Gauge-discontinuity contributions to Chern-Simons orbital magnetoelectric coupling

Jianpeng Liu[1,2] and David Vanderbilt[2]

[1]*Kavli Institute for Theoretical Physics, University of California, Santa Barbara, CA 93106, USA*
[2]*Department of Physics and Astronomy, Rutgers University, Piscataway, NJ 08854-8019, USA*

(Dated: November 5, 2015)

We propose a new method for calculating the Chern-Simons orbital magnetoelectric coupling, conventionally parametrized in terms of a phase angle θ. According to previous theories, θ can be expressed as a 3D Brillouin-zone integral of the Chern-Simons 3-form defined in terms of the occupied Bloch functions. Such an expression is valid only if a smooth and periodic gauge has been chosen in the entire Brillouin zone, and even then, convergence with respect to the $\mathbf{k}$-space mesh density can be difficult to obtain. In order to solve this problem, we propose to relax the periodicity condition in one direction (say, the k_z direction) so that a gauge discontinuity is introduced on a 2D $\mathbf{k}$ plane normal to k_z. The total θ response then has contributions from both the integral of the Chern-Simons 3-form over the 3D bulk BZ and the gauge discontinuity expressed as a 2D integral over the $\mathbf{k}$ plane. Sometimes the boundary plane may be further divided into subregions by 1D "vortex loops" which make a third kind of contribution to the total θ, expressed as a combination of Berry phases around the vortex loops. The total θ thus consists of three terms which can be expressed as integrals over 3D, 2D and 1D manifolds. When time-reversal symmetry is present and the gauge in the bulk BZ is chosen to respect this symmetry, both the 3D and 2D integrals vanish; the entire contribution then comes from the vortex-loop integral, which is either 0 or π corresponding to the $\mathbb{Z}_2$ classification of 3D time-reversal invariant insulators. We demonstrate our method by applying it to the Fu-Kane-Mele model with an applied staggered Zeeman field.

PACS numbers: 03.65.Vf, 75.85.+t, 71.15.Rf

I. INTRODUCTION

Magnetoelectric coupling is an interesting but complicated phenomenon that can occur in some insulating solids when an electric polarization $\mathbf{P}$ is linearly induced by an external magnetic field $\mathbf{B}$, or conversely, when a magnetization $\mathbf{M}$ is generated by an applied electric field $\mathbf{E}$. The linear ME coupling coefficient is a rank-2 tensor defined as

$$\alpha_{ab} = \frac{\partial M_b}{\partial E_a}\Big|_{\mathbf{E}=0} = \frac{\partial P_a}{\partial B_b}\Big|_{\mathbf{B}=0} \tag{1}$$

where $a,\ b = \{x, y, z\}$ denote the directions in real space. ME phenomena have contributions from both electronic and lattice degrees of freedom, where the electronic contribution refers to the ME response when the ions are completely frozen, while the lattice contribution takes into account the response that is mediated by ionic displacements. Moreover, depending on the origin of the $\mathbf{E}$-induced magnetization, each of the two contributions can be further decomposed into spin and orbital components.[1,2]

The spin contribution to the ME response (from both electronic and lattice degrees of freedom) has been thoroughly studied with well established theoretical methods in typical magnetoelectrics such as Cr_2O_3.[3–6] On the other hand, the orbital ME response is theoretically more challenging and intriguing. It has been shown that the frozen-ion orbital ME coupling consists of two terms. One term can be expressed as a standard linear response of the Bloch functions to external electric or magnetic fields, denoted as the "Kubo term", while the other, known as the Chern-Simons term, is isotropic and is completely determined by the unperturbed ground-state wavefunctions.[2,7]

The Chern-Simons orbital ME coupling has drawn significant attention recently due to the interest in topological phases in condensed-matter physics. Not surprisingly, in the presence of either time-reversal ($\mathcal{T}$) or inversion ($\mathcal{P}$) symmetry, the ME responses coming from the spin terms and from the Kubo-like orbital terms all vanish. However, there can still be an exotic isotropic ME response, which vanishes in an ordinary insulator but takes values of $\pm e^2/2h$ in $\mathcal{T}$-respecting strong topological insulators[8,9] and in $\mathcal{P}$-respecting axion insulators,[10,11] arising from the Chern-Simons term.[11–13]

This Chern-Simons coupling is conventionally parametrized by a dimensionless phase angle θ via

$$\alpha_{ab}^{\mathrm{CS}} = \frac{\theta e^2}{2\pi h}\delta_{ab}\ , \tag{2}$$

where θ is expressed as an integral of the Chern-Simons 3-form over the 3D Brillouin zone (BZ),

$$\theta = -\frac{1}{4\pi}\int d^3k\ \varepsilon_{abc}\,\mathrm{Tr}\left[A_a\partial_b A_c - \frac{2}{3}iA_aA_bA_c\right]\ . \tag{3}$$

Here A_a, A_b and A_c are the Berry connection matrices of the occupied Bloch bands, and the trace is taken over the occupied bands (see Sec. II A). For TIs and axion insulators, $\theta = \pm\pi$. In the more general cases that $\mathcal{T}$ and $\mathcal{P}$ are both broken, θ is no longer quantized as $\pm\pi$, and other components of the ME response contribute as well.

The Chern-Simons ME coupling has several interesting properties. First, a material with a non-zero Chern-Simons ME coupling can be considered as a medium exhibiting axion electrodynamics,[14] where an additional term $\Delta\mathcal{L} = \alpha^{\mathrm{CS}}\mathbf{E}\cdot\mathbf{B}$ is added to the conventional Lagrangian of electromagnetic fields in media. The electrodynamics with such an axion coupling turns out to be invariant under $\theta \to \theta + 2\pi$.[14]

Secondly, θ is physically measurable only if it varies in space or time.[7] In particular, for a time-independent crystal

with a surface truncation, the presence of the bulk Chern-Simons coupling manifests itself as a surface anomalous Hall effect, where the anomalous Hall conductance is proportional to θ through $\sigma_{xy} = \theta e^2/(2\pi h)$. The connection between the surface anomalous Hall effect and the bulk Chern-Simons ME coupling provides an intuitive explanation of the ambiguity of θ as follows. Suppose an insulating quantum anomalous Hall (QAH) layer with non-zero Chern number C is wrapped around a 3D crystallite having an original bulk value of θ, such that it interacts only weakly with all of the surfaces. Then the the new surface anomalous Hall conductance would be $\sigma_{xy} = \theta e^2/(2\pi h) + Ce^2/h$, which we can be interpreted as a change $\theta \to \theta + 2\pi C$. Thus, such a freedom to coat the surfaces with Chern layers implies the need for a 2π ambiguity in defining θ. The ambiguity in θ is closely analogous to the ambiguity in the definition of the bulk electric polarization, which can be regarded as being due to the freedom of adding or removing an integer number of charges per surface unit cell, as by filling or emptying a surface band.[15]

Despite these intriguing properties, up to now it has remained challenging to calculate θ accurately using Eq. (3) for many systems of interest. For example, as reported in Ref. 16, the calculated θ on an $11\times11\times11$ first-principles $\mathbf{k}$ mesh for Bi_2Se_3, one of the prototype TIs, is only $\sim 35\%$ of π. Similarly, in Ref. 13, the authors calculated the ME response of the Fu-Kane-Mele model with applied staggered Zeeman field. As the system approaches the TI phase, however, the authors switched to some indirect methods to compute θ, because a direct numerical implementation of Eq. (3) became difficult to converge. In other words, despite its theoretical importance, Eq. (3) has not been straightforward to calculate in practice.

The essential problem is that the integrand in Eq. (3) is gauge-dependent. As a result, in order to implement Eq. (3) numerically on a discrete $\mathbf{k}$ mesh, one has to adopt a smooth and periodic gauge over the entire 3D BZ. On the other hand, as is well known, nontrivial topological indices usually bring some obstructions against constructing a smooth and periodic gauge in the BZ. For example, for a 2D quantum anomalous Hall (QAH) insulator (such as the Haldane model[17]) with non-zero Chern number, it is simply impossible to construct a smooth and periodic gauge in the entire 2D BZ. This implies that Eq. (3) would completely break down for a 3D analogue of a 2D QAH insulator,[18] so we regard these cases as beyond the scope of the present work. For 2D and 3D $\mathbb{Z}_2$ TIs, it is impossible to construct a smooth and periodic gauge respecting $\mathcal{T}$ symmetry throughout the BZ,[19,20] although in principle a smooth and periodic gauge breaking $\mathcal{T}$ symmetry is allowed.[20] As a result, for $\mathbb{Z}_2$ TIs (and for $\mathcal{T}$-broken systems close to a $\mathbb{Z}_2$-odd phase) the constraint of being both smooth and periodic is typically too strong, forcing the gauge to be strongly twisted in the BZ to satisfy both conditions. This makes the numeric implementation of Eq. (3) difficult.

In this paper we propose a new method to compute the Chern-Simons orbital ME coefficient. The general idea is to relax the periodicity condition on the gauge in one direction, say the k_z direction, thus introducing some gauge discontinuity on a 2D $\mathbf{k}$ plane (normal to k_z), denoted by $\mathcal{S}$. Then the total θ has one contribution from the bulk-BZ integral of Eq. (3) plus a second one arising from the gauge discontinuity. Furthermore, as will be shown in Sec. IV, $\mathcal{S}$ may also be divided into subregions by 1D "vortex loops" (Sec. IV A), each of which makes a contribution to the total θ in the form of an average of two Berry phases computed around the loop. The total θ can then be expressed as the sum of the 3D integral over the bulk BZ (θ_{BK}), the 2D integral over the gauge-discontinuity plane (θ_{GD}), and the 1D integral(s) over the vortex loop(s) (θ_{VL}).

This method can be generalized to situations where the BZ is divided into multiple subvolumes, with these subvolumes meeting at multiple 2D surface patches where the gauge discontinuities reside. Furthermore, the 2D surface patches may meet at some 1D curves, which again have to be treated as as vortex lines in general. And again, the subvolumes, surface patches, and vortex lines all make contributions to the total θ. However, the definition of a vortex line becomes trickier in this more generalized case, which we therefore leave for future study.

The advantage of our method is that the gauge can be made smoother in the bulk BZ because the periodicity condition is relaxed, so that it becomes much easier to get numeric convergence using Eq. (3). The loss of periodicity is then compensated by contributions from the gauge discontinuities, and possibly from vortex loops as well. We will show that the formulas for the gauge discontinuity and vortex terms take simple forms and can be implemented efficiently in practical numerical calculations.

This paper is organized as follows. In Sec. II we review the definitions of the Berry connection and curvature and introduce the bulk formula for θ. We also put the main idea into a more specific context and make a formal statement of the problem. In Sec. III we derive a formula for θ_{GD}, which is expressed as a 2D integral over the boundary where the gauge discontinuity resides, and discuss the properties of this formula. In Sec. IV we discuss why the vortex-loop term is needed and derive a formula for it. We also show that the quantized θ in TIs is completely determined by the vortex-loop term when a $\mathcal{T}$-symmetric gauge is chosen in the bulk BZ. In Sec. V, we demonstrate the method by applying it to the Fu-Kane-Mele model with a staggered Zeeman field. Finally, we summarize in Sec. VI.

II. PRELIMINARIES

In this section, we first review the definitions of some basic quantities, such as Berry curvatures and Berry connections, that will be used frequently in the paper. We also rewrite the bulk formula for θ, Eq. (3), in a more explicit form. Finally we explain the main idea in more detail and make a formal statement of the problem and the goals.

A. Definitions

We adopt the following definitions. The Berry connection matrix is

$$A_{a,mn}(\mathbf{k}) = i\langle u_{m\mathbf{k}}|\partial_a|u_{n\mathbf{k}}\rangle \ , \tag{4}$$

where $u_{n\mathbf{k}}(\mathbf{r}) = e^{-i\mathbf{k}\cdot\mathbf{r}}\psi_{n\mathbf{k}}(\mathbf{r})$ are the cell-periodic Bloch functions, and a and b run over the three primitive reciprocal lattice directions with $\partial_a \equiv \partial/\partial k_a$. Indices m and n run over the occupied Block bands, possibly after the application of a gauge transformation $U_{nm}(\mathbf{k})$ to smoothen them in $\mathbf{k}$-space. The wavevector components k_x etc. are rescaled to run over $[0, 2\pi]$, and correspondingly the real-space coordinates x etc. run over $[0, 1]$. We shall start dropping the explicit $\mathbf{k}$ arguments and subscripts, keeping in mind that everything is a function of $\mathbf{k}$. Then the non-covariant Berry curvature tensor is

$$\Omega_{ab,mn} = i\,\langle\partial_a u_m|\partial_b u_n\rangle - i\,\langle\partial_b u_m|\partial_a u_n\rangle \ , \tag{5}$$

while

$$\widetilde{\Omega}_{ab,mn} = \Omega_{ab,mn} - i[A_a, A_b]_{mn} \tag{6}$$

is the covariant one (that is, unlike $\Omega_{ab,mn}$, it transforms in the standard way under a gauge transformation).

The Chern-Simons coupling θ has been defined in Eq. (3), where the trace is over the occupied band indices. Using the cyclic property of the trace, Eq. (3) can be written in the more explicit form

$$\theta = -\frac{1}{4\pi}\int d^3k\ \mathrm{Tr}\left[A_x\Omega_{yz}+A_y\Omega_{zx}+A_z\Omega_{xy}-2i[A_x,A_y]A_z\right] . \tag{7}$$

We can also choose to replace one of the non-covariant Berry curvatures with a covariant one to get

$$\theta = -\frac{1}{4\pi}\int d^3k\ \mathrm{Tr}\left[A_x\Omega_{yz}+A_y\Omega_{zx}+A_z\widetilde{\Omega}_{xy}-i[A_x,A_y]A_z\right] , \tag{8}$$

which turns out to be convenient for the derivation of θ_{GD} as will be shown in Sec. III.

B. Statement of the problem

Assume that the gauge has been chosen such that it is smooth and periodic in the k_x and k_y directions and smooth in $k_z \in [-\pi, \pi]$, but not periodic in k_z. (The k_z location of the boundary can easily be generalized.) From now on $\mathbf{k} = (k_x, k_y)$ denotes a point in the 2D slice at $k_z = \pm\pi$, and $|u^{(0)}\rangle$ and $|u^{(1)}\rangle$ denote the wavefunctions just below and above the discontinuity plane respectively. For this reason we refer to $|u^{(0)}\rangle$ and $|u^{(1)}\rangle$ as associated with the "bottom" and "top" planes, even though these are obtained from the top and bottom of the original BZ, respectively. The corresponding Berry potentials are $A_x^{(0)}$ and $A_y^{(0)}$ on the bottom plane and $A_x^{(1)}$ and $A_y^{(1)}$ on the top plane. The states at the top and bottom are physically identical, so we can define a unitary matrix $U(\mathbf{k})$ relating them via

$$|\psi^{(1)}_{m\mathbf{k}}\rangle = \sum_n |\psi^{(0)}_{n\mathbf{k}}\rangle\, U_{nm}(\mathbf{k}) \tag{9}$$

for the original Bloch functions or

$$|u^{(1)}_{m\mathbf{k}}\rangle = e^{i2\pi z}\sum_n |u^{(0)}_{n\mathbf{k}}\rangle\, U_{nm}(\mathbf{k}) \tag{10}$$

for the cell-periodic Bloch functions. Our goal is to calculate the contribution θ_{GD} coming from this gauge discontinuity, such that if we add this contribution to the bulk volume integral θ_{BK} as in Eq. (7), we get the correct total θ. Later, we shall see that there may also be a contribution θ_{VL} from vortex loops around which the gauge discontinuity circulates by an integer multiple of 2π, so that the total axion coupling is given by

$$\theta = \theta_{\mathrm{BK}} + \theta_{\mathrm{GD}} + \theta_{\mathrm{VL}} \ , \tag{11}$$

i.e., a sum of contributions evaluated on 3D, 2D, and 1D manifolds.

III. CALCULATION OF θ_{GD} ON A PLANAR SURFACE

In this section, we derive a formula for θ_{GD} and discuss various properties of the formula. We assume, as above, that the gauge discontinuity occurs on the $k_z = \pm\pi$ plane as schematically shown in Fig. 1, and is described by the unitary matrices $U_{\mathbf{k}}$ as a function of $\mathbf{k}$ lying in the 2D plane. We let

$$U(\mathbf{k}) = e^{-iB(\mathbf{k})} \tag{12}$$

where $B(\mathbf{k})$ is a Hermitian matrix that varies smoothly with $\mathbf{k}$ in the 2D plane. Note that $B(\mathbf{k})$ is basically just $i\ln(U(\mathbf{k}))$, but a set of branch choices is involved in picking a particular B. That is, in the representation that diagonalizes B, we can add $2\pi n_j$ to the j'th eigenvalue without changing U (n_j is an arbitrary integer). For now we insist that the branch choice is made in such a way that $B(\mathbf{k})$ is continuous, with no 2π discontinuities in any of its eigenvalues throughout the 2D $\mathbf{k}$ plane, but this condition will be relaxed in Sec. IV.

A. Formalism

Our strategy is to introduce a parameter λ and define $|\psi_{m\mathbf{k}}(\lambda)\rangle$ in such a way that it smoothly interpolates from one gauge to the other as shown in Fig. 1, i.e.,

$$|\psi_{m\mathbf{k}}(\lambda)\rangle = \sum_n |\psi^{(0)}_{n\mathbf{k}}\rangle\, W_{nm}(\mathbf{k},\lambda) \tag{13}$$

where

$$W(\mathbf{k},\lambda) = e^{-i\lambda B(\mathbf{k})} \tag{14}$$

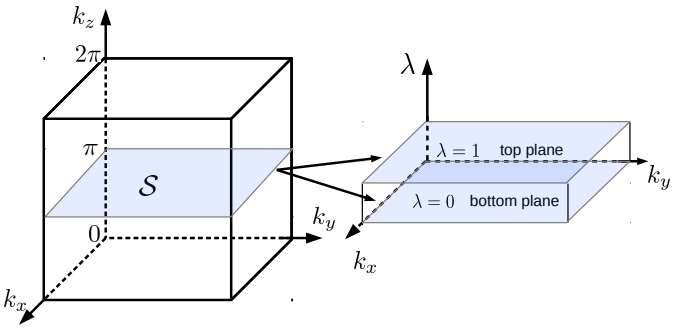

FIG. 1. A planar gauge discontinuity $\mathcal{S}$ in the 3D BZ can be expanded into a fictitious slab whose thickness dimension is described by a parameter $\lambda \in [0,1]$ that interpolates smoothly between the gauge just below ($\lambda=0$) and just above ($\lambda=1$) the plane $\mathcal{S}$.

where $W(\mathbf{k},\lambda)$ is a unitary matrix defined so that $W(\mathbf{k},0) = 1$ and $W(\mathbf{k},1) = U(\mathbf{k})$. Note that $W(\mathbf{k},\lambda)$ commutes with $B(\mathbf{k})$. We shall again begin dropping the $\mathbf{k}$ labels, and will frequently use W and B below.

We then calculate the gauge-discontinuity contribution to θ, denoted by $\theta_{\rm GD}$, by integrating Eq. (8) over the region $\lambda \in [0,1]$, where Eq. (8) is applied in (k_x, k_y, λ) space instead of (k_x, k_y, k_z) space. A straightforward set of calculations shows that the Berry connections in the k_x, k_y, and λ directions are respectively

$$A_x(\lambda) = W^\dagger(\lambda)\, A_x^{(0)}\, W(\lambda) + iW^\dagger(\lambda)\,\partial_x W(\lambda)\ , \quad (15)$$

$$A_y(\lambda) = W^\dagger(\lambda)\, A_y^{(0)}\, W(\lambda) + iW^\dagger(\lambda)\,\partial_y W(\lambda)\ , \quad (16)$$

$$A_\lambda(\lambda) = B\ , \quad (17)$$

where $A^{(0)}_{x(y)}$ is the Berry connection evaluated at the bottom plane as defined earlier. We also write

$$\theta_{\rm GD} = -\frac{1}{4\pi}\int d^2k\ G(\mathbf{k})\ , \quad (18)$$

where

$$G = \int_0^1 d\lambda\ {\rm Tr}\left[A_x\Omega_{y\lambda} - A_y\Omega_{x\lambda} + A_\lambda\widetilde{\Omega}_{xy} - i[A_x,A_y]A_\lambda\right] \quad (19)$$

is the contribution from a particular $\mathbf{k}$. Then G can be written as the sum of three contributions, $G = G_1 + G_2 + G_3$, where

$$G_1(\mathbf{k}) = \int_0^1 d\lambda\ {\rm Tr}\,[B\,\widetilde{\Omega}_{xy}]\ , \quad (20)$$

$$G_2(\mathbf{k}) = \int_0^1 d\lambda\ {\rm Tr}\,[A_x\Omega_{y\lambda} - A_y\Omega_{x\lambda}]\ , \quad (21)$$

$$G_3(\mathbf{k}) = \int_0^1 d\lambda\ {\rm Tr}\left[-i\,[A_x,A_y]\,B\right]\ . \quad (22)$$

The G_1 term is easily evaluated. Because $\widetilde{\Omega}_{xy}$ is gauge-covariant, it follows that $\widetilde{\Omega}_{xy}(\lambda) = W^\dagger(\lambda)\,\widetilde{\Omega}^{(0)}_{xy}\,W(\lambda)$. But $[B, W(\lambda)] = 0$, so that the integrand is independent of λ, and it follows that

$$G_1(\mathbf{k}) = {\rm Tr}\,[B\,\widetilde{\Omega}^{(0)}_{xy}]\ . \quad (23)$$

Here no λ integration is needed.

In order to evaluate G_2 and G_3, we need to evaluate objects such as $\partial_x W(\lambda)$ in Eq. (15), which can be done by noting that the derivative of an exponential of a matrix can be written as

$$\partial_x e^{-i\lambda M} = -i\int_0^\lambda d\mu\ e^{-i\,\mu M}(\partial_x M)e^{-i\,(\lambda-\mu)M}\ . \quad (24)$$

This motivates us to define

$$\overline{B}_a(\lambda) = \int_0^\lambda d\mu\ e^{-i\mu B}\,B_a\,e^{i\mu B}\ , \quad (25)$$

where $B_a \equiv \partial_a B$. Then Eq. (14) gives

$$\partial_a W(\lambda) = \partial_a e^{-i\lambda B} = -i\overline{B}_a(\lambda)W(\lambda) \quad (26)$$

where $a = \{x, y\}$, and Eqs. (15-16) become

$$A_a(\lambda)\ = W^\dagger\,\mathcal{A}_a\,W\ , \quad (27)$$

where

$$\mathcal{A}_a = A_a^{(0)} + \overline{B}_a\ . \quad (28)$$

The dependence on λ is implicit.

Now for the G_2 term we need to compute terms like $\partial_\lambda A_x$. Using Eq. (27) and $\partial_\lambda\, W(\lambda) = -iBW(\lambda)$, it becomes

$$\partial_\lambda A_x\ = iW^\dagger\,[B,\mathcal{A}_x]\,W + B_x\ . \quad (29)$$

Recalling that $\Omega_{x\lambda} = \partial_x A_\lambda - \partial_\lambda A_x$ and $\partial_x A_\lambda = B_x$, we get a nice cancellation, and can write

$$\Omega_{x\lambda} = -i\,W^\dagger\,[B,\mathcal{A}_x]\,W\ , \Omega_{y\lambda} = -i\,W^\dagger\,[B,\mathcal{A}_y]\,W\ .$$

Substituting these expressions into Eq. (21) then gives

$$G_2(\mathbf{k}) = \int_0^1 d\lambda\ {\rm Tr}\left[2i\,B\,[\mathcal{A}_x,\mathcal{A}_y]\right]\ . \quad (30)$$

As it happens, this is almost the same as the expression for G_3 in Eq. (22). Since B commutes with W, we can use the representation-invariance and cyclic properties of the trace to write it as

$$G_3(\mathbf{k}) = \int_0^1 d\lambda\ {\rm Tr}\left[-i\,B\,[\mathcal{A}_x,\mathcal{A}_y]\right]\ . \quad (31)$$

Thus, this term cancels half of G_2.

Restoring the explicit λ dependencies, we get

$$G = {\rm Tr}\left[B\left(\widetilde{\Omega}^{(0)}_{xy} + i\int_0^1 d\lambda\,[\mathcal{A}_x(\lambda),\mathcal{A}_y(\lambda)]\right)\right]\ , \quad (32)$$

which is a remarkably simple result in the end. Using Eq. (28), this can be written explicitly as

$$G(\mathbf{k}) = {\rm Tr}\left[B\Big(\Omega^{(0)}_{xy} + \overline{\overline{B}}_{[x,y]} + i[\,\overline{\overline{B}}_x, A^{(0)}_y\,] - i[\,\overline{\overline{B}}_y, A^{(0)}_x\,]\Big)\right]\ , \quad (33)$$

where

$$\overline{\overline{B}}_x = \int_0^1 d\lambda\, \overline{B}_x(\lambda)\ , \tag{34}$$

$$\overline{\overline{B}}_y = \int_0^1 d\lambda\, \overline{B}_y(\lambda)\ , \tag{35}$$

$$\overline{\overline{B}}_{[x,y]} = i\int_0^1 d\lambda\, [\,\overline{B}_x(\lambda), \overline{B}_y(\lambda)\,]\ . \tag{36}$$

Eq. (33) is one of the central results of this paper.

We would like to make some remarks on the formula for $\theta_{\rm GD}$. First, the results are almost independent of the actual states at the top and bottom of the gauge discontinuity. The only way these come in is through the Berry potentials $A_x^{(0)}$ and $A_y^{(0)}$ and the Berry curvature $\Omega_{xy}^{(0)}$ defined on one of the planes. Second, it can easily be shown that the results are the same whether one uses the "bottom" surface in Fig. 1 as a reference and integrates up in λ, as done above, or chooses the "top" surface as a reference and integrates down. Third, the integration over the λ axis can be carried out analytically in the basis that locally diagonalizes $B(\mathbf{k})$, as detailed in Appendix B. Therefore only a 2D discrete integration over the $\mathbf{k}$ plane is needed, which is numerically efficient. Lastly, in the single-band case all quantities such as W, B_x and $\overline{B}_x$ obviously commute with each other, leaving $G = G_1 = \mathrm{Tr}\left[B\Omega_{xy}^{(0)}\right]$.[21]

In the following subsection, we discuss the properties of the $\theta_{\rm GD}$ formula in the presence of $\mathcal{T}$ symmetry, showing that if a TR-symmetric gauge has been chosen in the bulk BZ and assuming that $B(\mathbf{k})$ varies smoothly in the 2D $\mathbf{k}$ plane, both $\theta_{\rm BK}$ and $\theta_{\rm GD}$ must vanish.

B. Time-reversal symmetry

Consider the situation in which the system has $\mathcal{T}$ symmetry and is topologically normal, and a gauge respecting $\mathcal{T}$ symmetry has been chosen smoothly throughout the bulk BZ for the $2N$ occupied bands. For such a system we can construct $2N$ localized Wannier functions (WFs) which fall into N $\mathcal{T}$-symmetric pairs,

$$\begin{aligned}\mathcal{T}|w_{n\mathbf{R},1}\rangle &= -|w_{n\mathbf{R},2}\rangle\ ,\\ \mathcal{T}|w_{n\mathbf{R},2}\rangle &= |w_{n\mathbf{R},1}\rangle,\end{aligned} \tag{37}$$

where $1 \le n \le N$ is the index of a $\mathcal{T}$-symmetric pair and $\mathbf{R}$ denotes a real-space lattice vector. Typically $|w_{n\mathbf{R},1}\rangle$ and $|w_{n\mathbf{R},2}\rangle$ are chosen to diagonalize the S_z operator in their two-dimensional subspace,[22] so that "1" and "2" can be interpreted roughly as "spin indices." The Fourier transform of the $\mathcal{T}$-symmetric WF pairs leads to a smooth gauge respecting $\mathcal{T}$ symmetry in the bulk BZ,

$$\begin{aligned}\mathcal{T}|\psi_{n\mathbf{k},1}\rangle &= -|\psi_{n-\mathbf{k},2}\rangle\\ \mathcal{T}|\psi_{n\mathbf{k},2}\rangle &= |\psi_{n-\mathbf{k},1}\rangle\end{aligned} \tag{38}$$

and

$$\begin{aligned}\mathcal{T}|u_{n\mathbf{k},1}\rangle &= -|u_{n-\mathbf{k},2}\rangle\\ \mathcal{T}|u_{n\mathbf{k},2}\rangle &= |u_{n-\mathbf{k},1}\rangle\end{aligned} \tag{39}$$

where the indices "1" and "2" are again the "spin indices", even if the directions of the spin expectation values can have some variations with $\mathbf{k}$. Note that the states in Eq. (38) are of Bloch form, but in general are not the eigenstates of the Hamiltonian.

Henceforth we shall say that a gauge that obeys Eqs. (38-39) is a $\mathcal{T}$-symmetric gauge. However, in general a gauge obeying these equations is not necessarily periodic. For example, there may be a gauge discontinuity located at some boundary plane in the 3D BZ. When the $\mathbb{Z}_2$ index of the system is even, such a gauge discontinuity can typically be removed by smoothening the gauge without breaking the $\mathcal{T}$ symmetry. When the $\mathbb{Z}_2$ index is odd, however, the gauge discontinuity can never be eliminated without breaking the $\mathcal{T}$ symmetry in the gauge. For, if it could, one could again construct $\mathcal{T}$-respecting WFs, which is known to be impossible for $\mathbb{Z}_2$-odd insulators.

If the gauge in the bulk BZ satisfies Eq. (39), it follows that the Berry curvatures and Berry connections obey

$$\begin{aligned}A_a(\mathbf{k}) &= \sigma_y \left(A_a(-\mathbf{k})\right)^T \sigma_y,\\ \Omega_{ab}(\mathbf{k}) &= -\sigma_y \left(\Omega_{ab}(-\mathbf{k})\right)^T \sigma_y\ ,\end{aligned} \tag{40}$$

where a and b run over the reciprocal-lattice directions. All the quantities in Eq. (40) are $2N \times 2N$ matrices. In particular, σ_y denotes the outer product between the 2×2 Pauli matrix τ_y and the $N \times N$ identity matrix, and the superscript "T" refers to matrix transpose for the $2N \times 2N$ matrices. Since the Berry curvature is odd in $\mathbf{k}$, while the Berry connections behave as even functions of $\mathbf{k}$, it is easy to show that both $\mathrm{Tr}\left[\,A_a(\mathbf{k})\,\Omega_{bc}(\mathbf{k})\,\right]$ and $\mathrm{Tr}\left[\,iA_a(\mathbf{k})\,[A_b(\mathbf{k}), A_c(\mathbf{k})]\,\right]$ are canceled by their time-reversal partners at $-\mathbf{k}$. Therefore, the bulk integral $\theta_{\rm BK}$ in Eq. (7) vanishes if a smooth $\mathcal{T}$-respecting gauge is constructed in the bulk BZ.

In particular, at the boundary plane where the gauge discontinuity is located, the wavefunctions at the bottom and top planes (say, $k_z = \pm\pi$) are connected via $\mathcal{T}|u^{(0)}_{n\mathbf{k},1}\rangle = -|u^{(1)}_{n-\mathbf{k},2}\rangle$ and $\mathcal{T}|u^{(0)}_{n\mathbf{k},2}\rangle = |u^{(1)}_{n-\mathbf{k},1}\rangle$, where $\mathbf{k}$ is now understood to be a wavevector in the 2D plane. With such a $\mathcal{T}$-respecting gauge choice, the B matrix, the Berry connections, and the Berry curvature satisfy the following relationships:

$$B(\mathbf{k}) = \sigma_y B(-\mathbf{k})^T \sigma_y\ , \tag{41}$$

$$A_x^{(0)}(\mathbf{k}) = \sigma_y \left(A_x^{(1)}(-\mathbf{k})\right)^T \sigma_y\ , \tag{42}$$

$$A_y^{(0)}(\mathbf{k}) = \sigma_y \left(A_y^{(1)}(-\mathbf{k})\right)^T \sigma_y\ , \tag{43}$$

$$\Omega_{xy}^{(0)}(\mathbf{k}) = -\sigma_y \left(\Omega_{xy}^{(1)}(-\mathbf{k})\right)^T \sigma_y\ . \tag{44}$$

Again, superscripts "(0)" and "(1)" refer to the quantities evaluated at $\lambda = 0$ and $\lambda = 1$ respectively. We now show that if Eqs. (41)-(44) are satisfied, and if all the quantities involved in the Eqs. (32)-(33) vary smoothly in the 2D plane, then $\theta_{\rm GD}$ must vanish.

First of all, it is straightforward to show that the first term in Eq. (32) vanishes due to $\mathcal{T}$ symmetry. As the gauge-

covariant Berry curvature on the top plane ($\lambda = 1$) is connected the one on the bottom plane ($\lambda = 0$) via $\widetilde{\Omega}^{(1)}_{xy} = U^\dagger \widetilde{\Omega}^{(0)}_{xy} U$, and $U = e^{-iB}$ commutes with B, it follows that $\mathrm{Tr}\left[B(\mathbf{k})\widetilde{\Omega}^{(1)}_{xy}(\mathbf{k})\right] = \mathrm{Tr}\left[B(\mathbf{k})\widetilde{\Omega}^{(0)}_{xy}(\mathbf{k})\right]$. On the other hand, from Eq. (41) and Eq. (44) we know that $\mathrm{Tr}\left[B(\mathbf{k})\widetilde{\Omega}^{(0)}_{xy}(\mathbf{k})\right] = -\mathrm{Tr}\left[B(-\mathbf{k})\widetilde{\Omega}^{(1)}_{xy}(-\mathbf{k})\right]$, which leads to an exact cancellation for the first term.

The second term in Eq. (32) is trickier. First, from the representation-invariance of the trace and the fact that $W = e^{-i\lambda B}$ commutes with B, we know that $\mathrm{Tr}\left[i\,B\,[\mathcal{A}_x, \mathcal{A}_y]\right] = \mathrm{Tr}\left[i\,B\,[A^{(\lambda)}_x, A^{(\lambda)}_y]\right]$. Then we claim that the Berry connection matrix at $(\mathbf{k}, \lambda)$ is connected to the one at $(-\mathbf{k}, 1-\lambda)$ via a $\mathcal{T}$ transformation

$$A^{(\lambda)}_a(\mathbf{k}) = \sigma_y \left(A^{(1-\lambda)}_a(-\mathbf{k})\right)^T \sigma_y \ , \tag{45}$$

where $A^{(\lambda)}_a \equiv A_a(\lambda)$ with $a = \{x, y\}$ as defined in Eq. (15)-(16). Eq. (45) will be proved properly in Appendix C, but if one considers λ as the third wavevector component, Eq. (45) is indeed very intuitive. Combing Eq. (45) and Eq. (27), it follows that

$$\eta(\mathbf{k}, \lambda) = -\eta(-\mathbf{k}, 1-\lambda) \ , \tag{46}$$

where

$$\begin{aligned} \eta(\mathbf{k}, \lambda) &= \mathrm{Tr}\left[\, i\,B(\mathbf{k})\,[\mathcal{A}^{(\lambda)}_x(\mathbf{k}), \mathcal{A}^{(\lambda)}_y(\mathbf{k})]\,\right] \\ &= \mathrm{Tr}\left[\, i\,B(\mathbf{k})\,[A^{(\lambda)}_x(\mathbf{k}), A^{(\lambda)}_y(\mathbf{k})]\,\right] \end{aligned} \tag{47}$$

is exactly the second term in Eq. (32). Therefore, that term also vanishes due to the cancellation between the integrands at $(\mathbf{k}, \lambda)$ and $(-\mathbf{k}, 1-\lambda)$. It thus follows that θ_{GD} has to vanish for a $\mathcal{T}$-respecting gauge choice.

IV. VORTEX-LOOP CONTRIBUTION

In the previous section, we derived a formula for the gauge discontinuity contribution θ_{GD}, as expressed in Eq. (18) and Eqs. (32)-(33). We also demonstrated that for a system with $\mathcal{T}$ symmetry, if a $\mathcal{T}$-respecting gauge is constructed in the bulk BZ, and if the branch choice is made in such a way that $B(\mathbf{k})$ varies smoothly over the entire 2D $\mathbf{k}$ plane, then both θ_{BK} and θ_{GD} must vanish.

However, it is well known that $\theta = \pi$ for $\mathbb{Z}_2$ TIs, so one may wonder where the quantized θ can come from? The answer is that, in the $\mathbb{Z}_2$-odd case, it is topologically impossible to insist on a branch choice such that B remains smooth throughout the (k_x, k_y) plane. In other words, the 2D $\mathbf{k}$ plane has to be subdivided such that one or more of the eigenvalues of B change by an integer multiple of 2π when crossing from one subregion to another. We denote the boundaries of such 2D subregions as "vortex loops." It turns out that the vortex-loop contribution is exactly π for a $\mathbb{Z}_2$ TI.

In this section, we introduce such vortex loops and discuss their contribution to the θ coupling. We first propose a formal definition of a vortex loop in Sec. IV A, and then derive a formula for the vortex-loop contribution θ_{VL} in Sec. IV B. This formula turns out to be rather simple, involving two Berry phases that are accumulated as one traverses the vortex loop, one associated with the electronic Bloch-like functions and the other with the eigenvectors of $B(\mathbf{k})$. In Sec. IV C we discuss several properties of our formula for θ_{VL}. In particular, we show that in systems with $\mathcal{T}$ symmetry, and for which the gauge also respects $\mathcal{T}$ symmetry, θ_{VL} must be either 0 or π, corresponding to the $\mathbb{Z}_2$ classification of 3D $\mathcal{T}$-invariant insulators.

A. What is a vortex loop

In Sec. II B we suggested that the complete formula for θ should include three kinds of contributions, as expressed by Eq. (11). Here we review the philosophy of the calculation, explaining why the third vortex-loop contribution θ_{VL} may be needed.

First, we choose a smooth gauge in the 3D bulk BZ, but the periodicity condition in the k_z direction is relaxed. Hence some gauge discontinuity is introduced at a 2D boundary plane normal to k_z. The 3D bulk integral of Eq. (3) (excluding the boundary plane) is the θ_{BK} term in Eq. (11).

Next, we identify the 2D boundary as $\mathcal{S}$. Let us define $\mathcal{S}$ as a directed area with surface normal $\hat{\mathbf{n}}$. In order to compute the integral over the 2D plane $\mathcal{S}$, $\hat{\mathbf{n}}$ is chosen in such a way that $\hat{\mathbf{x}}$-$\hat{\mathbf{y}}$-$\hat{\mathbf{n}}$ form a right-handed coordinate triad. The gauge discontinuity in the $\hat{\mathbf{n}}$ direction is given by a unitary matrix $U(\mathbf{k}) = e^{-iB(\mathbf{k})}$ which varies smoothly with $\mathbf{k}$ lying in the 2D plane. Since the Hermitian matrix $B(\mathbf{k}) = i \ln U(\mathbf{k})$ is involved in the formula for θ_{GD} (Eq. (33)), a branch choice for B has to be made. If possible we make a branch choice so that $B(\mathbf{k})$ is smooth and continuous over the entire $\mathbf{k}$ plane, but this may not always be possible or desirable. In that case $\mathcal{S}$ is divided into subregions within each of which $B(\mathbf{k})$ is smooth and continuous. For example, Fig. 2 shows $\mathcal{S}$ divided into two subregions $\mathcal{S}_{\mathrm{GD}}$ and $\overline{\mathcal{S}}_{\mathrm{GD}}$ separated by a boundary loop $\mathcal{C}$, which we refer to as a "vortex loop." The 2D contribution θ_{GD} is then computed by integrating over all subregions of $\mathcal{S}$ using Eq. (32)-(33) of Sec. III.

Since the B and U matrices have the same eigenvectors, the eigenvalues of B may exhibit abrupt 2π jumps as they vary from one subregion to another (from $\mathcal{S}_{\mathrm{GD}}$ to $\overline{\mathcal{S}}_{\mathrm{GD}}$ in Fig. 2), even though U remains smooth throughout the (k_x, k_y) plane. The behavior of B is thus singular when crossing the vortex loops. The vortex-loop contributions cannot be computed from the formula for θ_{GD}; a new formula is needed to account for them.

In more general cases, a 3D BZ may be divided into multiple subvolumes, and these subvolumes can meet on multiple 2D surface patches with gauge discontinuities. These surface patches may further meet at one or more 1D lines or curves, which may behave as vortex loops. For such cases the definition of a vortex loop would need to be generalized, since

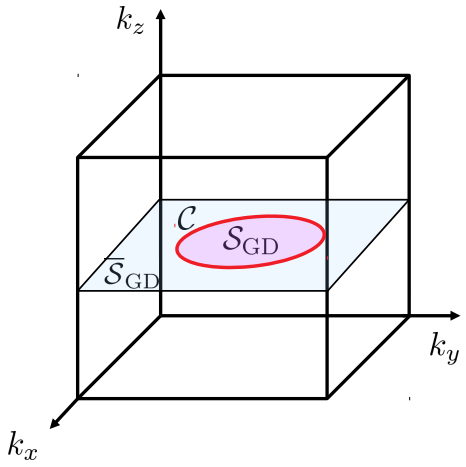

FIG. 2. Schematic illustration of a 3D BZ containing a 2D plane of gauge discontinuity that is divided into two patches $\mathcal{S}_{\rm GD}$ and $\overline{\mathcal{S}}_{\rm GD}$ by a vortex loop (red line).

the $U(\mathbf{k})$ matrices obtained by approaching the meeting line from different patches are in general no longer consistent, and may not even commute with one another. We leave this more complicated situation to a future study.

The presence or absence of vortex loops clearly depends on how the branch choice of B is made in the 2D $\mathbf{k}$ plane. We normally try to make this choice so as to avoid vortices. If the system does not have $\mathcal{T}$ symmetry (and assuming vanishing Chern numbers), then it is usually straightforward to do this, since the eigenvalues of B typically remain non-degenerate throughout the 2D $\mathbf{k}$ plane (degeneracies in a general Hermitian matrix are of codimension three, and so do not occur without special tuning in a 2D $\mathbf{k}$ plane).

However, when the system is topologically nontrivial, this may become impossible; a topological obstruction may force the existence of at least one vortex loop. In particular, if $\mathcal{T}$ symmetry is present, there must be a degeneracy between two different eigenvalues of B at the four time-reversal invariant momenta (TRIM) in the 2D $\mathbf{k}$ plane.[23] As a result, the topological properties of the bulk Hamiltonian become closely related to the number of vortex loops. In the same vein as the $\mathbb{Z}_2$ classification based on the number of surface Dirac cones,[24] when there is an odd number of vortex loops, the system is $\mathbb{Z}_2$-odd, corresponding to a $\mathcal{T}$-respecting topological insulator. Otherwise when the number of vortex loops is even, the system is topologically trivial. In the topologically nontrivial case, it is impossible to insist on the smoothness of all of the eigenvalues of B throughout the 2D $\mathbf{k}$ plane. In principle the last vortex loop can be made infinitesimally small by shrinking it around one of the TRIM, but the symmetry-protected degeneracy at the TRIM prevents it from being removed completely. Therefore, we must consider the contribution from vortex loops in such topologically nontrivial phases.

On the other hand, vortex loops may be present even in topologically trivial cases unless one makes a proper branch choice to remove them. In realistic calculations, for example, one usually adopts some default branch choice for the eigenvalues of B (e.g., from -π to π), which is not necessarily the one that makes B globally smooth. In such cases one has to consider both $\theta_{\rm GD}$ and $\theta_{\rm VL}$. In this regard, it would be useful to have a formula for the vortex loop contribution, so that one can evaluate the gauge-discontinuity contribution to θ for an arbitrary branch choice.

In the remainder of this section we will derive a formula for $\theta_{\rm VL}$ and discuss properties of the formula. We will also show that in the presence of $\mathcal{T}$ symmetry in both the Hamiltonian and the gauge, the vortex-loop contribution alone determines whether the system is $\mathbb{Z}_2$-odd ($\theta_{\rm VL} = \pm\pi$) or $\mathbb{Z}_2$-even ($\theta_{\rm VL} = 0$).

B. The formula for $\theta_{\rm VL}$

Let us first consider the topologically trivial case in which we can always find a proper branch choice such that B remains smooth throughout the 2D plane. Assuming this has been done, now shift the nth eigenvalue of B by $2\pi\nu(n)$ within subregion $\mathcal{S}_{\rm GD}$, thus creating a vortex loop $\mathcal{C}$ whose interior is $\mathcal{S}_{\rm GD}$ as shown in Fig. 2. The above operation is equivalent to making a different branch choice. However, a physical quantity should be independent of the branch choice, so θ should remain invariant after such an operation. Letting $\theta_{\rm shift}$ be the change in $\theta_{\rm GD}$ arising from this redefinition of B in the interior region $\mathcal{S}_{\rm GD}$, it follows that we must have

$$\theta_{\rm VL} = -\theta_{\rm shift}\ . \tag{48}$$

We begin by considering a simple case in which only one of the eigenvalues of B is shifted by 2π within $\mathcal{S}_{\rm GD}$. We make the decomposition $B = B_0 + \Delta B$, where B_0 is the original smooth part and ΔB is the change arising from the 2π shift. We then choose to connect the states at the bottom and top planes in two steps. In the first step,

$$|\psi_m^{(\lambda)}\rangle = \sum_{n=1}^{N} |\psi_n^{(0)}\rangle\, (e^{-i\lambda\Delta B})_{nm}, \quad \lambda \in [0,1). \tag{49}$$

In the second step,

$$|\psi_m^{(\lambda)}\rangle = \sum_{n=1}^{N} |\psi_n^{(1)}\rangle\, (e^{-i(\lambda-1)B_0})_{nm}, \quad \lambda \in [1,2]. \tag{50}$$

Note that the states at the top plane are now denoted as $|\psi_m^{(2)}\rangle$ instead of $|\psi_m^{(1)}\rangle$. In the second step, B_0 is smooth over the entire 2D BZ; one can define $\lambda' = \lambda - 1$ with $\lambda' \in [0,1]$, and the formula for $\theta_{\rm GD}$ derived in Sec. III applies. Thus, $\theta_{\rm shift} = -\theta_{\rm VL}$ is just the contribution to $\theta_{\rm GD}$ coming from the gauge twist of Eq. (49) in the loop interior $\mathcal{S}_{\rm GD}$.

We assume without loss of generality that the first eigenvalue of B (denoted by b_1) jumps by 2π in the subregion $\mathcal{S}_{\rm GD}$. Then ΔB can be written as

$$\Delta B = \begin{cases} V\,\Delta_1\,V^\dagger & \text{if } \mathbf{k} \in \mathcal{S}_{\rm GD} \\ 0 & \text{otherwise} \end{cases} \tag{51}$$

where Δ_1 is an $N \times N$ matrix (N is the number of occupied bands), with $(\Delta_1)_{11} = 2\pi$ and all the remaining matrix elements vanishing. Here $V = (v_1, v_2, ..., v_N)$ is the unitary matrix whose n'th column v_n is the n'th eigenvector of B_0. Plugging this expression for ΔB into the expression for G in Eq. (33), one obtains a formula for θ_{shift}, and θ_{VL} is simply the opposite of θ_{shift}. After some considerable algebra, which we defer to Appendix D, it turns out that many terms cancel, and one obtains the surprisingly simple formula

$$\theta_{\text{VL}} = -\theta_{\text{shift}} = \Big[\, \phi_1(\mathcal{C}) + \xi_1(\mathcal{C}) \,\Big]/2 \,. \tag{52}$$

Here ϕ_1 and ξ_1 are two different, but related, Berry phases that need to be computed around loop $\mathcal{C}$ (taking the positive sense of circulation with respect to the unit normal to $\mathcal{S}_{\text{GD}}$). The second term $\xi_1(\mathcal{C})$ is easier to describe; it is just the Berry phase of the N-component vector v_1 (the first column of V) as it is taken around the loop $\mathcal{C}$. To understand the first term, note that the elements of V can be used to build the linear combinations

$$|\bar{\psi}^{(0)}_{n\mathbf{k}}\rangle = \sum_{m=1}^{N} |\psi^{(0)}_{m\mathbf{k}}\rangle \, V_{mn} \tag{53}$$

out of the Bloch functions $|\psi^{(0)}_{m\mathbf{k}}\rangle$ at the bottom plane ($\lambda = 0$), such that

$$|\bar{\psi}^{(0)}_{1\mathbf{k}}\rangle = \sum_{m=1}^{N} v_{1,m} \, |\psi^{(0)}_{m\mathbf{k}}\rangle \tag{54}$$

is precisely the state whose phase is shifted by 2π, while the other $N-1$ states are unaffected by ΔB. Then the first term $\phi_1(\mathcal{C})$ in Eq. (52) is just the Berry phase of $|\bar{\psi}^{(0)}_{1\mathbf{k}}\rangle$ as it is carried around the loop $\mathcal{C}$. The gauge-invariance and other properties of this formula will be discussed further in the next subsection.

In the most general case, there may be multiple vortex loops $\{\mathcal{C}_i, i = 1, ...L\}$ in the 2D $\mathbf{k}$ plane, and inside the ith vortex loop the nth eigenvalue of B may be shifted by $2\pi\nu_n(i)$ with $\nu_n(i)$ being an integer. Then Eq. (52) can be generalized in a straightforward manner to

$$\theta_{\text{VL}} = \sum_i \sum_n \frac{\phi_n(\mathcal{C}_i) + \xi_n(\mathcal{C}_i)}{2} \, \nu_n(i) \tag{55}$$

where $\phi_n(\mathcal{C}_i)$ and $\xi_n(\mathcal{C}_i)$ are the Berry phases around the loop $\mathcal{C}_i$ of the nth Bloch-like state $|\bar{\psi}^{(0)}_{n\mathbf{k}}\rangle$ (Eq. (53)) and the nth eigenvector of B respectively. Eq. (52), together with its generalized form Eq. (55), is the other central result of this paper.

C. Discussion

We discuss the properties of Eq. (55) in this subsection. We first show that Eq. (55) is indeed gauge invariant modulo 2π, which is consistent with the 2π ambiguity of θ. Secondly we prove that Eq. (55) remains unchanged by interchanging the two steps corresponding to Eqs. (49) and (50). Lastly we discuss the case of $\mathcal{T}$ symmetry and conclude that as long as a gauge respecting $\mathcal{T}$ symmetry is used, $\theta_{\text{VL}} = \pm\pi$ or 0 depending on whether the system is $\mathbb{Z}_2$-odd or $\mathbb{Z}_2$-even, respectively.

1. *Gauge invariance*

Eq. (55) is rather unexpected, as it involves the average of two Berry phases in a manner that, to our knowledge, has not been encountered before. Nevertheless, it is easy to confirm that it obeys one important property, namely, that it is well-defined modulo 2π, as required for any plausible formula for θ. To prove this, we first note that the only gauge freedom in Eq. (55) is a $U(1)$ gauge transformation acting on v_n, i.e., $v_n \to v_n e^{i\beta}$ ($\mathbf{k}$-dependence is implicit). On the other hand, since $|\bar{u}^{(0)}_n\rangle = \sum_{m=1}^{N} |u^{(0)}_m\rangle \, v_{n,m}$, the same gauge transformation must also be applied to $|\bar{u}^{(0)}_n\rangle$, i.e., $|\bar{u}^{(0)}_n\rangle \to |\bar{u}^{(0)}_n\rangle e^{i\beta}$. As a result, if the gauge transformation has a non-zero winding number J, such that ξ_n is changed by $2\pi J$, then ϕ_n must change by $2\pi J$ as well. It follows that Eq. (55) is gauge invariant modulo 2π.

2. *Order of the two steps*

In Sec.IV B we decomposed B into two parts, $B = B_0 + \Delta B$, where B_0 is the smooth part and ΔB is the contribution from the 2π shift (equal and opposite to the vortex-loop contribution). Then, in Eqs. (49) and (50), B was treated in two steps in the fictitious λ space. The first step ($0 < \lambda < 1$) dealt with ΔB, while the second step ($1 < \lambda < 2$) treated the smooth part B_0. Here we would like to show that Eqs. (52) and (55) remain correct regardless of the order of the λ integrations.

If the order is reversed, it is straightforward to show that Eq. (52) remains unchanged, but the first term ϕ_1 is interpreted as the Berry phase of $|\bar{u}^{(1)}_1\rangle$, where $|\bar{u}^{(1)}_1\rangle = \sum_n |u^{(1)}_n\rangle \, v_{1,n} = \sum_{n,m} |u^{(0)}_m\rangle \, (e^{-iB_0})_{mn} \, v_{1,n}$. The Berry phases of $|\bar{u}^{(1)}_1\rangle$ and $|\bar{u}^{(0)}_1\rangle$ around the vortex loop $\mathcal{C}$ are exactly the same, because $|\bar{u}^{(1)}_1\rangle = \sum_{m,n,l} |\bar{u}^{(0)}_l\rangle \, v^*_{l,m} \, (e^{-iB_0})_{mn} \, v_{1,n} = |\bar{u}^{(0)}_1\rangle \, e^{-ib_1}$, where b_1 is the first eigenvalue of B_0. Since b_1 is smooth and single-valued everywhere in the 2D $\mathbf{k}$-plane, the Berry phase would not change under such a single-band transformation. Therefore, Eq. (52) and Eq. (55) remain valid even if the order of Eq. (49) and Eq. (50) is reversed.

3. *Time-reversal symmetry*

We proceed to prove that θ_{VL} must be either $\pm\pi$ or 0 for $\mathcal{T}$-invariant systems when the gauge in the bulk BZ is chosen to respect $\mathcal{T}$ symmetry. Again, let us consider the simple case that there is only one vortex loop $\mathcal{C}$ in the 2D $\mathbf{k}$ plane, and that only the first eigenvalue of B is shifted by 2π inside the vortex loop. Suppose that a smooth gauge respecting $\mathcal{T}$ symmetry has been constructed in the bulk BZ, so that

both the bulk integral $\theta_{\rm BK}$ and the surface integral $\theta_{\rm GD}$ vanish as discussed in Sec. III B. Due to the $\mathcal{T}$-respecting gauge of Eq. (39), the B matrix must satisfy Eq. (41), with two eigenvalues being degenerate at each of the four TRIM, i.e., $(0,0)$, $(0,\pi)$, $(\pi,0)$ and (π,π). As a result, the vortex loop $\mathcal{C}$ has to be a "$\mathcal{T}$-symmetric" loop centered at one of the TRIM, which means that for any $\mathbf{k}$ on the loop $\mathcal{C}$, $-\mathbf{k}$ must also lie on the loop. Then it is well known that the Berry phase around such a $\mathcal{T}$-symmetric loop enclosing a degeneracy point is $\pm\pi$, as has been demonstrated in the surface states of TIs and in $\mathcal{T}$-invariant systems with giant Rashba spin-orbit splitting.[25,26] It follows that $\xi_1 = \pm\pi$ in Eq. (52)

It can be further shown that ϕ_1 in Eq. (52) is exactly the same as ξ_1 as a result of the $\mathcal{T}$ symmetry. Let us first make a branch choice such that the vortex loop is negligibly small, then the Berry connection of $|\bar{u}_1^{(0)}\rangle$ can be expressed as

$$\begin{aligned}\bar{A}^{(0)}_{a,11} &= i\,\langle \bar{u}_1^{(0)} | \partial_a \bar{u}_1^{(0)} \rangle \\ &= i \sum_{m,n=1}^{N} V^*_{m1} \langle u_m^{(0)} | \partial_a u_n^{(0)} \rangle\, V_{n1} + i \sum_{n=1}^{N} V^*_{n1} \partial_x V_{n1} \\ &= (V^\dagger A_a^{(0)}\, V)_{11} + C_{a,11} \end{aligned} \tag{56}$$

where N is the number of occupied bands, $A_a^{(0)}$ is the Berry-connection matrix in the bottom-plane gauge with $a = \{x, y\}$, and

$$C_{a,mn} = i \sum_{l=1}^{N} V^*_{lm} \partial_a V_{ln}\ , \tag{57}$$

may be interpreted as the "Berry connection" in the gauge space. As the vortex loop is chosen to be vanishingly small, the variation of $|u_1^{(0)}\rangle$ within the vortex loop is negligible. Therefore $A_a^{(0)} = 0$, which means $\bar{A}^{(0)}_{a,11}$ comes only from the gauge twist, i.e., $\bar{A}^{(0)}_{a,11} = C_{a,11}$. It follows that $\xi_1 = \phi_1 = \pm\pi$ for such a special branch choice, and $\theta_{\rm VL} = \pm\pi$ according to Eq. (52).

Now suppose the loop is enlarged while preserving $\mathcal{T}$ symmetry in the shape of the loop. We showed at the end of Sec. IV C 3 that contributions to $\theta_{\rm GD}$ coming from $\mathbf{k}$ and $-\mathbf{k}$ always cancel when there is a $\mathcal{T}$-respecting gauge in the bulk, so $\theta_{\rm GD}$ continues to vanish as the loop is enlarged. By the argument given around Eq. (48), this means $\theta_{\rm shift}$, and therefore $\theta_{\rm VL}$, cannot change as the loop is enlarged, even if the variation of $|u_1^{(0)}\rangle$ is no longer negligible. In other words, given a $\mathcal{T}$-respecting gauge in the bulk BZ, $\theta_{\rm VL}$ must be quantized as $\pm\pi$ in the $\mathbb{Z}_2$-odd case regardless of the size of the vortex loop.

We can generalize the discussion to a more general case with multiple vortex loops. Obviously when there is an odd number of vortex loops, θ is still quantized as $\pm\pi$ (modulo 2π). If there is an even number of vortex loops, they can either enclose an even number of TRIM or fall into $\mathcal{T}$ partners without enclosing any TRIM, and θ has to vanish (modulo 2π) in either case.

V. APPLICATIONS

In this section, we apply our method to the Fu-Kane-Mele (FKM) model,[24] which is a 4-band tight-binding model of s electrons on the diamond lattice. The model Hamiltonian is

$$H = \sum_{\langle i,j\rangle} t_{ij} c_i^\dagger c_j + i8\lambda_{\rm SO} \sum_{\langle\langle i,j\rangle\rangle} c_i^\dagger\, \mathbf{s}\cdot(\mathbf{d}^1_{ij} \times \mathbf{d}^2_{ij})\, c_j\ , \tag{58}$$

where t_{ij} is the first-neighbor spin-independent hopping and $\lambda_{\rm SO}$ is the strength of the second-neighbor spin-dependent hopping generated by spin-orbit coupling (SOC); $\mathbf{d}^1_{ij}$ and $\mathbf{d}^2_{ij}$ are the two first-neighbor bond vectors connecting the two second-neighbor sites i and j; and $\mathbf{s} = (s_x, s_y, s_z)$ are Pauli matrices representing the electronic spin. Hereafter we only consider the case of half filling, i.e., two occupied bands. Setting $t_{ij} = t_0 = 1$ and $\lambda_{\rm SO} = 0.125$, it is easy to check that the system is a semimetal with gap closures at the three equivalent X points in the BZ when the diamond-lattice symmetry is preserved. An energy gap can be opened up if an appropriate symmetry-lowering perturbation is added. For example, when the first-neighbor bond along the [111] direction is distorted, the system can be either a trivial insulator or a topological insulator depending on the strength of the bond distortion.

In order to validate our method, we need to consider the general case without $\mathcal{T}$ symmetry. Following Ref. 13, we modify the system by applying a staggered Zeeman field with amplitude h, direction $\hat{\mathbf{h}}$ along [111], and opposite signs on the A and B sublattices. Moreover, the [111] first-neighbor bond is distorted by changing the corresponding hopping amplitude from t_0 to $3t_0 + \delta$. We work in polar coordinates in the (δ, h) parameter space, i.e., $\delta = m\cos\beta$ and $h = m\sin\beta$. The Hamiltonian then becomes

$$\begin{aligned} H(\beta) = &\sum_{\langle i,j\rangle = [111]} (3t_0 + m\cos\beta)\, c_i^\dagger c_j + \sum_{\langle i,j\rangle \neq [111]} t_0\, c_i^\dagger c_j \\ &+ i8\lambda_{\rm SO} \sum_{\langle\langle i,j\rangle\rangle} c_i^\dagger\, \mathbf{s}\cdot(\mathbf{d}^1_{ij} \times \mathbf{d}^2_{ij})\, c_j \\ &+ m\sin\beta \sum_i c_i^\dagger\, \hat{\mathbf{h}}\cdot\mathbf{s}\,\tau_z\, c_i \end{aligned} \tag{59}$$

where τ_z is the Pauli matrix defined in the space of the two sublattices. When $\beta = 0$ and π, the Zeeman field vanishes so that $\mathcal{T}$ symmetry is restored, but the topological index reverses between these two cases. As β increases from 0 to π, the system varies smoothly from a trivial to a topological insulator along a $\mathcal{T}$-breaking path without closing the bulk energy gap.

Setting $t_0 = 1$, $\lambda_{\rm SO} = 0.125$, and $m = 0.5$, we first study the behavior of the B matrix of Eq. (12) in the (k_x, k_y) plane with the branch choice $(-7\pi/4, \pi/4]$ for the eigenvalues of $B(\mathbf{k})$. As shown in Fig. 3(a), when the system is in the $\mathbb{Z}_2$-odd phase ($\beta = \pi$) there is a single vortex loop surrounding the TRIM at (π,π). Within the loop, one of the eigenvalues of B (shown in cyan) is shifted by 2π, while the other eigenvalue remains continuous. Moreover, as a result of $\mathcal{T}$ symmetry, the two eigenvalues of B are degenerate at each TRIM, leading to quantized Berry phases as discussed in Sec. IV C. Fig. 3(b) shows what happens if $\mathcal{T}$ symmetry is broken by setting $\beta =$

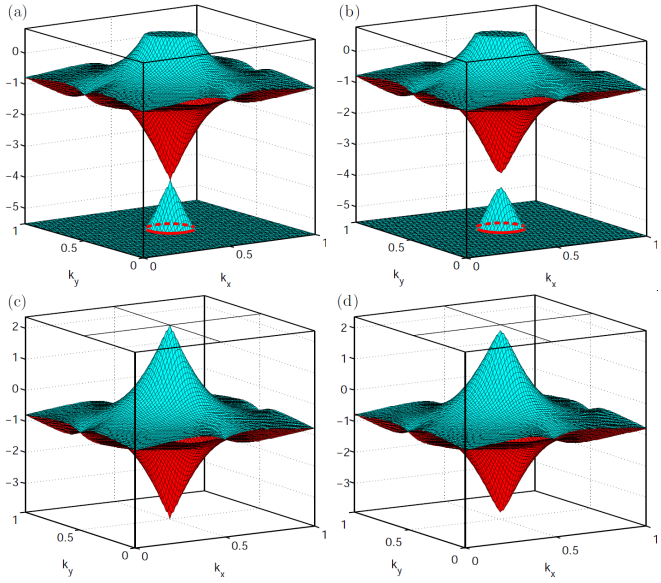

FIG. 3. 3D plots of the two eigenvalues (colored red and cyan) of $B(k_x, k_y) = i \ln U(k_x, k_y)$ for the Fu-Kane-Mele model at half filling: (a) when the system is a TI, i.e., $\beta = \pi$, with the branch choice taken as $(-7\pi/4, \pi/4]$; (b) when $\beta = 0.95\pi$, with the branch choice $(-7\pi/4, \pi/4]$; (c) when $\beta = \pi$, with the branch choice $(-5\pi/4, 3\pi/4]$; and (d) when $\beta = 0.95\pi$, with the branch choice $(-5\pi/4, 3\pi/4]$. The wavevectors k_x and k_y are reduced such that $k_j \in [0, 1]$ generates the 2D BZ.

0.95π. Even though the vortex loop is still present for this value of β, the two eigenvalues of B are no longer degenerate at the TRIM.

As discussed in Sec. IV A, for the $\mathbb{Z}_2$-odd case a vortex loop has to be present regardless of the branch choice. The best one can do is to compress the vortex loop to one of the TRIM in the 2D plane. This is illustrated in Fig. 3(c), where the system of Fig. 3(a) is reanalyzed using a branch choice of $(-5\pi/4, 3\pi/4]$. Now the vortex loop is compressed to the point (π, π) in the (k_x, k_y) plane. On the other hand, using the same branch choice, the vortex loop can be completely removed when $\beta = 0.95\pi$, as shown in Fig. 3(d).

Using the methods developed in Secs. III and IV, we have calculated the total axion response θ along the path from $\beta = 0$ to π by taking the sum of $\theta_{\rm BK}$, $\theta_{\rm GD}$ and $\theta_{\rm VL}$. We first explain the procedures for these calculations before discussing any specific results. The parallel-transport technique,[27] which is detailed in Appendix A, is heavily used in the gauge construction. As discussed earlier, the basic idea is that we first construct a smooth gauge in the bulk BZ that is periodic only in the k_x and k_y directions. Then we can extract the unitary matrix $U(k_x, k_y)$ describing the gauge discontinuity (Eq. (9)) by calculating the overlap matrix between the Bloch states in the top-plane and bottom-plane gauges. The logarithm of $U(k_x, k_y)$, taken with a given branch choice, is the B matrix. We also need to calculate the Berry curvature and Berry connections either in the top-plane gauge or in the bottom-plane gauge. Then all the formulas derived in previous sections can be applied.

To be specific, we first need to construct a smooth and periodic gauge on an arbitrary (k_x, k_y) plane. For definiteness suppose this is the $k_z = 0$ plane. We start by constructing the "1D maxloc" gauge (see Appendix A) along the k_y direction at $k_x = 0$, then make a set of separate parallel transports from $k_x = 0$ to $k_x = 2\pi$ at each k_y, leaving some gauge discontinuity at the line $k_x = 2\pi$ denoted by $Y(k_y) = e^{-iD(k_y)}$. We then apply a local (in $\mathbf{k}$ space) unitary transformation $R(k_x, k_y) = e^{ik_x D(k_y)/2\pi}$ to the occupied states at each point in the 2D plane to smooth out this discontinuity. In the above operation, we have maintained the smoothness of the gauge because the R matrix is defined so as to be smooth in the interior of the 2D plane. Furthermore, the gauge discontinuity at the boundary line $k_x = 2\pi$ has been removed. After these operations, we have successfully constructed a smooth and periodic gauge in the chosen $k_z = 0$ plane.

Taking this gauge in the $k_z = 0$ plane as a "reference gauge," at each (k_x, k_y) we further carry out two sets of parallel transports along the positive and negative k_z directions from $k_z = 0$ to $k_z = \pm\pi$. However, now the periodicity condition in k_z is relaxed so that the states are as aligned to each other as possible in the interval $k_z \in (-\pi, \pi)$. This makes the numeric convergence of the bulk integral, Eq. (7), much easier. The overall result is a gauge that is smooth everywhere in the bulk BZ and periodic only in the k_x and k_y directions. Some gauge discontinuity is left at the plane $k_z = \pm\pi$, which is described by the U matrix introduced in Sec. II B. We are now prepared to apply the formulas derived in Sec. III and Sec. IV to our system of interest.

The above procedures have to be implemented with caution if the system is in the $\mathbb{Z}_2$-odd phase. In this case, it is desirable to construct a bulk gauge respecting $\mathcal{T}$ symmetry, so that both $\theta_{\rm BK}$ and $\theta_{\rm GD}$ vanish, and the remaining contribution from $\theta_{\rm VL}$ is quantized as $\pm\pi$. For a 3D strong TI, however, the 2D $\mathbb{Z}_2$ indices for the $k_z = 0$ plane and the $k_z = \pi$ plane must be opposite. Since it is impossible to construct a smooth and periodic $\mathcal{T}$-symmetric gauge in the $\mathbb{Z}_2$-odd plane,[19] one has to select the $\mathbb{Z}_2$-even plane for the construction of the reference gauge. Since standard methods for computing $\mathbb{Z}_2$ indices are now available,[28,29] even in the absence of $\mathcal{P}$ symmetry, the selection of the $\mathbb{Z}_2$-even plane should be straightforward.

The axion response θ for the FKM model is shown as blue circles in Fig. 4. As β increases from 0 to π, the system evolves from a $\mathbb{Z}_2$-even to a $\mathbb{Z}_2$-odd phase without closing the bulk energy gap, and θ increases smoothly from 0 to π. When β is below $\sim 0.85\pi$, a conventional 3D numeric integral using a fully smooth and periodic gauge throughout the BZ is still practical, and the results obtained from our method are perfectly consistent with those from the conventional method in this regime. Nevertheless, it is much easier to reach numerical convergence using our method. For example, when $\beta = 0.85\pi$, the conventional method requires a $120\times120\times120$ $\mathbf{k}$ mesh to reduce the numerical error to within 1%, while only an $80 \times 80 \times 80$ $\mathbf{k}$ mesh is needed to obtain the same numerical convergence using our method. When β exceeds 0.85π, it becomes impractical to get the expected convergence using the conventional method, and the advantage of our method becomes more obvious. For example, when $\beta = 0.9\pi$, the bulk

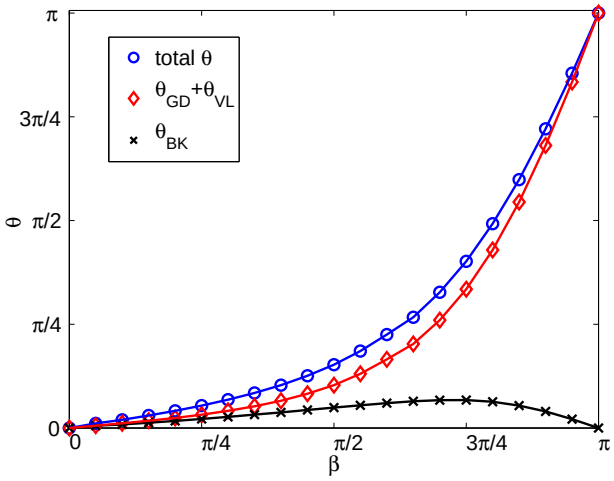

FIG. 4. Axion response θ for the Fu-Kane-Mele model. Blue circles denote the total response. Red diamonds indicate the contribution from the gauge discontinuity, including both the 2D surface integral $\theta_{\rm GD}$ and the 1D vortex-loop integral $\theta_{\rm VL}$. Black crosses represent the contribution from the bulk integral without enforcing periodicity in the k_z direction.

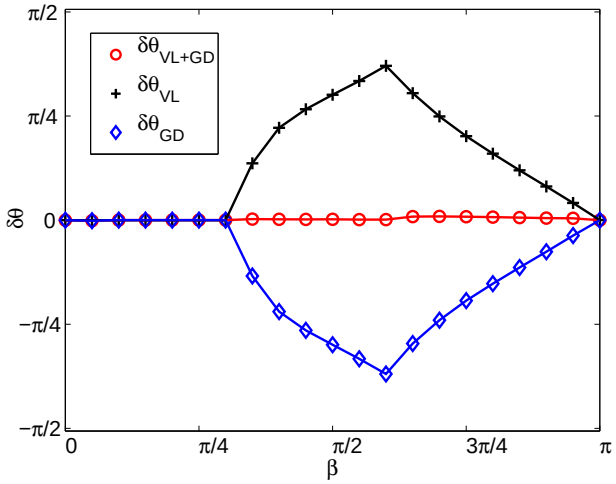

FIG. 5. Difference between the θ values calculated with two different branch choices (see text) for the Fu-Kane-Mele model. The blue diamonds, black pluses and red circles denote the differences for $\theta_{\rm GD}$, $\theta_{\rm VL}$ and $\theta_{\rm VL}+\theta_{\rm GD}$ respectively.

integral using the conventional method (enforcing periodicity in all three directions) does not converge to the expected value even for a $200\times200\times200$ **k** mesh,[30] while it converges easily for a $100\times100\times80$ **k** mesh for $\theta_{\rm BK}$ in our method. The 2D integral $\theta_{\rm GD}$ also converges with a 100×100 2D **k** mesh after the bulk gauge is constructed. The convergence for the vortex-loop integral ($\theta_{\rm VL}$) is even easier; discretizing the loop into $\sim$40 **k** points would typically be enough to get converged values of Berry phases (a 100×100 2D **k** mesh discretizes the vortex loop into 41 **k** points when $\beta=0.9\pi$ with the branch choice $(-7\pi/4,\pi/4]$). Summing over all three terms $\theta_{\rm BK}$, $\theta_{\rm GD}$ and $\theta_{\rm VL}$ eventually leads to the results indicated by blue circles in Fig. 4.

Note that the axion coupling of the FKM model has been calculated previously using other methods. In Ref. 13, when β approaches π, Essin *et al.* switched to some indirect methods such as calculating the total polarization of a finite sample subject to a weak external magnetic field; while in Ref. 31 Taherinejad *et al.* calculated θ in the "hybrid-Wannier-function" basis. The results obtained from our method also agree very well with these previous results when β is close to π.

As shown in Fig. 4, it is helpful to decompose the total θ into the bulk-BZ integral $\theta_{\rm BK}$ and the remainder $\theta_{\rm GD}+\theta_{\rm VL}$, which are indicated by black crosses and red diamonds respectively. One finds that as β increases, $\theta_{\rm GD}+\theta_{\rm VL}$ becomes more and more dominant. Eventually when $\beta=\pi$, θ comes entirely from by the vortex-loop term, which equals π, because both $\theta_{\rm GD}$ and $\theta_{\rm BK}$ vanish due to the $\mathcal{T}$-symmetric bulk gauge.

It should be noted that none of the three terms $\theta_{\rm BK}$, $\theta_{\rm GD}$ or $\theta_{\rm VL}$, is independently gauge invariant. As the size of the vortex loop is dependent on the branch choice, in general both $\theta_{\rm VL}$ and $\theta_{\rm GD}$ are branch-choice dependent, but the sum of them should remain invariant if the bulk gauge is fixed.

The above statement is verified by computing $\theta_{\rm VL}$ and $\theta_{\rm GD}$ using different branch choices for a given gauge in the bulk BZ as shown in Fig. 5, where the blue diamonds (black plus signs) denote the difference between the values of $\theta_{\rm GD}$ ($\theta_{\rm VL}$) calculated using the two different branch choices $(-2\pi,0]$ and $(-7\pi/4,\pi/4]$. For the first branch choice $(-2\pi,0]$, a vortex loop appears when $\beta=0.35\pi$ and then grows as β increases, while for the other branch choice the B matrix remains continuous throughout the 2D **k** plane until $\beta=0.65\pi$. It is clearly seen from Fig. 5 that both $\theta_{\rm VL}$ and $\theta_{\rm GD}$ depend on the branch choice. On the other hand, the red circles in Fig. 5 represent the difference of the total $\theta_{\rm GD}+\theta_{\rm VL}$ computed for the two different branch choices. The difference remains vanishingly small throughout the adiabatic path, thus numerically confirming that the sum of $\theta_{\rm VL}$ and $\theta_{\rm GD}$ remains branch-choice-invariant.

Besides the branch choice, there is still the freedom to choose the gauge in the bulk BZ; both $\theta_{\rm BK}$ and $\theta_{\rm GD}+\theta_{\rm VL}$ depend on this gauge choice. However, since the bulk gauge was chosen in such a way as to align the states with each other as much as possible in the k_z direction, the bulk integral $\theta_{\rm BK}$ is typically small, explaining why $\theta_{\rm GD}+\theta_{\rm VL}$ dominates over $\theta_{\rm BK}$ in Fig. 4.

VI. SUMMARY

To summarize, we have developed a new method for computing the Chern-Simons axion coupling θ. The basic idea is to relax the periodicity condition of the gauge in one of the **k** directions, thus introducing a gauge discontinuity residing at a 2D **k** plane in the BZ. The total θ then has both a bulk contribution $\theta_{\rm BK}$, obtained as a conventional 3D integral over the interior of the bulk BZ, and a gauge-discontinuity contribution $\theta_{\rm GD}$, which is expressed as a 2D integral over

the gauge-discontinuity plane as given by Eqs. (18) and (33). Moreover, it may happen that discontinuities are introduced for a given branch choice of $B(\mathbf{k})$, the logarithm of the unitary connection matrix describing the gauge discontinuity; this is sometimes done for convenience, but may also be required depending on the topological properties of the system. In such cases the gauge-discontinuity plane is further divided into subregions by 1D vortex loops, and one must also consider the vortex-loop contribution as expressed in Eq. (55). The total θ is then $\theta = \theta_{\rm BK} + \theta_{\rm GD} + \theta_{\rm VL}$.

Since the periodicity condition in one of the $\mathbf{k}$ directions (e.g., the k_z direction) is relaxed, the gauge in the bulk BZ does not twist as strongly as in the case when both periodicity and smoothness are required. This leads to improved numerical convergence of the 3D bulk integral of Eq. (7). The loss of periodicity is compensated by extra contributions from the gauge discontinuity ($\theta_{\rm GD}$) and possible vortex loops ($\theta_{\rm VL}$). The formulas for both terms turn out to be fairly simple and can be implemented numerically without difficulty.

It is interesting to note that if a $\mathcal{T}$-respecting gauge has been constructed in the bulk BZ for a $\mathcal{T}$-invariant system, then both $\theta_{\rm BK}$ and $\theta_{\rm GD}$ must vanish. The only surviving term $\theta_{\rm VL}$ is then either 0 or π, corresponding to the $\mathbb{Z}_2$ classification of 3D $\mathcal{T}$-invariant insulators. Our theory thus provides a new interpretation to the formally quantized magnetoelectric response in TIs.

We have applied our method to the Fu-Kane-Mele model with staggered Zeeman field. We calculated the axion response for the model along a $\mathcal{T}$-breaking path connecting the $\mathbb{Z}_2$-even and $\mathbb{Z}_2$-odd phases. Our results agree well with the previous results obtained from other methods.[13,31] In particular, we find that the gauge-discontinuity contribution $\theta_{\rm GD} + \theta_{\rm VL}$ becomes increasingly dominant as the system approaches the $\mathbb{Z}_2$-odd phase. In the TI phase, as mentioned above, θ is completely determined by the vortex-loop term for a $\mathcal{T}$-symmetric gauge in the bulk BZ, and the π quantization of θ is due to the π quantization of the Berry phase around a single vortex loop.

Our method may be generalized to the case that the 3D BZ is divided into multiple subvolumes. These subvolumes may meet each other at multiple 2D surface patches, each with its own gauge discontinuity. The surface patches may further meet at 1D lines or curves, which may be vortex loops. In such more complicated cases, the formula for $\theta_{\rm GD}$ still applies, but the definition of a vortex line has to be generalized to the situation that the U matrices obtained by approaching the vortex loop from different surface patches may no longer commute with each other. Thus the formula for $\theta_{\rm VL}$ may need to be modified. We leave this problem for future study.

From a theoretical point of view, the results presented in this paper provide a step forward in understanding the axion coupling in 3D insulators. We introduced the gauge-discontinuity and vortex-loop contributions to θ, and found that the latter can be expressed in an unusual way as the sum of two closely related Berry phases. From the perspective of first-principles calculations, we have proposed a numerically efficient method for computing the Chern-Simons orbital ME coupling in solids. Our method can be implemented straightforwardly in standard first-principles code packages. This makes it possible to compute the orbital ME coefficients efficiently for realistic materials, thus facilitating the search for functional materials with enhanced ME couplings.

ACKNOWLEDGMENTS

This work was supported by NSF Grant DMR-14-08838.

Appendix A: Parallel transport

In this section, we discuss how to carry out the parallel transport operation and construct the "1D maxloc" gauge starting from a set of occupied eigenstates with an arbitrary gauge on a given $\mathbf{k}$ path. The basic idea is to recursively make the (periodic part of) the Bloch states at each $\mathbf{k}$ point on the path as aligned as possible with the states at the immediately previous $\mathbf{k}$ point. If the $\mathbf{k}$ path is chosen to be a closed loop, the states obtained at the end of the loop may differ from those at the start by some phase factors after the parallel-transport operation, with the mismatch giving exactly the Berry phases of the Bloch states around the loop. The Berry phases accumulated along the $\mathbf{k}$ path can be smeared out by smoothly distributing the phase to the states at every $\mathbf{k}$ point on the loop. After such an operation, one obtains a gauge which is both smooth and periodic along the path (loop), and which we refer to as a 1D maxloc gauge.[27]

To be specific, let us consider a set of occupied bands $|u_{n\mathbf{k}}\rangle$, with $n = 1, ...N$, which are isolated from other bands (in energy) everywhere in the BZ. Let us take a closed $\mathbf{k}$ path along k_z running from 0 to 2π, with the path sampled by J discrete points, so that $\mathbf{k}_j = 2\pi(j-1)\hat{z}/J$. Assume that the eigenstates with some arbitrary gauge $|u^0_{n\mathbf{k}_j}\rangle$ are given for $j = 1, ...J$, and a periodic gauge is chosen at the $(J+1)$th point so that $|u^0_{n\mathbf{k}_{J+1}}\rangle = e^{-i2\pi z}|u^0_{n\mathbf{k}_1}\rangle$. To carry out the parallel transport, we need to insist that each overlap matrix between the occupied states at $\mathbf{k}_{j+1}$ and $\mathbf{k}_j$, i.e., $M_{mn}(j) = \langle u^0_{m\mathbf{k}_j}|u^0_{n\mathbf{k}_{j+1}}\rangle$, is positive-definite Hermitian. This can be done as follows. At each $\mathbf{k}_j$, make a singular-value decomposition to the overlap matrix $M_j = V_j\Sigma_j W_j^\dagger$, where V and W are unitary and Σ is real positive diagonal. Then apply a unitary transformation $L_j = W_j V_j^\dagger$ to $|u^0_{n\mathbf{k}_{j+1}}\rangle$, $|\widetilde{u}_{n\mathbf{k}_{j+1}}\rangle = \sum_{m=1}^N L_{j,mn}|u^0_{m\mathbf{k}_{j+1}}\rangle$, so that the overlap matrix between the unitarily transformed states at neighboring $\mathbf{k}$ points becomes positive-definite Hermitian. If one repeats such an operation from $j = 1$ to $j = J$, the states will become as aligned to each other as possible at all $\mathbf{k}$ points on the path. However, there is still some gauge discontinuity left at the boundary, $|\widetilde{u}_{n\mathbf{k}_{J+1}}\rangle = e^{-i2\pi z}\sum_m \Lambda_{mn}|\widetilde{u}_{m\mathbf{k}_1}\rangle$, where Λ is a unitary matrix. The logarithms of the eigenvalues of Λ, $\beta_n = -i\ln\lambda_n$, are then identified as the non-Abelian Berry phases (also known as Wilson loop eigenvalues) of the Bloch states.

The gauge obtained from the parallel-transport operation is smooth along the $\mathbf{k}$ path, but not periodic. To restore the pe-

riodicity, we need to rotate all the states on the $\mathbf{k}$ path to the basis that diagonalizes Λ, i.e., $|u'_{n\mathbf{k}_j}\rangle = \sum_{m=1}^{N} |\widetilde{u}_{m\mathbf{k}_j}\rangle\, L_{mn}$, where L is the eigenvector matrix of Λ. Then we gradually smear out the discontinuity by applying the following phase twist to the states at $\mathbf{k}_j$: $|u_{n\mathbf{k}_j}\rangle = e^{-i(j-1)\beta_n/J}|u'_{n\mathbf{k}_j}\rangle$. This results in a "1D maxloc gauge" that is both smooth and periodic along the $\mathbf{k}$ path.

Appendix B: Integration over λ in the formula for $\theta_{\rm GD}$

In deriving the expression in Eq. (18) for the gauge discontinuity contribution $\theta_{\rm GD}$ in Sec. III A, we arrived at Eq. (33) involving the quantities $\overline{\overline{B}}_x$, $\overline{\overline{B}}_y$, and $\overline{\overline{B}}_{[x,y]}$ which were expressed as integrals over λ in Eqs. (34-36). We show here that these three quantities can all be computed analytically in the sense that the λ integral does not have to be discretized.

The plan is as follows. Suppose that $A_x^{(0)}(\mathbf{k})$, $A_y^{(0)}(\mathbf{k})$, $\Omega_{xy}^{(0)}(\mathbf{k})$, $B(\mathbf{k})$, $B_x(\mathbf{k})$, and $B_y(\mathbf{k})$ are known. The first term in Eq. (33) is independent of λ and is trivial. For the remaining terms, at each $\mathbf{k}$, locally diagonalize $B(\mathbf{k})$, then transform all of the matrices $A_x^{(0)}(\mathbf{k})$, $A_y^{(0)}(\mathbf{k})$, $B_x(\mathbf{k})$, and $B_y(\mathbf{k})$ to the basis that locally diagonalizes $B(\mathbf{k})$, i.e.,

$$\begin{aligned} &B(\mathbf{k}) \to V^\dagger(\mathbf{k})\, B(\mathbf{k})\, V(\mathbf{k}),\\ &B_a(\mathbf{k}) \to V^\dagger(\mathbf{k})\, B_a(\mathbf{k})\, V(\mathbf{k}),\\ &A_a(\mathbf{k}) \to V^\dagger(\mathbf{k})\, A_a(\mathbf{k})\, V(\mathbf{k}),\\ &\Omega_{xy}(\mathbf{k}) \to V^\dagger(\mathbf{k})\, \Omega_{xy}(\mathbf{k})\, V(\mathbf{k})\ , \end{aligned} \tag{B1}$$

where $V(\mathbf{k})$ is the eigenvector matrix of $B(\mathbf{k})$, and $a = \{x, y\}$. Then one can compute the trace in this basis. Letting $B_{mn} = b_n\, \delta_{mn}$, we find

$$\begin{aligned} \overline{B}_{x,mn}(\lambda) &= \int_0^\lambda d\mu\ e^{-i\mu b_m}\, B_{x,mn}\, e^{i\mu b_n}\\ &= g_{mn}(\lambda)\, B_{x,mn}\ , \end{aligned} \tag{B2}$$

where

$$g_{mn}(\lambda) = \frac{e^{-i\lambda(b_m-b_n)}-1}{-i(b_m-b_n)}\ . \tag{B3}$$

Then

$$\begin{aligned} \overline{\overline{B}}_{x,mn} &= \Big(\int_0^1 g_{mn}(\lambda)\, d\lambda\Big)\ B_{x,mn}\\ &= \Big(\frac{e^{i(b_n-b_m)}-1}{-(b_n-b_m)^2} - \frac{1}{i(b_n-b_m)}\Big)\ B_{x,mn} \end{aligned} \tag{B4}$$

and

$$\begin{aligned} \overline{\overline{B}}_{[x,y]mn} =& i\sum_l \Big(\int_0^1 g_{ml}(\lambda)\, g_{ln}(\lambda)\, d\lambda\Big)\ \times\\ &\Big(B_{x,ml}\, B_{y,ln} - B_{y,ml}\, B_{x,ln}\Big)\ . \end{aligned} \tag{B5}$$

Because we are interested in the trace of $B\,\overline{\overline{B}}_{[x,y]}$ in the basis that B is locally diagonal, only the diagonal matrix elements of $\overline{\overline{B}}_{[x,y]}$ are relevant. After carrying out the integral in Eq. (B5) one obtains the following expression:

$$\begin{aligned} \overline{\overline{B}}_{[x,y]nn} =& i\sum_m \Big(\frac{2}{(b_n-b_m)^2} - \frac{2\sin(b_m-b_n)}{(b_m-b_n)^3}\Big)\ \times\\ &\Big(B_{x,nm}B_{y,mn} - B_{y,nm}B_{x,mn}\Big)\ . \end{aligned} \tag{B6}$$

If two eigenvalues b_m and b_n are degenerate, one needs to take the limit $(b_n - b_m) \to 0$. It turns out that both quantities are finite:

$$\lim_{b_n\to b_m} \overline{\overline{B}}_{x,mn} = B_{x,mn}/2\ , \tag{B7}$$

and

$$\lim_{b_m\to b_n} \overline{\overline{B}}_{[x,y]nn} = \frac{i}{3}\Big(B_{x,nm}B_{y,mn} - B_{y,nm}B_{x,mn}\Big)\ . \tag{B8}$$

Of course the entire calculation still has to be done on a discretized mesh on the $\mathbf{k}$ plane, with finite-difference expressions used to evaluate objects like $A_x^{(0)}(\mathbf{k})$, so it it not "exact". However, it is convenient that we don't have to discretize the λ axis, instead doing all λ integrals analytically.

Appendix C: Derivation of Eq. (45)

In Sec. III B we considered the effect of time-reversal symmetry on the gauge discontinuity on the boundary plane and showed that $\theta_{\rm GD}$ has to vanish for a $\mathcal{T}$-respecting choice of gauge. The demonstration rested on the use of Eq. (45), which was only introduced heuristically there.

Here we prove it properly. From Eqs. (27), (28) and (25), we know that

$$A_a^{(\lambda)} = \widetilde{A}_a^{(\lambda)} + \Gamma_a(0,\lambda) \tag{C1}$$

where

$$\widetilde{A}_a^{(\lambda)} = W^\dagger(\lambda)\, A_a^{(0)}\, W(\lambda)\ , \tag{C2}$$

and the function $\Gamma_a(\lambda_1,\lambda_2)$ is defined as

$$\Gamma_a(\lambda_1,\lambda_2) = \int_{\lambda_1}^{\lambda_2} d\mu\, W^\dagger(\mu)\, B_a\, W(\mu)\ . \tag{C3}$$

Letting $\lambda = 1$, we get the expression

$$A_a^{(1)} = \widetilde{A}_a^{(1)} + \Gamma_a(0,1), \tag{C4}$$

Applying a unitary transformation $W(1-\lambda)$ to the matrix $A_a^{(1)}$, one obtains

$$\begin{aligned} W(1-\lambda)\, A_a^{(1)}\, W^\dagger(1-\lambda) &= \widetilde{A}_a^{(\lambda)} + \Gamma_a(\lambda-1,\lambda)\ ,\\ &= A_a^{(\lambda)} + \Gamma_a(\lambda-1,0)\ , \end{aligned} \tag{C5}$$

where a variable transformation $(\lambda+\nu-1)\to\mu$ has been made to obtain the second term $\Gamma_a(\lambda-1,\lambda)$ on the RHS of the first line in Eq. (C5). The integral from $\lambda-1$ to λ in $\Gamma_a(\lambda-1,\lambda)$ is further divided into two integrals: one from $\lambda-1$ to 0, and the other from 0 to λ. $A_a^{(\lambda)}$ in the second line is then obtained by combining the integral from 0 to λ together with $W^\dagger(\lambda)\,A_a^{(0)}\,W(\lambda)$ (Eq. (C1)). Therefore

$$A_a^{(\lambda)} = W(1-\lambda)\,A_a^{(1)}\,W^\dagger(1-\lambda)\ -\Gamma_a(\lambda-1,0) \quad \text{(C6)}$$

and it immediately follows that

$$\begin{aligned} A_a^{(1-\lambda)} &= W(\lambda)\,A_a^{(1)}\,W^\dagger(\lambda) - \Gamma_a(-\lambda,0) \\ &= W(\lambda)\,A_a^{(1)}\,W^\dagger(\lambda) - \int_0^\lambda d\mu\, W(\mu)\,B_a\,W^\dagger(\mu)\ , \end{aligned} \quad \text{(C7)}$$

where we let $\mu\to-\mu$ in going from the first to the second line in Eq. (C7). Equation (45 then follows by combining Eq. (42)-(43) and Eq. (C7):

$$\begin{aligned} A_a^{(1-\lambda)}(-\mathbf{k}) &= e^{-i\lambda B(-\mathbf{k})}\,A_a^{(1)}(-\mathbf{k})\,e^{i\lambda B(-\mathbf{k})} - \int_0^\lambda d\mu\, e^{-i\mu B(-\mathbf{k})}\,B_a(-\mathbf{k})\,e^{i\mu B(-\mathbf{k})} \\ &= \sigma_y\,e^{-i\lambda B^T(\mathbf{k})}\left(A_a^{(0)}(\mathbf{k})\right)^T e^{i\lambda B^T(\mathbf{k})}\,\sigma_y + \sigma_y\int_0^\lambda d\mu\, e^{-i\mu B^T(\mathbf{k})}\,B_a^T(\mathbf{k})\,e^{i\mu B^T(\mathbf{k})}\,\sigma_y \\ &= \sigma_y\left(W^\dagger(\lambda)\,A_a^{(0)}(\mathbf{k})\,W(\lambda) + \int_0^\lambda d\mu\, W^\dagger(\mu)\,B_a(\mathbf{k})\,W(\mu)\right)^T\sigma_y\ . \end{aligned} \quad \text{(C8)}$$

The last line in Eq. (C8) is simply $\sigma_y\left(A_a^{(\lambda)}(\mathbf{k})\right)^T\sigma_y$, thus proving Eq. (45) and thereby confirming that $\theta_{\rm GD}$ vanishes for a TR-invariant gauge.

Appendix D: Derivation of Eq. (52)

In Sec. IV B we proposed a formula for the vortex-loop contribution as expressed in Eq. (52). We only explained the main idea there, and the formula was introduced without proof. Here we provide a rigorous derivation.

To derive Eq. (52), it is convenient to decompose $G(\mathbf{k})$ into four terms G_1, G_2, G_3 and G_4 corresponding to the four terms on the right-hand side (RHS) of Eq. (33):

$$G_1 = B\,\Omega_{xy}^{(0)}\ , \quad \text{(D1)}$$
$$G_2 = iB\,[\,\overline{B}_x(\lambda),\overline{B}_y(\lambda)\,]\ , \quad \text{(D2)}$$
$$G_3 = iB\,[\,A_x^{(0)},\overline{B}_y(\lambda)\,]\ , \quad \text{(D3)}$$
$$G_4 = iB\,[\,\overline{B}_x(\lambda),A_y^{(0)}\,]\ . \quad \text{(D4)}$$

Since all the quantities such as Ω_{xy} and $A_{x(y)}$ are defined in the bottom-plane gauge, we will drop the superscript "(0)" (indicating the bottom-plane gauge) in later steps. Recalling that the change ΔB in the interior region was expressed in Eq. (51) as $V\Delta_1 V^\dagger$, where V is the unitary matrix that diagonalizes B and Δ is diagonal with 2π-integer entries, we can transform the needed matrices to the B-diagonal representation via

$$A_a' = V^\dagger\,A_a\,V\ , \quad \text{(D5)}$$
$$\Omega_{xy}' = V^\dagger\,\Omega_{xy}\,V\ , \quad \text{(D6)}$$
$$\overline{B}_a' = V^\dagger\,\overline{B}_a\,V\ . \quad \text{(D7)}$$

We will prove Eq. (52) by explicitly calculating the four terms in Eqs. (D1)-(D4).

1. The G_1 term

Plugging Eq. (51) first into the expression for G_1 in Eq. (D1), one obtains

$$\begin{aligned} \mathrm{Tr}\left[\,G_1\,\right] &= \mathrm{Tr}\left[\,V\,\Delta_1\,V^\dagger\Omega_{xy}\,V\,V^\dagger\,\right] \\ &= \mathrm{Tr}\left[\,\Delta_1\,V^\dagger\Omega_{xy}\,V\,\right]\ . \end{aligned} \quad \text{(D8)}$$

Note that $\Omega_{xy}' = V^\dagger\,\Omega_{xy}\,V$ is associated with the Berry curvature of the Bloch states in the bottom-plane gauge that are unitarily transformed by V: $|\bar{u}_n^{(0)}\rangle = \sum_{m=1}^N |u_m^{(0)}\rangle\,V_{mn}$. One can express the Berry curvature of $|\bar{u}_n^{(0)}\rangle$ (denoted by $\bar{\Omega}_{xy}$) in terms of A_x, A_y, Ω_{xy}, V and the partial derivatives of V,

$$\bar{\Omega}_{xy} = \Omega_{xy}' + \Lambda_{xy} + i[\,C_x,A_y'\,] - i[\,C_y,A_x'\,]\ , \quad \text{(D9)}$$

where C_x and C_y are defined in Eq. (57), and

$$\Lambda_{xy} = \partial_x C_y - \partial_y C_x \quad \text{(D10)}$$

can be considered as the Berry curvature in the "gauge space."

From Eq. (D9) it immediately follows that

$$\mathrm{Tr}\Big[\,G_1\,\Big] = \mathrm{Tr}\Big[\,\Delta_1\,\bar{\Omega}_{xy} - \Delta_1\,\Lambda_{xy} - i\Delta_1\,[\,C_x, A'_y\,] + i\Delta_1\,[\,C_y, A'_x\,]\Big]\,. \tag{D11}$$

Before further simplifying Eq. (D11), let us go to the other terms and come back to G_1 later.

2. The G_3 and G_4 terms

Let us deal with the G_3 and G_4 terms. Since B_x and B_y are involved in G_3 and G_4, let us first evaluate these two terms.

$$\begin{aligned} B_x &= \partial_x(\,V\,\Delta_1\,V^\dagger\,) \\ &= \partial_x V\,\Delta_1\,V^\dagger + V\,\Delta_1\,\partial_x V^\dagger \\ &= iV\,[\,\Delta_1, C_x\,]\,V^\dagger\,. \end{aligned} \tag{D12}$$

Similarly, $B_y = iV\,[\,\Delta_1, C_y\,]\,V^\dagger$. Plugging the expressions for B_x and B_y into Eq. (25), one immediately obtains

$$\overline{B}_a(\lambda) = \int_0^\lambda du\,V\,e^{-iu\Delta_1}\,i[\,\Delta_1, C_a\,]\,e^{iu\Delta_1}\,V^\dagger\,. \tag{D13}$$

We now evaluate $\mathrm{Tr}\,[G_3]$ by carrying out the trace in the basis that diagonalizes B using Eqs. (D5-D7). We find

$$\begin{aligned} \mathrm{Tr}\Big[\,G_3\,\Big] &= \mathrm{Tr}\Big[\,i\Delta_1\,[\,A'_x,\,\overline{B}'_y,]\Big] \\ &= \mathrm{Tr}\Big[\,\int_0^\lambda du\,i\Delta_1\,[\,A'_x,\,e^{-iu\Delta_1}\,i[\,\Delta_1, C_y\,]\,e^{iu\Delta_1}\,]\Big] \\ &= \mathrm{Tr}\Big[\,\int_0^\lambda du\,iA'_x\,[\,e^{-iu\Delta_1}\,i[\,\Delta_1, C_y\,]\,e^{iu\Delta_1},\,\Delta_1\,]\Big] \\ &= \mathrm{Tr}\Big[\,iA'_x\int_0^\lambda du\,\partial_u\Big(\,e^{-iu\Delta_1}\,[\,\Delta_1, C_y\,]\,e^{iu\Delta_1}\Big)\Big] \\ &= \mathrm{Tr}\Big[\,iA'_x\,e^{-i\lambda\Delta_1}\,[\,\Delta_1, C_y\,]\,e^{i\lambda\Delta_1} - iA'_x\,[\,\Delta_1, C_y\,]\Big]\,, \end{aligned} \tag{D14}$$

where we have used the equation

$$[\,e^{-iu\Delta_1}\,i[\,\Delta_1, C_y\,]\,e^{iu\Delta_1},\,\Delta_1\,] = \partial_u\Big(\,e^{-iu\Delta_1}\,[\,\Delta_1, C_y\,]\,e^{iu\Delta_1}\Big) \tag{D15}$$

when going from the third to the fourth line in Eq. (D14). Making use of the cyclic property of trace, one immediately realizes that the second term in the last line of Eq. (D14) cancels the last term on the RHS of Eq. (D11), which will be dropped in later steps. Therefore,

$$\begin{aligned} \int_0^1 d\lambda\,\mathrm{Tr}\Big[\,G_3\,\Big] &= \int_0^1 d\lambda\,\mathrm{Tr}\Big[\,iA'_x\,e^{-i\lambda\Delta_1}\,[\,\Delta_1, C_y\,]\,e^{i\lambda\Delta_1}\,\Big] \\ &= \int_0^1 d\lambda\,\mathrm{Tr}\Big[\,-A'_x\,\partial_\lambda(\,e^{-i\lambda\Delta_1}\,C_y\,e^{i\lambda\Delta_1}\,)\,\Big] \\ &= \mathrm{Tr}\Big[\,-A'_x\,(\,e^{-i\lambda\Delta_1}\,C_y\,e^{i\lambda\Delta_1}\,)|_{\lambda=0}^{\lambda=1}\Big] \\ &= 0\,, \end{aligned} \tag{D16}$$

where the following equation has been used to go from the second to the third line in Eq. (D16):

$$ie^{-i\lambda\Delta_1}\,[\,\Delta_1, C_y\,]\,e^{i\lambda\Delta_1} = -\partial_\lambda(\,e^{-i\lambda\Delta_1}\,C_y\,e^{i\lambda\Delta_1}\,)\,. \tag{D17}$$

Similar derivations can be applied to the G_4 term, i.e.,

$$\mathrm{Tr}\Big[\,G_4\,\Big] = \mathrm{Tr}\Big[\,iA'_y\,e^{-i\lambda\Delta_1}\,[\,C_x, \Delta_1\,]\,e^{i\lambda\Delta_1} - iA'_y\,[\,C_x, \Delta_1\,]\Big]\,. \tag{D18}$$

The second term on the RHS of Eq. (D18) cancels the third term on the RHS of Eq. (D11). Dropping the second term in Eq. (D18) and integrating over λ, one obtains $\int_0^1 d\lambda\,\mathrm{Tr}\Big[\,G4\,\Big] = 0$.

3. The G_2 term

In the basis that locally diagonalizes B,

$$\mathrm{Tr}\Big[\,G_2\,\Big] = \mathrm{Tr}\Big[\,i\Delta_1\,[\,\overline{B}'_x, \overline{B}'_y\,]\,\Big]\,. \tag{D19}$$

On the other hand, combining Eq. (D13), Eq. (D7) and Eq. (D17), we get

$$\begin{aligned} \overline{B}'_a &= -\int_0^\lambda d\mu\partial_u(\,e^{-iu\Delta_1}\,C_a\,e^{iu\Delta_1}\,) \\ &= C_a - \widetilde{C}_a\,, \end{aligned} \tag{D20}$$

where $\widetilde{C}_a = e^{-i\lambda\Delta_1}\,C_a\,e^{i\lambda\Delta_1}$. It follows that

$$\mathrm{Tr}\Big[\,G_2\,\Big] = \mathrm{Tr}\Big[\,i\Delta_1\,[\,\widetilde{C}_x - C_x\,,\,\widetilde{C}_y - C_y\,]\,\Big]\,. \tag{D21}$$

If one expands the RHS of Eq. (D21), one would obtain four commutators between C_a and $\widetilde{C}_b$ $(a, b = x, y)$. Since $e^{\pm i\lambda\Delta_1}$ commute with Δ_1, the term involving $[\,\widetilde{C}_x, \widetilde{C}_y\,]$ is equal to the term with $[\,C_x, C_y\,]$. Therefore,

$$\mathrm{Tr}\Big[\,G_2\,\Big] = \mathrm{Tr}\Big[\,\Delta_1\,\Big(\,2i[\,C_x, C_y\,] - i[\,C_x, \widetilde{C}_y\,] - i[\,\widetilde{C}_x, C_y\,]\Big)\Big]\,. \tag{D22}$$

The second term on the RHS of Eq. (D22) can be written as a total derivative of λ as

$$\mathrm{Tr}\Big[\,-i\Delta_1\,[\,C_x, \widetilde{C}_y\,]\,\Big] = -\mathrm{Tr}\Big[\partial_\lambda(\,e^{-i\lambda\Delta_1}\,C_y\,e^{i\lambda\Delta_1}\,C_x\,)\Big]\,. \tag{D23}$$

We need to use $\Delta_1\,e^{\pm i\lambda\Delta_1} = \mp i\partial_\lambda(e^{\pm i\lambda\Delta_1})$ to obtain the above equation. Integrating Eq. (D23) over λ, one obtains zero. Similarly, after integrating over λ, the third term on the RHS of Eq. (D22) also vanishes. Therefore,

$$\int_0^1 d\lambda\,\mathrm{Tr}\Big[\,G_2\,\Big] = \mathrm{Tr}\Big[\,2\Delta_1\,i[\,C_x, C_y\,]\,\Big]\,. \tag{D24}$$

Note that the gauge-covariant Berry curvature defined in the gauge space $\widetilde{\Lambda}_{xy} = \Lambda_{xy} - i[\,C_x, C_y\,]$ has to vanish (Λ_{xy} defined in Eq. (D10)), because $\widetilde{\Lambda}_{xy}$ is the Berry curvature projected onto the unoccupied subspace, which is zero. Therefore, $\Lambda_{xy} = i[\,C_x, C_y\,]$. It can also be shown by explicitly

writing out the commutator of C_x and C_y:

$$\begin{aligned} i[\,C_x, C_y\,] &= i(-V^\dagger\,\partial_x V\,V^\dagger \partial_y V + V^\dagger\,\partial_y V\,V^\dagger \partial_x V) \\ &= i(V^\dagger\,V\,\partial_x V^\dagger \partial_y V - V^\dagger\,V\,\partial_y V^\dagger \partial_x V) \\ &= i\partial_x(V^\dagger\,\partial_y V) - i\partial_y(V^\dagger\,\partial_x V) \\ &= \Lambda_{xy}\ . \end{aligned} \tag{D25}$$

We have used the fact that $V\,V^\dagger = 1$ and $\partial_a(V\,V^\dagger) = 0$ in the above derivations. Therefore,

$$\int_0^1 d\lambda\,\mathrm{Tr}\Big[\,G_2\,\Big] = \mathrm{Tr}\Big[\,2\Delta_1\,\Lambda_{xy}\,\Big]\ . \tag{D26}$$

Combining Eqs. (D11), (D14), (D18) and (D26), we get

$$\begin{aligned} \theta_{\text{shift}} &= \frac{-1}{4\pi}\int dk_x dk_y \int_0^1 d\lambda\,\mathrm{Tr}\Big[\,G_1+G_2+G_3+G_4\,\Big] \\ &= \frac{-1}{4\pi}\int_{\mathcal{S}} dk_x dk_y\,\mathrm{Tr}\Big[\,\Delta_1\Omega'_{xy} + \Delta_1\Lambda_{xy}\,\Big] \\ &= \frac{-1}{4\pi}\int_{\mathcal{S}} dk_x dk_y\,\Big(\,2\pi\,(\Omega'_{xy})_{11} + 2\pi\,(\Lambda_{xy})_{11}\,\Big) \\ &= -\Big[\,\phi_1(\mathcal{C}) + \xi_1(\mathcal{C})\,\Big]/2\ . \end{aligned} \tag{D27}$$

This completes the derivation of Eq. (52), demonstrating that θ_{VL} is just the average of the two Berry phases $\phi_1(\mathcal{C})$ and $\xi_1(\mathcal{C})$ appearing above.

[1] M. Fiebig, Journal of Physics D: Applied Physics **38**, R123 (2005).
[2] A. Malashevich, I. Souza, S. Coh, and D. Vanderbilt, New Journal of Physics **12**, 053032 (2010).
[3] J. Íñiguez, Phys. Rev. Lett. **101**, 117201 (Sep 2008).
[4] E. Bousquet, N. A. Spaldin, and K. T. Delaney, Phys. Rev. Lett. **106**, 107202 (Mar 2011).
[5] A. Malashevich, S. Coh, I. Souza, and D. Vanderbilt, Phys. Rev. B **86**, 094430 (Sep 2012).
[6] M. Ye and D. Vanderbilt, Phys. Rev. B **89**, 064301 (Feb 2014).
[7] A. M. Essin, A. M. Turner, J. E. Moore, and D. Vanderbilt, Phys. Rev. B **81**, 205104 (May 2010).
[8] M. Z. Hasan and C. L. Kane, Rev. Mod. Phys. **82**, 3045 (Nov 2010).
[9] X.-L. Qi and S.-C. Zhang, Rev. Mod. Phys. **83**, 1057 (Oct 2011).
[10] T. L. Hughes, E. Prodan, and B. A. Bernevig, Phys. Rev. B **83**, 245132 (Jun 2011).
[11] A. M. Turner, Y. Zhang, R. S. K. Mong, and A. Vishwanath, Phys. Rev. B **85**, 165120 (Apr 2012).
[12] X. L. Qi, T. L. Hughes, and S. C. Zhang, Phys. Rev. B **78**, 195424 (2008).
[13] A. M. Essin, J. E. Moore, and D. Vanderbilt, Phys. Rev. Lett. **102**, 146805 (Apr 2009).
[14] F. Wilczek, Phys. Rev. Lett. **58**, 1799 (May 1987).
[15] D. Vanderbilt and R. King-Smith, Phys. Rev. B **48**, 4442 (1993).
[16] S. Coh, D. Vanderbilt, A. Malashevich, and I. Souza, Phys. Rev. B **83**, 085108 (Feb 2011).
[17] F. D. M. Haldane, Phys. Rev. Lett. **61**, 2015 (Oct 1988).
[18] A 3D QAH insulator is defined as a 3D insulator with the property that for at least one orientation of 2D $\mathbf{k}$ slices through the BZ, the Chern number of these slices is non-zero. Such a system is adiabatically connected to one made by stacking QAH layers in the third spatial dimension.
[19] L. Fu and C. L. Kane, Phys. Rev. B **74**, 195312 (2006).
[20] A. A. Soluyanov and D. Vanderbilt, Phys. Rev. B **85**, 115415 (Mar 2012).
[21] A similar simplification occurs in the multiband case if B is globally diagonal (i.e., at all $\mathbf{k}$), but this cannot normally be expected.
[22] In general the spin quantization axis can be chosen to be different for different $\mathcal{T}$-symmetric pairs.
[23] Let us consider the gauge-discontinuity plane as an isolated 2D BZ without worrying about its k_z value.
[24] L. Fu, C. L. Kane, and E. J. Mele, Phys. Rev. Lett. **98**, 106803 (2007).
[25] S.-Q. Shen, Phys. Rev. B **70**, 081311 (Aug 2004).
[26] H. Murakawa, M. Bahramy, M. Tokunaga, Y. Kohama, C. Bell, Y. Kaneko, N. Nagaosa, H. Hwang, and Y. Tokura, Science **342**, 1490 (2013).
[27] N. Marzari and D. Vanderbilt, Phys. Rev. B **56**, 12847 (1997).
[28] A. A. Soluyanov and D. Vanderbilt, Phys. Rev. B **83**, 235401 (JUN 2 2011), ISSN 1098-0121.
[29] R. Yu, X. L. Qi, A. Bernevig, Z. Fang, and X. Dai, Phys. Rev. B **84**, 075119 (Aug 2011).
[30] The convergence of θ using the conventional method is trapped into some local minimum when $\beta > 0.85\pi$. For example, when $\beta = 0.9\pi$, the converged value for θ with a $200\times200\times200$ mesh is 0.819, which is about 38.5% of the value obtained from our method.
[31] M. Taherinejad and D. Vanderbilt, Phys. Rev. Lett. **114**, 096401 (Mar 2015).