Second-order nonlinear optical response of graphene

Yongrui Wang, Mikhail Tokman, and Alexey Belyanin

Phys. Rev. B **94**, 195442 — Published 30 November 2016

DOI: 10.1103/PhysRevB.94.195442

Second-order nonlinear optical response of graphene

Yongrui Wang,[1] Mikhail Tokman,[2] and Alexey Belyanin[1]

[1] *Department of Physics and Astronomy,*
Texas A&M University, College Station, TX, 77843 USA

[2] *Institute of Applied Physics, Russian Academy of Sciences*

(Dated: November 7, 2016)

Abstract

Although massless Dirac fermions in graphene constitute a centrosymmetric medium for in-plane excitations, their second-order nonlinear optical response is nonzero if the effects of spatial dispersion are taken into account. Here we present a rigorous quantum-mechanical theory of the second-order nonlinear response of graphene beyond the electric dipole approximation, which includes both intraband and interband transitions. The resulting nonlinear susceptibility tensor satisfies all symmetry and permutation properties, and can be applied to all three-wave mixing processes. We obtain useful analytic expressions in the limit of a degenerate electron distribution, which reveal quite strong second-order nonlinearity at long wavelengths, Fermi-edge resonances, and unusual polarization properties.

I. INTRODUCTION

Nonlinear optical properties of graphene have attracted considerable interest in the community. The magnitude of the matrix element of the interaction Hamiltonian describing coupling of massless Dirac electrons to light scales in proportion to $v_F/\omega \propto \lambda$, i.e. it grows more rapidly with wavelength λ than in conventional materials with parabolic energy dispersion, where it scales roughly as $\sqrt{\lambda}$. This promises a strong nonlinear response at long wavelengths. Unfortunately, graphene is also a centrosymmetric medium for low-energy in-plane excitations, which suppresses second-order nonlinear response in the electric dipole approximation. Therefore, most of the effort was concentrated on the third-order nonlinear processes that are electric dipole-allowed. Recent theoretical proposals and some experiments include third-harmonic generation[1,2], four-wave mixing[3–5], current-induced second-harmonic generation (SHG)[6–8] and SHG in biased bilayer graphene[9]. In few-layer graphene, SHG arising from the interactions between layers, which breaks the inversion symmetry, has been observed[10,11]. Multiphoton excitation of electron-hole pairs in monolayer and bilayer graphene was theoretically studied in[12,13].

The aim of this paper is to show that monolayer graphene does demonstrate quite significant second-order nonlinearity at long wavelengths despite its inversion symmetry. Here and throughout the paper, we will discuss only the 2D (surface) nonlinearity due to in-plane motion of electrons. Like any surface, graphene exhibits anisotropy between in-plane and out-of-plane electron motion. However, the corresponding second-order nonlinearity is very small and we will not discuss it here.

We develop the full quantum-mechanical theory of the in-plane second-order nonlinear response beyond the electric dipole approximation. In this case one has to consider oblique or in-plane propagation of electromagnetic waves. A non-zero in-plane second-order susceptibility $\chi^{(2)}$ of monolayer graphene appears when one includes the dependence of $\chi^{(2)}$ on the in-plane photon wave vectors, i.e. the *spatial dispersion.* Physically, this means that the inversion symmetry of graphene is broken by the wave vector direction. The spatial dispersion in momentum space is of course equivalent to the nonlocal response in real space. Spatial dispersion effects turn out to be quite large because of a large magnitude of the electron velocity v_F. A non-zero value of the nonlocal $\chi^{(2)}$ has been pointed out before for second-harmonic generation[14–16] (which only included intraband transitions in a free-carrier model), difference-frequency generation[17], and parametric frequency down-conversion[18]. The latter two papers developed a quantum theory including both intraband and interband transitions and applied it to the nonlinear generation of surface plasmons. In the recent experiment[19], evidence for the difference-frequency generation of surface plasmons in graphene was reported. Here we provide a systematic derivation of the second-order nonlinear conductivity tensor, valid for all second-order processes, all frequencies and doping densities, as long as the massless Dirac fermion approximation for a single-particle Hamiltonian is applicable. For graphene, this means the range of frequencies from zero (more precisely, from inverse scattering time) to the near-infrared. Our approach can be applied to any system of massless chiral Dirac fermions, for example surface states in topological insulators such as $\mathrm{Bi_2Se_3}$. The resulting nonlinear susceptibility tensor satisfies all symmetry and permutation properties, and predicts unusual polarization properties of the nonlinear signal. We also summarize main properties of the linear current as a necessary step in deriving the nonlinear response functions, and present a detailed discussion of its gauge properties and regularization.

II. BASIC EQUATIONS

Consider a 2D quantum system which in the absence of external fields can be described by the Dirac Hamiltonian

$$\hat{H}_0(\hat{\boldsymbol{p}}) = v_F \hat{\boldsymbol{\sigma}} \cdot \hat{\boldsymbol{p}}, \tag{1}$$

where $\hat{\boldsymbol{p}} = \boldsymbol{x}_0 \hat{p}_x + \boldsymbol{y}_0 \hat{p}_y$, $\hat{p}_{x,y} = -i\hbar \frac{\partial}{\partial x, \partial y}$, $\hat{\sigma} = \boldsymbol{x}_0 \hat{\sigma}_x + \boldsymbol{y}_0 \hat{\sigma}_y$, where $\hat{\sigma}_{x,y}$ are Pauli matrices. The spinor eigenfunctions $\boldsymbol{\Psi} = \begin{pmatrix} \Psi_1 \\ \Psi_2 \end{pmatrix}$ of the Hamiltonian (1) are

$$\boldsymbol{\Psi}_{\boldsymbol{k},s}(\boldsymbol{r}) \equiv \langle \boldsymbol{r} | \boldsymbol{k}, s \rangle = \frac{e^{i\boldsymbol{k}\cdot\boldsymbol{r}}}{\sqrt{2A}} \begin{pmatrix} s \\ e^{i\theta(\boldsymbol{k})} \end{pmatrix}, \tag{2}$$

and the eigenenergies are $E = s\hbar v_F k$ where $s = \pm 1$ for conduction and valence bands, respectively; $\boldsymbol{k} = \boldsymbol{x}_0 k_x + \boldsymbol{y}_0 k_y$, $\theta(\boldsymbol{k})$ is the angle between the electron momentum $\boldsymbol{k}$ and x-axis, and A is the normalization area. This description is valid for carriers in monolayer graphene up to the energies of order 1 eV. For higher energies, quadratic and trigonal warping corrections become non-negligible.

Consider the most general light-matter interaction Hamiltonian utilizing both vector and scalar potentials: $\boldsymbol{E} = -\nabla\varphi - c^{-1}\dot{\boldsymbol{A}}$ and $\boldsymbol{B} = \nabla \times \boldsymbol{A}$. Following a standard procedure[20,21], we replace $\hat{\boldsymbol{p}} \Rightarrow \hat{\boldsymbol{p}} + \frac{e}{c}\boldsymbol{A}$ in the unperturbed Hamiltonian $\hat{H}_0(\hat{\boldsymbol{p}})$ and add the potential energy operator $-e\varphi$ assuming a particle with the charge $-e$. This gives

$$\hat{H} = \hat{H}_0 + \hat{H}_{int}^{opt}, \qquad \hat{H}_{int}^{opt} = \frac{ev_F}{c}\hat{\boldsymbol{\sigma}} \cdot \boldsymbol{A} - e\varphi \cdot \hat{1}, \tag{3}$$

where $\hat{1}$ is a unit 2×2 matrix. The Hamiltonian in Eq. (3) leads to the von Neumann equation for the density matrix:

$$i\hbar\frac{\partial}{\partial t}\rho_{mn} = (E_m - E_n)\rho_{mn} + \sum_l \left[\left(\hat{H}_{int}^{opt}\right)_{ml}\rho_{ln} - \rho_{ml}\left(\hat{H}_{int}^{opt}\right)_{ln}\right], \tag{4}$$

where $|n\rangle = |\boldsymbol{k}, s\rangle$.

We will consider a monochromatic electromagnetic field in plane of graphene,

$$\boldsymbol{E} = \frac{1}{2}\left[\boldsymbol{x}_0 E_x(\omega) + \boldsymbol{y}_0 E_y(\omega)\right] e^{-i\omega t + iqx} + \text{C.C.} \tag{5}$$

or its bichromatic combinations. The field component $\boldsymbol{z}_0 E_z$ can be ignored because neither this field component itself nor the magnetic field it generates can affect the 2D carrier motion. Furthermore, the component of the vector potential $\boldsymbol{z}_0 A_z$ which generates the z-component of the electric field $\boldsymbol{z}_0 E_z$ does not enter the Hamiltonian (3) because $\hat{\boldsymbol{\sigma}} \cdot \boldsymbol{z}_0 = 0$. The field described by Eq. (5) corresponds to the electromagnetic potentials

$$\begin{aligned} \varphi &= \frac{1}{2}\phi(\omega)e^{-i\omega t + iqx} + \text{C.C.}, \\ \boldsymbol{A} &= \frac{1}{2}\left[\boldsymbol{x}_0 A_x(\omega) + \boldsymbol{y}_0 A_y(\omega)\right] e^{-i\omega t + iqx} + \text{C.C.} \end{aligned} \tag{6}$$

Note that the P-polarized radiation can be defined through both the scalar potential,

$$\varphi = \frac{1}{2}\frac{iE_x(\omega)}{q}e^{-i\omega t + iqx} + \text{C.C.}, \tag{7}$$

and the vector potential:

$$\boldsymbol{A}_P = \frac{1}{2}\boldsymbol{x}_0\frac{cE_x(\omega)}{i\omega}e^{-i\omega t + iqx} + \text{C.C.} \tag{8}$$

At the same time, the S-polarized radiation, can be defined only through the vector potential:

$$\boldsymbol{A}_S = \frac{1}{2}\boldsymbol{y}_0\frac{cE_y(\omega)}{i\omega}e^{-i\omega t + iqx} + \text{C.C.} \tag{9}$$

It is convenient to represent the surface current density generated in response to a harmonic field as a sum over spatial harmonics: $\boldsymbol{j}(\boldsymbol{r}) = 2^{-1}\sum_{\boldsymbol{q}} \boldsymbol{j}^{(q)} e^{i\boldsymbol{q}\cdot\boldsymbol{r}} + \text{C.C.}$, where $2^{-1}\boldsymbol{j}^{(q)} = S^{-1}\int_S \boldsymbol{j}(\boldsymbol{r}) e^{-i\boldsymbol{q}\cdot\boldsymbol{r}} d^2\boldsymbol{r}$; the set of in-plane photon wave vectors $\boldsymbol{q}$ is specified by appropriate conditions on the boundary of a large area $S \gg A$. It is also convenient to choose the area S to be a multiple of the normalization area A, so that

$$\frac{1}{2A}\int_A \boldsymbol{\Psi}_n^*(\boldsymbol{r})\boldsymbol{\Psi}_m(\boldsymbol{r}) d^2r = \frac{1}{2S}\int_S \boldsymbol{\Psi}_n^*(\boldsymbol{r})\boldsymbol{\Psi}_m(\boldsymbol{r}) d^2r. \tag{10}$$

After calculating the matrix elements $\boldsymbol{j}_{nm}^{(q)}$ of the current density operator and solving independently the master equations (4), one can calculate the average amplitude of a given current density harmonic, which could be used as a source in Maxwell's equations or to determine the conductivity tensor:

$$\boldsymbol{j}^{(q)} = \sum_{mn} \boldsymbol{j}_{nm}^{(q)} \rho_{mn}. \tag{11}$$

In order to evaluate $\boldsymbol{j}_{nm}^{(q)}$ we determine the velocity operator $\hat{\boldsymbol{v}} = i\hbar^{-1}\left[\hat{H}, \hat{\boldsymbol{r}}\right]$ and define the current density operator as $\hat{\boldsymbol{j}} = -e\hat{\boldsymbol{v}}$:

$$\hat{\boldsymbol{j}} = -ev_F\hat{\boldsymbol{\sigma}}. \tag{12}$$

Next, we take into account a standard expression for the current density operator in a second-quantized formalism[20]: $\hat{\boldsymbol{j}}(\boldsymbol{r}) = \hat{\boldsymbol{\Psi}}^\dagger \cdot \hat{\boldsymbol{j}} \cdot \hat{\boldsymbol{\Psi}}$, where $\hat{\boldsymbol{\Psi}} = \sum_n \hat{a}_n \boldsymbol{\Psi}_n(\boldsymbol{r})$ and $\hat{\boldsymbol{\Psi}}^\dagger = \sum_m \hat{a}_m^\dagger \boldsymbol{\Psi}_m^\dagger(\boldsymbol{r})$ are second-quantized operators, and $\hat{a}_m^\dagger$ and $\hat{a}_n$ are fermion creation and annihilation operators. Treating $\hat{a}_m^\dagger$ and $\hat{a}_n$ as Heisenberg operators and using $\boldsymbol{j}(\boldsymbol{r}) = \langle \hat{\boldsymbol{j}}(\boldsymbol{r})\rangle$, $\langle \hat{a}_m^\dagger(t)\hat{a}_n(t)\rangle = \rho_{mn}(t)$, we arrive at $2^{-1}\boldsymbol{j}^{(q)} = \sum_{mn}\left(e^{-i\boldsymbol{q}\cdot\boldsymbol{r}}\hat{\boldsymbol{j}}\right)_{nm}\rho_{mn}$, which gives

$$2^{-1}\boldsymbol{j}_{nm}^{(q)} = \langle n|e^{-i\boldsymbol{q}\cdot\boldsymbol{r}}\hat{\boldsymbol{j}}|m\rangle. \tag{13}$$

To calculate the matrix elements $\boldsymbol{j}_{mn}^{(q)}$ and $\left(\hat{H}_{int}^{opt}\right)_{mn}$ we will need the following useful relationships:

$$\left(e^{iqx}\right)_{mn} = \frac{1}{2}\left(s_m s_n + e^{i(\theta_n - \theta_m)}\right)\delta_{\boldsymbol{k}_m,\boldsymbol{k}_n+\boldsymbol{q}}, \tag{14}$$

$$\left(\hat{\boldsymbol{\sigma}}e^{iqx}\right)_{mn} = \frac{1}{2}\left[(\boldsymbol{x}_0 - i\boldsymbol{y}_0)s_m e^{i\theta_n} + (\boldsymbol{x}_0 + i\boldsymbol{y}_0)s_n e^{-i\theta_m}\right]\delta_{\boldsymbol{k}_m,\boldsymbol{k}_n+\boldsymbol{q}}. \tag{15}$$

The above general equations should allow one to calculate the conductivity in any order with respect to the external optical field. There is however a complication related to the fact that the model described by the effective Hamiltonian Eq. (1) contains a "bottomless" valence band with electrons occupying all states to $k \to \infty$. Therefore, only the converging integrals make sense:

$$\sum_{mn} \boldsymbol{j}^{(q)}_{nm}\rho_{mn} \Rightarrow g\sum_{ss'}\int_{\infty'}\frac{d^2k'}{4\pi^2}\int_{\infty}\frac{d^2k}{4\pi^2}\boldsymbol{j}^{(q)}_{\boldsymbol{k}'\boldsymbol{k}s's}\rho_{\boldsymbol{k}\boldsymbol{k}'ss'}, \tag{16}$$

where g is the degeneracy factor. Otherwise the optical response could be determined by the electron dispersion far from the Dirac point where the effective Hamiltonian Eq. (1) is no longer valid. It turns out that the convergence of the linear current depends on the choice of the gauge, whereas for the second-order nonlinear current the integral in Eq. (16) converges for any gauge. The divergence of the linear response can be regularized as discussed in the next section. In addition, the gauge dependence of the linear response violates gauge invariance, which is a consequence of the fact that the density matrix corresponding to the bottomless Hamiltonian in Eq. (1) has an infinite trace. In the next section we discuss this issue in more detail.

III. THE LINEAR RESPONSE OF MASSLESS DIRAC FERMIONS

The perturbation expansion of the nonlinear response functions implies that the second-order nonlinear terms depend on the first-order linear response. Therefore, in this section we outline the derivation of the linear current. The nontrivial aspect of this derivation is an apparent violation of gauge invariance and divergence of the linear current. We address these issues in this section and related Appendix sections.

The solution of the density matrix equation (4) in the linear approximation with respect to the field is

$$\rho^{(1)}_{nm}(\omega) = \frac{1}{2}\frac{\left[\hat{V}(\omega)e^{iqx}\right]_{nm}(\rho_{mm}-\rho_{nn})}{\hbar\omega-(E_n-E_m)}, \tag{17}$$

where we defined $\hat{H}^{opt}_{int} = 2^{-1}\left[\hat{V}(\omega)e^{-i\omega t+iqx} + \text{H.C.}\right]$.
Here $\hat{V}(\omega) = -e\phi(\omega)\cdot\hat{1} + \frac{ev_F}{c}\left[\hat{\sigma}_x A_x(\omega) + \hat{\sigma}_y A_y(\omega)\right]$ and $\rho^{(1)}_{nm}(\omega)$ is a complex-valued amplitude of the linear perturbation $\propto e^{-i\omega t}$ of the density matrix. For a monochromatic current $\boldsymbol{j} = 2^{-1}\boldsymbol{j}^{(q)}(\omega)e^{-i\omega t+iqx} + \text{C.C.}$ we have

$$\boldsymbol{j}^{(q)}(\omega) = \sum_{mn}\boldsymbol{j}^{(q)}_{mn}\rho^{(1)}_{nm}(\omega). \tag{18}$$

The expression (18) is evaluated in Appendix A. The most straightforward derivation is for a P-polarized field defined through a scalar potential, Eq. (7), since in this case the integral (16) converges. If we keep only the terms of the lowest order in q (i.e. the linear terms since $E_x = -iq\phi$), the resulting 2D (surface) conductivity tensor is independent of q. In the limit of strong degeneracy or low temperatures, the relevant terms are

(i) intraband conductivity, which has a Drude-like form:

$$\sigma_{xx}^{(intra)}(\omega) = \frac{ie^2 v_F k_F}{\pi\hbar(\omega + i\gamma)}, \tag{19}$$

(ii) and the interband term:

$$\sigma_{xx}^{(inter)}(\omega) = \frac{ie^2}{4\pi\hbar} \ln\left[\frac{2v_F k_F - (\omega + i\gamma)}{2v_F k_F + (\omega + i\gamma)}\right]. \tag{20}$$

Here $k = k_F$ is Fermi momentum, and we also added the relaxation terms by replacing $\omega \to \omega + i\gamma$ in Eq. (17); in the limit $\gamma \to +0$ one can obtain from Eq. (20) the well known result for the interband conductivity[22]: $\mathrm{Re}\sigma_{xx}^{(inter)} = \frac{e^2}{4\hbar}\Theta(\omega - 2v_F k_F)$, where $\Theta(x)$ is the Heaviside step function.

If we define the optical field with a vector potential, the same calculation will lead to divergent integrals. In this case the finite, and at the same time gauge-invariant, expression for the linear current at frequency ω can be obtained by subtracting the same current evaluated at zero frequency[23]:

$$\boldsymbol{j}^{(q)}(\omega) = \sum_{mn} \boldsymbol{j}_{mn}^{(q)} \left[\rho_{nm}^{1,A}(\omega) - \rho_{nm}^{1,A}(\omega \to 0)\right]. \tag{21}$$

Here $\rho_{nm}^{1,A}(\omega)$ is Eq. (17) with $\phi(\omega) = 0$ in the interaction Hamiltonian. This prescription cancels the divergent term and leads to the Kubo formula for the linear response. In our case Eq. (21) is equal to the sum of Eqs. (19) and (20) for the diagonal conductivities $\sigma_{yy} = \sigma_{xx}$, and gives $\sigma_{xy} = 0$. The procedure in Eq. (21) can be justified by considering the graphene Hamiltonian with a small quadratic term in the energy dispersion:

$$E = s\hbar v_F k + \epsilon\frac{\hbar^2 k^2}{2}, \tag{22}$$

where ϵ is a small parameter. Adding this term provides a bottom to the valence band. As shown in Appendix B, the linear current for such a system approaches Eq. (21) when $\epsilon \to 0$.

For a P-polarized field which can be represented through both scalar and vector potentials the renormalization procedure in Eq. (21) is equivalent to the gauge transformation of the density matrix from the A-gauge (8) to the φ-gauge (7). Indeed, let the function $\rho^{1,A_P}_{nm}(\omega)$ correspond to the solution of Eq. (17) for the field defined in the gauge given by Eq. (8), whereas the function $\rho^{1,\varphi}_{nm}(\omega)$ correspond to the gauge of Eq. (7). Since we just found that the sum $\sum_{mn} \dot{\boldsymbol{j}}^{(q)}_{mn}\rho^{(1,\varphi)}_{nm}(\omega)$ is finite, it makes sense to try the transformation $\rho^{1,A_P}_{nm} \Rightarrow \rho^{1,\varphi}_{nm}$. The gauge transformation from $\boldsymbol{A}$ and φ to $\tilde{\boldsymbol{A}}$ and $\tilde{\varphi}$ corresponds to the unitary transformation of the density matrix (see Appendix C)

$$\tilde{\rho}_{nm} = \sum_{qp} \left(e^{-\frac{ief}{\hbar c}}\right)_{nq} \rho_{qp} \left(e^{+\frac{ief}{\hbar c}}\right)_{pm}, \tag{23}$$

where the scalar function $f(t, \boldsymbol{r})$ determines the gauge transformation of the potentials

$$\tilde{\boldsymbol{A}} = \boldsymbol{A} + \nabla f(t, \boldsymbol{r}), \qquad \tilde{\varphi} = \varphi - \frac{1}{c}\frac{\partial f(t, \boldsymbol{r})}{\partial t}. \tag{24}$$

In particular, the transformation from the vector potential (8) to scalar potential (7) is

$$\nabla f = -\boldsymbol{A}_P. \tag{25}$$

Within the linear approximation with respect to f we obtain from Eq. (23):

$$\rho^{1,A_P}_{nm} \Rightarrow \rho^{1,A_P}_{nm} - \frac{ie}{\hbar c} f_{nm}(\rho_{mm} - \rho_{nn}). \tag{26}$$

Next, we will use the general relationship (see e.g.[24])

$$f_{nm} = \frac{-i\hbar}{E_n - E_m} \left(\frac{\nabla f \cdot \hat{\boldsymbol{v}} + \hat{\boldsymbol{v}} \cdot \nabla f}{2}\right)_{nm}, \tag{27}$$

from which we obtain from $\hat{\boldsymbol{v}} = v_F \hat{\boldsymbol{\sigma}}$ that

$$f_{nm} = \frac{-i\hbar v_F \left(\hat{\boldsymbol{\sigma}} \cdot \nabla f\right)_{nm}}{E_n - E_m}. \tag{28}$$

As a result, from Eqs. (26), (28) and (25) one gets

$$\rho^{1,A_P}_{nm}(\omega) \Rightarrow \rho^{1,A_P}_{nm}(\omega) + \frac{ev_F}{c} \frac{\left[\hat{\sigma}_x A_x(\omega)\right]_{nm}(\rho_{mm} - \rho_{nn})}{E_n - E_m}. \tag{29}$$

Taking into account Eq. (17), Eq. (29) can be represented as $\rho^{1,A_P}_{nm}(\omega) \Rightarrow \rho^{1,A_P}_{nm}(\omega) - \rho^{1,A_P}_{nm}(\omega \to 0)$, which is identical to Eq. (21).

The structure of transformation (23) makes it clear why the density matrix with an infinite trace can give rise to the divergent current. Consider the density matrix in the form $\rho_{nm} = \rho_{mm}\delta_{nm} + \xi_{n\neq m}$, where ξ is a small perturbation. The sum $\sum_{mn} \boldsymbol{j}_{mn}\xi_{nm}$ can converge in a certain gauge even if the trace $\sum_m \rho_{mm}$ diverges. However, the transformation (23) to a different gauge projects the diagonal of the matrix with an infinite trace onto off-diagonal elements, which can lead to the divergence in Eq. (16). The inverse is also true: the divergence can be eliminated by the transformation (23) as we have just shown above.

It is also clear that the separation of the response into intraband and interband components depends generally on the choice of the gauge since the transformation (23) mixes different contributions. At the same time, a correctly defined current has to be gauge-invariant.

IV. SECOND-ORDER NONLINEAR RESPONSE

Now we consider the second-order nonlinear response to the bichromatic field which we will represent through the vector potential in order to describe both P- and S-polarized fields with the same formalism. We will write the in-plane field components at frequencies $\omega_{1,2}$ directed along unit vectors $\eta_{1,2}$ as

$$\boldsymbol{A} = \frac{1}{2}\boldsymbol{\eta}_1 A(\omega_1) e^{i(\boldsymbol{q}_1\cdot\boldsymbol{r}_\parallel - \omega_1 t)} + \frac{1}{2}\boldsymbol{\eta}_2 A(\omega_2) e^{i(\boldsymbol{q}_2\cdot\boldsymbol{r}_\parallel - \omega_2 t)} + \text{c.c.} \tag{30}$$

We need to calculate the perturbation of the density matrix at the sum frequency $\omega_1 + \omega_2$. The term quadratic with respect to the field can be written as

$$\begin{aligned}
&\rho^{(2)}_{mn}(\omega_1+\omega_2) = \left(\frac{e}{2c}\right)\frac{1}{\hbar(\omega_1+\omega_2) - (\epsilon_m - \epsilon_n)} \\
&\times \sum_{l\neq m,n}\left[\left((\hat{\boldsymbol{v}}\cdot\boldsymbol{\eta}_1)e^{i\boldsymbol{q}_1\cdot\boldsymbol{r}}\right)_{ml} A(\omega_1)\rho^{(1)}_{ln}(\omega_2) - \rho^{(1)}_{ml}(\omega_1)\left((\hat{\boldsymbol{v}}\cdot\boldsymbol{\eta}_2)e^{i\boldsymbol{q}_2\cdot\boldsymbol{r}}\right)_{ln} A(\omega_2)\right] + \{1\leftrightarrow 2\} \\
&= \frac{1}{2}\left(\frac{e}{c}\right)^2 \frac{A(\omega_1)A(\omega_2)}{\hbar(\omega_1+\omega_2) - (\epsilon_m - \epsilon_n)} \times \sum_{l\neq m,n}\left((\hat{\boldsymbol{v}}\cdot\boldsymbol{\eta}_1)e^{i\boldsymbol{q}_1\cdot\boldsymbol{r}}\right)_{ml}\left((\hat{\boldsymbol{v}}\cdot\boldsymbol{\eta}_2)e^{i\boldsymbol{q}_2\cdot\boldsymbol{r}}\right)_{ln} \\
&\times\left[\frac{(\rho_{nn} - \rho_{ll})}{\hbar\omega_2 - (\epsilon_l - \epsilon_n)} - \frac{(\rho_{ll} - \rho_{mm})}{\hbar\omega_1 - (\epsilon_m - \epsilon_l)}\right] + \{1\leftrightarrow 2\}\,.
\end{aligned} \tag{31}$$

The trace of the corresponding Fourier harmonic of the induced current can be then calculated as

$$\boldsymbol{j}^{(q_1+q_2)}(\omega_1+\omega_2) = \sum_{mn}\boldsymbol{j}^{(q_1+q_2)}_{nm}\rho^{(2)}_{mn}(\omega_1+\omega_2). \tag{32}$$

The second-order response at the difference frequency, $\rho^{(2)}_{mn}(\omega_1-\omega_2)$ can be obtained by replacing

$$\omega_2 \Rightarrow -\omega_2,\ q_2 \Rightarrow -q_2,\ A(\omega_2) \Rightarrow A^*(\omega_2). \tag{33}$$

Next, we transform from summation to integration over k-states, introduce the corresponding occupation numbers $f(s,k)$ of the momentum states in each band, apply the momentum conservation in a three-wave mixing process, and take into account spin and valley degeneracy. Note that the integral over the electron momenta converges, as opposed to the linear response calculations where one needs to regularize the integral by either subtracting the contribution at zero frequency or adding a k^2 term to the Hamiltonian, as discussed above. The result is

$$\begin{aligned}
&\boldsymbol{j}^{(q_1+q_2)}(\omega_1+\omega_2)\\
&= -\frac{e^3 v_F^3}{16\pi^2 c^2 \hbar^2} A(\omega_1)A(\omega_2) \sum_{s_m,s_n,s_l} \int d^2\boldsymbol{k} \frac{1}{(\omega_1+\omega_2) - v_F(s_m|\boldsymbol{k}+\boldsymbol{q}_1| - s_n|\boldsymbol{k}-\boldsymbol{q}_2|)}\\
&\times \left[\frac{f(s_n,|\boldsymbol{k}-\boldsymbol{q}_2|) - f(s_l,|\boldsymbol{k}|)}{\omega_2 - v_F(s_l|\boldsymbol{k}| - s_n|\boldsymbol{k}-\boldsymbol{q}_2|)} - \frac{f(s_l,|\boldsymbol{k}|) - f(s_m,|\boldsymbol{k}+\boldsymbol{q}_1|)}{\omega_1 - v_F(s_m|\boldsymbol{k}+\boldsymbol{q}_1| - s_l|\boldsymbol{k}|)}\right]\\
&\times \left[(\eta_{1x} - i\eta_{1y})s_m e^{i\theta(\boldsymbol{k})} + (\eta_{1x} + i\eta_{1y})s_l e^{-i\theta(\boldsymbol{k}+\boldsymbol{q}_1)}\right]\\
&\times \left[(\eta_{2x} - i\eta_{2y})s_l e^{i\theta(\boldsymbol{k}-\boldsymbol{q}_2)} + (\eta_{2x} + i\eta_{2y})s_n e^{-i\theta(\boldsymbol{k})}\right]\\
&\times \left[(\boldsymbol{x}_0 - i\boldsymbol{y}_0)s_n e^{i\theta(\boldsymbol{k}+\boldsymbol{q}_1)} + (\boldsymbol{x}_0 + i\boldsymbol{y}_0)s_m e^{-i\theta(\boldsymbol{k}-\boldsymbol{q}_2)}\right]\\
&+ \{1 \leftrightarrow 2\}\,.
\end{aligned} \tag{34}$$

This equation can be integrated numerically for any given geometry of incident fields and electron distribution. We consider the limit of the Fermi distribution with a strong degeneracy, direct all in-plane photon wave vectors along x-axis, and expand the integrand in Eq. (34) in powers of q_1, q_2. The integral over the term of zeroth-order in q vanishes, as expected from symmetry. We will keep the terms linear in q. Also we have to evaluate separately the intraband contribution $s_l = s_m = s_n$ and all types of mixed interband-intraband contributions: $s_m = s_n = -s_l$, $s_m = s_l = -s_n$, and $s_n = s_l = -s_m$. After performing this procedure, we find that the xxx, xyy, yxy, and yyx components of the second-order nonlinear conductivity tensor are nonzero, while all other components are zero. Their expressions are bulky, so we give them in Appendix D and plot them in the figures below.

A sketch of the second-order nonlinear process for an obliquely incident light of mixed polarization is shown in Fig. 1. Note that when both pump fields have either S- or P-polarization, the generated nonlinear current has only the x-component (along the in-plane direction of propagation of the pumps). When the polarizations are mixed, the y-component of the nonlinear current appears due to yxy and yyx components of the nonlinear conductivity (they are different only by permutation of indices 1 and 2 referring to the two pump fields).

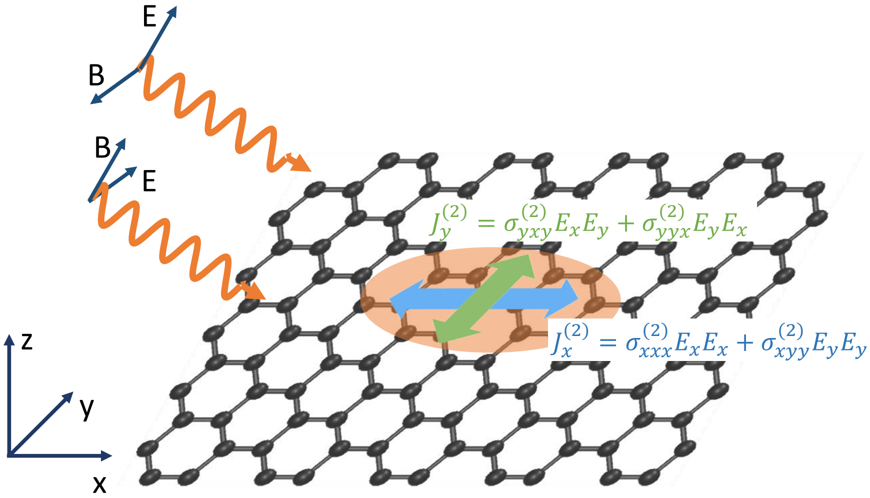
E
B
B
E
$J_y^{(2)} = \sigma^{(2)}_{yxy}E_xE_y + \sigma^{(2)}_{yyx}E_yE_x$
$J_x^{(2)} = \sigma^{(2)}_{xxx}E_xE_x + \sigma^{(2)}_{xyy}E_yE_y$
z
y
x

FIG. 1. A sketch of the second order nonlinear current generation in the graphene plane for obliquely incident light.

Apparent "non-reciprocity" of the expressions for $\sigma^{(2)}_{yxx} = 0$ (P-in, S-out channel) and $\sigma^{(2)}_{xyy} \neq 0$ (S-in, P-out channel) has a simple physical explanation: a P-polarized incident field cannot create a current in the y-direction orthogonal to the electric field, whereas an incident S-polarized field creates such a current in x-direction via the magnetic field component B_z normal to the layer. Note that any cross-components of the *linear* conductivity such as σ_{xy} have to vanish since the linear response was calculated in Sec. III neglecting any spatial variation of the field, i.e. for isotropic graphene.

The expressions for the nonlinear conductivity tensor that we obtained pass all symmetry and gauge invariance tests. Indeed, one can verify that the value of $\sigma^{(2)}_{xxx}$ agrees with the one derived using scalar potential in the interaction Hamiltonian. Furthermore, after converting the nonlinear conductivity to the nonlinear susceptibility according to

$$\chi^{(2)}_{ijk}(\omega_1+\omega_2;\omega_1,\omega_2) = \frac{i\sigma^{(2)}_{ijk}(\omega_1+\omega_2;\omega_1,\omega_2)}{\omega_1+\omega_2},$$

one can verify that all components of the nonlinear susceptibility tensor satisfy proper permutation relations; see e.g. Ch. 2.9 in[25]:

$$\chi^{(2)}_{ijk}(\omega_3=\omega_1+\omega_2) = \chi^{(2)}_{jik}(-\omega_1=-\omega_3+\omega_2) = \chi^{(2)}_{kji}(-\omega_2=-\omega_3+\omega_1), \tag{35}$$

where in-plane wave vectors have to be permuted together with frequencies.

The second-order response goes to zero when the Fermi energy ϵ_F goes to zero, and has maxima at resonances when one of the three frequencies involved in three-wave mixing is close to $2\epsilon_F/\hbar = 2v_F k_F$. Far from these resonances and for high frequencies or low doping, $2v_F k_F \ll \omega_1, \omega_2, \omega_1+\omega_2$, expressions for the nonlinear conductivity are greatly simplified (we will give only the expressions for $\sigma^{(2)}_{xxx}$ and $\sigma^{(2)}_{xyy}$ for brevity):

$$\sigma^{(2)}_{xxx} = s(\epsilon_F)\frac{2e^3 v_F^2}{\pi\hbar^2}\frac{v_F^4 k_F^4\left[q_1\omega_2^3(2\omega_1+\omega_2)+q_2\omega_1^3(2\omega_2+\omega_1)\right]}{\omega_1^4\omega_2^4(\omega_1+\omega_2)^3}, \tag{36}$$

$$\sigma^{(2)}_{xyy} = -s(\epsilon_F)\frac{2e^3 v_F^2}{\pi\hbar^2}\frac{v_F^2 k_F^2\,(q_1\omega_2^2+q_2\omega_1^2)}{\omega_1^3\omega_2^3(\omega_1+\omega_2)}. \tag{37}$$

Here $s(\epsilon_F)=\pm 1$ depending on whether the Fermi level is in the conduction or valence band. An interesting and surprising result contained in these expressions is that the nonlinear frequency conversion of S-polarized radiation into P-polarized radiation is much more efficient at high frequencies as compared to the P-in, P-out channel: $\sigma^{(2)}_{xyy}/\sigma^{(2)}_{xxx} \propto \frac{\omega^2}{v_F^2 k_F^2} \gg 1$. In particular, for the second-harmonic generation process $\omega_1=\omega_2=\omega$ and $q_1=q_2=q$, and the dominant component of the nonlinear conductivity tensor is simply

$$\sigma^{(2)}_{xyy}(2\omega;\omega,\omega) = -s(\epsilon_F)\frac{e^3}{\pi\hbar^2}\frac{v_F^4 k_F^2 q}{\omega^5}. \tag{38}$$

Although it is expected that xyy and xxx components of the nonlinear conductivity should scale differently because the magnetic field contributes only to the xyy component, we are not aware of any simple argument that would predict their particular ratio in the high-frequency limit. It is clear that the contribution from each three-wave mixing channel in Eq. (34) will be at least linear in k_F, because the differences in the occupation numbers contribute $q\cos\theta\delta(k-k_F)$ when expanded in powers of q. After performing integration $\int k\,dk$ a factor of k_F is present in every term. The subsequent integration over the angles and summation over all three-wave mixing channels cancels many terms. The cancellation is different between xyy and xxx components, so that the leading nonzero order in the xyy component is k_F^2, whereas the leading term in the xxx component scales as k_F^4.

In the opposite limit of low frequencies or high doping, $2v_F k_F \gg \omega_1, \omega_2, \omega_1 + \omega_2$, we also obtain simplified expressions:

$$\sigma_{xxx}^{(2)} = s(\epsilon_F) \frac{e^3 v_F^2}{8\pi\hbar^2 \omega_1 \omega_2} \left(\frac{q_1 + q_2}{\omega_1 + \omega_2} + \frac{q_1}{\omega_1} + \frac{q_2}{\omega_2} \right), \tag{39}$$

$$\sigma_{xyy}^{(2)} = s(\epsilon_F) \frac{e^3 v_F^2}{8\pi\hbar^2 \omega_1 \omega_2} \left[\frac{q_1 + q_2}{\omega_1 + \omega_2} + \frac{\omega_1 - \omega_2}{\omega_1 + \omega_2} \left(\frac{q_1}{\omega_1} - \frac{q_2}{\omega_2} \right) \right]. \tag{40}$$

We verified that Eqs. (39) and (40) can be derived independently from the single-band kinetic equation, i.e. in the quasiclassical approximation described in Appendix E. This provides another test of our general expressions, since one should indeed expect that the single-band physics emerges in the limit of a strong doping and low frequencies, when all interband transitions become suppressed by Pauli blocking. Note that although Eqs. (39) and (40) do not depend on k_F, they are valid only in the high-k_F limit and are completely inapplicable for undoped graphene. In fact, exact expressions (D1) and (D2) give $\sigma_{xxx}^{(2)} = 0$ and $\sigma_{xyy}^{(2)} = 0$ for $k_F = 0$, since in this case the nonlinear currents due to interband and intraband transitions cancel each other. This can be viewed as a manifestation of the electron-hole symmetry in graphene.

The nonlinear conductivity components (D1) and (D2) diverge when one or more of the three frequencies involved in three-wave mixing is close to $2\epsilon_F/\hbar = 2v_F k_F$. Close to resonance with $2\epsilon_F/\hbar$ one has to include the imaginary part of the frequency which comes from the omitted relaxation term $-\gamma\rho_{mn}$ in the density-matrix equations. This amounts to substituting $\omega_1 \to \omega_1 + i\gamma_1$, $\omega_2 \to \omega_2 + i\gamma_2$, $\omega_1 + \omega_2 \to \omega_1 + \omega_2 + i\gamma_3$. Note that if we flip the sign of ω_2 to describe the difference frequency generation process, the sign of $+i\gamma_2$ remains the same, i.e. $\omega_2 \to -\omega_2 + i\gamma_2$.

Even if dissipation is included, we can still use Eqs. (35) to derive the components of the nonlinear susceptibility tensor from other components. In order to do that, one needs to use Eqs. (35) in the absence of dissipation and then add imaginary parts of frequencies. Of course the resulting expressions after adding dissipation won't satisfy the permutation relation Eqs. (35).

Figures 2-4 illustrate the above properties of the nonlinear conductivity for the processes of the second-harmonic generation (SHG), difference-frequency generation (DFG), and sum-

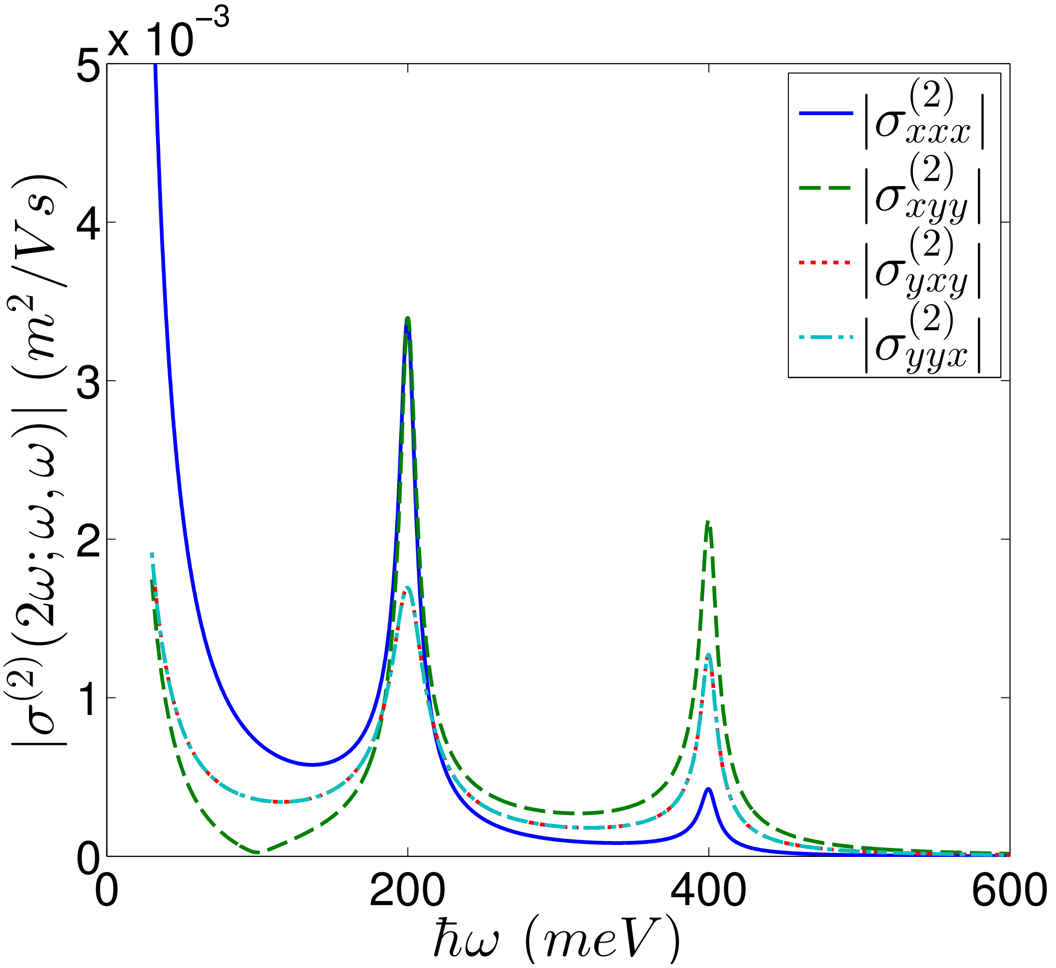

FIG. 2. Nonzero components of the second order nonlinear conductivity tensor for the process of SHG as a function of the fundamental frequency. The pump is incident at 45 degrees. The Fermi energy is 200 meV and all resonances are broadened by the same factor γ equal to 5 meV.

frequency generation (SFG). We used SI units in the figures for easier comparison of the values with known materials. In Fig. 2, absolute values of nonzero components of the nonlinear conductivity tensor for the SHG process $\omega + \omega \Rightarrow 2\omega$ are plotted as a function of the fundamental frequency ω, assuming that the Fermi energy is 200 meV and all resonances are broadened by the same half-width factor γ equal to 5 meV in energy units. The plots for $\sigma^{(2)}_{yxy}$ and $\sigma^{(2)}_{yyx}$ are identical as they should be. There are two prominent resonances at $\hbar\omega = 2\epsilon_F = 400$ meV and $2\hbar\omega = 2\epsilon_F$. At high frequencies, the xxx component falls off much faster than the xyy component. At low frequencies, both components diverge as $1/\omega^2$. Our treatment, however, becomes invalid in the static limit $\omega \leq \gamma$ when any of the frequencies

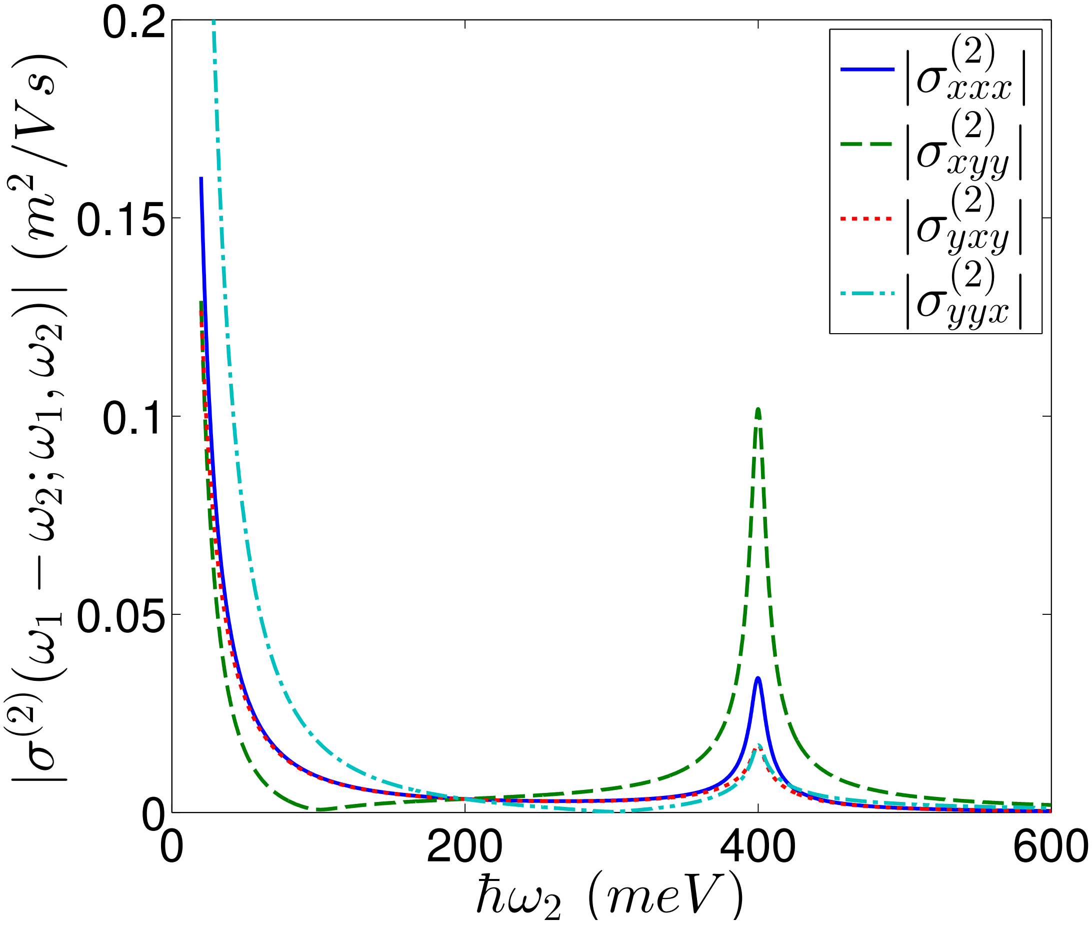

FIG. 3. Nonzero components of the second order nonlinear conductivity tensor for the process of DFG as a function of one of the pump frequencies (ω_2). Frequency ω_1 is fixed at 400 meV. Both pumps are incident in the (xz)-plane at 45 degrees. The Fermi energy is 200 meV and all resonances are broadened by the same factor γ equal to 5 meV.

becomes lower than the scattering rate; that is why the plots are truncated at $\omega = 20$ meV. The quasi-classical method of the kinetic equation has the same applicability limit.

Figure 3 shows absolute values of the nonzero components of the nonlinear conductivity tensor for the DFG process for the same values of ϵ_F and γ, as a function of ω_2. The second frequency $\hbar\omega_1$ is fixed to be 400 meV. The same qualitative behavior is observed: there is a double resonance when both ω_1 and ω_2 are equal to $2k_Fv_F$. Note that there is no divergence at $\omega_1 - \omega_2 \to 0$ because the same factor $\omega_1 - \omega_2$ appears in the numerator. There is divergence when $\omega_2 \to 0$ which should be truncated at $\omega_2 \sim \gamma$.

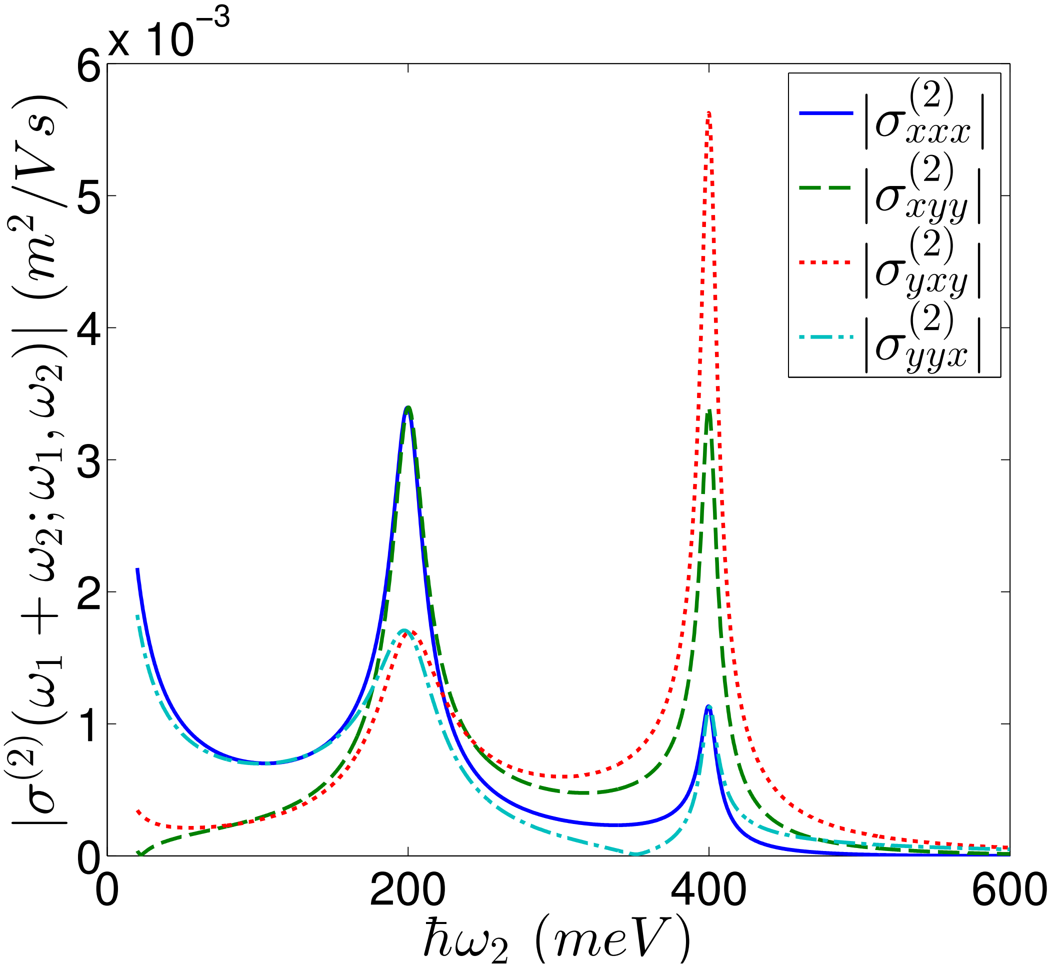

FIG. 4. Nonzero components of the second order nonlinear conductivity tensor for the process of SFG as a function of one of the pump frequencies (ω_2). Frequency ω_1 is fixed at 200 meV. Both pumps are incident in the (xz)-plane at 45 degrees. The Fermi energy is 200 meV and all resonances are broadened by the same factor γ equal to 5 meV.

In Fig. 4, the nonzero components of the nonlinear conductivity tensor for the SFG process are shown as a function of ω_2. The second frequency $\hbar\omega_1$ is fixed at 200 meV. As expected, all components show strong resonances when one of the frequencies or their sum is equal to $2\epsilon_F = 400$ meV.

The magnitude of the nonlinear response generally increases rapidly when one or more of the frequencies is decreased, as is obvious also from analytic expressions. For the DFG process, the magnitude of the nonlinear conductivity components is two orders of magnitude higher as compared to SHG or SFG. As one of the frequencies goes to zero, the treatment becomes invalid, but one could get an order of magnitude estimate of the maximum nonlinear conductivity by putting this frequency equal to γ. Using the same value of $\gamma = 5$ meV, one gets the nonlinear conductivity for DFG of the order of several $\mathrm{m}^2/(\mathrm{Vs})$ in the THz range. This is a 2D conductivity. Purely for the sake of comparison with known bulk nonlinear materials, we can convert it to the bulk nonlinear susceptibility dividing by the frequency and the monolayer thickness of 0.3 nm, to arrive at $|\chi^{(2)}_{3D}| \sim 10^{-3}$ m/V. This is a huge value as compared to 1-100 pm/V values for most materials. As we already mentioned in the introduction, the reason is a large magnitude of the matrix element of the interaction Hamiltonian which scales in proportion to $v_F/\omega \propto \lambda$, i.e. it grows more rapidly with wavelength λ than in conventional materials with parabolic energy dispersion, where it scales as $\sqrt{\lambda}$. Of course, only the 2D values of the graphene conductivity or susceptibility enter all physical results such as the intensity of the generated nonlinear signal[17,18] or the parametric gain[18]. Still, combination of intrinsically large nonlinear conductivity of graphene and a surface plasmon resonance for the nonlinear signal may lead to quite significant efficiency of the nonlinear processes, as emphasized in the theoretical proposals[17,18].

Massless metallic surface states in the topological insulator $\mathrm{Bi_2Se_3}$ can be described by a low-energy effective Hamiltonian $\hat{H}_0(\hat{\boldsymbol{p}}) = v'_F\,(\hat{\sigma}_x\hat{p}_y - \hat{\sigma}_y\hat{p}_x)$, (see for example[26] and references therein) where $v'_F \simeq 5\times 10^7$ cm/s and the chirality is now associated with real spin as opposed to pseudospin in graphene. The eigenenergies of this Hamiltonian have the same form as for graphene. Also, the matrix elements of the velocity operator and interaction Hamiltonian are the same as in graphene, see Eqs. (14) and (15). Therefore, expressions for the nonlinear conductivity tensor are also the same, except for a four times lower degeneracy per surface and about two times lower Fermi velocity. In fact, the Hamiltonian for the surface states of $\mathrm{Bi_2Se_3}$ can be written in the form equivalent to that of graphene by choosing the representation of the Pauli matrices as $(-\hat{\sigma}_y, \hat{\sigma}_x)$, which still satisfies the required commutation relation.

In conclusion, we developed the full quantum-mechanical theory of the in-plane second-order nonlinear response of graphene beyond the electric dipole approximation. We provided a systematic derivation of the second-order nonlinear conductivity tensor, valid for all second-order processes, all frequencies and doping densities, as long as the massless Dirac fermion approximation for a single-particle Hamiltonian is applicable. Our approach can be applied to any system of massless chiral Dirac fermions, for example surface states in topological insulators such as $\mathrm{Bi_2Se_3}$. We derived useful analytic expressions for the components of the nonlinear conductivity tensor, which satisfy all symmetry and permutation properties, and have a correct quasi-classical limit. We also summarized main features of the linear response, with emphasis on its gauge properties and regularization.

ACKNOWLEDGMENTS

This work has been supported by the Air Force Office for Scientific Research through grants FA9550-15-1-0153 and FA9550-14-1-0376. M.T. acknowledges support from the Russian Foundation for Basic Research Grant No. 14-22-02034 and thanks I.D. Tokman and V.Ya. Aleshkin for helpful discussions.

Appendix A: Evaluation of the linear current

To calculate the current in the linear approximation with respect to the electromagnetic (EM) field, we will use Eqs. (13), (14), (15), (17), and (18). Assuming that the photon wave vector is much smaller than typical wave vectors of electrons, $q \ll k$, we calculate the following quantities in the zeroth and first order in q:

$$n_{(\boldsymbol{k}+\boldsymbol{q})(s=+1)} - n_{\boldsymbol{k}(s=+1)} \approx q\cos\theta(\boldsymbol{k})\frac{\partial n_{\boldsymbol{k}(+1)}}{\partial k}, \tag{A1}$$

$$E_{(\boldsymbol{k}+\boldsymbol{q})(s=+1)} - E_{\boldsymbol{k}(s=+1)} \approx \hbar v_F q\cos\theta(\boldsymbol{k}), \tag{A2}$$

$$\frac{1}{2}\boldsymbol{j}^{(q)}_{\boldsymbol{k}(\boldsymbol{k}+\boldsymbol{q})(+1)(+1)} \approx -ev_F\left[\boldsymbol{x}_0\cos\theta(\boldsymbol{k}) + \boldsymbol{y}_0\sin\theta(\boldsymbol{k})\right], \tag{A3}$$

$$\frac{1}{2}\boldsymbol{j}^{(q)}_{\boldsymbol{k}(\boldsymbol{k}+\boldsymbol{q})(+1)(-1)} \approx -iev_F\left[\boldsymbol{x}_0\sin\theta(\boldsymbol{k}) - \boldsymbol{y}_0\cos\theta(\boldsymbol{k})\right]. \tag{A4}$$

Consider first the EM field determined through a scalar potential. In this case we can replace in Eq. (17)

$$\left[\hat{V}(\omega)e^{iqx}\right]_{(\boldsymbol{k}+\boldsymbol{q})\boldsymbol{k}ss'} \approx -\frac{e\phi(\omega)}{4}\left[i\frac{q}{k}\sin\theta(\boldsymbol{k}) + 1 + ss'\right]. \tag{A5}$$

The summation in Eq. (18), can be replaced by integration using Eq. (16). Keeping the terms of the first order in q in the conduction band, the integral can be transformed as $\int_0^\infty(\partial n_{k(+1)}/\partial k)kdk = -\int_0^\infty n_{k(+1)}dk = -k_F$. Introducing relaxation through the substitution $\omega \to \omega + i\gamma$, we arrive at Eqs. (19) and (20).

Now we determine the EM field through the vector potential, in which case we should substitute the following in Eq. (17):

$$\left[\hat{V}(\omega)e^{iqx}\right]_{(\boldsymbol{k}+\boldsymbol{q})\boldsymbol{k}ss'} \approx \frac{ev_F}{4c}\left[A_x\left(se^{i\theta(\boldsymbol{k})}+s'e^{-i\theta(\boldsymbol{k})}\right)-iA_y\left(se^{i\theta(\boldsymbol{k})}-s'e^{-i\theta(\boldsymbol{k})}\right)\right]. \tag{A6}$$

After exactly the same steps as in the case of a scalar potential, we arrive at

$$\boldsymbol{j}^{(q)}_{(intra)}(\omega) = -\frac{gv_F^2e^2(\boldsymbol{x}_0A_x+\boldsymbol{y}_0A_y)}{4\pi^2\hbar c}\int_0^{2\pi}\frac{q\cos^3\theta d\theta}{\omega-v_Fq\cos\theta}\int_0^{k_F}n_{k(+1)}dk, \tag{A7}$$

$$\boldsymbol{j}^{(q)}_{(inter)}(\omega) = \frac{gv_F^2e^2(\boldsymbol{x}_0A_x+\boldsymbol{y}_0A_y)}{4\pi\hbar c}\int_0^{\infty}\left(\frac{1}{\omega+2kv_F}-\frac{1}{\omega-2kv_F}\right)\left(n_{k(-1)}-n_{k(+1)}\right)kdk. \tag{A8}$$

Note that $\boldsymbol{j}^{(q)}_{(inter)}(\omega)\to\infty$ when $\int_0^\infty n_{k(-1)}kdk\to\infty$. Therefore the current needs to be renormalized. Applying the renormalization Eq. (21), we obtain

$$\begin{aligned}\boldsymbol{j}^{(q)}_{(intra)}(\omega) &= -\frac{gv_Fe^2\omega(\boldsymbol{x}_0A_x+\boldsymbol{y}_0A_y)}{4\pi^2\hbar c}\int_0^{2\pi}\frac{\cos^2\theta d\theta}{\omega-v_Fq\cos\theta}\int_0^{k_F}n_{k(+1)}dk\\ &\approx -\frac{gv_Fe^2(\boldsymbol{x}_0A_x+\boldsymbol{y}_0A_y)}{4\pi\hbar c}\int_0^{k_F}n_{k(+1)}dk,\end{aligned} \tag{A9}$$

$$\boldsymbol{j}^{(q)}_{(inter)}(\omega) = -\frac{gv_Fe^2\omega(\boldsymbol{x}_0A_x+\boldsymbol{y}_0A_y)}{8\pi\hbar c}\int_0^{\infty}\left(\frac{1}{\omega+2kv_F}-\frac{1}{\omega-2kv_F}\right)\left(n_{k(-1)}-n_{k(+1)}\right)dk, \tag{A10}$$

which again yields the expressions given in Sec. III.

If we choose the carrier distribution limited not only in the conduction band but also in the valence band, i.e. $n_{k(-1)}=0$ for $k>k_{max;(-1)}$, then for the P-polarized field that can be defined through both scalar and vector potentials the sum $\boldsymbol{j}^{(q)}_{(intra;+1)}+\boldsymbol{j}^{(q)}_{(intra;(-1))}+\boldsymbol{j}^{(q)}_{(inter)}$ is invariant and finite without regularization with Eq. (21). This corroborates our conclusion that for massless Dirac fermions the need in renormalization (21) is due to the bottomless valence band filled with electrons to infinite energies and wave vectors, which is an artifact of the model Hamiltonian (1).

Appendix B: How to correctly define current in a system with a massless Dirac spectrum

The prescription Eq. (21) for renormalization of the diverging linear current in a system of massless Dirac fermions can be justified if we consider a system with small deviation from the massless conical spectrum, for which the current becomes finite, and then let the deviation go to zero. Of course, the actual electron spectrum of graphene does deviate from the massless conical spectrum at high electron energies. However, it is reasonable to expect that at low enough energies any correction to the Hamiltonian (1) becomes small, and all essential physics including the linear response is dominated by massless fermions. Therefore, it is important, at least from the methodological perspective, to provide physical justification of Eq. (21).

Let's modify the Hamiltonian (1) by adding a quadratic correction to the massless Dirac spectrum $E = s\hbar v_F k$:

$$\hat{H}_0(\hat{\boldsymbol{p}}) = v_F \hat{\boldsymbol{\sigma}} \cdot \hat{\boldsymbol{p}} + \epsilon \frac{\hat{\boldsymbol{p}}^2}{2} \cdot \hat{1}. \tag{B1}$$

This Hamiltonian leads to the energy spectrum given by Eq. (22), whereas the eigenstates Eq. (2) remain the same. We will also assume that the change in the energy spectrum in the conduction band is insignificant, since

$$\epsilon \hbar k_F \ll v_F. \tag{B2}$$

At the same time, the spectrum of Eq. (B1) creates a "bottom" of the valence band at $k = K$, where

$$\epsilon \hbar K = v_F. \tag{B3}$$

Therefore, the integral over k-states in the valence band has now finite limits.

In the presence of an EM field given by the vector potential $\boldsymbol{A}$, one needs to replace $\hat{\boldsymbol{p}} \Rightarrow \hat{\boldsymbol{p}} + \frac{e}{c}\boldsymbol{A}$ in the Hamiltonian:

$$\hat{H}_0(\hat{\boldsymbol{p}}) = v_F \hat{\boldsymbol{\sigma}} \cdot \left(\hat{\boldsymbol{p}} + \frac{e}{c}\boldsymbol{A}\right) + \epsilon \frac{\left(\hat{\boldsymbol{p}} + \frac{e}{c}\boldsymbol{A}\right)^2}{2} \cdot \hat{1}. \tag{B4}$$

The resulting velocity operator,

$$\hat{\boldsymbol{v}} = \frac{i}{\hbar}\left[\hat{H}, \hat{\boldsymbol{r}}\right] = v_F \hat{\boldsymbol{\sigma}} + \epsilon \left(\hat{\boldsymbol{p}} + \frac{e}{c}\boldsymbol{A}\right) \cdot \hat{1}, \tag{B5}$$

and the current operator,

$$\hat{\boldsymbol{j}} = -e\hat{\boldsymbol{v}} = -e\left[v_F \hat{\boldsymbol{\sigma}} + \epsilon \left(\hat{\boldsymbol{p}} + \frac{e}{c}\boldsymbol{A}\right) \cdot \hat{1}\right] \tag{B6}$$

acquire a component which depends on the vector potential:

$$\delta\hat{\boldsymbol{j}} = -\epsilon \frac{e^2}{c}\boldsymbol{A} \cdot \hat{1}. \tag{B7}$$

Consider for definiteness an EM field given by the second of Eq. (6) with $A_x = 0$, and also keep only the solution in zeroth order in q/k.

A new, $\boldsymbol{A}$-dependent component of the current operator $\delta\hat{\boldsymbol{j}}$ gives rise to an additional component of the linear current (see, e.g.,[24]):

$$\delta j_y = -\frac{\epsilon e^2 A_y e^{-i\omega t}}{2c} \sum_{\boldsymbol{k}} n_{\boldsymbol{k}(s=-1)} + \text{C.C.}, \tag{B8}$$

where

$$\sum_{\boldsymbol{k}} n_{\boldsymbol{k}(s=-1)} = \frac{g}{4\pi^2} \int_0^{2\pi} d\theta \int_0^K n_{\boldsymbol{k}(-1)} k dk, \tag{B9}$$

and the value of K is determined by Eq. (B3). In the limit of Eq. (B2) we can keep only the contribution of the valence band to the current component δj_y. This gives (in the limit of strong degeneracy)

$$\delta j_y = -\frac{A_y e^{-i\omega t}}{2c} \frac{g v_F e^2}{4\pi\hbar} \int_0^K n_{\boldsymbol{k}(-1)} dk + \text{C.C.} \tag{B10}$$

Equation (B10) can be represented as a sum of two terms:

$$\begin{aligned} &-\frac{g v_F e^2}{4\pi\hbar} \frac{A_y e^{-i\omega t}}{2c} \int_0^K n_{\boldsymbol{k}(-1)} dk \\ &= -\frac{g v_F^2 e^2}{4\pi\hbar} \frac{A_y e^{-i\omega t}}{2c} \int_0^K \left(\frac{1}{2kv_F} - \frac{1}{-2kv_F} \right) (n_{\boldsymbol{k}(-1)} - n_{\boldsymbol{k}(+1)}) k dk \\ &+ \frac{g v_F^2 e^2}{4\pi^2\hbar} \frac{A_y e^{-i\omega t}}{2c} \int_0^{2\pi} \frac{q \cos^2\theta \cos\theta d\theta}{-v_F q \cos\theta} \int_0^{k_F} n_{\boldsymbol{k}(+1)} dk, \end{aligned} \tag{B11}$$

where for a degenerate electron gas $n_{\boldsymbol{k}(+1)} = 0$ for $k > k_F$. Let us now compare this current component with the expressions (A7) and (A8) for the linear current that we derived in Appendix A for a massless Dirac current ($\hat{\boldsymbol{j}} = -ev_F\hat{\boldsymbol{\sigma}}$), namely,

$$j_y^{(intra)} = -\frac{g v_F^2 e^2}{4\pi^2\hbar} \frac{A_y e^{-i\omega t}}{2c} \int_0^{2\pi} \frac{q \cos^2\theta \cos\theta d\theta}{\omega - v_F q \cos\theta} \int_0^{k_F} n_{\boldsymbol{k}(+1)} dk + \text{C.C.}, \tag{B12}$$

$$j_y^{(inter)} = \frac{g v_F^2 e^2}{4\pi\hbar} \frac{A_y e^{-i\omega t}}{2c} \int_0^K \left(\frac{1}{\omega + 2kv_F} - \frac{1}{\omega - 2kv_F} \right) (n_{\boldsymbol{k}(-1)} - n_{\boldsymbol{k}(+1)}) k dk + \text{C.C.} \tag{B13}$$

From comparing (B10) with (B12), (B13), it is obvious that $-\delta j_y = j_y^{(intra)}(\omega \to 0) + j_y^{(inter)}(\omega \to 0)$, i.e., adding this current component to the total current as $j_y^{(intra)} + j_y^{(inter)} + \delta j_y$ is completely equivalent to the renormalization given by Eq. (21) in the limit $K \to \infty$ which corresponds to the limit $\epsilon \to 0$. Note also that the current component $\hat{\boldsymbol{j}} = -\epsilon e \hat{\boldsymbol{p}}$ which we neglected in Eq. (B6) becomes negligible as compared to $j_y^{(intra)} + j_y^{(inter)}$ in the same limit $\epsilon \to 0$. Actually this term vanishes since the distributions $n_{\boldsymbol{k}(-1)}$ and $n_{\boldsymbol{k}(+1)}$ don't depend on the direction of $\boldsymbol{k}$.

Appendix C: Gauge transformation properties for massless Dirac systems

We start from the Schrödinger equation

$$i\hbar\frac{\partial\boldsymbol{\Psi}}{\partial t} = \hat{H}(\boldsymbol{A},\varphi)\boldsymbol{\Psi} \tag{C1}$$

with the Hamiltonian of Eq. (3). Consider a gauge transformation of the field potentials from $(\boldsymbol{A},\varphi)$ to $(\tilde{\boldsymbol{A}},\tilde{\varphi})$. This transformation is determined by Eqs. (24) through a scalar function $f(\boldsymbol{r},t)$. Let $\tilde{\boldsymbol{\Psi}}$ be the solution of Eq. (C1) for $\hat{H}(\tilde{\boldsymbol{A}},\tilde{\varphi})$. One can see by direct substitution that the spinor $\boldsymbol{\Psi}$ is transformed is the same way as a scalar state function: $\tilde{\boldsymbol{\Psi}} = e^{-i\frac{e}{\hbar c}f}\boldsymbol{\Psi}$[20] (we consider a particle with negative charge $-e$). This transformation conserves the quantum-mechanical average current $\boldsymbol{j} = -ev_F\langle\boldsymbol{\Psi}|\hat{\boldsymbol{\sigma}}|\boldsymbol{\Psi}\rangle = -ev_F\langle\tilde{\boldsymbol{\Psi}}|\hat{\boldsymbol{\sigma}}|\tilde{\boldsymbol{\Psi}}\rangle$.

To obtain gauge transformation rules for the density matrix, it is convenient to use its coordinate representation as $\hat{\rho}(\boldsymbol{r},\boldsymbol{r}')$[24]. Following the standard procedure[20], we obtain

$$\hat{\rho}(\boldsymbol{r},\boldsymbol{r}') = \sum_{mn}\rho_{mn}\left[\boldsymbol{\Psi}_m(\boldsymbol{r})\boldsymbol{\Psi}_n^*(\boldsymbol{r}')\right], \tag{C2}$$

where the expression $[\boldsymbol{\Psi}_m(\boldsymbol{r})\boldsymbol{\Psi}_n^*(\boldsymbol{r}')]$ is a matrix formed by the elements of spinors $\boldsymbol{\Psi}_m(\boldsymbol{r})$ and $\boldsymbol{\Psi}_n^*(\boldsymbol{r}')$. Therefore, the operator $\hat{\rho}(\boldsymbol{r},\boldsymbol{r}')$ is a matrix with elements dependent on the pair of arguments $(\boldsymbol{r},\boldsymbol{r}')$:

$$\hat{\rho}(\boldsymbol{r},\boldsymbol{r}') = \begin{pmatrix} \rho_{11}(\boldsymbol{r},\boldsymbol{r}') & \rho_{12}(\boldsymbol{r},\boldsymbol{r}') \\ \rho_{21}(\boldsymbol{r},\boldsymbol{r}') & \rho_{22}(\boldsymbol{r},\boldsymbol{r}'). \end{pmatrix} \tag{C3}$$

The equation of motion for the operator $\hat{\rho}(\boldsymbol{r},\boldsymbol{r}')$ has a standard form, which follows directly from Eq. (C1):

$$i\hbar\frac{\partial\hat{\rho}(\boldsymbol{r},\boldsymbol{r}')}{\partial t} = \hat{H}\hat{\rho}(\boldsymbol{r},\boldsymbol{r}') - \hat{\rho}(\boldsymbol{r},\boldsymbol{r}')\overleftarrow{\hat{H}'}, \tag{C4}$$

where the operator $\hat{H}$ acts only on the arguments $\boldsymbol{r}$, whereas $\hat{H}'$ acts only on $\boldsymbol{r}'$, and the arrow above it means acting from right to left. The quantum-mechanical average of any operator $\hat{\boldsymbol{\Theta}}$ can be written in the matrix representation as $\boldsymbol{\Theta} = \sum_{mn}\boldsymbol{\Theta}_{nm}\rho_{mn}$, and in the coordinate representation as $\boldsymbol{\Theta} = \int_A d^2\boldsymbol{r}\int_{A'} d^2\boldsymbol{r}'\left\{\delta(\boldsymbol{r}-\boldsymbol{r}')\left[\hat{\boldsymbol{\Theta}}\hat{\rho}(\boldsymbol{r},\boldsymbol{r}')\right]\right\}$, where it is assumed that the operator $\hat{\boldsymbol{\Theta}}$ acts only on $\boldsymbol{r}$.

Let $\tilde{\hat{\rho}}(\boldsymbol{r},\boldsymbol{r}')$ be the solution of Eq. (C4) for the Hamiltonian $\hat{H}(\tilde{\boldsymbol{A}},\tilde{\varphi})$ given by Eq. (3). Then, following Ref.[24], from Eq. (C4) one can obtain

$$\tilde{\hat{\rho}}(t,\boldsymbol{r},\boldsymbol{r}') = \hat{\rho}(t,\boldsymbol{r},\boldsymbol{r}')e^{-iu(t,\boldsymbol{r},\boldsymbol{r}')}, \qquad u(t,\boldsymbol{r},\boldsymbol{r}') = \frac{e}{\hbar c}\left[f(t,\boldsymbol{r}) - f(t,\boldsymbol{r}')\right]. \tag{C5}$$

Taking into account $\rho_{mn} = \langle \boldsymbol{\Psi}_m(\boldsymbol{r})|\hat{\rho}(\boldsymbol{r},\boldsymbol{r}')|\boldsymbol{\Psi}_n(\boldsymbol{r}')\rangle$ which follows from Eq. (C2), we arrive at Eq. (23).

Note that gauge transformation of the density matrix equation includes an appropriate transformation of the relaxation operator[24]. The simplest approach which allows one to avoid complicated transformations is to neglect dissipation first, and then to replace $\omega \to \omega + i\gamma$ in the resulting expression for the dissipationless current. Of course, this approach works only for the simplest form of the relaxation operator in the relaxation time approximation.

Appendix D: Components of the second-order nonlinear conductivity tensor

Here we give analytic expressions for the nonzero components of the second-order nonlinear conductivity tensor, obtained by integrating Eq. (34) in the limit of strong degeneracy:

$$
\begin{aligned}
&\sigma^{(2)}_{xxx}(\omega_1+\omega_2;\omega_1,\omega_2)\\
&= -s(\epsilon_F)\frac{e^3 v_F^2}{2\pi\hbar^2}\frac{1}{\omega_1^2\omega_2^2(\omega_1+\omega_2)}\frac{1}{(\omega_1^2-4v_F^2k_F^2)(\omega_2^2-4v_F^2k_F^2)((\omega_1+\omega_2)^2-4v_F^2k_F^2)}\\
&\times\left[-4v_F^4k_F^4(q_1\omega_2^3(2\omega_1+\omega_2)+q_2\omega_1^3(\omega_1+2\omega_2))+16v_F^6k_F^6(q_1\omega_2(2\omega_1+\omega_2)+q_2\omega_1(\omega_1+2\omega_2))\right],
\end{aligned}
\tag{D1}
$$

$$
\begin{aligned}
&\sigma^{(2)}_{xyy}(\omega_1+\omega_2;\omega_1,\omega_2)\\
&= -s(\epsilon_F)\frac{e^3 v_F^2}{2\pi\hbar^2}\frac{1}{\omega_1^2\omega_2^2(\omega_1+\omega_2)}\frac{1}{(\omega_1^2-4v_F^2k_F^2)(\omega_2^2-4v_F^2k_F^2)((\omega_1+\omega_2)^2-4v_F^2k_F^2)}\\
&\times\big[4(v_Fk_F)^2\omega_1\omega_2(\omega_1+\omega_2)^2(q_1\omega_2^2+q_2\omega_1^2)\\
&\quad+4(v_Fk_F)^4(q_1\omega_2^4-(6q_1+4q_2)\omega_1\omega_2^3-8(q_1+q_2)\omega_1^2\omega_2^2-(4q_1+6q_2)\omega_1^3\omega_2+q_2\omega_1^4)\\
&\quad+16(v_Fk_F)^6(q_1\omega_2(2\omega_1-\omega_2)+q_2\omega_1(2\omega_2-\omega_1))\big],
\end{aligned}
\tag{D2}
$$

$$
\begin{aligned}
&\sigma^{(2)}_{yxy}(\omega_1+\omega_2;\omega_1,\omega_2)\\
&= -s(\epsilon_F)\frac{e^3 v_F^2}{2\pi\hbar^2}\frac{1}{\omega_1^2\omega_2^2(\omega_1+\omega_2)}\frac{1}{(\omega_1^2-4v_F^2k_F^2)(\omega_2^2-4v_F^2k_F^2)((\omega_1+\omega_2)^2-4v_F^2k_F^2)}\\
&\times\big[4(v_Fk_F)^2\omega_1^2\omega_2(\omega_1+\omega_2)(q_1\omega_2^2-q_2\omega_1(\omega_1+2\omega_2))\\
&\quad+4(v_Fk_F)^4(q_2\omega_1(\omega_1+2\omega_2)^3-q_1\omega_2(4\omega_1^3+4\omega_1^2\omega_2+2\omega_1\omega_2^2+3\omega_2^3))\\
&\quad+16(v_Fk_F)^6(q_1\omega_2(2\omega_1+3\omega_2)-q_2\omega_1(\omega_1+2\omega_2))\big],
\end{aligned}
\tag{D3}
$$

$$
\begin{aligned}
&\sigma^{(2)}_{yyx}(\omega_1+\omega_2;\omega_1,\omega_2)\\
&= -s(\epsilon_F)\frac{e^3 v_F^2}{2\pi\hbar^2}\frac{1}{\omega_1^2\omega_2^2(\omega_1+\omega_2)}\frac{1}{(\omega_1^2-4v_F^2k_F^2)(\omega_2^2-4v_F^2k_F^2)((\omega_1+\omega_2)^2-4v_F^2k_F^2)}
\end{aligned}
$$

$$\times\,\Big[4(v_Fk_F)^2\omega_1\omega_2^2(\omega_1+\omega_2)(q_2\omega_1^2-q_1\omega_2(2\omega_1+\omega_2))$$
$$+\,4(v_Fk_F)^4(q_1\omega_2(2\omega_1+\omega_2)^3-q_2\omega_1(3\omega_1^3+2\omega_1^2\omega_2+4\omega_1\omega_2^2+4\omega_2^3))$$
$$+\,16(v_Fk_F)^6(q_2\omega_1(3\omega_1+2\omega_2)-q_1\omega_2(2\omega_1+\omega_2))\Big]\,. \tag{D4}$$

Here $s(\epsilon_F)=\pm1$ depending on whether the Fermi level is in the conduction or valence band.

Appendix E: Quasiclassical approximation

Here we provide the derivation of the quasiclassical equations of motion which allow one to derive Eqs. (39,40) of the main text in the single-band approximation of low frequencies and high Fermi energy, when the contribution of interband transitions can be neglected.

In the absence of external fields ($\boldsymbol{A}=0$,$\varphi=0$) the solution of the Schrödinger equation with Hamiltonian (1) for a fixed energy of a quasiparticle can be written as

$$\boldsymbol{\varPsi}=\begin{pmatrix}\varPsi_1\\ \varPsi_2\end{pmatrix}=\mathrm{const}\times\frac{e^{i\boldsymbol{k}\cdot\boldsymbol{r}-i\frac{E(\boldsymbol{k})}{\hbar}t}}{\sqrt{2}}\begin{pmatrix}s\\ e^{i\theta(\boldsymbol{k})}\end{pmatrix}, \tag{E1}$$

where $s=\pm1$, $E=s\hbar v_F|\boldsymbol{k}|$, $\theta(\boldsymbol{k})$ is an angle between the wave vector $\boldsymbol{k}$ and the x-axis. In the presence of the field, consider the solution of the Schrödinger equation with Hamiltonian (3) in the WKB approximation. Treating $\hbar$ as a small parameter, we seek the solution in the form close to (E1):

$$\boldsymbol{\varPsi}(\boldsymbol{A},\varphi)=e^{\frac{i}{\hbar}S(t,\boldsymbol{r})}\left\{\begin{pmatrix}\varPsi_1^{(0)}(t,\boldsymbol{r})\\ \varPsi_2^{(0)}(t,\boldsymbol{r})\end{pmatrix}+\hbar\begin{pmatrix}\varPsi_1^{(1)}(t,\boldsymbol{r})\\ \varPsi_2^{(1)}(t,\boldsymbol{r})\end{pmatrix}+\hbar^2...\right\}. \tag{E2}$$

First consider the terms of zeroth order with respect to $\hbar$:

$$\begin{aligned}&(-\partial_tS+e\varphi)\varPsi_1^{(0)}+v_F\left[\left(-\partial_xS-\tfrac{e}{c}A_x\right)+i\left(\partial_yS+\tfrac{e}{c}A_y\right)\right]\varPsi_2^{(0)}=0,\\ &v_F\left[\left(-\partial_xS-\tfrac{e}{c}A_x\right)-i\left(\partial_yS+\tfrac{e}{c}A_y\right)\right]\varPsi_1^{(0)}+(-\partial_tS+e\varphi)\varPsi_2^{(0)}=0.\end{aligned} \tag{E3}$$

From (E3) we derive

(**i**) the eikonal equation:

$$(-\partial_tS+e\varphi)^2=v_F^2\left(\frac{e}{c}\boldsymbol{A}+\nabla S\right)^2, \tag{E4}$$

(**ii**) the relationship between the spinor components:

$$\frac{\varPsi_1^{(0)}}{\varPsi_2^{(0)}}=\pm\left[\cos\theta\left(\frac{e}{c}\boldsymbol{A}+\nabla S\right)-i\sin\theta\left(\frac{e}{c}\boldsymbol{A}+\nabla S\right)\right]=\pm e^{-i\theta\left(\frac{e}{c}\boldsymbol{A}+\nabla S\right)}, \tag{E5}$$

where $\theta\left(\frac{e}{c}\boldsymbol{A}+\nabla S\right)$ is the angle between vector $\frac{e}{c}\boldsymbol{A}+\nabla S$ and the x-axis. Equation (E5) allows one to represent the WKB solution in the form

$$\boldsymbol{\Psi}(\boldsymbol{A},\varphi)=\frac{\Phi(t,\boldsymbol{r})e^{\frac{i}{\hbar}S(t,\boldsymbol{r})}}{\sqrt{2}}\begin{pmatrix} s \\ e^{i\theta\left(\frac{e}{c}\boldsymbol{A}+\nabla S\right)}. \end{pmatrix} \tag{E6}$$

The expression for the factor Φ can be obtained by requiring that there exist the nontrivial solution to the next order term $\hbar\begin{pmatrix}\Psi_1^{(1)}(t,\boldsymbol{r})\\ \Psi_2^{(1)}(t,\boldsymbol{r})\end{pmatrix}$; this approach is used for example, in order to find the normal modes in anisotropic media[27]. However, it is much easier to use the conservation of the probability flux, which in our case is given by

$$\frac{\partial}{\partial t}\left(\boldsymbol{\Psi}^*\boldsymbol{\Psi}\right)=-v_F\nabla\cdot\left[\boldsymbol{\Psi}^*\hat{\boldsymbol{\sigma}}\boldsymbol{\Psi}\right]. \tag{E7}$$

From here,

$$\frac{\partial|\Phi|^2}{\partial t}=-\nabla\cdot\left[\frac{sv_F\left(\frac{e}{c}\boldsymbol{A}+\nabla S\right)}{\left|\frac{e}{c}\boldsymbol{A}+\nabla S\right|}|\Phi|^2\right]. \tag{E8}$$

Now consider the solution to the eikonal equation (E4), which we will interpret as a Hamilton-Jacobi equation[28], corresponding to the Hamiltonian $H(\boldsymbol{P},\boldsymbol{r},t)$:

$$\begin{aligned}&\partial_t S+H(\boldsymbol{P},\boldsymbol{r},t)=0, \qquad \nabla S=\boldsymbol{P},\\ &H(\boldsymbol{P},\boldsymbol{r},t)=-e\varphi(\boldsymbol{r},t)+sv_F\sqrt{\left(P_x+\frac{e}{c}A_x\right)^2+\left(P_y+\frac{e}{c}A_y\right)^2}.\end{aligned} \tag{E9}$$

The canonical equations of motion for this Hamiltonian are

$$\begin{aligned}\dot{\boldsymbol{r}}&=\frac{\partial H(\boldsymbol{P},\boldsymbol{r},t)}{\partial\boldsymbol{P}}=sv_F\frac{\boldsymbol{P}+\frac{e}{c}\boldsymbol{A}}{\left|\boldsymbol{P}+\frac{e}{c}\boldsymbol{A}\right|},\\ \dot{\boldsymbol{P}}&=-\frac{\partial H(\boldsymbol{P},\boldsymbol{r},t)}{\partial\boldsymbol{r}}=e\nabla\varphi-\frac{e}{c}\left[\dot{\boldsymbol{r}}\times(\nabla\times\boldsymbol{A})+(\dot{\boldsymbol{r}}\cdot\nabla)\boldsymbol{A}\right].\end{aligned} \tag{E10}$$

Introducing the kinematic momentum $\boldsymbol{p}=\boldsymbol{P}+\frac{e}{c}\boldsymbol{A}$, for which $\dot{\boldsymbol{p}}=\dot{\boldsymbol{P}}+\frac{e}{c}\left[\frac{\partial\boldsymbol{A}}{\partial t}+(\dot{\boldsymbol{r}}\cdot\nabla)\boldsymbol{A}\right]$, we obtain from Eqs. (E10) the quasiclassical equations of motion:

$$\dot{\boldsymbol{r}}=sv_F\frac{\boldsymbol{p}}{|\boldsymbol{p}|}, \qquad \dot{\boldsymbol{p}}=-e\boldsymbol{E}-\frac{e}{c}\left(\dot{\boldsymbol{r}}\times\boldsymbol{B}\right), \tag{E11}$$

where $\boldsymbol{E}=-\nabla\varphi-\frac{1}{c}\frac{\partial\boldsymbol{A}}{\partial t}$, $\boldsymbol{B}=\nabla\times\boldsymbol{A}$. Equations of motion (E11) correspond to the kinetic equation for quasiparticles:

$$\frac{\partial f(\boldsymbol{r},\boldsymbol{p},t)}{\partial t}+sv_F\frac{\boldsymbol{p}}{|\boldsymbol{p}|}\frac{\partial f(\boldsymbol{r},\boldsymbol{p},t)}{\partial\boldsymbol{r}}-e\left[\boldsymbol{E}+\frac{sv_F}{c}\left(\frac{\boldsymbol{p}}{|\boldsymbol{p}|}\times\boldsymbol{B}\right)\right]\frac{\partial f(\boldsymbol{r},\boldsymbol{p},t)}{\partial\boldsymbol{p}}=\mathrm{St}[f(\boldsymbol{r},\boldsymbol{p},t)], \tag{E12}$$

where $\mathrm{St}[f(\boldsymbol{r},\boldsymbol{p},t)]$ is the collision integral. Quasiclassical equations (E11) or (E12) were the starting point for evaluation of both linear and nonlinear optical response in a number of works; see, e.g.[14–16,29,30]. As we have already discussed in the previous section, this approach can be justified only in the limit of low photon frequencies and large Fermi energies, when the contribution of interband transitions can be neglected.

[1] N. Kumar, J. Kumar, C. Gerstenkorn, R. Wang, H.-Y. Chiu, A. L. Smirl, and H. Zhao, Phys. Rev. B **87**, 121406 (2013).

[2] S.-Y. Hong, J. I. Dadap, N. Petrone, P.-C. Yeh, J. Hone, and R. M. Osgood, Phys. Rev. X **3**, 021014 (2013).

[3] E. Hendry, P. J. Hale, J. Moger, A. K. Savchenko, and S. A. Mikhailov, Phys. Rev. Lett. **105**, 097401 (2010).

[4] T. Gu, N. Petrone, J. F. McMillan, A. van der Zande, M. Yu, G.-Q. Lo, D.-L. Kwong, J. Hone, and C. W. Wong, Nature Photonics **6**, 554 (2012).

[5] D. Sun, C. Divin, J. Rioux, J. E. Sipe, C. Berger, W. A. de Heer, P. N. First, and T. B. Norris, Nano Letters **10**, 1293 (2010).

[6] M. Glazov and S. Ganichev, Physics Reports **535**, 101 (2014).

[7] A. Y. Bykov, T. V. Murzina, M. G. Rybin, and E. D. Obraztsova, Phys. Rev. B **85**, 121413 (2012).

[8] J. L. Cheng, N. Vermeulen, and J. E. Sipe, Opt. Express **22**, 15868 (2014).

[9] S. J. Brun and T. G. Pedersen, Phys. Rev. B **91**, 205405 (2015).

[10] J. J. Dean and H. M. van Driel, Applied Physics Letters **95**, 261910 (2009).

[11] J. J. Dean and H. M. van Driel, Phys. Rev. B **82**, 125411 (2010).

[12] H. K. Avetissian, A. K. Avetissian, G. F. Mkrtchian, and K. V. Sedrakian, Phys. Rev. B **85**, 115443 (2012).

[13] H. K. Avetissian, G. F. Mkrtchian, K. G. Batrakov, S. A. Maksimenko, and A. Hoffmann, Phys. Rev. B **88**, 165411 (2013).

[14] S. A. Mikhailov, Phys. Rev. B **84**, 045432 (2011).

[15] M. M. Glazov, JETP Letters **93**, 366 (2011).

[16] D. A. Smirnova, I. V. Shadrivov, A. E. Miroshnichenko, A. I. Smirnov, and Y. S. Kivshar,

Phys. Rev. B **90**, 035412 (2014).

[17] X. Yao, M. Tokman, and A. Belyanin, Phys. Rev. Lett. **112**, 055501 (2014).

[18] M. Tokman, Y. Wang, I. Oladyshkin, A. R. Kutayiah, and A. Belyanin, Phys. Rev. B **93**, 235422 (2016).

[19] T. J. Constant, S. M. Hornett, D. E. Chang, and E. Hendry, Nature Physics **12**, 124 (2016).

[20] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics: Non-Relativistic Theory*, Teoreticheskaia fizika (Elsevier Science, 2013).

[21] V. F. Gantmakher and I. B. Levinson, *Carrier scattering in metals and semiconductors* (North-Holland, Amsterdam, 1987).

[22] R. R. Nair, P. Blake, A. N. Grigorenko, K. S. Novoselov, T. J. Booth, T. Stauber, N. M. R. Peres, and A. K. Geim, Science **320**, 1308 (2008).

[23] A. L. Falkovsky and A. A. Varlamov, The European Physical Journal B **56**, 281 (2007).

[24] M. D. Tokman, Phys. Rev. A **79**, 053415 (2009).

[25] Y. A. Il'inskii and L. V. Keldysh, *Electromagnetic Response of Material Media* (Springer US, 1994).

[26] X. Yao, M. Tokman, and A. Belyanin, Opt. Express **23**, 795 (2015).

[27] V. L. Ginzburg, *The propagation of electromagnetic waves in plasmas*, International series of monographs on electromagnetic waves (Pergamon Press, 1970).

[28] H. Goldstein, *Classical mechanics*, Addison-Wesley series in physics (Addison-Wesley, 1980).

[29] S. A. Mikhailov and K. Ziegler, Journal of Physics: Condensed Matter **20**, 384204 (2008).

[30] M. D. Tokman, M. A. Erukhimova, and A. Belyanin, JETP Letters **100**, 390 (2014).