcal{E}_j$ itself. Note that F_{0n} means the amplitude of the component related to the nth quantized exciton when we expand $\mathcal{E}_j$, for $1 \leqslant j \leqslant N$, in terms of eigenfunctions of c.m. motion of excitons in the film, namely, as

$$\mathcal{E}_j = \sum_{n=1}^{N} \sqrt{\frac{2}{N+1}} \sin\left(\frac{n\pi}{N+1} j\right) F_{0n}, \tag{46}$$

which can be understood from equation (24). In the present model, an explicit expression for X_{0n} is

$$X_{0n} = \sum_{j} \sqrt{\frac{2}{N+1}} \sin\left(\frac{n\pi}{N+1} j\right) \mathcal{E}_j (E_1(n) - \hbar\omega - \mathrm{i}\Gamma)^{-1}. \tag{47}$$

In figure 2, we show the energy dependence of $|F_{0n}|^2$ ($n = 1$, 2, and 3) and $|X_{0n}|^2$ ($n = 1$, 2, and 3). We see, for each n, that X and F are resonantly enhanced at particular energy positions. The extents of the enhancement are different for different thicknesses of the film. As we can understand from the fact that $\det \mathbf{S}$ in equation (25) becomes zero at $\{\mathrm{Re}[\Omega_n]\}$, X_{0n} is resonantly enhanced at $\{\mathrm{Re}[\Omega_n]\}$. It should be recalled that $\{\mathrm{Re}[\Omega_n]\}$ are eigenenergies of the radiation–exciton coupled system including the radiative shift. On the other hand, the energy positions where the $\{F\}$ are enhanced are not just at $\{\mathrm{Re}[\Omega_n]\}$ though they are very near $\{\mathrm{Re}[\Omega_n]\}$. The extent of the enhancement of X is determined through the competition between the enhancement of F and the off-resonance due to the radiative shift, i.e., $E_n(1) - \mathrm{Re}[\Omega_n]$, appearing in the energy denominator of X. They both increase with increase of the film thickness in the initial stage. As we see in figure 2, X_{02} shows a sharp enhancement. However, no prominent enhancement of X_{01} is seen. This is because $|E_1(1) - \mathrm{Re}[\Omega_1]|$ is much larger than $|E_2(1) - \mathrm{Re}[\Omega_2]|$ in the size region considered for the present model. (This situation depends greatly on the geometry of the sample.) It should be remarked that the above situation leads to a trade-off relationship between the magnitude of the nonlinearity and that of the radiation–matter coupling through the same relationship as between the radiative shift and the enhancement of X.

4.2. Pump–probe measurements

In this section, we consider the situation where the system is pumped at the frequency ω_2 and probed at the frequency ω_1; that is, we require $\omega_1 = \omega_p + \omega_q + \omega_s$ and pick up the contribution of the most (triply) resonant terms. If the probe beam is much weaker than the pump beam, we can make the following approximation: in Maxwell's equation for the frequency ω_2, there exist terms proportional to the cube of $F(\omega_2)$ and ones proportional to $[F(\omega_1)]^2 F(\omega_2)$, and the latter are negligible as compared with the former. Therefore, the $F(\omega_2)$s can be calculated independently from $F(\omega_1)$s. On the other hand, in the equation for the frequency ω_1, the terms proportional to the cube of $F(\omega_1)$ are negligible compared with the terms proportional to $[F(\omega_2)]^2 F(\omega_1)$. Therefore, inserting the values of the $F(\omega_2)$s into Maxwell's equation for the frequency ω_1, we obtain a linear equation for determining the $F(\omega_1)$s with a renormalized linear susceptibility due to the pumping. This approximation is valid under the usual condition of third-order pump–probe measurement, which enables us to make a remarkable saving in computing time.

In the nonlinear calculation, we introduce the term

$$\mathcal{H}_{\mathrm{ex-ex}} = -\bar{\delta} \sum_{j=0}^{N} a_j^\dagger a_{j+1}^\dagger a_j a_{j+1}, \tag{48}$$

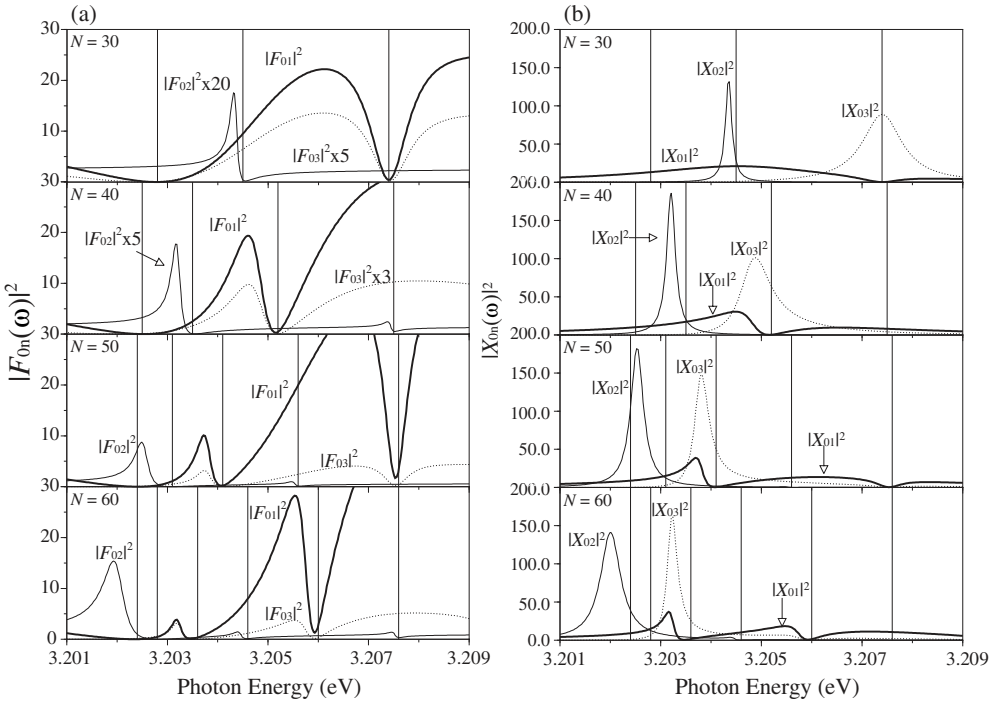

Figure 2. (a) Spectra of $|F_{0n}|^2$ ($n = 1, 2, 3$) for various thicknesses N. (b) Spectra of $|X_{0n}|^2$ ($n = 1, 2, 3$) for various thicknesses N.

i.e., the nearest neighbour exciton–exciton interaction, which allows bound two-exciton states in addition to the Hamiltonian (41). We treat $\bar{\delta}$ as a free parameter to see the effects in the presence of the biexciton resonance. If we restrict ourselves to normal incidence, the contribution to the optical response arises only from the $\mathbf{k} = 0$ one-exciton states and the $\mathbf{K} = 0$ two-exciton states consisting of $+\mathbf{k}$ and $-\mathbf{k}$ one-exciton states. In the following treatment, we consider only the contribution from the $\mathbf{k} = 0$ subspace as the two-exciton states. This assumption corresponds to the neglect of the lateral interaction between two excitons.

Since it is not easy to obtain two-exciton wavefunctions in the closed form for this model, we prepare the two-exciton states by expanding them in the products of a complete set of one-exciton wavefunctions. We take a set of $\{|m, n)\}$ $(=\{b_m^\dagger b_n^\dagger|0\rangle\})$ as a base, where the case $m = n$ is excluded. Using this base, we expand the two-exciton states as

$$|\mu\rangle = \sum_{n<m} C_{n,m}^{(\mu)} |m, n). \tag{49}$$

Inserting (41), (48), and (49) into the Schrödinger equation, we solve the eigenvalue problem, and determine $\{C_{n,m}^{(\mu)}\}$ and the eigenvalues of the two-exciton states $\{\bar{E}_\mu\}$ numerically. By using these results together with the explicit forms of the one-exciton states, the dipole matrix elements between the ground and the one-exciton states, and the one- and two-exciton states can be calculated. In this stage, we introduce the phenomenological damping constant in the usual way as in [31]—that is, the population decay constant γ, the phase decay constant Γ for between the ground state and one-exciton states, and 2Γ for between the ground state and two-exciton states. However, we do not include nonlinear damping effects such as the

excitation induced dephasing [43] because they do not have a qualitatively essential role in our mechanism considered for the weak excitation limit.

In the following numerical calculation, we consider the pump frequency to be near the lowest excitonic level $E_1(1)$, and the probe frequency scans around the energy of transition between the biexciton states and one-exciton states where the sample is transparent in the absence of the pump beam, and we see a pump induced change in the transmittance of the probe beam. Among the many terms in $P^{(3)}$, a few terms make the main contribution to the nonlinear signal in the present conditions, and they are proportional to factors that can be written as

$$\frac{(\omega_2 - \mathrm{i}\Gamma - E_1(n))|X_n(\omega_2)|^2}{(\omega_1 + \mathrm{i}\Gamma - \tilde{E}_{\mu n})(\omega_1 + \omega_2 + 2\mathrm{i}\Gamma - \bar{E}_\mu)} F_m(\omega_1) \tag{50}$$

and

$$\frac{-2\Gamma}{\gamma} \frac{|X_n(\omega_2)|^2}{(\omega_1 + \mathrm{i}\Gamma - \tilde{E}_{\mu n})} F_m(\omega_1), \tag{51}$$

where

$$\tilde{E}_{\mu n} = \bar{E}_\mu - E_1(n). \tag{52}$$

(We denote F_{0n} and X_{0n} simply as F_n and X_n hereafter.) The former is known as a term contributing to the two-photon absorption where the sum frequency of two photons (ω_1 and ω_2 in the present case) becomes identical to the energy of transition between the ground state and the two-exciton states. The latter represents the pump induced transition from the one-exciton states $|K_n\rangle$ to the biexciton states $|\mu\rangle$ caused by a probe beam. The forms of these expressions indicate that the nonlinear signal should strongly reflect the behaviour of $|X_n(\omega_2)|^2$.

The spectrum of the nonlinear signal is a function of the thickness N and the pump frequency (ω_2). First, we observe the pump frequency dependence by fixing the thickness. The pump energy scans the region where $|X_2|^2$ and $|X_3|^2$ take peak values. In this calculation, the same parameter values as in (45) are used. Besides those, we choose $\bar{\delta} = 195.0$ meV for the attractive energy of two excitons, 7.2×10^4 and 7.2×10^2 V m^{-1} for the amplitudes of the incident pump and probe beams, and 0.02 meV for the longitudinal damping constant (γ).

Figure 3 shows the change in the transmittance spectrum of the probe beam $\delta T(\omega_1)$ when $N = 50$. There appear three main structures near the constant values 3.1684 eV ($[A_1]$), 3.1679 eV ($[A_2]$), and 3.1672 eV ($[A_3]$) of the probe energy. These probe energies correspond to the $E_1(n) \to \bar{E}_\mu$ transition energies for $\mu = 1$, 2, and 3, respectively. Since the peak positions on the ω_1-axis do not depend on ω_2, we know that these signals mainly come from the terms in (51). In addition, a small peak structure ($[A_0]$) due to the two-photon absorption satisfying $\omega_1 + \omega_2 = \bar{E}_1$ can be seen. This signal comes from the terms in (50). It should be remarked that the spectral structure of X_n is strongly reflected in these nonlinear signals. Especially for the peaks indicated by $[A_1]$–$[A_3]$, the X_n spectra are directly reflected in the ω_2 dependence of the nonlinear signal; that is, the remarkable enhancements of the peaks $[A_2]$ and $[A_3]$ are caused by the enhancements of $|X_2|^2$ and $|X_3|^2$, respectively. As shown in figure 2, $|X_2|^2$ and $|X_3|^2$ are greatly enhanced around this thickness, exceeding the magnitude of $|X_1|^2$. It should also be noted that the energy positions of the pump beam where the $[A_n]$ peak values take their maxima are shifted from the uncoupled excitonic levels $E_1(n)$ due to their radiative shifts. The reason that the enhanced $|X_n|^2$ leads to an induced transition from $E_1(n)$ to $\bar{E}_n$, not to $\bar{E}_m$ ($m \neq n$), is that the probe beam energy ω_1 is in the transparent region in the linear response regime. This means dominance of the uniform component in its spatial variation.

The effect of the resonant enhancement of $|X_n|^2$ also leads to anomalous size dependence of $\delta T(\omega_1)$. In figure 4, the N dependence of $\delta T(\omega_1)$ is given, where we fix the pump energy

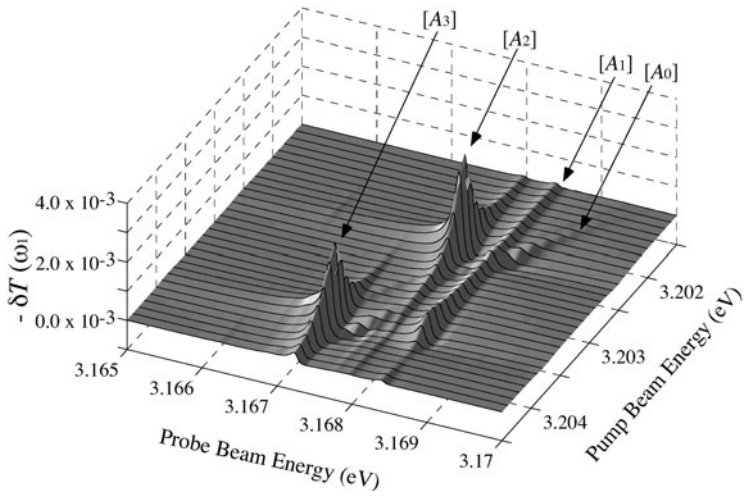

Figure 3. The pump induced change in the transmittance $(-\delta T)$ as a function of ω_1 and ω_2. $N = 50$, $(\Gamma, \gamma) = (0.06, 0.02)$ meV

at the lowest excitonic level for each thickness, i.e., $\omega_2 = E_1(1)$. Two distinct peak structures and an additional small structure can be seen in figure 4(a). We denote them by $[B_0]$, $[B_1]$ and $[B_2]$, as shown in the figure. Their peak positions in energy are $\omega_1 = \bar{E}_2 - E_1(1)$ for $[B_0]$, $\omega_1 = \bar{E}_1 - E_1(1)$ for $[B_1]$, and $\omega_1 = \bar{E}_2 - E_1(2)$ for $[B_2]$. The small signal $[B_0]$ which is conspicuous around $N = 50$ is due to the two-photon absorption where the total energy of the pump and the probe beam is identical to the eigenenergy of the second biexciton level $\bar{E}_2$. The signal $[B_1]$, whose size dependence is moderate, is due to two kinds of superimposed nonlinear process, namely pump induced absorption by biexcitons where the transition between $E_1(1)$ and $\bar{E}_1$ occurs, and two-photon absorption where the total energy of the pump and the probe beam is identical to $\bar{E}_1$. Thus, the signal $[B_1]$ contains the both processes, terms (50) and (51), and both of them always go through a complete resonance with respect to all three factors in the denominators. In contrast with these signals, the signal $[B_2]$ shows a remarkable size dependence, and the maximum of the peak very much exceeds that of signal $[B_1]$ around size $N = 52$, though the energy denominator of this process (51) does not show complete resonance. The enhancement of this nonlinear signal $[B_2]$ is due to the size-resonant enhancement of $|X_2|^2$, and the size dependence of $[B_2]$ clearly reflects that of $|X_2|^2$ in figure 2. At $N = 52$, Ω_2 becomes equal to $E_1(1)$ and $|X_2|^2$ is enhanced at this energy. Since the pump beam is tuned to $E_1(1)$, an enhancement of the non-dipole-type spatial pattern relating to $|K_2\rangle$ in the induced polarization and internal field occurs.

The result for the larger damping constants, $(\gamma, \Gamma) = (0.2, 0.6)$ meV, is shown in figure 4(b). The distinction between $[B_1]$ and $[B_2]$ is not very clear in this case and the size-resonant enhancement of $[B_2]$ cannot be seen. This means that coherence of the excitonic wavefunction is essential for the emergence of size-resonant enhancement of the nonlinear signal.

5. Observation of anomalous size dependence of the DFWM signal

A similar enhancement effect to that explained in the above section has been predicted also for the DFWM [44, 32]. In the presence of the pump beam with the frequency and wavevector

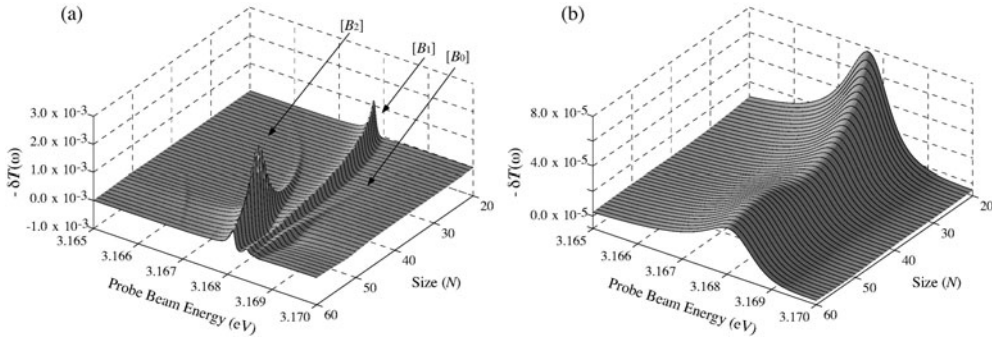

Figure 4. The pump induced change in the transmittance $(-\delta T)$ as a function of ω_1 and thickness N. $\omega_2 = E_1(1)$. (a) $(\Gamma, \gamma) = (0.06, 0.02)$ meV, (b) $(\Gamma, \gamma) = (0.6, 0.2)$ meV.

$(\omega_2, \mathbf{k}_2)$ and the test beam with $(\omega_1, \mathbf{k}_1)$, there arises a signal with $(\omega_s = 2\omega_2 - \omega_1, \mathbf{k}_s = 2\mathbf{k}_2 - \mathbf{k}_1)$. This process is called four-wave mixing (FWM). In the degenerate case, that is $\omega_2 = \omega_1$, the signal appears as a phase conjugate of the test beam. This nonlinear signal is observed in the direction where no signal exists within the linear response; hence it is easier to observe the signal as compared with the pump induced change of the transmittance. One more attractive point as regards studying this nonlinear process is that an enhancement of the internal field is expected to occur not only for the incident (pump and test) beams but also for the signal beam itself. Therefore, we can expect much larger size enhancement of the nonlinear signal than that of the pump–probe process. This effect has a good analogy with the mechanism of surface enhanced Raman scattering where the double enhancements of both the incident and scattered lights play an important role in enhancing the signal intensity [45]. A size-resonant enhancement of the nonlinear response has been actually observed for the GaAs double-heterostructure layers [19]. Also, systematic measurements of the thickness dependent nonlinear signal and the theoretical analysis with an appropriate model of this material system have confirmed the predictions of nonlocality induced size-resonant enhancement of the nonlinear response [20]. In this section, we present the experimental results for high quality of the GaAs double-heterostructure layers and their analysis.

5.1. The DFWM signal in the experiments

As we saw in the previous section, sample quality is essential for the emergence of a strong size-resonant enhancement of the nonlinear response. GaAs is a promising material from this point of view because of the technique for its growth being most advanced. The sample preparation has been carried out with special attention paid to realizing ultrahigh quality so that the coherence of the excitonic c.m. motion is well maintained over a whole sample. The layers are grown on a GaAs(100) substrate by molecular beam epitaxy making double heterostructures consisting of three layers of GaAs with equal thickness L. 5 nm thick $Al_{0.3}Ga_{0.7}As$ layers divide GaAs layers. The sample structure is shown in figure 5(a).

The reflectance spectrum near the Brewster angle for each sample is taken for sample characterization. Figure 5(b) shows the spectrum for $L = 110$ nm in the vicinity of the 1s exciton resonance. We can see very sharp peaks attributed to the quantized c.m. levels of excitons. The linear response has been calculated with the nonlocal theory (figure 6(c)) by using the model of confined excitons explained in the next subsection. Through the fitting of the spectral shape, it has been clarified that the quality of the top layer is much better than those

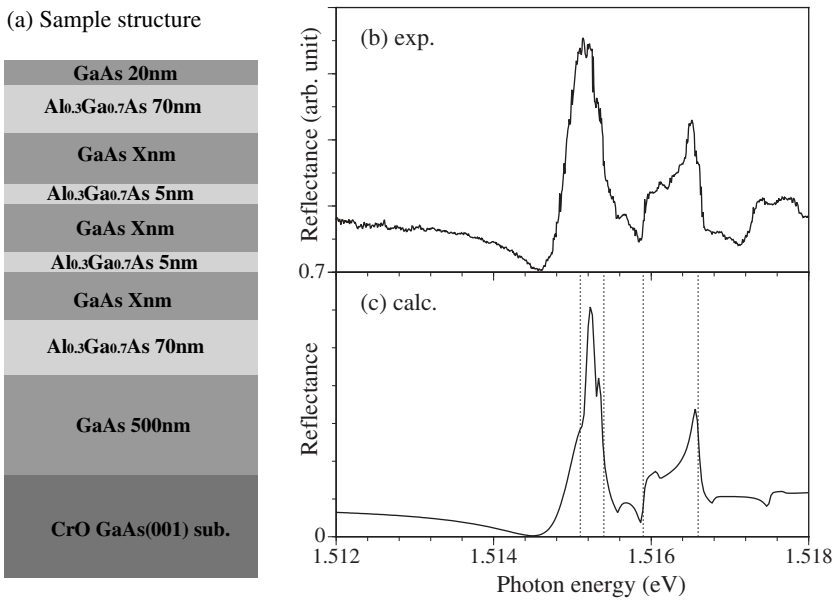

Figure 5. (a) The sample structure. (b) The observed reflectance spectrum for $L = 110$ nm. (c) The calculated reflectance spectrum for $L = 110$ nm. The vertical dotted lines indicate the quantized excitonic levels determined in the theory.

of the other layers, and the spectral shape is almost completely determined by the contribution of the top layer. The best fit is obtained with $\Gamma = 0.03$ meV for the top layer, which means very high quality of this layer.

With several samples prepared with different values of L and with almost the same damping constant Γ, the thickness dependence of the nonlinear signal has been examined. The two-pulse self-diffraction configuration has been used to measure the DFWM signal, where the pulse width is 2 ps and its peak excitation intensity is set below 5 kW cm^{-2} where the detected DFWM signal intensity shows a cubic dependence on the excitation intensity. The sample was kept at 5 K during measurements. Figure 6(a) shows the energy dependence of the DFWM signal intensity for several values of L. It should be noted that the signal is enhanced at $L = 110$ nm. This enhanced value is 25 times larger than that of the bulk (1 μm thick) sample [19]. Figure 6(b) shows the comparison between the signal intensities obtained by a collinear measurement and a cross-linear one. The contributions from the one-exciton resonance and the biexciton resonance are both contained in the former, whereas only the biexciton resonance contributes in the latter. This result indicates a small contribution from the biexciton resonance in the present nonlinear process.

5.2. Model of calculation

In the analysis of the above experiment, we have treated the contributions from two-exciton states in a simple way because they are not essential in the size dependence of DFWM near one-exciton resonance. As for biexcitons, we consider their c.m. motion similarly to that of single excitons. The magnitude of the matrix element arising from the relative motion is determined by comparing the signal intensity of the collinear measurement with that of the cross-linear measurement [20]. We do not treat the scattering two-exciton states in an explicit

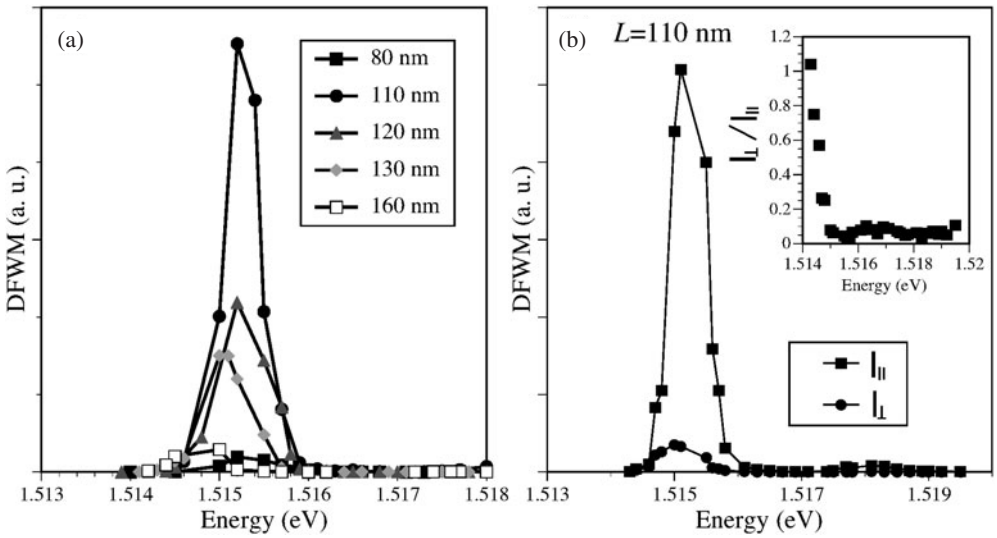

Figure 6. (a) The energy dependence of the DFWM intensity for several values of L. The peak excitation intensity is set below 5 kW cm^{-2}. (b) The energy dependence of the DFWM intensity obtained by cross-linear and collinear measurements for $L = 110$ nm. The inset indicates their ratio. The peak excitation intensity is set at 5 kW cm^{-2}.

way because the cancellation between the contributions from (ground state–one-exciton state) transitions and that from (one-exciton–two-exciton state) transitions does not strongly affect the thickness dependence of the nonlinear signal in the thickness region considered [13, 14]. If the absolute signal intensity needs to be discussed, this effect should be treated seriously.

As a model of confined excitons in excitonic active layers of GaAs, we consider the contributions from the heavy hole excitons alone because the experimental result for the linear response shows little contribution of the light hole excitons to the main peak structures near the lowest exciton level. Limiting ourselves to normal incidence, we deal only with the freedom of Z components of the excitonic c.m. motion, where Z is the coordinate of the surface normal direction. The eigenenergies and eigenfunctions of excitons in a film for $K_\parallel = 0$ can be written as

$$E_n = \hbar\omega_{\mathrm{T}} + \hbar^2 k_n^2 / 2M_{\mathrm{ex}} \tag{53}$$

and

$$\phi_n(Z) = \sqrt{2/(\bar{L})}\sin(k_n Z), \tag{54}$$

respectively, where $\hbar\omega_{\mathrm{T}}$ is the transverse exciton energy in the bulk, M_{ex} is the excitonic total mass, $\bar{L}$ is the thickness of the excitonic active region, and the allowed values of k_n are $n\pi/\bar{L}$ $(n = 1, 2, \ldots)$. For this model, the explicit form of the numerator of the quantity X_n defined in equation (22) is $\sqrt{1/\bar{L}}\int \mathrm{d}Z \sin(k_n Z)E(Z)$. For each GaAs layer, we consider the homogeneous dead layers (HDL) [46] near the interfaces where the excitons are not active because of the image force.

Since the contribution from the biexciton resonance is small in the present case, the terms giving the main contributions to the nonlinear signal among those in equation (30) are the ones relating only to one-exciton states. Those terms are proportional to factors that can be written in the following form when $\omega_2 = \omega_1$:

$$-\frac{4\Gamma}{\gamma} X_n(\omega_2) X_n(\omega_2) X_n(\omega_1)^*. \tag{55}$$

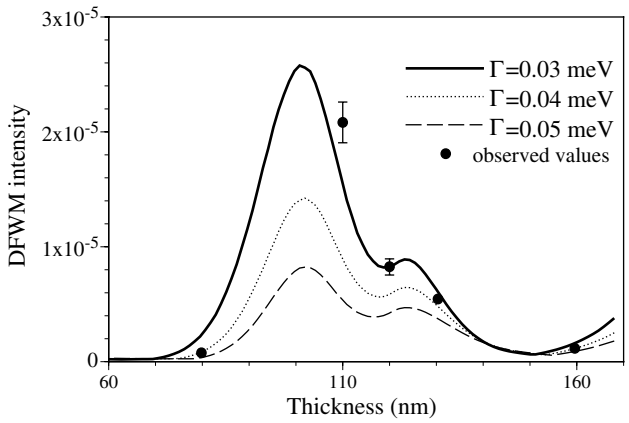

Figure 7. The thickness dependence of the peak value of the DFWM intensity. Closed circles are measured values in arbitrary units. The solid curve shows calculated values normalized by the incident probe intensity. The extent of the enhancement is affected by the nonradiative damping constant.

(In the actual calculation, we consider all kinds of terms including those relating to biexciton states.) Also for DFWM, $\{X_n\}$ is a fundamental quantity for understanding the size and frequency dependence of the nonlinear response. Since X_n is resonantly enhanced at $\mathrm{Re}[\Omega_n]$ depending on the size, $\{\Omega_n\}$ and $\{X_n\}$ should be examined in order to understand the nonlinear response beyond the LWA, which is in contrast with the oscillator strength and the uncoupled excitonic levels being the essential quantities for knowing the magnitude and the spectral structure of the nonlinear response in the LWA.

5.3. Analysis of the DFWM signal

Here, we discuss the observed size dependence of the DFWM signal showing the calculated result and examining $\{X_n\}$ for the present model of GaAs layers. In figure 7, we show the calculated result together with the observed data. The thickness dependence with a size-resonant enhancement is very well reproduced by the theory. This size dependence can be understood from the dependence of X_n on the frequency and thickness.

As explained in the previous section, X_n is enhanced at $\mathrm{Re}[\Omega_n]$. Therefore, the condition of size and frequency where the nonlinear signal is enhanced for each excitonic level can be seen by calculating $\mathrm{Re}[\Omega_n]$ as a function of thickness. On the other hand, to consider how greatly they are enhanced, we need to see the radiative shift, namely $\mathrm{Re}[\Omega_n] - E_1(n)$, which shows the extent of the off-resonance at the energy position where X_n takes a maximum value. Figure 8(a) shows the thickness dependence of $\mathrm{Re}[\Omega_n]$ where we assume that only the top layer is excitonic active. Because of the relatively large coupling between the radiation and the lowest excitonic level, the (positive) shift of this level is large as compared with that of the second ($n = 2$) level. The $\{X_n\}$ are resonantly enhanced and take maximum values at $\mathrm{Re}[\Omega_n]$. Similarly the $\{F_n\}$, which are in the numerator in $\{X_n\}$, are also enhanced near $\mathrm{Re}[\Omega_n]$. As the thickness increases, the $\{F_n\}$ and the radiative shift increase, and as a result of the competition between the increase of $\{F_n\}$ and the off-resonance, the $\{X_n\}$ take a maximum value at a certain thickness. The maximum values of $|X_n|^2$ ($n = 1, 2$) in the spectrum are plotted as functions of thickness in figure 8(b). As we can see in figure 8(a), $\mathrm{Re}[\Omega_2]$ is nearer the uncoupled excitonic state as compared with the case of $\mathrm{Re}[\Omega_1]$. Because of this, the maximum value

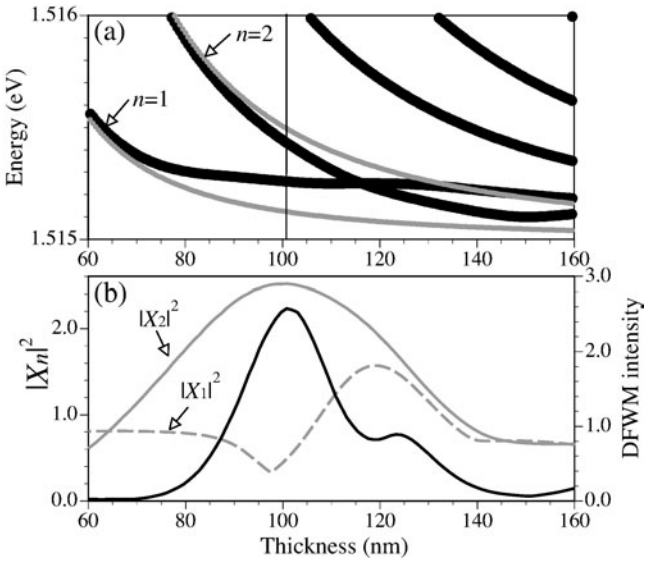

Figure 8. (a) The thickness dependence of $\mathrm{Re}[\Omega_n]$ (black curves) and the uncoupled excitonic levels (grey curves). The vertical line shows the thickness where the DFWM signal takes a maximum value. (b) The thickness dependence of the maximum values of $|X_n|^2$ ($n = 1, 2$) in units of $a_0|E_0|^2/\Gamma^2$, where a_0 is the lattice constant. The black solid curve is the same as that for $\Gamma = 0.03$ meV in figure 7 but normalized by $|E_0|^2 \times 10^{-5}$.

of $|X_2|^2$ is enhanced by thickness, while the enhancement of $|X_1|^2$ is not remarkable. This leads to the situation being quite different from that in the LWA; that is, the $n = 2$ exciton, whose wavefunction has a non-dipole-type spatial structure with one node in the surface normal direction, provides a much larger nonlinear signal than the lowest excitonic state.

We can say that this effect is a result of a double resonance, i.e., the resonance in the internal field due to the nanoscale Fabry–Pérot interference of exciton–polaritons and that with poles in the energy denominator. Furthermore, it should be remarked that the internal field of the signal itself is also enhanced due to the Fabry–Pérot interference in the degenerate case, i.e., $\omega_2 = \omega_1$. This means that we observe the signal enhanced through the multiple resonance affecting the internal field for both the incident and signal light beams in the medium. On the other hand, the weight of the contribution from the lowest exciton becomes smaller because the off-resonance effect appears as the third power in the energy denominator for DFWM.

6. Ultrafast radiative decay of non-dipole-type excitonic states

The enhancement of the nonlinear signal of $n = 2$ excitons provides the opportunity to observe the radiative decay process of 'non-dipole-type' excitonic states. Actually, Akiyama *et al* [21] has demonstrated the very fast optical switching, 1.5 ps, of $n = 2$ excitons confined in GaAs layers, where the pump–probe method in a cross-Nicol configuration has been used. This switching time is dominated by the radiative decay process, and is much shorter than the usual recombination time for the bulk excitons. It is obvious that the analysis of the radiative lifetime based on the oscillator strength cannot be applied to the present case where the LWA does not hold. In this section, the radiative width of confined excitons in a thin layer is theoretically discussed, revealing the peculiar thickness dependence of the radiative width in the thickness regime beyond the LWA.

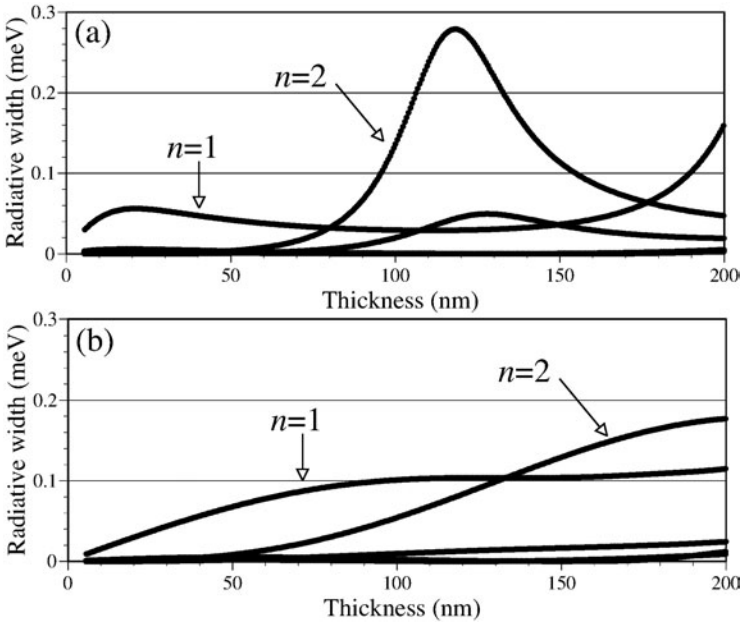

Figure 9. The radiative width of the confined excitons as a function of the film thickness. In (b), the terms arising from the multiple reflections in the Green function are omitted in the calculation.

In the present calculation of $\mathrm{Im}[\Omega_n]$, basically the same model as in the previous section is used, but we assume a free standing film without HDL for simplicity. As for the Green function, we use that for the film geometry [47] and assume $K_\parallel = 0$. For homogeneous space, the Green function is proportional to the factor $e^{iq|Z-Z'|}$. Since we treat a film geometry here, we add the terms containing the different coordinate dependences expressed by $e^{\pm iq(Z+Z'-L)}$ and $e^{\pm iq(Z-Z'-d)}$ that represent the effect of the reflection at both surfaces. Figure 8(a) shows the thickness dependence of the radiative width ($\mathrm{Im}[\Omega_n]$) for each excitonic state. When the thickness L is very much less than the light wavelength $2\pi/q$, the width of the lowest (dipole-type) excitonic state increases linearly with L. This effect corresponds to the size-linear enhancement of the oscillator strength in the LWA regime. However, in the present case, it turns out to decrease with further increase of L. This means that the spatial structure of the radiation is no longer negligible and the LWA breaks down. Then the width of the second (non-dipole-type) state increases and greatly exceeds that of the lowest one, and it takes a maximum value at a certain thickness. This is due to the spatial structure of the radiation field. It should be noted that the thickness where $\mathrm{Im}[\Omega_2]$ takes a maximum value is much less than $2\pi/q$. This is due to a cavity-like effect caused by the large reflectivity at surfaces. This effect can be understood if we consider $\mathrm{Im}[\Omega_n]$ calculated by considering the $e^{iq|Z-Z'|}$ type of coordinate dependence alone in $G(Z, Z')$ (see figure 8(b)). In this case, the thickness where $\mathrm{Im}[\Omega_2]$ takes a maximum value almost coincides with the light wavelength $2\pi/q$. This result indicates that the dielectricity of the surrounding medium is also an essentially important factor in determining the radiative coupling of confined excitons. Interestingly, for the present material, this effect makes the optimum thickness for the radiative width of the $|K_2\rangle$ state very near the optimum thickness for the enhancement of the DFWM signal.

This peculiar thickness dependence of the radiative lifetime of $n = 2$ excitons has also observed experimentally [48]. With a sample structure similar to those in the previous section,

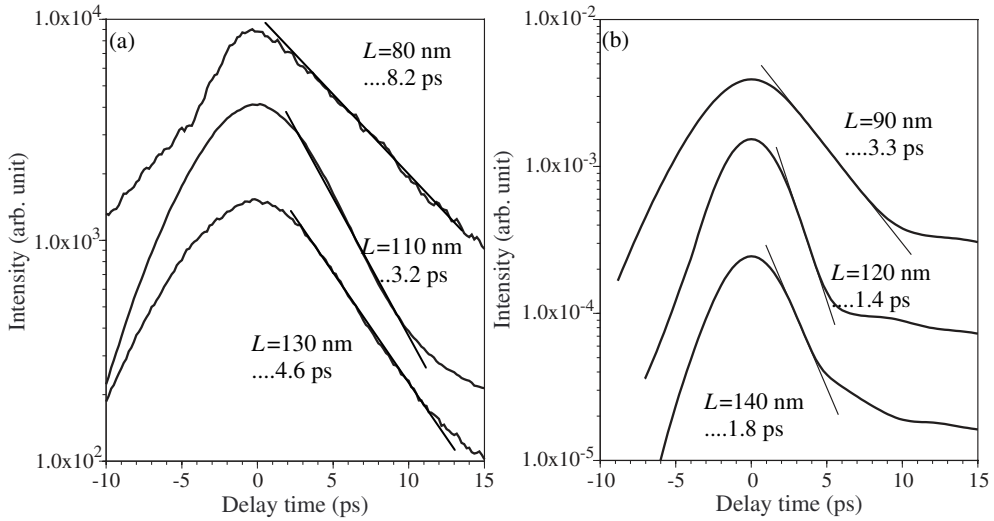

Figure 10. (a) The measured DFWM intensity as a function of the delay time τ_{23} at $\tau_{12} = 0$. (b) The calculated DFWM intensity as a function of the delay time τ_{23} at $\tau_{12} = 0$. The centre frequency is tuned to the peak position in the DFWM spectrum for each thickness. A stand-alone film of GaAs is assumed. In both (a) and (b), the values indicated are the thickness and the decay time.

we chose three different values of the thickness, namely $L = 80$, 110, 130 nm. Picosecond pulses from a mode-locked Ti:sapphire laser with a linewidth of 0.7 meV were used as an excitation source. The temperature of the sample was kept at 5 K during the measurement and the signal is obtained in a weak excitation condition (2.5 kW cm^{-2}). We prepare three beams with the wavevectors k_1, k_2, and k_3. The signal in the backward diffraction geometry $k_s = k_1 + k_2 - k_3$ is detected. In figure 9(a), we show the measured temporal profiles of the DFWM signals; that is, the signal intensities as a function of the delay time τ_{23} at $\tau_{12} = 0$ are plotted. The results clearly show the very fast decay of the $|K_2\rangle$ state. The coherent decay times for $L = 80$, 110, 130 nm are 8.2, 3.2, and 4.6 ps, respectively. It has been verified that the radiative decay time for the last two samples is much shorter than that for the lowest state. As a reference, we show in figure 9(b) the calculated results obtained by the nonlocal theory for the transient pulse response [49, 50], where the same model for the calculation as in figure 8, i.e., a stand-alone film of GaAs, is assumed. Although the result in figure 9(b) should not be compared directly with the above experimental one because the sample structures for the two cases are different, it should be remarked that the present experimental results indicate the following important points:

(1) Beyond the LWA, it is possible that the non-dipole-type excitonic states show much faster radiative decay than the lowest state.
(2) The thickness dependence of the radiative decay rate is not monotonic, but shows an enhancement with size.

Also we can say that this experimental result, which is the first observation of anomalous size dependence of the radiative lifetime beyond the LWA, strongly supports the above theoretically explained mechanism.

The systematic study of the radiation–matter coupling beyond the LWA using more realistic models is in progress in our group; the possibility of wide range control of the radiative

lifetime and nonlinearity through manipulation of the interplay between the spatial structures of the excitonic wavefunction and the radiation field is being revealed. We hope that this type of study will lead to a new guiding principle for the development of highly efficient nonlinear materials and devices free from the trade-off relationship between nonlinearity and response speed.

7. Summary and conclusions

After a brief review of the conventional discussion of the size dependent nonlinear response based on the LWA where the oscillator strength is a fundamental quantity for describing the radiation–matter coupling strength, we explained (1) how the LWA breaks down with increase of the size and (2) the microscopic nonlocal response theory that can describe the optical response beyond the LWA. Then, we introduced the nonlocal theory for the nonlinear response which is an extended version of that for the linear response.

For the unified description of the optical response from the microscale to the bulk system, a nonlocal treatment is necessary in principle. For the size regime beyond the LWA, this treatment is indispensable because the spatial structure of the internal field plays an essential role. Such situations typically appear when the excitonic centre-of-mass (c.m.) motion is coherently extended over the whole volume of the sample. Since the coherent length of the c.m. motion can be very long, the peculiar size dependence of the optical response appears over a wide range of size.

The calculations of the electromagnetic field in nanostructures have revealed the following points: since the motion of the Maxwell field should be determined self-consistently with that of the induced polarization, the nanoscale spatial structure is reflected in the response internal field. Depending on the system size and frequency, a particular spatial pattern similar to that of the resonance excitonic state is dominantly enhanced. Therefore, the LWA breaks down in an unexpected small size region. The theoretical demonstration of the pump–probe measurement has predicted this size-resonant enhancement of the internal field to lead to an anomalous size dependence of the nonlinear signal. A particularly noteworthy point is an enhancement of the signal from a non-dipole-type excitonic state, which is in striking contrast with optical responses described in the LWA. The size and frequency dependence of the nonlinear response can be explained on the basis of the behaviour of the component of the induced polarization X_n relevant to each confined exciton that is resonantly enhanced at the eigenenergies including the radiative shift. A similar effect has been predicted also for degenerate four-wave mixing (DFWM), and an experiment using GaAs heterolayers has confirmed the theoretical prediction. Interestingly, the size-resonantly enhanced nonlinear signal exhibits an ultrafast radiative decay of a few picoseconds. The analysis of the radiative width has explained this short radiative decay time of non-dipole-type excitonic states, and revealed the peculiar thickness dependence of the radiative width beyond the LWA.

Owing to the recent development of nanofabrication technologies, a lot of novel nanostructures with highly coherent electronic systems are appearing and, hence, study based on the nonlocal framework will be much more important, and will lead to the finding of novel optical effects. We believe that such study is also important for establishing a new guiding principle for developing highly efficient optical materials and devices, where the nanoscale spatial variation is a new degree of freedom in manipulating the radiation–matter interaction.

Acknowledgments

This review is dedicated to Professor Kikuo Cho on the occasion of his retirement from Osaka University. The author is grateful to him for collaboration, and continued inspiration through

discussions and encouragement. The experiments presented in this review were performed by the group at Mitsubishi Electric Corporation managed by Dr T Isu. The author also thanks the members of his group, Dr T Isu, Mr K Akiyama, Dr N Tomita, and Dr Y Nomura, for useful discussions throughout the collaboration. He also thanks his former student Mr Amakata for his collaboration in the calculation of the DFWM signal, and thanks Dr Ajiki for useful discussions. The work presented in this review was supported in part by Grants-in-Aid for Scientific Research (14540301) and for COE Research (10CE2004) from the Ministry of Education, Science, Sports and Culture of Japan.

References

[1] See, for example, Del Sole R, D'Andrea A and Lapiccirella A (ed) 1988 *Excitons in Confined Systems* vol 25 (Berlin: Springer)
Gaponenko S V 1998 *Optical Properties of Semiconductor Nanocrystals* (Cambridge: Cambridge University) and references therein
[2] Shinada M and Sugano S 1966 *J. Phys. Soc. Japan* **21** 32
[3] Miller R C, Kleinmen D A, Tsang W T and Grossard A C 1981 *Phys. Rev.* B **24** 1134
[4] Takagahara T 1987 *Phys. Rev.* B **36** 9293
[5] Kayanuma Y 1988 *Phys. Rev.* B **38** 9797
[6] D'Andrea A and Del Sole R 1990 *Phys. Rev.* B **41** 1413
[7] Rashba E I and Gurgenishvili G E 1962 *Fiz. Tverd. Tela* **4** 1029
Rashba E I and Gurgenishvili G E 1962 *Sov. Phys.—Solid State* **4** 759 (Engl. Transl.)
[8] Henry C H and Nassau K 1970 *Phys. Rev.* B **1** 1628
[9] Hanamura E 1988 *Phys. Rev.* B **37** 1273
[10] Takagahara T 1989 *Phys. Rev.* B **39** 10206
[11] Spano F C and Mukamel S 1989 *Phys. Rev.* A **40** 5783
[12] Spano F C and Mukamel S 1991 *Phys. Rev. Lett.* **66** 1197
[13] Ishihara H and Cho K 1990 *Phys. Rev.* B **42** 1724
[14] Ishihara H and Cho K 1991 *J. Nonlinear Opt. Phys.* **1** 287
[15] Ohfuti Y and Cho K 1995 *Phys. Rev.* B **51** 14379
[16] Cho K and Arakawa T 1997 *Mater. Sci. Eng.* B **48** 71
[17] Ishihara H and Cho K 1996 *Phys. Rev.* B **53** 15823
[18] Ishihara H and Cho K 2002 *Phys. Rev.* B **65** 035305
[19] Akiyama K, Tomita N, Nomura Y and Isu T 1999 *Appl. Phys. Lett.* **75** 475
[20] Ishihara H, Cho K, Akiyama K, Tomita N, Nomura Y and Isu T 2002 *Phys. Rev. Lett.* **89** 017402
[21] Akiyama K, Tomita N, Nishimura T, Nomura Y and Isu T 2000 *7th Int. Workshop on Femtosecond Technology (FST, Tsukuba, 2000)* p 69 (Abstracts)
[22] Isu T, Akiyama K, Tomita N, Nomura Y, Ishihara H and Cho K 2003 *Ultrafast Phenomena in Semiconductors VII: Proc. SPIE (San Jose, 2003)* (*SPIE Proceedings Series* vol 4992) p 165
[23] Masumoto Y, Yamazaki M and Sugawara H 1988 *Appl. Phys. Lett.* **53** 1527–9
[24] Kataoka T, Tokizaki T and Nakamura A 1993 *Phys. Rev.* B **48** 2815
[25] Banyai L, Hu Y Z, Lindberg M and Koch S W 1988 *Phys. Rev.* B **37** 8142
[26] Ishihara H and Amakata T 2001 *Int. J. Mod. Phys.* B **15** 3809
[27] Ishihara H and Nakatani T 2003 *QELS Technical Digest* QTuG19
[28] Cho K 1991 *Prog. Theor. Phys.* **106** (Suppl.) 225
[29] Cho K 2003 *Optical Response of Nanostructures: Microscopic Nonlocal Theory* (New York: Springer)
[30] Ishihara H and Cho K 1993 *Phys. Rev.* B **48** 7960
[31] See, for example, Shen Y R 1984 *The Principle of Nonlinear Optics* (New York: Wiley)
[32] Ishihara H, Amakata T and Cho K 2002 *Phys. Rev.* B **65** 035305
[33] Ishihara H 2003 *Phys. Rev.* B **67** 113302
[34] Kiselev V A, Razbirin B S and Uraltsev I N 1975 *Phys. Status Solidi* b **72** 161
[35] Makarenko I V, Uraltsev I N and Kiselev V A 1980 *Phys. Status Solidi* b **98** 161
[36] Mita T and Nagasawa N 1982 *Solid State Commun.* **44** 1003
[37] Kiselev V A, Makarenko V, Razbirin B S and Uraltsev I N 1977 *Fiz. Tverd. Tela* **19** 2348
Kiselev V A, Makarenko V, Razbirin B S and Uraltsev I N 1977 *Sov. Phys.—Solid State* **19** 1374 (Engl. Transl.)
[38] Cho K and Kawata M 1985 *J. Phys. Soc. Japan* **54** 4431

[39] Cho K, D'Andrea A, Del Sole R and Ishihara H 1990 *J. Phys. Soc. Japan* **59** 1853
[40] D'Andrea A and Del Sole R 1982 *Phys. Rev.* B **25** 3714
[41] Cho K and Ishihara H 1990 *J. Phys. Soc. Japan* **59** 754
[42] Ishihara H and Cho K 1989 unpublished result
[43] Wang H, Ferrio K, Steel D G, Hu Y, Binder R and Koch S 1993 *Phys. Rev. Lett.* **71** 1261
[44] Ishihara H, Amakata T and Cho K 2000 *J. Lumin.* **87–89** 850
[45] Jha S S, Kirtley J R and Tsang J C 1980 *Phys. Rev.* B **22** 3937
[46] Hopfield J J and Thomas D G 1963 *Phys. Rev.* **132** 563
[47] Chew W C 1990 *Waves and Fields in Inhomogeneous Media* (Princeton, NJ: Van Nostrand-Reinhold)
[48] Ishihara H, Cho K, Akiyama K, Tomita N and Isu T 2002 *Phys. Status Solidi* a **190** 849
[49] Ishihara H 1997 *Mater. Sci. Eng.* B **48** 75
[50] Ishihara H 2000 *J. Lumin.* **87–89** 905