Aalborg Universitet

AALBORG UNIVERSITY

DENMARK

Gauge invariance of excitonic linear and nonlinear optical response

Taghizadeh, Alireza; Pedersen, Thomas Garm

Gauge invariance of excitonic linear and nonlinear optical response

Alireza Taghizadeh[1,*] and T. G. Pedersen[1,2]

[1]*Department of Physics and Nanotechnology, Aalborg University, DK-9220 Aalborg Øst, Denmark*
[2]*Center for Nanostructured Graphene (CNG), DK-9220 Aalborg Øst, Denmark*

We study the equivalence of four different approaches to calculate the excitonic linear and nonlinear optical response of multiband semiconductors. These four methods derive from two choices of gauge, i.e., length and velocity gauges, and two ways of computing the current density, i.e., direct evaluation and evaluation via the time-derivative of the polarization density. The linear and quadratic response functions are obtained for all methods by employing a perturbative density-matrix approach within the mean-field approximation. The equivalence of all four methods is shown rigorously, when a correct interaction Hamiltonian is employed for the velocity gauge approaches. The correct interaction is written as a series of commutators containing the unperturbed Hamiltonian and position operators, which becomes equivalent to the conventional velocity gauge interaction in the limit of infinite Coulomb screening and infinitely many bands. As a case study, the theory is applied to hexagonal boron nitride monolayers, and the linear and nonlinear optical response found in different approaches are compared.

I. INTRODUCTION

The optical response of crystals provides valuable information about material properties, e.g.,,, important features of the band structure [1,2]. The response can be characterized by the linear response as well as diverse nonlinear ones, e.g., second/third harmonic generation, optical rectification, etc. [1]. Theoretically, accurate estimates of optical response functions based on the material band structure are highly desirable, since they can offer important insights for experiments and device applications. Nowadays, perturbative calculations of linear and nonlinear optical response functions are routinely performed in the independent-particle approximation (IPA), in which the electron-hole interaction is simply ignored, e.g., see Refs. [3–11] (and references therein). However, it is well known that including the electron-hole interaction, i.e., excitonic effects, can have a significant influence on the optical response of solids [12–16]. In particular, excitons dramatically modify the optical response of low-dimensional systems, including carbon nanotubes [17,18], and two-dimensional (2D) materials such as hexagonal boron nitride (hBN) [19–22] and transition-metal dichalcogenides [23–27] due to the reduced screening and enhanced confinement of electrons. Typically, excitons affect the linear response by introducing strong resonances inside the band gap and renormalizing the continuum part of the spectrum [19,28]. Regarding the nonlinear optical response, a few theoretical studies have been done both on bulk [29,30] and 2D materials [19,23,31], which show that complex modifications of the spectra occur due to excitons.

From a theoretical point of view, the optical response should, in principle, be independent of the chosen electromagnetic gauge. However, in practice, the choice of electromagnetic gauge, e.g., the so-called length and velocity gauges, influences the results due to various approximations [32–41]. Formally, it is straightforward to show that the wave functions in the length and velocity gauges are related via a time-dependent unitary transformation [32]. The unitary transformation converts the length gauge (LG) Hamiltonian to its equivalent velocity gauge (VG) counterpart. This Hamiltonian includes the unperturbed Hamiltonian plus an interaction part, which is given by a series of commutators between the position and unperturbed Hamiltonian [33,42]. In the IPA limit, the series can be truncated to the first two terms if the canonical commutator relation between position $\hat{\mathbf{r}}$ and momentum $\hat{\mathbf{p}}$ is used, i.e., $[\hat{\mathbf{r}},\hat{\mathbf{p}}] = i\hbar\mathbf{I}$, where $\mathbf{I}$ denotes the unit tensor [33]. This leads to the conventional VG (CVG) interaction given by $e\boldsymbol{\mathcal{A}} \cdot \hat{\mathbf{p}}/m + e^2\boldsymbol{\mathcal{A}}^2/2m$, where $\boldsymbol{\mathcal{A}}$ is the vector potential. For an infinite periodic system, the calculation using this interaction has the advantage over the LG that it avoids using the ill-defined matrix elements of position and uses only the well-defined momentum ones. Nonetheless, the price paid for calculating the optical response in the CVG is that a large number of bands is typically required to obtain an acceptable result [40]. Thus, if an insufficient number of bands is used in the calculations, the response functions computed in the CVG may suffer from the well-known zero-frequency divergences [3], or even become entirely incorrect. For instance, the even-order responses obtained within a two-band model are identically zero due to the time-reversal symmetry [9,40]. In contrast, if the exact series of commutators is used for the VG interaction, the LG and VG results become identical regardless of the number of bands in the calculations [42]. However, one of the most important strengths of the VG, which is its simple implementation, is utterly lost.

Including the electron-hole interaction complicates the optical response calculation dramatically, which compels us to introduce additional approximations such as the mean-field approximation (MFA) to solve the many-body problem. Within the MFA, the Hamiltonian becomes effectively nonlocal due to

*ata@nano.aau.dk

the Coulomb term. Despite the nonlocality of the Hamiltonian, the MFA theory has been shown to be formally invariant under the gauge transformation [32]. Nonetheless, the meaning of the gauge freedom in the excitonic optical response calculation is not fully understood. For instance, using the CVG interaction, as in Refs. [31] and [43], results in an incorrect excitonic optical response regardless of the basis completeness, as will be demonstrated in Sec. II C. In addition to the gauge freedom, it is well known that the optical response can be computed in two ways: direct evaluation of the current density and via the time derivative of the polarization density [19,40]. Similarly to the equivalence of the two gauges, these two approaches should, in principle, be equivalent but may generate different results when excitons are considered [19]. Therefore, four formally equivalent but computationally different methods for obtaining the optical response are available, which consist of the combinations of two choices of gauge, i.e., LG and VG, and two ways of computing the current density, directly and indirectly via the polarization density.

In the present work, we investigate systematically the influence of the chosen gauge and observable on the excitonic linear and nonlinear optical response. We develop a practical framework for calculating the multiband semiconductor response by adopting a many-body density-matrix approach within the MFA. Employing a perturbative solution for the density matrix, the expressions of the first- and second-order response functions are derived. Using this framework, we show that (1) the excitonic responses obtained using the four above-mentioned alternatives are identical regardless of the number of bands in the calculations if the VG interaction is written as the commutator series of the position and unperturbed Hamiltonian; (2) despite the equivalence of the four methods, it is simpler to derive the conductivity expressions and perform the calculations in the LG approaches compared to their VG counterparts, particularly in the case of nonlinear responses; (3) the excitonic response computed using the CVG interaction is not reliable in the MFA even if a complete basis is employed for the calculations, since the MFA Hamiltonian includes an effective nonlocal potential; (4) the response generated by the CVG and VG becomes identical if a complete basis set is used and the Coulomb interaction is neglected. We apply the proposed theory to hBN monolayers as a case study and confirm the validity of the theoretical framework through numerical simulations.

II. THEORETICAL FRAMEWORK

In this section, we present the theoretical framework for calculating the excitonic optical response of periodic systems. We begin by introducing the equation of motion for the density matrix. Then, using a perturbative solution of the dynamical equations, the linear and second-order optical conductivities are derived for all methods. Finally, the equivalence of response functions obtained by these different methods is discussed. It should noted that throughout the text, all vectors and tensors are indicated by bold letters, and the single-particle/many-body operators and matrix elements are denoted by lower-case/uppercase letters, respectively.

A. Dynamical equation

The many-body Hamiltonian of a system of electrons under the influence of an external perturbation is written in second quantization as

$$
\hat{H} \equiv \hat{H}_0 + \hat{\mathcal{V}} + \hat{U}(t) \equiv \hat{\mathcal{H}}_0 + \hat{U}(t) \equiv \sum_k \varepsilon_k^0 \hat{c}_k^\dagger \hat{c}_k
$$
$$
+ \frac{1}{2} \sum_{klmn} \mathcal{V}_{klmn} \hat{c}_k^\dagger \hat{c}_l^\dagger \hat{c}_n \hat{c}_m + \sum_{kl} u_{kl}(t) \hat{c}_k^\dagger \hat{c}_l, \quad (1)
$$

where $\hat{c}$ and $\hat{c}^\dagger$ are the fermionic annihilation and creation operators, respectively, and $\hat{H}_0$, $\hat{\mathcal{V}}$, and $\hat{U}(t)$ are the single-electron, Coulomb potential, and time-dependent interaction parts, respectively. $\hat{\mathcal{H}}_0 \equiv \hat{H}_0 + \hat{\mathcal{V}}$ is the total unperturbed Hamiltonian. For simplicity, the spin-orbit coupling is neglected here. Note that the Hamiltonian $\hat{H}$ here is written in the single-particle basis $|n\rangle$, i.e., $\hat{h}_0|n\rangle = \varepsilon_n^0|n\rangle$, with $\hat{h}_0$ the unperturbed Hamiltonian of a single electron, and the matrix elements of the external potential read $u_{kl}(t) \equiv \langle k|\hat{u}(t)|l\rangle$, where $\hat{u}(t)$ is the interaction potential of an individual electron.

The dynamical behavior of the system is then studied by employing a density-matrix approach, for which we follow the procedure outlined in Ref. [19] and explained in Appendix A. Within the MFA, the equation of motion for the density matrix, $\rho_{ji}(t) \equiv \langle \psi_0|\hat{c}_i^\dagger \hat{c}_j|\psi_0\rangle$, is derived as shown in Eq. (A1). Here, $|\psi_0\rangle$ is the many-body ground state, in which all valence states are occupied. For the special case of periodic systems, the single-particle basis states are of the Bloch form $\langle \mathbf{r}|n\mathbf{k}\rangle = A^{-1/2} e^{i\mathbf{k}\cdot\mathbf{r}} \varphi_{n\mathbf{k}}(\mathbf{r})$, where A, $\varphi_{n\mathbf{k}}(\mathbf{r})$, n, and $\mathbf{k}$ are the crystal volume, cell-periodic part, band index, and wave vector, respectively. For a general single-particle operator $\hat{o}$, we denote single-particle matrix elements by $o_{nm\mathbf{k}} \equiv \langle n\mathbf{k}|\hat{o}|m\mathbf{k}\rangle$, such as $\mathbf{p}_{nm\mathbf{k}}$ for the momentum. In the Bloch basis, the equation of motion for the density matrix $\rho_{ji\mathbf{k}}$ is given in Eq. (A2), which can be solved perturbatively up to any required order of perturbation. The solutions for the first and second order are presented in Eqs. (A7a)–(A7e). These expressions are obtained in the so-called Tamm-Dancoff approximation [14,15], where the coupling between off-diagonal elements $\rho_{cv\mathbf{k}}$ and $\rho_{vc\mathbf{k}}$ is ignored (the indices c and v imply conduction and valence bands, respectively). Upon determining the density matrix, the expectation value of any observable of the system is found using Eqs. (A9a) and (A9b) in Appendix A.

B. Linear and quadratic optical response

The optical response is calculated as the induced current density inside the material owing to the interaction with an external electromagnetic field. Throughout this work, the electric field $\boldsymbol{\mathcal{E}}$ is decomposed into its harmonic components,

$$
\boldsymbol{\mathcal{E}}(t) = \frac{1}{2} \sum_p \boldsymbol{\mathcal{E}}(\omega_p) e^{-i\omega_p t}, \quad (2)
$$

where the p summation is performed over both positive and negative frequencies. Note that we neglect the spatial variation of the field, i.e., a long-wavelength regime is assumed. On the other hand, the form of time-dependent interaction $\hat{U}(t)$ depends on the choice of gauge. In the LG, $\hat{U}_{\mathrm{LG}}(t) =$

TABLE I. Four equivalent methods ($\mathcal{A}$–$\mathcal{D}$) for computing the current-density response and their respective labels. For comparison purposes, the conventional velocity gauge method is also shown and labeled by $\mathcal{C}'$. Here, $\hat{\mathbf{R}}$, $\hat{\mathbf{P}}$, $\mathcal{E}$, and $\mathcal{A}$ represent the many-body position operator, many-body momentum operator, electric field, and vector potential, respectively, $g = 2$ for spin degeneracy, and gN is the total number of electrons. $\hat{\mathbf{\Pi}}$ denotes the Heisenberg momentum operator defined as $\hbar\hat{\mathbf{\Pi}} \equiv im[\hat{\mathcal{H}}_0, \hat{\mathbf{R}}]$, where $\hat{\mathcal{H}}_0$ is the unperturbed Hamiltonian.

Label	$\hat{U}(t) \propto$	$\mathbf{J}(t) \propto$
$\mathcal{A}$	$\hat{\mathbf{R}} \cdot \mathcal{E}$	$\langle \hat{\mathbf{\Pi}} \rangle$
$\mathcal{B}$	$\hat{\mathbf{R}} \cdot \mathcal{E}$	$\partial \langle \hat{\mathbf{R}} \rangle / \partial t$
$\mathcal{C}$	$\mathcal{A} \cdot \hat{\mathbf{\Pi}} + e[\mathcal{A} \cdot \hat{\mathbf{R}}, \mathcal{A} \cdot \hat{\mathbf{\Pi}}]/2i\hbar$	$\langle \hat{\mathbf{\Pi}} + e[\mathcal{A} \cdot \hat{\mathbf{R}}, \hat{\mathbf{\Pi}}]/i\hbar \rangle$
$\mathcal{D}$	$\mathcal{A} \cdot \hat{\mathbf{\Pi}} + e[\mathcal{A} \cdot \hat{\mathbf{R}}, \mathcal{A} \cdot \hat{\mathbf{\Pi}}]/2i\hbar$	$\partial \langle \hat{\mathbf{R}} \rangle / \partial t$
$\mathcal{C}'$	$\mathcal{A} \cdot \hat{\mathbf{P}} + egN\mathcal{A}^2/2$	$\langle \hat{\mathbf{P}} + egN\mathcal{A} \rangle$

$e\hat{\mathbf{R}} \cdot \mathcal{E}(t)$, where $\hat{\mathbf{R}}$ denotes the many-body position operator. In the CVG, the interaction reads $\hat{U}_{\mathrm{CVG}}(t) = e(\mathcal{A} \cdot \hat{\mathbf{P}} + egN\mathcal{A}^2/2)/m$, where gN and $\hat{\mathbf{P}}$ are the total number of electrons and many-body momentum operator, respectively, and $g = 2$ accounts for the spin degeneracy. Moreover, the vector potential $\mathcal{A}$ is mapped to $\mathcal{E}$ via $\mathcal{E} = -\partial\mathcal{A}/\partial t$. When the electron-hole interaction is considered in the MFA, an effective nonlocal potential is introduced in the unperturbed Hamiltonian, which puts the validity of $\hat{U}_{\mathrm{CVG}}(t)$ into question [32,33]. In this case, it has been shown in Refs. [44] and [33] that the (correct) VG interaction should instead be written as a series of commutators. Up to the second order in $\mathcal{A}$, the interaction reads $\hat{U}_{\mathrm{VG}}(t) = e(\mathcal{A} \cdot \hat{\mathbf{\Pi}} + e[\mathcal{A} \cdot \hat{\mathbf{R}}, \mathcal{A} \cdot \hat{\mathbf{\Pi}}]/2i\hbar)/m$, where $\hbar\hat{\mathbf{\Pi}} \equiv im[\hat{\mathcal{H}}_0, \hat{\mathbf{R}}] \neq \hbar\hat{\mathbf{P}}$. We refer to $\hat{\mathbf{\Pi}}$ as the Heisenberg momentum operator, since it is proportional to the time derivative of the position operator (or velocity [33]) in the Heisenberg picture, i.e., $d\hat{\mathbf{R}}/dt = i[\hat{\mathcal{H}}_0, \hat{\mathbf{R}}]/\hbar$.

In addition to the gauge freedom, it is possible to calculate the optical response either by evaluating directly the expectation value of the current-density operator, $\mathbf{J}(t) \equiv \langle \hat{\mathbf{J}} \rangle$, or by computing the time derivative of the expectation value of the polarization density operator, $\mathbf{J}(t) \equiv \partial\mathcal{P}(t)/\partial t = \partial\langle\hat{\mathcal{P}}\rangle/\partial t$ [19,40]. The many-body current and polarization density operators read $\hat{\mathbf{J}} = -e\hat{\mathbf{\Pi}}/(mA)$ and $\hat{\mathcal{P}} = -e\hat{\mathbf{R}}/A$, respectively, and the expectation values of the operators are determined by employing the density matrix as discussed in Sec. II A. Note that the current-density operator in the VG includes an extra diamagnetic term and reads $\hat{\mathbf{J}} = -e(\hat{\mathbf{\Pi}} - e[\mathcal{A} \cdot \hat{\mathbf{R}}, \hat{\mathbf{\Pi}}]/i\hbar)/(mA)$. Hence, a total of four alternatives for computing the optical response are possible, which are formed by the combination of two gauges and two ways of evaluating the current-density response as labeled in Table I. For comparison purposes, we include the CVG labeled by $\mathcal{C}'$, where $\hat{U}_{\mathrm{CVG}}(t)$ is used as the interaction Hamiltonian. Note that hereafter, the normalized position operators $\hat{x}$ and $\hat{\mathbf{X}}$ are used for convenience, which are defined as $\hat{x} \equiv m\hat{r}/\hbar$, and similarly for the many-body operator $\hat{\mathbf{X}} \equiv m\hat{\mathbf{R}}/\hbar$. Using the normalized position, the many-body canonical commutator relation becomes $[\hat{\mathbf{X}}, \hat{\mathbf{P}}] = igNm\mathbf{I}$.

Without loss of generality, the first-order current density $\mathbf{J}^{(1)}(t)$ reads

$$\mathbf{J}^{(1)}(t) = \frac{1}{2}\sum_p \boldsymbol{\sigma}^{(1)}(\omega_p)\mathcal{E}(\omega_p)e^{-i\omega_p t}, \tag{3}$$

where the optical conductivity (OC) tensors $\boldsymbol{\sigma}^{(1)}(\omega_p)$ for the five methods of Table I are given by

$$\boldsymbol{\sigma}^{\mathcal{A}(1)} = -C_e \sum_n \left[\frac{\mathbf{\Pi}_n \mathbf{X}_n^*}{\hbar\omega_p - E_n} - \frac{\mathbf{\Pi}_n^* \mathbf{X}_n}{\hbar\omega_p + E_n} \right], \tag{4a}$$

$$\boldsymbol{\sigma}^{\mathcal{B}(1)} = C_e(i\hbar\omega_p) \sum_n \left[\frac{\mathbf{X}_n \mathbf{X}_n^*}{\hbar\omega_p - E_n} - \frac{\mathbf{X}_n^* \mathbf{X}_n}{\hbar\omega_p + E_n} \right], \tag{4b}$$

$$\boldsymbol{\sigma}^{\mathcal{C}(1)} = \frac{iC_e}{\hbar\omega_p} \left\{ \sum_n \left[\frac{\mathbf{\Pi}_n \mathbf{\Pi}_n^*}{\hbar\omega_p - E_n} - \frac{\mathbf{\Pi}_n^* \mathbf{\Pi}_n}{\hbar\omega_p + E_n} \right] + mN\mathbf{L} \right\}, \tag{4c}$$

$$\boldsymbol{\sigma}^{\mathcal{D}(1)} = [\boldsymbol{\sigma}^{\mathcal{A}(1)}]^T, \tag{4d}$$

$$\boldsymbol{\sigma}^{\mathcal{C}'(1)} = \frac{iC_e}{\hbar\omega_p} \left\{ \sum_n \left[\frac{\mathbf{P}_n \mathbf{P}_n^*}{\hbar\omega_p - E_n} - \frac{\mathbf{P}_n^* \mathbf{P}_n}{\hbar\omega_p + E_n} \right] + mN\mathbf{I} \right\}. \tag{4e}$$

Here, $C_e \equiv ge^2\hbar/(m^2 A)$, $\mathbf{L} \equiv 2i\sum_n \mathbf{\Pi}_n \mathbf{X}_n^*/mN$, and T denotes transposition. E_n is the exciton energy obtained by solving the Bethe-Salpeter equation (BSE), i.e., $H_{eh}|\psi^{(n)}\rangle = E_n|\psi^{(n)}\rangle$, with H_{eh} given in Eq. (A5). In addition, $\mathbf{\Pi}_n = -iE_n\mathbf{X}_n$ due to the definition of Heisenberg momentum $\hat{\mathbf{\Pi}}$, and the excitonic momentum $\mathbf{P}_n$ and position $\mathbf{X}_n$ are defined as

$$\mathbf{P}_n \equiv \sum_{cv\mathbf{k}} \psi_{cv\mathbf{k}}^{(n)} \mathbf{p}_{vc\mathbf{k}}, \quad \mathbf{X}_n \equiv \sum_{cv\mathbf{k}} \psi_{cv\mathbf{k}}^{(n)} \mathbf{x}_{vc\mathbf{k}}, \tag{5}$$

where $\psi_{cv\mathbf{k}}^{(n)} = \langle v\mathbf{k}\rightarrow c\mathbf{k}|\psi^{(n)}\rangle$ is the exciton projection onto a singlet band-to-band transition, and the summation over $\mathbf{k}$ implies an integral over the first Brillouin zone (BZ), i.e., $(2\pi)^D \sum_{\mathbf{k}} \rightarrow A \int_{\mathrm{BZ}} d^D\mathbf{k}$ ($D = 2$ for 2D materials). Note that $\mathbf{P}_n$, $\mathbf{X}_n$, and $\mathbf{\Pi}_n$ are indeed the matrix elements of the many-body momentum, position, and Heisenberg momentum operators between the ground state $|\psi_0\rangle$ and excited state $|\psi^{(n)}\rangle$, respectively.

Similarly, the quadratic current-density response reads

$$\mathbf{J}^{(2)}(t) = \frac{1}{4}\sum_{p,q} \boldsymbol{\sigma}^{(2)}(\omega_p, \omega_q)\mathcal{E}(\omega_p)\mathcal{E}(\omega_q)e^{-i(\omega_p+\omega_q)t}, \tag{6}$$

where $\boldsymbol{\sigma}^{(2)}(\omega_p, \omega_q)$ are rank-3 conductivity tensors given for the five methods in Eqs. (B1a)–(B1e). The conductivities are written in terms of matrix elements for transition between two excitons n and m denoted by $O_{nm} = \langle\psi^{(n)}|\hat{O}|\psi^{(m)}\rangle$ such as $\mathbf{X}_{nm}$. We note that expressions for optical susceptibilities χ can readily be derived from their corresponding conductivities by $\epsilon_0\chi^{(1)} \equiv i\boldsymbol{\sigma}^{(1)}/\omega_p$ and $\epsilon_0\chi^{(2)} \equiv i\boldsymbol{\sigma}^{(2)}/(\omega_p + \omega_q)$. The derivation of the quadratic conductivity tensor in the CVG, i.e., Eq. (B1e), is a rather straightforward problem, since it only contains the well-defined matrix elements of momentum. In contrast, the intraband part of the single-particle position operator appearing in $\mathbf{X}_{nm}$, Eq. (B2b), is ill defined inherently for infinite periodic systems.

In spite of the problems associated with the position operator, it has been shown in Ref. [4] that the optical response can be computed by separating formally the interband and intraband parts of the position matrix elements, i.e., $\langle n\mathbf{k}|\hat{\mathbf{x}}|m\mathbf{k}\rangle = \langle n\mathbf{k}|\hat{\mathbf{x}}^{(i)}|m\mathbf{k}\rangle\delta_{nm} + \langle n\mathbf{k}|\hat{\mathbf{x}}^{(e)}|m\mathbf{k}\rangle(1-\delta_{nm})$. The interband part is simply related to the momentum matrix element [9], whereas the intraband block is handled by employing a commutator relation [4]:

$$\frac{\hbar}{m}\langle n\mathbf{k}|[\hat{\mathbf{x}}^{(i)},\hat{o}]|m\mathbf{k}\rangle = i(o_{nm\mathbf{k}})_{;\mathbf{k}}, \tag{7a}$$

$$(o_{nm\mathbf{k}})_{;\mathbf{k}} \equiv \nabla_{\mathbf{k}}o_{nm\mathbf{k}} - i[\boldsymbol{\Omega}_{nn\mathbf{k}} - \boldsymbol{\Omega}_{mm\mathbf{k}}]o_{nm\mathbf{k}}. \tag{7b}$$

Here, $(o_{nm\mathbf{k}})_{;\mathbf{k}}$ is the generalized derivative written in terms of the Berry connections $\boldsymbol{\Omega}_{nm\mathbf{k}} \equiv iA_{uc}^{-1}\int_{uc}\varphi_{n\mathbf{k}}^*(\mathbf{r})\nabla_{\mathbf{k}}\varphi_{m\mathbf{k}}(\mathbf{r})d^D\mathbf{r}$ (A_{uc} is the unit-cell volume). By employing this technique, the interband/intraband parts of the position operator in $\mathbf{X}_{nm}$ are separated. This separation leads to $\mathbf{X}_{nm} = \mathbf{Y}_{nm} + m\mathbf{Q}_{nm}/\hbar$, where $\mathbf{Y}_{nm}$ [see Eq. (B3a)] and $\mathbf{Q}_{nm}$ [see Eq. (B3b)] are the interband and intraband parts of excitonic position matrix elements, respectively. Therefore, the quadratic conductivities in Eqs. (B1a)–(B1d) consist of two distinct blocks: an interband contribution (terms including $\mathbf{Y}_{nm}$) and an intraband part (terms containing $\mathbf{Q}_{nm}$). Despite the seemingly distinct appearance of Eqs. (B1a)–(B1d), they are equivalent as illustrated analytically in Sec. II C and numerically in Sec. III. We note that the VG quadratic conductivities computed using $\hat{U}_{VG}$ require the evaluation of $\mathbf{Q}_{nm}$, i.e., the generalized derivative, which is in contrast to the CVG using $\hat{U}_{CVG}$. Hence, the main advantage of performing computation in the VG, which is the absence of the generalized derivative, is lost if $\hat{U}_{VG}$ is used. In fact, computing the nonlinear conductivities in the VG with the correct interaction Hamiltonian, i.e., tensors labeled by $\mathcal{C}$ and $\mathcal{D}$, is more complicated than the LG approaches due to the presence of several extra terms in the conductivity expressions [cf. Eqs. (B1c) and (B1d)]. This becomes even more difficult in higher-order nonlinear responses due to additional terms in the interaction Hamiltonian and observable. Finally, as will be emphasized in Sec. III, a dense $\mathbf{k}$-vector grid is typically essential in order to eliminate the apparent zero-frequency divergence of $\boldsymbol{\sigma}^{\mathcal{C}(2)}$ and $\boldsymbol{\sigma}^{\mathcal{D}(2)}$.

C. Gauge invariance

It is straightforward to show that the initial dynamical equation for the density matrix in the MFA, i.e., Eq. (A1), behaves in a gauge-independent manner. Therefore it is expected that the ultimate expressions for the linear and nonlinear optical response, i.e., Eqs. (4) and (B1), are equivalent. Indeed, we demonstrate in this section that the expressions obtained by methods $\mathcal{A}$–$\mathcal{D}$ are equivalent. Regarding the CVG approach using $\hat{U}_{CVG}$, i.e., tensors labeled by $\mathcal{C}'$, we show that they are generally different from the rest.

First, let us focus on the tensors labeled $\mathcal{A}$–$\mathcal{D}$. Beginning with the linear response, it is obvious that the OC tensors obtained using methods $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{D}$ are indeed identical, since $\boldsymbol{\Pi}_n = -iE_n\mathbf{X}_n$. For the method $\mathcal{C}$, the denominators in Eq. (4c) are decomposed using a partial fraction expansion and rewritten as $1/\hbar\omega(\hbar\omega \pm E_n) = \pm 1/E_n(\hbar\omega \pm E_n) \mp 1/\hbar\omega E_n$. The terms due to $1/E_n(\hbar\omega \pm E_n)$ form a conductivity identical

to the $\boldsymbol{\sigma}^{\mathcal{B}(1)}$, whereas the remaining terms are canceled by the diamagnetic contribution. Therefore, $\boldsymbol{\sigma}^{\mathcal{C}(1)} = \boldsymbol{\sigma}^{\mathcal{B}(1)}$, and all four methods become equivalent.

Proceeding to the second-order response, hereafter for simplicity we limit our analysis to the second-harmonic generation (SHG) process, i.e., $\omega_p = \omega_q \equiv \omega$. However, the conclusions are generally valid for other second-order processes. The SHG conductivities obtained in the LG, i.e., $\boldsymbol{\sigma}^{\mathcal{A}(2)}$ and $\boldsymbol{\sigma}^{\mathcal{B}(2)}$, are related to each other via

$$\boldsymbol{\sigma}^{\mathcal{B}(2)} = \boldsymbol{\sigma}^{\mathcal{A}(2)} + C_{ee}\sum_{nm}\frac{\mathbf{X}_n\boldsymbol{\Pi}_{nm}\mathbf{X}_m^*}{(\hbar\omega + E_n)(\hbar\omega + E_m)}. \tag{8}$$

This is seen by rewriting the frequency-dependent terms in $\boldsymbol{\sigma}^{\mathcal{B}(2)}$, e.g., $2\hbar\omega/(2\hbar\omega - E_n) = 1 + E_n/(2\hbar\omega - E_n)$. The extra term on the right-hand side of Eq. (8) can be shown to vanish by exchanging the dummy indices, $m \leftrightarrow n$, and noticing that $\boldsymbol{\Pi}_{nm} = \boldsymbol{\Pi}_{mn}^* = -\boldsymbol{\Pi}_{mn}$ due to the time-reversal symmetry (see Appendix B). For the SHG conductivities in the VG, i.e., $\boldsymbol{\sigma}^{\mathcal{C}(2)}$ and $\boldsymbol{\sigma}^{\mathcal{D}(2)}$, it is straightforward to show that

$$\boldsymbol{\sigma}^{\mathcal{D}(2)} = \boldsymbol{\sigma}^{\mathcal{C}(2)} + \frac{iC_{ee}}{(\hbar\omega)^2}\sum_{nm}\boldsymbol{\Pi}_n\mathbf{X}_{nm}\boldsymbol{\Pi}_m^*$$

$$\times \frac{(2\hbar\omega + E_m - E_n)}{(\hbar\omega + E_n)(\hbar\omega + E_m)}, \tag{9a}$$

$$\boldsymbol{\sigma}^{\mathcal{D}(2)} = \boldsymbol{\sigma}^{\mathcal{B}(2)} + 2C_{ee}\sum_{nm}\frac{\mathbf{X}_n\boldsymbol{\Pi}_{nm}\mathbf{X}_m^*}{(\hbar\omega + E_n)(\hbar\omega + E_m)}. \tag{9b}$$

To derive these relations, the frequency-dependent fractions of Eqs. (B1c) and (B1d) have been rewritten using the same technique as employed for deriving Eq. (8). For Eq. (9a), as well as Eq. (9b), the second term on the right-hand side vanishes due to the time-reversal symmetry as in Eq. (8). So, despite the fact that $\boldsymbol{\sigma}^{\mathcal{A}(2)}$, $\boldsymbol{\sigma}^{\mathcal{B}(2)}$, $\boldsymbol{\sigma}^{\mathcal{C}(2)}$, and $\boldsymbol{\sigma}^{\mathcal{D}(2)}$ differ in form, they are equivalent regardless of the number of bands used in the calculation. In particular, we note that the zero-frequency divergences of $\boldsymbol{\sigma}^{\mathcal{C}(2)}$ and $\boldsymbol{\sigma}^{\mathcal{D}(2)}$ are only apparent.

Now, let us focus on the CVG, i.e., tensors labeled by $\mathcal{C}'$. One can show that the conductivity tensors obtained by method $\mathcal{C}'$ using Eqs. (4e) and (B1e) include several additional nonvanishing terms compared to the other four methods. Here, we demonstrate this fact for the linear response function, Eq. (4e), but the same conclusion can be drawn for the quadratic response, Eq. (B1e). Using Eq. (A5a), it is straightforward to show that

$$\boldsymbol{\Pi}_n = -iE_n\mathbf{X}_n = \mathbf{P}_n - i\mathbf{F}_n, \tag{10a}$$

$$\mathbf{F}_n \equiv \sum_{cv\mathbf{k}}\psi_{cv\mathbf{k}}^{(n)}\mathbf{f}_{cv\mathbf{k}}, \tag{10b}$$

where $\mathbf{f}_{cv\mathbf{k}} \equiv \sum_{c'v'\mathbf{k}'}W_{c'v'\mathbf{k}',cv\mathbf{k}}\mathbf{x}_{v'c'\mathbf{k}'}$. The value of $\mathbf{F}_n$ depends on the strength of the electron-hole interaction, and, hence, vanishes when excitonic effects are ignored. Thus, if $\mathbf{P}_n$ in Eq. (4e) is replaced by $\boldsymbol{\Pi}_n + i\mathbf{F}_n$, we obtain $\boldsymbol{\sigma}^{\mathcal{C}'(1)} = \boldsymbol{\sigma}^{\mathcal{C}(1)} +$ "extra term," where the "extra term" depends on the value of $\mathbf{F}_n$. We confirm numerically that this term is generally nonzero and contributes to the conductivity $\boldsymbol{\sigma}^{\mathcal{C}'(1)}$, which makes it different from the other four methods. Indeed, we will demonstrate numerically in Sec. III that by decreasing the effect of the Coulomb potential and including more bands in the

calculation the "extra term" contributes less and, hence, $\boldsymbol{\sigma}^{\mathcal{C}'(1)}$ converges toward $\boldsymbol{\sigma}^{\mathcal{A}(1)} - \boldsymbol{\sigma}^{\mathcal{D}(1)}$. The same behavior should follow for the nonlinear responses obtained using the CVG interaction. This is readily seen by noticing that $\boldsymbol{\Pi}_{nm} = i(E_n - E_m)\mathbf{X}_{nm} = \mathbf{P}_{nm} + i\mathbf{F}_{nm}$, where $\mathbf{F}_{nm}$ has a complicated form written in terms of $W_{c'v'\mathbf{k}',cv\mathbf{k}}$, analogous to Eq. (10). Hence, it is straightforward to confirm that $\boldsymbol{\sigma}^{\mathcal{C}(2)} = \boldsymbol{\sigma}^{\mathcal{C}(2)} + $ "extra term," where the nonvanishing "extra term" here is a function of both $\mathbf{F}_n$ and $\mathbf{F}_{nm}$.

III. NUMERICAL RESULTS

In this section, we apply the proposed theory to compute the excitonic optical response of hBN monolayers and compare the calculated OC and SHG spectra generated by the five methods of Table I. The single-particle band structure and required matrix elements are obtained from an empirical pseudopotential Hamiltonian [2]. This approach, which accurately reproduces the low-energy properties of hBN monolayers, allows us to have access to a large number of bands. The pseudopotential parametrization is reported in our previous work (see Ref. [40]). For the present numerical examples, we have used 85 reciprocal lattice vectors in the pseudopotential implementation, which generates a total of 85 bands including one valence (the band with lowest energy) and 84 conduction bands. In our numerical implementation, we assume that the eigenenergies obtained by the pseudopotential correspond to the quasiparticle energies, and the pseudopotential wave functions are used for computing all matrix elements. For instance, the interband position matrix elements are obtained using $i(\varepsilon_{n\mathbf{k}} - \varepsilon_{m\mathbf{k}})\langle n\mathbf{k}|\hat{\mathbf{x}}^{(e)}|m\mathbf{k}\rangle = \langle n\mathbf{k}|\hat{\mathbf{p}}|m\mathbf{k}\rangle$, where $\varepsilon_{n\mathbf{k}}$ and $|m\mathbf{k}\rangle$ are the pseudopotential energies and wave functions, respectively. Out of the 85 available bands, only the $N_b \geqslant 2$ lowest bands are included in the calculations. To ensure a proper convergence of the results, more than 11 000 $\mathbf{k}$ points
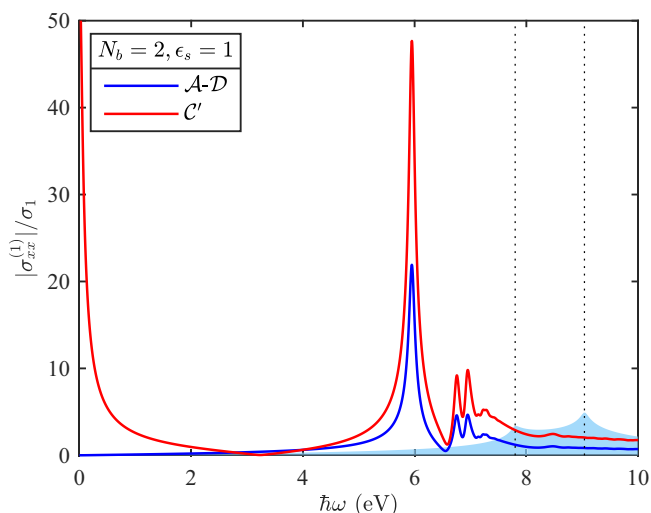

FIG. 2. Excitonic OC spectrum of hBN monolayer obtained by methods $\mathcal{A}$–$\mathcal{D}$ (blue) and $\mathcal{C}'$ (red) for different values of ϵ_s and N_b. Top panel: $\epsilon_s = 10$ (solid) and $\epsilon_s = 100$ (dashed) for $N_b = 2$. Bottom panel: $N_b = 2$ (solid) and $N_b = 5$ (dashed) for $\epsilon_s = 100$. Here, the line-shape broadening is set to $\eta = 0.1$ eV, and the OC values are normalized to σ_1 (see Fig. 1 caption). The black dotted lines mark $\hbar\omega = \{E_g, E_{\text{vH}}\}$.

FIG. 1. Excitonic OC spectrum of hBN monolayer obtained by methods $\mathcal{A}$–$\mathcal{D}$ (blue) and $\mathcal{C}'$ (red) for $N_b = 2$. The values are normalized to $\sigma_1 \equiv e^2/4\hbar = 6.0853 \times 10^{-5}$ S. For comparison purposes, the OC spectrum without excitonic effects found in method $\mathcal{A}$ is also plotted (filled light blue). The black dotted lines indicate $\hbar\omega = \{E_g, E_{\text{vH}}\}$.
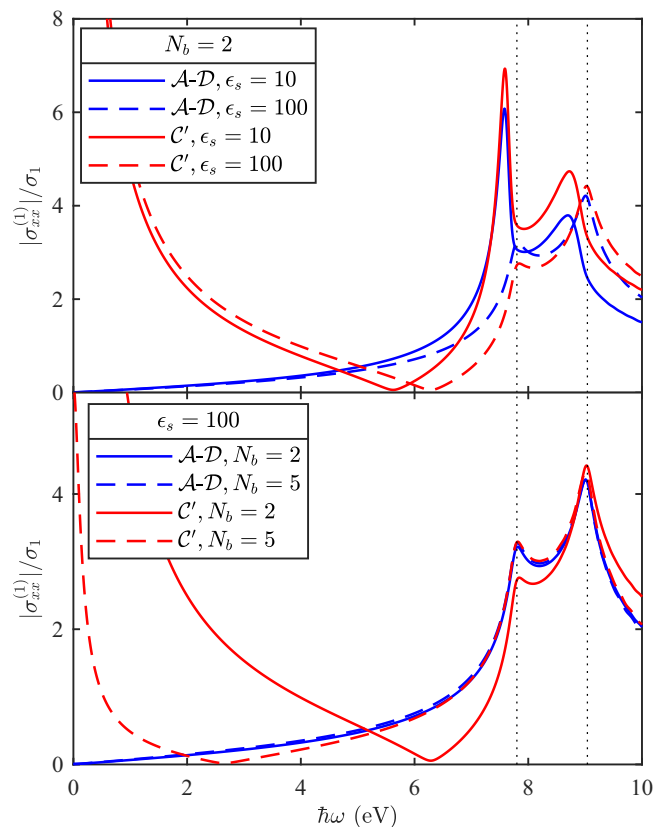
are used for discretizing the first BZ. A lattice constant of $a = 2.51$ Å is assumed, and the quasiparticle band gap and van Hove transition energies are $E_g = 7.78$ eV and $E_{\text{vH}} = 9.04$ eV, respectively. Due to the symmetry of the honeycomb lattice in hBN monolayers, it is sufficient to study only the diagonal components of the conductivity tensors, i.e., $\sigma_{xx}^{(1)}$ and $\sigma_{xxx}^{(2)}$ [9]. Finally, the line-shape broadening is accounted for by adding a small phenomenological imaginary part, $i\eta$, to the frequency, i.e., $\omega \rightarrow \omega + i\eta$. We set $\eta = 0.05$ eV for Figs. 1 and 3, whereas it is increased to $\eta = 0.1$ eV for Fig. 2 to ensure sufficiently smooth curves.

It is well known that for a realistic description of the exciton spectrum of 2D materials, the Coulomb potential should be accurately screened. However, the screening is not properly included in the MFA and, hence, it is introduced phenomenologically [45]. In the present work, we use the Keldysh potential for the direct Coulomb interaction, which is a widely accepted form of the screened potential for 2D materials [21,26,46–48]. In real space, the Keldysh potential is given by

$$\mathcal{V}^d(\mathbf{r}) = C_0 \frac{\pi}{2r_0}\left[\mathbb{H}_0\left(\frac{\epsilon_s r}{r_0}\right) - \mathbb{Y}_0\left(\frac{\epsilon_s r}{r_0}\right)\right], \quad (11)$$

where $C_0 \equiv e^2/4\pi\epsilon_0$, $r = |\mathbf{r}|$, and $\mathbb{H}_0$ and $\mathbb{Y}_0$ are the Struve function and Bessel function of the second type, respectively. The two parameters ϵ_s and r_0 are the substrate screening and screening length, respectively, which are set to $\epsilon_s = 1$ and $r_0 = 10$ Å for freely suspended hBN monolayers [21]. The Fourier transform of the Keldysh potential reads $\mathcal{V}^d(\mathbf{q}) \equiv C_0 2\pi/q(\epsilon_s + r_0 q)$, which is used for obtaining the Coulomb matrix elements according to Eq. (A3). The summation over $\mathbf{G}$ in Eq. (A3) is truncated to the seven smallest reciprocal vectors, since the impact of larger $\mathbf{G}$s becomes negligible. Regarding the exchange terms, we neglect them due to their minor impact on the results [48]. We note that the screening of the Coulomb potential influences the shape of the spectrum, yet our conclusions concerning gauge invariance are independent of the screening model.

Figure 1 shows $|\sigma_{xx}^{(1)}|$ of suspended hBN monolayers versus frequency obtained using Eqs. (4a)–(4e) for $N_b = 2$. For comparison purposes, we also plot the OC computed in the IPA limit simply by increasing the screening, i.e., $\epsilon_s \to \infty$. Without excitons, the response shows the expected features associated with the band gap and van Hove singularity [9]. In contrast, including the excitonic effects dramatically changes the spectrum by introducing a strong peak below the band gap at approximately 5.95 eV due to the fundamental exciton, while several other strong peaks are formed due to higher-order excitons. The excitonic OC spectrum is in good qualitative agreement with the previous results for hBN monolayers in Refs. [49], [50], and [19].

Now, let us focus on the differences between the excitonic responses computed by the five methods. The results in Fig. 1 confirm that the spectra generated by methods $\mathcal{A}$–$\mathcal{D}$ are numerically identical, whereas the spectrum obtained by method $\mathcal{C}'$ is considerably different. For instance, $\boldsymbol{\sigma}^{\mathcal{C}'(1)}$ suffers from a zero-frequency divergence, in contrast to the divergence-free $\boldsymbol{\sigma}^{\mathcal{A}(1)}$ to $\boldsymbol{\sigma}^{\mathcal{D}(1)}$. In addition, method $\mathcal{C}'$ overestimates the magnitude of the response function substantially over the whole frequency range. For any finite Coulomb screening ϵ_s, the differences between $\boldsymbol{\sigma}^{\mathcal{C}'(1)}$ and $\boldsymbol{\sigma}^{\mathcal{A}(1)}$-$\boldsymbol{\sigma}^{\mathcal{D}(1)}$ persist, and they do not disappear even for a complete basis set. Nonetheless, for a very large screening value, the disagreement between $\mathcal{C}'$ and $\mathcal{A}$–$\mathcal{D}$ diminishes by including more bands in the calculation, as discussed in Sec. II C. This is illustrated quantitatively in Fig. 2, where the OC spectra computed by methods $\mathcal{A}$−$\mathcal{D}$ and $\mathcal{C}'$ are displayed for two representative values of substrate screening, namely, $\epsilon_s = \{10, 100\}$, with $N_b = 2$ in the top panel. In the bottom panel of Fig. 2, we plot the same spectra for the larger screening value, i.e., $\epsilon_s = 100$, with $N_b = 2$ and $N_b = 5$. Increasing N_b from 2 to 5 barely influences the response generated by methods $\mathcal{A}$–$\mathcal{D}$, whereas the results of method $\mathcal{C}'$ differ considerably. Furthermore, the OC obtained by the CVG, i.e., $\boldsymbol{\sigma}^{\mathcal{C}'(1)}$, converges toward the results generated by other methods if both the screening and basis set size is increased, which is in agreement with the IPA results reported in Ref. [40].

Proceeding to the nonlinear response, Fig. 3 illustrates the SHG conductivities computed by methods $\mathcal{A}$–$\mathcal{D}$ and $\mathcal{C}'$ for two representative sizes of the basis set, $N_b = 2$ in the top panel and $N_b = 4$ in the bottom one. The SHG conductivities obtained in the IPA limit are also depicted for comparison. These responses agree with the results in Ref. [19]. Beginning with the IPA
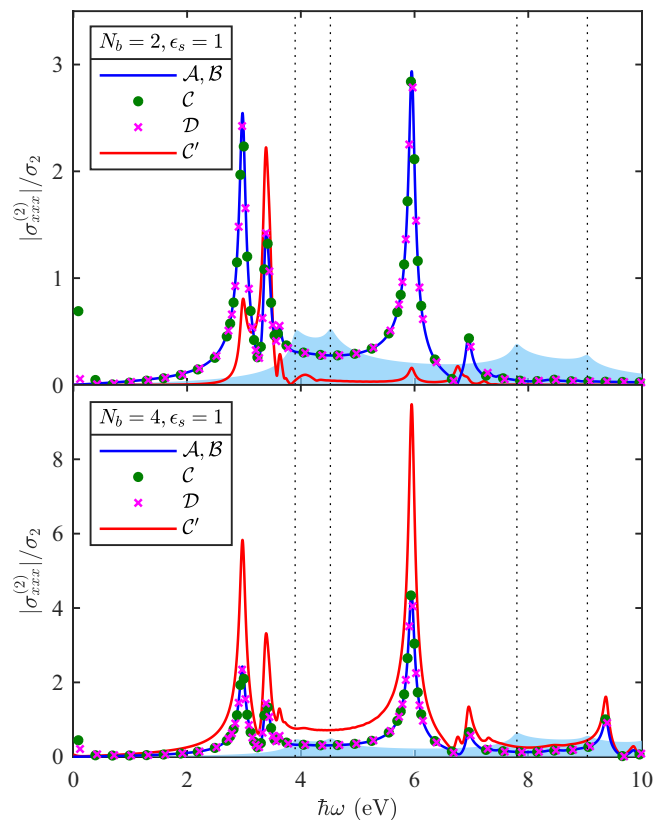

FIG. 3. Excitonic SHG spectrum of hBN monolayer obtained from methods $\mathcal{A}, \mathcal{B}$ (blue solid lines), $\mathcal{C}$ (green circles), $\mathcal{D}$ (magenta crosses), and $\mathcal{C}'$ (red solid lines) for $N_b = 2$ (top panel) and $N_b = 4$ (bottom panel). The values are normalized to $\sigma_2 \equiv e^3 a/4\gamma_0\hbar = 6.559 \times 10^{-15}$ S mV^{-1}, where we set $\gamma_0 = 2.33$ eV. For comparison purposes, the SHG conductivity spectrum of method $\mathcal{A}$ in the IPA (filled light blue) is also shown. The dotted lines from the left to right indicate $2\hbar\omega = E_g$, $2\hbar\omega = E_{vH}$, $\hbar\omega = E_g$, and $\hbar\omega = E_{vH}$, respectively.

result, the spectrum shows the features associated with $2\hbar\omega \approx \{E_g, E_{vH}\}$ and $\hbar\omega \approx \{E_g, E_{vH}\}$. Including more bands in the calculations barely changes the low-frequency resonances at $2\hbar\omega \approx \{E_g, E_{vH}\}$, whereas it enhances the high-frequency resonances mainly due to an interband contribution caused by higher conduction bands [40]. Adding excitons to the SHG response leads to a strong modification of the spectrum similar to the linear response, e.g., several strong resonances are formed by excitons at frequencies below $\{E_g/2, E_g\}$ [19,23].

Focusing on the excitonic SHG responses, the results show that the LG conductivities are numerically identical for both values of N_b, i.e., $\sigma_{xxx}^{\mathcal{A}(2)} = \sigma_{xxx}^{\mathcal{B}(2)}$. In addition, the conductivities computed in the VG with the correct interaction Hamiltonian, i.e., tensors labeled by $\mathcal{C}$ and $\mathcal{D}$, essentially agree with the calculations in the LG. The tiny differences between $\sigma_{xxx}^{\mathcal{C}(2)}/\sigma_{xxx}^{\mathcal{D}(2)}$ and $\sigma_{xxx}^{\mathcal{A}(2)}/\sigma_{xxx}^{\mathcal{B}(2)}$ at low frequencies are mainly due to the BZ discretization and diminish by using a finer $\mathbf{k}$ mesh. In particular, the zero-frequency divergences of methods $\mathcal{C}$ and $\mathcal{D}$ are only apparent. In contrast, the SHG responses found by method $\mathcal{C}'$ do not agree with the other four methods for both values of N_b. In particular, $\sigma_{xxx}^{\mathcal{C}'(2)}$ varies dramatically when more bands are included in the calculations, and the result for $N_b = 2$ is

highly inaccurate. In addition, even for $N_b = 4$, method $\mathcal{C}'$ overestimates the magnitudes of both ω and 2ω resonances by roughly a factor of 2. Nonetheless, method $\mathcal{C}'$ converges toward the conductivities computed by the other methods if $\epsilon_s \to \infty$ and $N_b \to \infty$, similarly to the OC responses discussed before. Summarizing, the excitonic SHG conductivities obtained using Eqs. (B1a)–(B1d) are equivalent regardless of the number of bands in the calculations, whereas the SHG response computed by the CVG, i.e., Eq. (B1e), does not agree with the rest even for a complete basis set.

IV. SUMMARY

In summary, we have theoretically investigated the gauge invariance of linear and nonlinear optical responses when excitonic effects are included. The expressions for conductivity tensors were derived rigorously in the density-matrix framework within the MFA for a multiband semiconductor. We have considered four distinct theoretical approaches derived from the combination of two choices of gauge and two ways of evaluating the current density, i.e., directly or via the polarization. We have shown both analytically and numerically that by using the correct interaction Hamiltonian and observable in

the VG, both the linear and quadratic responses obtained by the four methods become identical. The correct interaction in VG should be written in terms of the Heisenberg momentum $\hat{\mathbf{\Pi}}$, defined as the commutator of the unperturbed Hamiltonian and position operators. Despite the equivalence of the four methods, computing the conductivities in the LG, i.e., tensors labeled with $\mathcal{A}$ and $\mathcal{B}$, is more straightforward than the VG, i.e., tensors labeled with $\mathcal{C}$ and $\mathcal{D}$. Finally, the excitonic optical responses generated by the CVG interaction, i.e., tensors labeled with $\mathcal{C}'$, do not agree with the other methods, since $\hat{\mathbf{\Pi}}$ is generally different from the momentum operator $\hat{\mathbf{P}}$ when electron-hole interaction is included. The present formalism can readily be extended to generate gauge-invariant responses for higher-order nonlinear processes.

APPENDIX A: EQUATION OF MOTION AND ITS PERTURBATIVE SOLUTION

Here, we review the derivation of the dynamical equation for the density matrix and present its perturbative solution up to the second order. Our starting point is the many-body Hamiltonian in second quantization, Eq. (1). In the Heisenberg picture and within the MFA, this Hamiltonian leads to the usual equation of motion (quantum Liouville) for the density matrix ρ_{ji} [19],

$$i\hbar \frac{\partial \rho_{ji}}{\partial t} - \varepsilon_{ji}\rho_{ji} - \sum_{lmn}(\mathcal{V}_{mlni} - \mathcal{V}_{lmni})(\rho_{nl} - \delta_{mi}\delta_{nl}\delta_{lv})\rho_{jm} - \sum_{lmn}(\mathcal{V}_{jlmn} - \mathcal{V}_{jlnm})(\rho_{nl} - \delta_{mj}\delta_{nl}\delta_{lv})\rho_{mi}$$

$$= \sum_l (u_{jl}\rho_{li} - u_{li}\rho_{jl}), \tag{A1}$$

where $\varepsilon_{ji} \equiv \varepsilon_j - \varepsilon_i$ and the quasiparticle energies $\varepsilon_i \equiv \varepsilon_i^0 + \sum_l (\mathcal{V}_{ilil} - \mathcal{V}_{illi})\delta_{lv}$ are introduced, with the Kronecker delta serving to count occupied states only.

For the special case of Bloch states, each index should run over both band index and wave vector. To proceed, we assume that the density matrix is diagonal with respect to the wave vector, i.e., $\rho_{j\mathbf{k}_j i\mathbf{k}_i} \equiv \rho_{ji\mathbf{k}_i}\delta_{\mathbf{k}_i,\mathbf{k}_j}$ [19], since the diagonal part of the density matrix is the dominant contribution to the system response. Hence, the dynamical equation for the density matrix in crystals reads

$$i\hbar \frac{\partial \rho_{ji\mathbf{k}}}{\partial t} - \varepsilon_{ji\mathbf{k}}\rho_{ji\mathbf{k}} - \frac{1}{A}\sum_{lmn\mathbf{k}'}\left[\mathcal{V}_{mlni}^d(\mathbf{k},\mathbf{k}') - g\mathcal{V}_{lmni}^x(\mathbf{k}',\mathbf{k})\right](\rho_{nl\mathbf{k}'} - \delta_{mi}\delta_{nl}\delta_{lv})\rho_{jm\mathbf{k}} - \frac{1}{A}\sum_{lmn\mathbf{k}'}\left[g\mathcal{V}_{jlmn}^x(\mathbf{k},\mathbf{k}') - \mathcal{V}_{jlnm}^d(\mathbf{k},\mathbf{k}')\right]$$

$$\times (\rho_{nl\mathbf{k}'} - \delta_{mj}\delta_{nl}\delta_{lv})\rho_{mi\mathbf{k}} = \sum_l (u_{jl\mathbf{k}}\rho_{li\mathbf{k}} - u_{li\mathbf{k}}\rho_{jl\mathbf{k}}), \tag{A2}$$

where $\varepsilon_{ji\mathbf{k}} \equiv \varepsilon_{j\mathbf{k}} - \varepsilon_{i\mathbf{k}}$, and the extra factors of g appear due to the spin degeneracy of singlet states [12]. The direct and exchange Coulomb matrix elements $\mathcal{V}_{abcd}^d$ and $\mathcal{V}_{abcd}^x$ read

$$\mathcal{V}_{abcd}^d(\mathbf{k},\mathbf{k}') = \sum_{\mathbf{G}} I_{a\mathbf{k},c\mathbf{k}'}(\mathbf{G})I_{b\mathbf{k}',d\mathbf{k}}(-\mathbf{G})\mathcal{V}^d(\mathbf{k} - \mathbf{k}' - \mathbf{G}), \tag{A3a}$$

$$\mathcal{V}_{abcd}^x(\mathbf{k},\mathbf{k}') = \sum_{\mathbf{G}\neq\mathbf{0}} I_{a\mathbf{k},c\mathbf{k}}(\mathbf{G})I_{b\mathbf{k}',d\mathbf{k}'}(-\mathbf{G})\mathcal{V}^x(-\mathbf{G}). \tag{A3b}$$

Here, the summation is performed over reciprocal vectors $\mathbf{G}$, and the Bloch overlaps $I_{a\mathbf{k},c\mathbf{k}'}(\mathbf{G}) \equiv A_{\text{uc}}^{-1}\int_{\text{uc}}\varphi_{a\mathbf{k}}^*(\mathbf{r})\varphi_{c\mathbf{k}'}(\mathbf{r})\exp(i\mathbf{G}\cdot\mathbf{r})\text{d}^D\mathbf{r}$ are introduced. In Eq. (A3), $\mathcal{V}^d$ and $\mathcal{V}^x$ on the right-hand side are the Fourier transforms of the direct and exchange Coulomb potential, respectively. Note that the long-range contribution of the exchange part, i.e., $\mathbf{G} = \mathbf{0}$, is removed [16,22,30].

The equation of motion for $\rho_{ji\mathbf{k}}(t)$, Eq. (A2), is solved perturbatively by iteration to any order of perturbation, i.e., $\rho_{ji\mathbf{k}}(t) = \sum_N \rho_{ji\mathbf{k}}^{(N)}(t)$. The unperturbed solution, i.e., $\rho_{ji\mathbf{k}}^{(0)}$, for the case of cold clean semiconductors is given by $\rho_{vv'\mathbf{k}}^{(0)} = \delta_{vv'}$ and $\rho_{cc'\mathbf{k}}^{(0)} =$

$\rho_{cv\mathbf{k}}^{(0)} = \rho_{vc\mathbf{k}}^{(0)} = 0$. To the first order, $\rho_{cc'\mathbf{k}}^{(1)} \approx 0$ and $\rho_{vv'\mathbf{k}}^{(1)} \approx 0$, i.e., the field-induced changes in the band occupation are negligible [19]. Furthermore, the equation of motion for $\rho_{cv\mathbf{k}}^{(1)}$ reads

$$i\hbar \frac{\partial \rho_{cv\mathbf{k}}^{(1)}}{\partial t} - \sum_{c'v'\mathbf{k}'} H_{cv\mathbf{k},c'v'\mathbf{k}'} \rho_{c'v'\mathbf{k}'}^{(1)} - \sum_{c'v'\mathbf{k}'} T_{cv\mathbf{k},c'v'\mathbf{k}'} \rho_{v'c'\mathbf{k}'}^{(1)} = u_{cv\mathbf{k}}^{(1)}(t), \tag{A4}$$

where $u_{cv\mathbf{k}}^{(1)}(t)$ is the first-order contribution of the perturbation, and $H_{cv\mathbf{k},c'v'\mathbf{k}'}$ and $T_{cv\mathbf{k},c'v'\mathbf{k}'}$ are defined as

$$H_{cv\mathbf{k},c'v'\mathbf{k}'} \equiv \varepsilon_{cv\mathbf{k}} \delta_{c,c'} \delta_{v,v'} \delta_{\mathbf{k},\mathbf{k}'} + \frac{1}{A} \big[g\mathcal{V}_{cv'vc'}^x(\mathbf{k},\mathbf{k}') - \mathcal{V}_{cv'c'v}^d(\mathbf{k},\mathbf{k}') \big] \equiv \varepsilon_{cv\mathbf{k}} \delta_{c,c'} \delta_{v,v'} \delta_{\mathbf{k},\mathbf{k}'} + W_{cv\mathbf{k},c'v'\mathbf{k}'}, \tag{A5a}$$

$$T_{cv\mathbf{k},c'v'\mathbf{k}'} \equiv \frac{1}{A} \big[g\mathcal{V}_{cc'vv'}^x(\mathbf{k},\mathbf{k}') - \mathcal{V}_{cc'v'v}^d(\mathbf{k},\mathbf{k}') \big]. \tag{A5b}$$

Similarly, the equation of motion for $\rho_{vc\mathbf{k}}^{(1)}$ is found by taking the complex conjugate of Eq. (A4). One may solve the full coupled set of equations for $\rho_{cv\mathbf{k}}^{(1)}$ and $\rho_{vc\mathbf{k}}^{(1)}$. However, the $T_{cv\mathbf{k},c'v'\mathbf{k}'}$ terms can be ignored due to their small magnitude when compared to $H_{cv\mathbf{k},c'v'\mathbf{k}'}$, because $|I_{c\mathbf{k},v\mathbf{k}'}| \ll |I_{c\mathbf{k},c'\mathbf{k}'}|, |I_{v\mathbf{k},v'\mathbf{k}'}|$. This leads to the decoupling of $\rho_{cv\mathbf{k}}^{(1)}$ and $\rho_{vc\mathbf{k}}^{(1)}$ equations, which is known as the Tamm-Dancoff approximation [14,15]. Going one step further, the dynamical equations for the second-order density matrix in the Tamm-Dancoff approximation read

$$i\hbar \frac{\partial \rho_{cv\mathbf{k}}^{(2)}}{\partial t} - \sum_{c'v'\mathbf{k}'} H_{cv\mathbf{k},c'v'\mathbf{k}'} \rho_{c'v'\mathbf{k}'}^{(2)} = \sum_{c'} u_{cc'\mathbf{k}}^{(1)}(t) \rho_{c'v\mathbf{k}}^{(1)} - \sum_{v'} u_{v'v\mathbf{k}}^{(1)}(t) \rho_{cv'\mathbf{k}}^{(1)} + u_{cv\mathbf{k}}^{(2)}(t), \tag{A6a}$$

$$i\hbar \frac{\partial \rho_{cc'\mathbf{k}}^{(2)}}{\partial t} - \varepsilon_{cc'\mathbf{k}} \rho_{cc'\mathbf{k}}^{(2)} - \sum_{v'} \sum_{c_1 v_1 \mathbf{k}_1} W_{c'v'\mathbf{k},c_1 v_1 \mathbf{k}_1} \rho_{v_1 c_1 \mathbf{k}_1}^{(1)} \rho_{cv'\mathbf{k}}^{(1)} + \sum_{v'} \sum_{c_1 v_1 \mathbf{k}_1} W_{cv'\mathbf{k},c_1 v_1 \mathbf{k}_1} \rho_{c_1 v_1 \mathbf{k}_1}^{(1)} \rho_{v'c'\mathbf{k}}^{(1)}$$

$$= \sum_{v'} u_{cv'\mathbf{k}}^{(1)}(t) \rho_{v'c'\mathbf{k}}^{(1)} - \sum_{v'} u_{v'c'\mathbf{k}}^{(1)}(t) \rho_{cv'\mathbf{k}}^{(1)}, \tag{A6b}$$

$$i\hbar \frac{\partial \rho_{vv'\mathbf{k}}^{(2)}}{\partial t} - \varepsilon_{vv'\mathbf{k}} \rho_{vv'\mathbf{k}}^{(2)} - \sum_{c'} \sum_{c_1 v_1 \mathbf{k}_1} W_{c'v'\mathbf{k},c_1 v_1 \mathbf{k}_1} \rho_{c_1 v_1 \mathbf{k}_1}^{(1)} \rho_{v'c'\mathbf{k}}^{(1)} + \sum_{c'} \sum_{c_1 v_1 \mathbf{k}_1} W_{c'v\mathbf{k},c_1 v_1 \mathbf{k}_1} \rho_{v_1 c_1 \mathbf{k}_1}^{(1)} \rho_{c'v'\mathbf{k}}^{(1)}$$

$$= \sum_{c'} u_{vc'\mathbf{k}}^{(1)}(t) \rho_{c'v'\mathbf{k}}^{(1)} - \sum_{c'} u_{c'v'\mathbf{k}}^{(1)}(t) \rho_{vc'\mathbf{k}}^{(1)}, \tag{A6c}$$

where $u_{cv\mathbf{k}}^{(2)}(t)$ is the second-order contribution to the perturbation. Similarly, the equation of motion for $\rho_{vc\mathbf{k}}^{(2)}(t)$ is obtained by taking the complex conjugate of Eq. (A6a).

The set of nonhomogeneous equations of motion for the density matrix, Eqs. (A4) and (A6), can be solved by employing the Green's functions as explained in Ref. [19]. This is done by diagonalizing the matrix H_{eh} given in Eq. (A5a), i.e., $H_{eh} |\psi^{(n)}\rangle = E_n |\psi^{(n)}\rangle$, which is essentially the well-known BSE. Here, E_n and $|\psi^{(n)}\rangle$ are the exciton energies and eigenstates, which are written in the basis of vertical transitions from valence to conduction bands, i.e., $|\psi^{(n)}\rangle = \sum_{cv\mathbf{k}} \psi_{cv\mathbf{k}}^{(n)} |v\mathbf{k} \to c\mathbf{k}\rangle$. To continue, we consider an interaction potential of the form $\hat{U}(t) \equiv \hat{U}^{(1)} S(t) + \hat{U}^{(2)} S^2(t)$, where $S(t)$ is given as a set of time-harmonic terms, $S(t) \equiv 1/2 \sum_p S(\omega_p) e^{-i\omega_p t}$. Thus, the solutions of Eqs. (A4) and (A6) read

$$\rho_{vv'\mathbf{k}}^{(1)}(t) \approx 0, \quad \rho_{cc'\mathbf{k}}^{(1)}(t) \approx 0, \tag{A7a}$$

$$\rho_{cv\mathbf{k}}^{(1)}(t) = \frac{1}{2} \sum_p S(\omega_p) e^{-i\omega_p t} \left\{ \sum_n \frac{\psi_{cv\mathbf{k}}^{(n)} U_n^*}{\hbar\omega_p - E_n} \right\} = \rho_{vc\mathbf{k}}^{(1)*}(t), \tag{A7b}$$

$$\rho_{cv\mathbf{k}}^{(2)}(t) = \frac{1}{4} \sum_{pq} S(\omega_p) S(\omega_q) e^{-i\omega_2 t} \left\{ \sum_{nm} \left[\frac{\psi_{cv\mathbf{k}}^{(n)} U_{nm} U_m^*}{(\hbar\omega_2 - E_n)(\hbar\omega_q - E_m)} \right] + \sum_n \left[\frac{\psi_{cv\mathbf{k}}^{(n)} \bar{U}_n^*}{\hbar\omega_2 - E_n} \right] \right\} = \rho_{vc\mathbf{k}}^{(2)*}(t), \tag{A7c}$$

$$\rho_{cc'\mathbf{k}}^{(2)}(t) = -\frac{1}{4} \sum_{pq} S(\omega_p) S(\omega_q) e^{-i\omega_2 t} \left\{ \sum_{nm} \frac{U_n U_m^*}{(\hbar\omega_q + E_n)(\hbar\omega_p - E_m)} \sum_{v_1} \psi_{c'v_1\mathbf{k}}^{(n)*} \psi_{cv_1\mathbf{k}}^{(m)} \right\}, \tag{A7d}$$

$$\rho_{vv'\mathbf{k}}^{(2)}(t) = \frac{1}{4} \sum_{pq} S(\omega_p) S(\omega_q) e^{-i\omega_2 t} \left\{ \sum_{nm} \frac{U_n U_m^*}{(\hbar\omega_q + E_n)(\hbar\omega_p - E_m)} \sum_{c_1} \psi_{c_1 v\mathbf{k}}^{(n)*} \psi_{c_1 v'\mathbf{k}}^{(m)} \right\}, \tag{A7e}$$

where $\omega_2 \equiv \omega_p + \omega_q$, and the excitonic matrix elements of the perturbation are defined as

$$U_n \equiv \sum_{cv\mathbf{k}} \psi_{cv\mathbf{k}}^{(n)} u_{vc\mathbf{k}}^{(1)}, \quad \bar{U}_n \equiv \sum_{cv\mathbf{k}} \psi_{cv\mathbf{k}}^{(n)} u_{vc\mathbf{k}}^{(2)}, \quad U_{nm} \equiv \sum_{cv\mathbf{k}} \psi_{cv\mathbf{k}}^{(n)*} \left[\sum_{c_1} \psi_{c_1 v\mathbf{k}}^{(m)} u_{cc_1\mathbf{k}}^{(1)} - \sum_{v_1} \psi_{cv_1\mathbf{k}}^{(m)} u_{v_1 v\mathbf{k}}^{(1)} \right]. \tag{A8}$$

Note that U_n, $\bar{U}_n$ are the matrix elements between the ground state and exciton eigenstates, i.e., $U_n = \langle \psi_0 | \hat{U}^{(1)} | \psi^{(n)} \rangle$, $\bar{U}_n = \langle \psi_0 | \hat{U}^{(2)} | \psi^{(n)} \rangle$, and U_{nm} corresponds to a matrix element between two exciton eigenstates, i.e., $U_{nm} = \langle \psi^{(n)} | \hat{U}^{(1)} | \psi^{(m)} \rangle$. The second-order density matrix oscillates at frequency ω_2, which describes various second-order processes such as SHG ($\omega_p = \omega_q$) or optical rectification ($\omega_p = -\omega_q$).

Upon obtaining the density matrix, the expectation value of any one-body operator, i.e., an operator that acts on individual electrons, is determined straightforwardly. In second quantization, a one-body operator is given by $\hat{O} = \sum_{kl} o_{kl} \hat{c}_k^\dagger \hat{c}_l$, and its expectation value reads $\langle \hat{O} \rangle = \sum_{kl} o_{kl} \rho_{lk} = \text{tr}\{\hat{o}\hat{\rho}\}$. The operator $\hat{O}$ is assumed to contain a time-independent part $\hat{O}^{(0)}$, and a part that is first order in the perturbative field $\hat{O}^{(1)} S(t)$, i.e., $\hat{O} \equiv \hat{O}^{(0)} + \hat{O}^{(1)} S(t)$. Thus, the first- and second-order macroscopic responses of a system measured by $\hat{O}$ read

$$O^{(1)}(t) \equiv \text{tr}\{\hat{O}^{(0)} \rho^{(1)}\} + \text{tr}\{\hat{O}^{(1)} \rho^{(0)}\} = \frac{1}{2} \sum_p S(\omega_p) e^{-i\omega_p t} \left\{ \sum_n \left[\frac{O_n U_n^*}{\hbar\omega_p - E_n} - \frac{O_n^* U_n}{\hbar\omega_p + E_n} \right] + \sum_{v\mathbf{k}} o_{vv\mathbf{k}}^{(1)} \right\}, \tag{A9a}$$

$$O^{(2)}(t) \equiv \text{tr}\{\hat{O}^{(0)} \rho^{(2)}\} + \text{tr}\{\hat{O}^{(1)} \rho^{(1)}\} = \frac{1}{4} \sum_{pq} S(\omega_p) S(\omega_q) e^{-i\omega_2 t} \left\{ \sum_{nm} \left[\frac{O_n U_{nm} U_m^*}{(\hbar\omega_2 - E_n)(\hbar\omega_q - E_m)} + \frac{O_n^* U_{nm}^* U_m}{(\hbar\omega_2 + E_n)(\hbar\omega_q + E_m)} \right. \right.$$

$$\left. \left. - \frac{U_n O_{nm} U_m^*}{(\hbar\omega_q + E_n)(\hbar\omega_p - E_m)} \right] + \sum_n \left[\frac{O_n \bar{U}_n^*}{\hbar\omega_2 - E_n} - \frac{O_n^* \bar{U}_n}{\hbar\omega_2 + E_n} \right] + \sum_n \left[\frac{\bar{O}_n U_n^*}{\hbar\omega_p - E_n} - \frac{\bar{O}_n^* U_n}{\hbar\omega_p + E_n} \right] \right\}, \tag{A9b}$$

where the matrix elements of many-body observables O_n, $\bar{O}_n$, and O_{nm} are defined analogous to their interaction counterpart, Eq. (A8), so that

$$O_n \equiv \sum_{cv\mathbf{k}} \psi_{cv\mathbf{k}}^{(n)} o_{vc\mathbf{k}}^{(0)}, \quad \bar{O}_n \equiv \sum_{cv\mathbf{k}} \psi_{cv\mathbf{k}}^{(n)} o_{vc\mathbf{k}}^{(1)}, \quad O_{nm} \equiv \sum_{cv\mathbf{k}} \psi_{cv\mathbf{k}}^{(n)*} \left[\sum_{c_1} \psi_{c_1 v\mathbf{k}}^{(m)} o_{cc_1\mathbf{k}}^{(0)} - \sum_{v_1} \psi_{cv_1\mathbf{k}}^{(m)} o_{v_1 v\mathbf{k}}^{(0)} \right]. \tag{A10}$$

We note that the last term in Eq. (A9a) is the matrix element of $\hat{O}^{(1)}$ with respect to the ground state, i.e., $\sum_{v\mathbf{k}} o_{vv\mathbf{k}}^{(1)} = \langle \psi_0 | \hat{O}^{(1)} | \psi_0 \rangle$.

APPENDIX B: QUADRATIC OPTICAL RESPONSE

The expressions for the second-order conductivities of the five methods in Table I are derived using Eq. (A9b) and given by

$$\sigma^{\mathcal{A}(2)} = -C_{ee} \sum_{nm} \left[\frac{\mathbf{\Pi}_n \mathbf{X}_{nm} \mathbf{X}_m^*}{(\hbar\omega_2 - E_n)(\hbar\omega_q - E_m)} + \frac{\mathbf{\Pi}_n^* \mathbf{X}_{nm}^* \mathbf{X}_m}{(\hbar\omega_2 + E_n)(\hbar\omega_q + E_m)} - \frac{\mathbf{X}_n \mathbf{\Pi}_{nm} \mathbf{X}_m^*}{(\hbar\omega_q + E_n)(\hbar\omega_p - E_m)} \right], \tag{B1a}$$

$$\sigma^{\mathcal{B}(2)} = +C_{ee}(i\hbar\omega_2) \sum_{nm} \left[\frac{\mathbf{X}_n \mathbf{X}_{nm} \mathbf{X}_m^*}{(\hbar\omega_2 - E_n)(\hbar\omega_q - E_m)} + \frac{\mathbf{X}_n^* \mathbf{X}_{nm}^* \mathbf{X}_m}{(\hbar\omega_2 + E_n)(\hbar\omega_q + E_m)} - \frac{\mathbf{X}_n \mathbf{X}_{nm} \mathbf{X}_m^*}{(\hbar\omega_q + E_n)(\hbar\omega_p - E_m)} \right], \tag{B1b}$$

$$\sigma^{\mathcal{C}(2)} = +\frac{C_{ee}}{(\hbar\omega_p)(\hbar\omega_q)} \sum_{nm} \left[\frac{\mathbf{\Pi}_n \mathbf{\Pi}_{nm} \mathbf{\Pi}_m^*}{(\hbar\omega_2 - E_n)(\hbar\omega_q - E_m)} + \frac{\mathbf{\Pi}_n^* \mathbf{\Pi}_{nm}^* \mathbf{\Pi}_m}{(\hbar\omega_2 + E_n)(\hbar\omega_q + E_m)} - \frac{\mathbf{\Pi}_n \mathbf{\Pi}_{nm} \mathbf{\Pi}_m^*}{(\hbar\omega_q + E_n)(\hbar\omega_p - E_m)} \right]$$

$$- \frac{C_{ee}}{2i(\hbar\omega_p)(\hbar\omega_q)} \sum_n \left[\frac{\mathbf{\Pi}_n \mathbf{A}_n^*}{\hbar\omega_2 - E_n} + \frac{\mathbf{\Pi}_n^* \mathbf{A}_n}{\hbar\omega_2 + E_n} \right] + \frac{C_{ee}}{i(\hbar\omega_p)(\hbar\omega_q)} \sum_n \left[\frac{\mathbf{A}_n \mathbf{\Pi}_n^*}{\hbar\omega_p - E_n} + \frac{\mathbf{A}_n^* \mathbf{\Pi}_n}{\hbar\omega_p + E_n} \right], \tag{B1c}$$

$$\sigma^{\mathcal{D}(2)} = -\frac{C_{ee}(i\hbar\omega_2)}{(\hbar\omega_p)(\hbar\omega_q)} \sum_{nm} \left[\frac{\mathbf{X}_n \mathbf{\Pi}_{nm} \mathbf{\Pi}_m^*}{(\hbar\omega_2 - E_n)(\hbar\omega_q - E_m)} + \frac{\mathbf{X}_n^* \mathbf{\Pi}_{nm}^* \mathbf{\Pi}_m}{(\hbar\omega_2 + E_n)(\hbar\omega_q + E_m)} - \frac{\mathbf{\Pi}_n \mathbf{X}_{nm} \mathbf{\Pi}_m^*}{(\hbar\omega_q + E_n)(\hbar\omega_p - E_m)} \right]$$

$$+ \frac{C_{ee}(\hbar\omega_2)}{2(\hbar\omega_p)(\hbar\omega_q)} \sum_n \left[\frac{\mathbf{X}_n \mathbf{A}_n^*}{\hbar\omega_2 - E_n} + \frac{\mathbf{X}_n^* \mathbf{A}_n}{\hbar\omega_2 + E_n} \right], \tag{B1d}$$

$$\sigma^{\mathcal{C}'(2)} = +\frac{C_{ee}}{(\hbar\omega_p)(\hbar\omega_q)} \sum_{n,m} \left[\frac{\mathbf{P}_n \mathbf{P}_{nm} \mathbf{P}_m^*}{(\hbar\omega_2 - E_n)(\hbar\omega_q - E_m)} + \frac{\mathbf{P}_n^* \mathbf{P}_{nm}^* \mathbf{P}_m}{(\hbar\omega_2 + E_n)(\hbar\omega_q + E_m)} - \frac{\mathbf{P}_n \mathbf{P}_{nm} \mathbf{P}_m^*}{(\hbar\omega_q + E_n)(\hbar\omega_p - E_m)} \right], \tag{B1e}$$

where $C_{ee} \equiv ge^3\hbar^2/(m^3A)$, $\mathbf{A}_n \equiv \langle\psi_0|[\hat{\mathbf{X}},\hat{\mathbf{\Pi}}]|\psi^{(n)}\rangle = \sum_m(\mathbf{X}_m\mathbf{\Pi}_{mn} - \mathbf{\Pi}_m\mathbf{X}_{mn})$, $\mathbf{\Pi}_{nm} = i(E_n - E_m)\mathbf{X}_{nm}$, and $\mathbf{P}_{nm}$ and $\mathbf{X}_{nm}$ are defined using Eq. (A8) as

$$\mathbf{P}_{nm} \equiv \sum_{cv\mathbf{k}} \psi_{cv\mathbf{k}}^{(n)*}\left[\sum_{c_1}\psi_{c_1v\mathbf{k}}^{(m)}\mathbf{p}_{cc_1\mathbf{k}} - \sum_{v_1}\psi_{cv_1\mathbf{k}}^{(m)}\mathbf{p}_{v_1v\mathbf{k}}\right], \tag{B2a}$$

$$\mathbf{X}_{nm} \equiv \sum_{cv\mathbf{k}} \psi_{cv\mathbf{k}}^{(n)*}\left[\sum_{c_1}\psi_{c_1v\mathbf{k}}^{(m)}\mathbf{x}_{cc_1\mathbf{k}} - \sum_{v_1}\psi_{cv_1\mathbf{k}}^{(m)}\mathbf{x}_{v_1v\mathbf{k}}\right]. \tag{B2b}$$

Evaluating the momentum matrix elements $\mathbf{P}_{nm}$ is rather straightforward, whereas the matrix elements of the ill-defined position operator in $\mathbf{X}_{nm}$ should be separated to its interband and intraband parts as outlined in Sec. II B. Hence, we split the summations in this expression into two distinct contributions: $\mathbf{X}_{nm} = \mathbf{Y}_{nm} + m\mathbf{Q}_{nm}/\hbar$, where $\mathbf{Y}_{nm}$ and $\mathbf{Q}_{nm}$ contain the interband ($c \neq c_1$ and $v \neq v_1$) and intraband ($c = c_1$ and $v = v_1$) components, respectively. So, $\mathbf{Y}_{nm}$ and $\mathbf{Q}_{nm}$ are given by

$$\mathbf{Y}_{nm} \equiv \sum_{cv\mathbf{k}} \psi_{cv\mathbf{k}}^{(n)*}\left[\sum_{c_1 \neq c}\psi_{c_1v\mathbf{k}}^{(m)}\mathbf{x}_{cc_1\mathbf{k}} - \sum_{v_1 \neq v}\psi_{cv_1\mathbf{k}}^{(m)}\mathbf{x}_{v_1v\mathbf{k}}\right], \tag{B3a}$$

$$\mathbf{Q}_{nm} \equiv \frac{\hbar}{m}\sum_{cv\mathbf{k}} \psi_{cv\mathbf{k}}^{(n)*}\left[\psi_{cv\mathbf{k}}^{(m)}\mathbf{x}_{cc\mathbf{k}} - \psi_{cv\mathbf{k}}^{(m)}\mathbf{x}_{vv\mathbf{k}}\right] = i\sum_{cv\mathbf{k}} \psi_{cv\mathbf{k}}^{(n)*}\left[\psi_{cv\mathbf{k}}^{(m)}\right]_{;\mathbf{k}}. \tag{B3b}$$

In the last line, the rule $(\mathbf{r}_{cc\mathbf{k}} - \mathbf{r}_{vv\mathbf{k}})\psi_{cv\mathbf{k}}^{(m)} = i[\psi_{cv\mathbf{k}}^{(m)}]_{;\mathbf{k}}$ has been used [40].

Time-reversal symmetry in periodic systems is extremely useful and allows one to choose the phase such that $\psi_{cv(-\mathbf{k})}^{(n)} = \psi_{cv\mathbf{k}}^{(n)*}$, $\mathbf{x}_{nm(-\mathbf{k})} = \mathbf{x}_{mn\mathbf{k}}^*$, and $\mathbf{p}_{nm(-\mathbf{k})} = -\mathbf{p}_{mn\mathbf{k}}^*$. With this choice of phase, one can show that $\mathbf{P}_n = -\mathbf{P}_n^*$, $\mathbf{\Pi}_n = -\mathbf{\Pi}_n^*$, $\mathbf{X}_n = \mathbf{X}_n^*$, $\mathbf{A}_n = -\mathbf{A}_n^*$, $\mathbf{Y}_{nm} = \mathbf{Y}_{nm}^*$, $\mathbf{Q}_{nm} = \mathbf{Q}_{nm}^*$, $\mathbf{P}_{nm} = -\mathbf{P}_{nm}^*$, and $\mathbf{\Pi}_{nm} = -\mathbf{\Pi}_{nm}^*$ [19]. These relations can be used to simplify the expressions of conductivity tensors, which are generally valid for any other phase choice, since all expressions should be independent of the chosen phase.

[1] R. W. Boyd, *Nonlinear Optics*, 3rd ed. (Elsevier Science Publishing Co., Inc., New York, 2008).

[2] P. Y. Yu and M. Cardona, *Fundamentals of Semiconductors: Physics and Materials Properties* (Springer, New York, 2010).

[3] J. E. Sipe and E. Ghahramani, Phys. Rev. B **48**, 11705 (1993).

[4] C. Aversa and J. E. Sipe, Phys. Rev. B **52**, 14636 (1995).

[5] S. N. Rashkeev and W. R. L. Lambrecht, Phys. Rev. B **63**, 165212 (2001).

[6] T. G. Pedersen and K. Pedersen, Phys. Rev. B **79**, 035422 (2009).

[7] V. A. Margulis, E. E. Muryumin, and E. Gaiduk, J. Phys.: Condens. Matter **25**, 195302 (2013).

[8] S. A. Mikhailov, Phys. Rev. B **93**, 085403 (2016).

[9] F. Hipolito, T. G. Pedersen, and V. M. Pereira, Phys. Rev. B **94**, 045434 (2016).

[10] F. Hipolito, A. Taghizadeh, and T. G. Pedersen, arXiv:1802.01430.

[11] J. Ibañez-Azpiroz, S. S. Tsirkin, and I. Souza, arXiv:1804.04030.

[12] K. Leung and K. B. Whaley, Phys. Rev. B **56**, 7455 (1997).

[13] S. Albrecht, L. Reining, R. Del Sole, and G. Onida, Phys. Rev. Lett. **80**, 4510 (1998).

[14] L. X. Benedict, E. L. Shirley, and R. B. Bohn, Phys. Rev. Lett. **80**, 4514 (1998).

[15] M. Rohlfing and S. G. Louie, Phys. Rev. B **62**, 4927 (2000).

[16] G. Onida, L. Reining, and A. Rubio, Rev. Mod. Phys. **74**, 601 (2002).

[17] T. G. Pedersen, Phys. Rev. B **67**, 073401 (2003).

[18] F. Wang, G. Dukovic, L. E. Brus, and T. F. Heinz, Science **308**, 838 (2005).

[19] T. G. Pedersen, Phys. Rev. B **92**, 235432 (2015).

[20] P. Cudazzo, L. Sponza, C. Giorgetti, L. Reining, F. Sottile, and M. Gatti, Phys. Rev. Lett. **116**, 066803 (2016).

[21] T. Galvani, F. Paleari, H. P. C. Miranda, A. Molina-Sánchez, L. Wirtz, S. Latil, H. Amara, and F. Ducastelle, Phys. Rev. B **94**, 125303 (2016).

[22] J. Koskelo, G. Fugallo, M. Hakala, M. Gatti, F. Sottile, and P. Cudazzo, Phys. Rev. B **95**, 035125 (2017).

[23] M. Grüning and C. Attaccalite, Phys. Rev. B **89**, 081102 (2014).

[24] M. L. Trolle, Y.-C. Tsao, K. Pedersen, and T. G. Pedersen, Phys. Rev. B **92**, 161409 (2015).

[25] S. Gao, Y. Liang, C. D. Spataru, and L. Yang, Nano Lett. **16**, 5568 (2016).

[26] D. Y. Qiu, F. H. da Jornada, and S. G. Louie, Phys. Rev. B **93**, 235435 (2016).

[27] T. Olsen, S. Latini, F. Rasmussen, and K. S. Thygesen, Phys. Rev. Lett. **116**, 056401 (2016).

[28] C. Attaccalite, E. Cannuccia, and M. Grüning, Phys. Rev. B **95**, 125403 (2017).

[29] E. K. Chang, E. L. Shirley, and Z. H. Levine, Phys. Rev. B **65**, 035205 (2001).

[30] R. Leitsmann, W. G. Schmidt, P. H. Hahn, and F. Bechstedt, Phys. Rev. B **71**, 195209 (2005).

[31] M. L. Trolle, G. Seifert, and T. G. Pedersen, Phys. Rev. B **89**, 235410 (2014).

[32] D. H. Kobe, Phys. Rev. A **19**, 205 (1979).

[33] R. Girlanda, A. Quattropani, and P. Schwendimann, Phys. Rev. B **24**, 2009 (1981).

[34] L. B. Madsen, Phys. Rev. A **65**, 053417 (2002).

[35] K. Rzążewski and R. W. Boyd, J. Mod. Opt. **51**, 1137 (2004).

[36] K. S. Virk and J. E. Sipe, Phys. Rev. B **76**, 035213 (2007).

[37] H. M. Dong, K. Han, and W. Xu, J. Appl. Phys. **115**, 063503 (2014).

[38] P. Földi, Phys. Rev. B **96**, 035112 (2017).

[39] G. B. Ventura, D. J. Passos, J. M. B. Lopes dos Santos, J. M. Viana Parente Lopes, and N. M. R. Peres, Phys. Rev. B **96**, 035431 (2017).

[40] A. Taghizadeh, F. Hipolito, and T. G. Pedersen, Phys. Rev. B **96**, 195413 (2017).

[41] V. S. Yakovlev and M. S. Wismer, Comput. Phys. Commun. **217**, 82 (2017).

[42] D. J. Passos, G. B. Ventura, J. M. Viana Parente Lopes, J. M. B. Lopes dos Santos, and N. M. R. Peres, arXiv:1712.04924.

[43] E. Ridolfi, C. H. Lewenkopf, and V. M. Pereira, Phys. Rev. B **97**, 205409 (2018).

[44] A. F. Starace, Phys. Rev. A **3**, 1242 (1971).

[45] W. Schäfer and M. Wegener, *Semiconductor Optics and Transport Phenomena* (Springer, New York, 2002).

[46] P. Cudazzo, I. V. Tokatly, and A. Rubio, Phys. Rev. B **84**, 085406 (2011).

[47] M. L. Trolle, T. G. Pedersen, and V. Véniard, Sci. Rep. **7**, 39844 (2017).

[48] K. S. Thygesen, 2D Mater. **4**, 022004 (2017).

[49] L. Wirtz, A. Marini, and A. Rubio, Phys. Rev. Lett. **96**, 126104 (2006).

[50] C. Attaccalite, M. Bockstedte, A. Marini, A. Rubio, and L. Wirtz, Phys. Rev. B **83**, 144115 (2011).