New Journal of Physics
The open-access journal for physics

Effective continuous model for surface states and thin films of three-dimensional topological insulators

Wen-Yu Shan, Hai-Zhou Lu and Shun-Qing Shen[1]
Department of Physics and Center of Theoretical and Computational Physics, University of Hong Kong, Hong Kong
E-mail: sshen@hkucc.hku.hk

New Journal of Physics **12** (2010) 043048 (23pp)
Received 19 January 2010
Published 28 April 2010
Online at http://www.njp.org/
doi:10.1088/1367-2630/12/4/043048

Abstract. Two-dimensional (2D) effective continuous models are derived for the surface states and thin films of a three-dimensional topological insulator (3DTI). Starting from an effective model for 3DTI based on first-principles calculations (Zhang *et al* 2009 *Nat. Phys.* **5** 438), we present solutions for both the surface states in a semi-infinite boundary condition and those in a thin film with finite thickness. The coupling between opposite topological surfaces and structure inversion asymmetry (SIA) gives rise to gapped Dirac hyperbolas with Rashba-like splittings in the energy spectrum. In addition, SIA leads to asymmetric distributions of wavefunctions for the surface states along the film growth direction, making some branches in the energy spectra much harder than others to probe by light. These features agree well with the recent angle-resolved photoemission spectra of Bi_2Se_3 films grown on SiC substrate (Zhang *et al* 2009 arXiv:0911.3706). More importantly, using the parameters fitted by experimental data, the result indicates that the thin film Bi_2Se_3 lies in the quantum spin Hall (QSH) region based on the calculation of the Chern number and Z_2 invariant. In addition, strong SIA always tends to destroy the QSH state.

[1] Author to whom any correspondence should be addressed.

Contents

1. Introduction

Topological insulators (TIs), which are band insulators with topologically protected edge or surface states, have attracted increasing attention recently [1][2]. A well-known TI paradigm is the quantum Hall effect, in which the cyclotron motion of electrons in a strong magnetic field gives rise to insulating bulk states but one-way conducting states propagating along edges of the system [2]. The idea was generalized to a graphene model with spin–orbit coupling, which exhibits the quantum spin Hall (QSH) state [3, 4]. Later, the realization of an existing QSH matter was predicted theoretically [5] and soon confirmed experimentally [6, 7] in two-dimensional (2D) HgTe/CdTe quantum wells. Furthermore, it was found that the QSH state can be induced even by disorders or impurities [8]–[10]. Meanwhile, the concept was also generalized for three-dimensional (3D) TIs, which are 3D band insulators surrounded by 2D conducting surface states with quantum spin texture [11]–[14]. Bi_xSb_{1-x}, an alloy with a complex structure of surface states, was first confirmed to be a three-dimensional topological insulator (3DTI) [15, 16]. Soon after that, it was verified by both experiments [17, 18] and first-principles calculations [19] that stoichiometric crystals Bi_2X_3 (X = Se, Te) are TIs with a well-defined single Dirac cone of surface states and extra large bandgaps comparable with

[2] For an introduction to TIs, see [1].

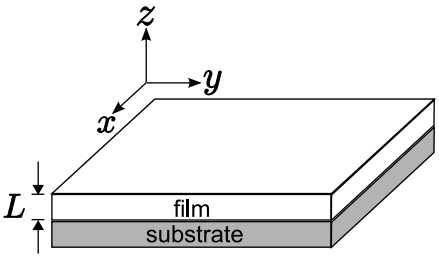

Figure 1. Schematic of a TI film grown on the substrate. The growth direction is defined as the z-axis. The thickness of the film is L.

room temperature. The Dirac fermions in the surface states of 3DTI obey the $2+1$ Dirac equations and reveal a lot of unconventional properties and possible applications, such as the topological magneto-electric effect [20] and Majorana fermions for fault-tolerant quantum computing [21]–[26].

Thanks to the state-of-the-art semiconductor technologies, low-dimensional structures of Bi_2X_3 can be routinely fabricated into ultrathin films [27, 28] and nanoribbons [29]. This has stimulated several theoretical works on the thin films of 3DTIs [30]–[32]. For further studies of the transport and optical properties of 3DTI films and their potential applications in spintronics and quantum information, it is desirable to establish an effective continuous model for thin films of TIs.

In this paper, we present an effective continuous model for the surface states and ultrathin film of TIs. Starting with a 3D effective low-energy model based on first-principles calculations [19], we first present the solutions for the surface states and the corresponding spectra for a semi-infinite boundary condition of gapless Dirac fermions and for the thin film of TIs. The finite size effect of spatial confinement in a thin film leads to a massive Dirac model that may exhibit the QSH effect. Within the same theoretical framework, a structure inversion asymmetry (SIA) term is further introduced in this work to account for the influence of substrate, providing a description of the Rashba-like energy spectra observed in the angle-resolved photoemission spectra (ARPES) in a recent experiment on Bi_2Se_3 films [28]. We derived the parameter conditions for the formation of the QSH effect in a thin film in the absence and presence of the SIA. By analyzing the fitting parameters with the help of the Chern number and Z_2 invariant, we identified the ultrathin films of Bi_2Se_3 in the QSH phase in the experiment.

The paper is organized as follows. In section 2, we introduce an anisotropic 3D Hamiltonian for 3DTI, which is a starting point of the present work. With this Hamiltonian, we present detailed solutions to the thin film in two different boundary conditions. In section 3, effective continuous models are established for the surface states and thin film of 3DTI. Within the framework of this effective continuous model, the SIA is taken into account and an effective Hamiltonian for SIA is derived in section 4. In section 5, we apply the model to newly fabricated thin film Bi_2Se_3 and demonstrate that thin films of Bi_2Se_3 are in the QSH regime. Finally, the conclusion is presented in section 6.

2. Model and general solutions for 3DTI

2.1. Model for 3DTI

As shown in figure 1, we will consider a thin film grown along the z-direction. The thickness of the film is L. We assume translational symmetry in the x–y plane so that the wave numbers

k_x and k_y are good quantum numbers. We start with the effective model proposed to describe the bulk states near the Γ point for bulk Bi_2Se_3 [19]. The states are mainly contributed by four hybridized states of Se and Bi p_z orbitals, denoted as $\{|p1_z^+,\uparrow\rangle, |p2_z^-,\uparrow\rangle, |p1_z^+,\downarrow\rangle, |p2_z^-,\downarrow\rangle\}$, where + (−) stands for even (odd) parity. The Hamiltonian is given by

$$H(\mathbf{k}) = \epsilon_0(\mathbf{k})I_{4\times4} + \begin{bmatrix} \mathcal{M}(\mathbf{k}) & -\mathrm{i}A_1\partial_z & 0 & A_2k_- \\ -\mathrm{i}A_1\partial_z & -\mathcal{M}(\mathbf{k}) & A_2k_- & 0 \\ 0 & A_2k_+ & \mathcal{M}(\mathbf{k}) & \mathrm{i}A_1\partial_z \\ A_2k_+ & 0 & \mathrm{i}A_1\partial_z & -\mathcal{M}(\mathbf{k}) \end{bmatrix}, \tag{1}$$

where $k_\pm = k_x \pm \mathrm{i}k_y$, $\epsilon_0(\mathbf{k}) = C - D_1\partial_z^2 + D_2k^2$, $\mathcal{M}(\mathbf{k}) = M + B_1\partial_z^2 - B_2k^2$, and $k^2 = k_x^2 + k_y^2$, with A_1, A_2, B_1, B_2, C, D_1, D_2 and M being the model parameters. This model has time-reversal symmetry and inversion symmetry. Although we start with a concrete model, the conclusion in this paper should be applicable to other TI films. We shall demonstrate that this model for bulk states can produce surface states with appropriate boundary conditions.

2.2. *General solutions of the surface states*

Following the method of Zhou *et al* [33], the general solution for either the bulk states or the surface states can be derived analytically. Despite the existence of time-reversal symmetry, the $A_2k_\pm$ term couples opposite spins in Hamiltonian (1), and one has to solve a 4×4 matrix, instead of the simplified 2×2 one in the 2D case [33]. By putting a four-component trial solution

$$\psi = \psi_\lambda \mathrm{e}^{\lambda z} \tag{2}$$

into the Schrödinger equation (E is the eigenvalue of energy)

$$H(k, -\mathrm{i}\partial_z)\psi = E\psi, \tag{3}$$

the secular equation

$$\det|H(k, -\mathrm{i}\lambda) - E| = 0 \tag{4}$$

gives four solutions of $\lambda(E)$, denoted as $\beta\lambda_\alpha(E)$, with $\alpha \in \{1, 2\}$, $\beta \in \{+, -\}$ and

$$\lambda_\alpha(E) = \left[-\frac{F}{2D_+D_-} + (-1)^{\alpha-1}\frac{\sqrt{R}}{2D_+D_-}\right]^{1/2}, \tag{5}$$

where for convenience we have defined

$$\begin{aligned} F &= A_1^2 + D_+(E - L_1) + D_-(E - L_2), \\ R &= F^2 - 4D_+D_-[(E - L_1)(E - L_2) - A_2^2k_+k_-], \\ D_\pm &= D_1 \pm B_1, \\ L_1 &= C + M + (D_2 - B_2)k^2, \\ L_2 &= C - M + (D_2 + B_2)k^2. \end{aligned} \tag{6}$$

Because of double degeneracy, each of the four $\beta\lambda_\alpha(E)$ corresponds to two linearly independent four-component vectors, found to be

$$\psi_{\alpha\beta 1} = \begin{bmatrix} D_+\lambda_\alpha^2 - L_2 + E \\ -\mathrm{i}A_1(\beta\lambda_\alpha) \\ 0 \\ A_2 k_+ \end{bmatrix}, \tag{7}$$

$$\psi_{\alpha\beta 2} = \begin{bmatrix} A_2 k_- \\ 0 \\ \mathrm{i}A_1(\beta\lambda_\alpha) \\ D_-\lambda_\alpha^2 - L_1 + E \end{bmatrix}. \tag{8}$$

The general solution should be a linear combination of these eight functions

$$\Psi(E, k, z) = \sum_{\alpha=1,2} \sum_{\beta=\pm} \sum_{\gamma=1,2} C_{\alpha\beta\gamma} \psi_{\alpha\beta\gamma} \mathrm{e}^{\beta\lambda_\alpha z} \tag{9}$$

with the superposition coefficients $C_{\alpha\beta\gamma}$ to be determined by boundary conditions. In the following, we will consider two different boundary conditions: one is semi-infinite focusing on only one surface at $z = 0$; the other includes two opposite surfaces at $z = \pm L/2$. In both cases, we assume open boundary conditions ($\Psi = 0$) for the surface states at the surfaces.

2.3. Solutions for the surface states with semi-infinite boundary conditions

The surface states have a finite distribution near the boundary. For a film thick enough that the states at opposite surfaces barely couple to each other, we can focus on just one surface. Without loss of generality, we study a system from $z = 0$ to $+\infty$. The boundary condition is given as

$$\Psi(z = 0) = 0 \quad \text{and} \quad \Psi(z \to +\infty) = 0. \tag{10}$$

The condition of $\Psi(z \to +\infty) = 0$ requires that Ψ contains only the four terms in which β is negative and that the real part of λ_α be positive.

Applying the boundary conditions of equation (10) to the general solution of equation (9), the secular equation of the nontrivial solution to the coefficients $C_{\alpha\beta\gamma}$ leads to

$$(\lambda_1 + \lambda_2)^2 = -\frac{A_1^2}{D_+ D_-}, \tag{11}$$

which along with equation (5) gives the dispersion of the surface states

$$E_\pm = C + \frac{D_1 M}{B_1} \pm A_2 \sqrt{1 - \left(\frac{D_1}{B_1}\right)^2}\, k + \left(D_2 - \frac{B_2 D_1}{B_1}\right) k^2. \tag{12}$$

Near the Γ point, the dispersion shows a massless Dirac cone in k space, with the Fermi velocity $v_\mathrm{F} = (A_2/\hbar)\sqrt{1 - (D_1/B_1)^2}$, instead of plain $A_2/\hbar$ as in [19].

The wavefunctions for $E_\pm$ are found to be

$$\Psi_+ = C_+^0 \begin{bmatrix} \frac{\mathrm{i}}{2}\sqrt{\frac{D_+}{B_1}} \\ -\frac{1}{2}\sqrt{\frac{-D_-}{B_1}} \\ -\frac{1}{2}\sqrt{\frac{D_+}{B_1}}\mathrm{e}^{\mathrm{i}\varphi} \\ \frac{\mathrm{i}}{2}\sqrt{\frac{-D_-}{B_1}}\mathrm{e}^{\mathrm{i}\varphi} \end{bmatrix} (\mathrm{e}^{-\lambda_2^+ z} - \mathrm{e}^{-\lambda_1^+ z}),$$

$$\Psi_- = C_-^0 \begin{bmatrix} -\frac{\mathrm{i}}{2}\sqrt{\frac{D_+}{B_1}} \\ \frac{1}{2}\sqrt{\frac{-D_-}{B_1}} \\ -\frac{1}{2}\sqrt{\frac{D_+}{B_1}}\mathrm{e}^{\mathrm{i}\varphi} \\ \frac{\mathrm{i}}{2}\sqrt{\frac{-D_-}{B_1}}\mathrm{e}^{\mathrm{i}\varphi} \end{bmatrix} (\mathrm{e}^{-\lambda_2^- z} - \mathrm{e}^{-\lambda_1^- z}), \tag{13}$$

where $\lambda_\alpha^\pm$ are short for $\lambda_\alpha(E = E_\pm)$ according to equation (5), $\tan\varphi \equiv k_y/k_x$, and $C_\pm^0$ are the normalization factors. The properties of the solution to λ_α determine the spatial distribution of the wavefunctions. Generally speaking, the edge states exist if λ_1 and λ_2 are both real or complex conjugate partners. For either case, there should be inequality relations

$$\begin{aligned} &\frac{M}{B_1} > 0, \\ &D_+ D_- < 0. \end{aligned} \tag{14}$$

The edge states distribute mostly near the surface ($z = 0$), with the scale of the decay length about $\lambda_{1,2}^{-1}$ for real $\lambda_{1,2}$ or $[\mathrm{Re}(\lambda_{1,2})]^{-1}$ for complex $\lambda_{1,2}$. In the former case, the wavefunctions decay exponentially and monotonously away from the surface (not from $z = 0$), whereas in the latter case, the decay is accompanied by a periodical oscillation, which can be easily seen from the wavefunctions in equation (13). In addition, there exist complex solutions to λ_α when

$$\frac{A_1^2}{-D_+ D_-} < \frac{4M}{B_1}. \tag{15}$$

2.4. Solutions for finite-thickness boundary conditions

When the thickness of the film is comparable with the characteristic length $1/\lambda_{1,2}$ of the surface states, there is coupling between the states on opposite surfaces. One has to consider the boundary conditions at both surfaces simultaneously. Without loss of generality, we will consider that the top surface is located at $z = L/2$ and the bottom surface at $-L/2$. The boundary conditions are given as

$$\Psi\left(z = \pm\frac{L}{2}\right) = 0. \tag{16}$$

In this case, the general solution consists of all eight linearly independent functions. Applying the boundary conditions in equation (16) to the general solution of equation (9), the

Table 1. Two sets of parameters for the 3D Dirac model. The first row is extracted from our effective model parameters for 4 QL Bi_2Se_3 film in table 2, and the second row is adopted from first-principles calculations [19].

M (eV)	A_1 (eV Å)	A_2 (eV Å)	B_1 (eV Å^2)	B_2 (eV Å^2)	C (eV)	D_1 (eV Å^2)	D_2 (eV Å^2)
0.28	3.3	4.1	1.5	−54.1	−0.0068	1.2	−30.1
0.28	2.2	4.1	10	56.6	−0.0068	1.3	19.6

secular equation of the nontrivial solution to the superposition coefficients $C_{\alpha\beta\gamma}$ leads to a transcendental equation

$$\frac{\frac{D_+D_-}{A_1^2}(\lambda_1^2-\lambda_2^2)^2+(\lambda_1^2+\lambda_2^2)}{\lambda_1\lambda_2}=\frac{\tanh\frac{\lambda_2 L}{2}}{\tanh\frac{\lambda_1 L}{2}}+\frac{\tanh\frac{\lambda_1 L}{2}}{\tanh\frac{\lambda_2 L}{2}}. \tag{17}$$

In a large L limit, $\tanh(\lambda_\alpha L/2)$ reduces to 1; then equation (17) can recover the result in equation (11). With the help of equation (5), equation (17) can be used to identify the energy spectra and the values of λ_α numerically.

Due to the finite size effect [33], the coupling between the states at the top and bottom surfaces will open an energy gap. We define the gap as $\Delta = E_+ - E_-$ at the Γ point, where E_+ and E_- are two solutions of equation (17). For $\lambda_\alpha L \gg 1$ and $\lambda_2 \gg \lambda_1$ (L can be finite), the approximate expression for Δ can be found. If λ_α is real, the gap can be approximated by

$$\Delta \simeq \frac{4|A_1 D_+ D_- M|}{\sqrt{B_1^3(A_1^2 B_1+4D_+D_-M)}}\mathrm{e}^{-\lambda_1 L}, \tag{18}$$

which decays exponentially as a function of L. Figure 2(a) shows the gap as a function of thickness, in which a set of model parameters used to fit the ARPES of 4 QL Bi_2Se_3 thin film is employed, as listed in the first row of table 1.

For some other materials there may exist complex $\lambda_1 = \lambda_2^*$ and we can define $\lambda_1 = a - \mathrm{i}b$ and $\lambda_2 = a + \mathrm{i}b$, where $a > 0$ and $b > 0$ according to equation (5). In this case, the gap is found to be

$$\Delta \simeq \frac{8|A_1 D_+ D_- M|}{\sqrt{-B_1^3(A_1^2 B_1+4D_+D_-M)}}\mathrm{e}^{-aL}\sin(bL) \tag{19}$$

with

$$a \backsimeq \frac{A_1}{2\sqrt{-D_+D_-}}, \tag{20}$$

$$b \backsimeq \sqrt{\frac{M}{B_1}+\frac{A_1^2}{4D_+D_-}}. \tag{21}$$

According to this result, the oscillation period of the gap π/b becomes $\pi\sqrt{B_1/M}$ when $A_1 = 0$, in accordance with the result obtained by Liu *et al* [32]. Figure 2(b) shows the gap oscillation by using the model parameters listed in the second entry of table 1. In addition, the sine function

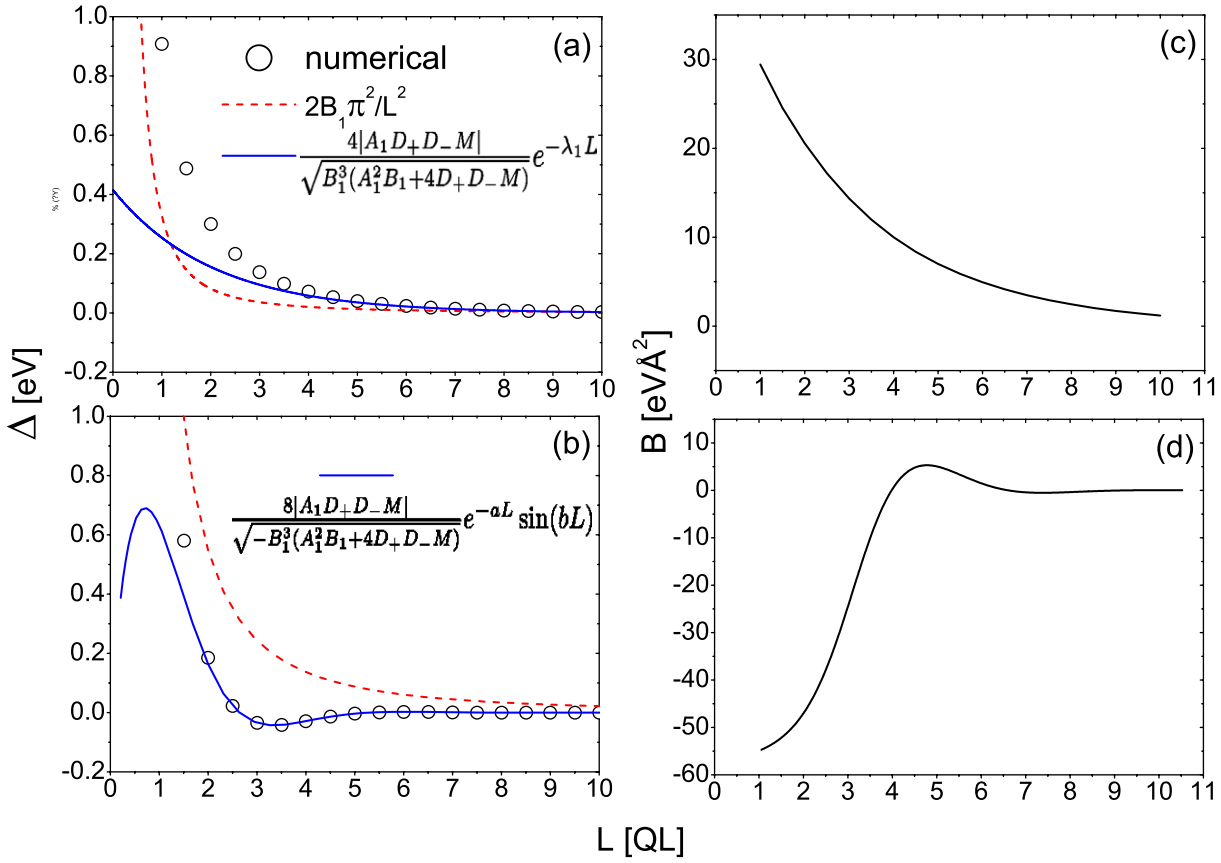

Figure 2. (a, b) The energy gap $\Delta \equiv E_+ - E_-$ and (c, d) the model parameter B (defined in equation (42)) as functions of the film thickness L. Circles correspond to the numerical results of the transcendental equations (30) and (31). Solid and dashed lines correspond to the approximate formulae to Δ when L is finite (equations (18) and (19)) or very small (equation (50)), respectively. All the parameters are adopted (a) by fitting experimental results of 4 QL Bi_2Se_3 and (b) from the numerical fitting for the first-principles calculation of Bi_2Se_3 [19], as listed in table 1.

implies that Δ may be negative. Later we will see that the sign of Δ can be found by solving E_+^0 and E_-^0 from equations (30) and (31), respectively.

3. Effective continuous models

The solutions of the surface states and thin film of 3DTI can be applied to calculate physical properties explicitly. For instance, we can see whether the ground state of a thin film exhibits QSHE or not by calculating the Chern number or Z_2 invariant. It is also desirable to establish an effective continuous model to explore the properties of these surface states, especially when other interactions have to be taken into account. For this purpose, in this section we derive effective low-energy and continuous models for the surface states and thin film of 3DTI.

Due to the low-energy long-wavelength nature of the Dirac cone of the surface electrons, we can use the solutions of the surface states at the Γ point as a basis to expand the Hamiltonian $H(k)$ in equation (1), which will be valid when the energy is limited within the bandgap between the conduction and valence bands. This is equivalent to a truncation approximation as we exclude the solutions for the bulk states in the basis. In this approach, the Hamiltonian in equation (1) can be expressed as

$$H(\vec{k}) = H_0(k=0) + \Delta H, \tag{22}$$

where

$$H_0 = \begin{bmatrix} h(A_1) & 0 \\ 0 & h(-A_1) \end{bmatrix}, \tag{23}$$

with

$$h(A_1) = \begin{bmatrix} -D_-\partial_z^2 + C + M & -\mathrm{i}A_1\partial_z \\ -iA_1\partial_z & -D_+\partial_z^2 + C - M \end{bmatrix}, \tag{24}$$

and

$$\Delta H = \begin{bmatrix} D_2k^2 - B_2k^2\sigma_z & A_2k_-\sigma_x \\ A_2k_+\sigma_x & D_2k^2 - B_2k^2\sigma_z \end{bmatrix}. \tag{25}$$

The first term can be solved exactly, and the last term describes the behaviors of electrons near the Γ point.

3.1. Basis states at the Γ *point*

H_0 in equation (23) is block-diagonal. Its solution can be found by solving each block separately, i.e. $h(A_1)\Psi_\uparrow = E\Psi_\uparrow$ and $h(-A_1)\Psi_\downarrow = E\Psi_\downarrow$. Because the lower block is the 'time' reversal of the upper block, the solutions satisfy $\Psi_\downarrow(z) = \Theta\Psi_\uparrow(z)$, where $\Theta = -\mathrm{i}\,\sigma_y\mathcal{K}$ is the time-reversal operator, with σ_y being the y-component of the Pauli matrices and $\mathcal{K}$ the complex conjugation operation. Equivalently, we can replace A_1 by $-A_1$ in all the results for the upper block, to obtain those for the lower block. Therefore, we only need to solve $h(A_1)$. Following the same approach as that in section 2, we put a two-component trial solution

$$\psi^\uparrow = \psi_\lambda^\uparrow \mathrm{e}^{\lambda z} \tag{26}$$

into

$$h(A_1, -\mathrm{i}\partial_z)\psi^\uparrow = E\psi^\uparrow, \tag{27}$$

and the secular equation for a nontrivial solution yields four roots of $\lambda(E)$, denoted as $\beta\lambda_\alpha$, with $\beta \in \{+,-\}$ and $a \in \{1,2\}$. Note that here λ_α is short for $\lambda_\alpha(k=0)$ in equation (5). Each $\beta\lambda_\alpha$ corresponds to a two-component vector

$$\psi_{\alpha\beta}^\uparrow = \begin{bmatrix} D_+\lambda_\alpha^2 - l_2 + E \\ -\mathrm{i}A_1(\beta\lambda_\alpha) \end{bmatrix}. \tag{28}$$

The general solution is a linear combination of the four linearly independent two-component vectors

$$\Psi_\uparrow = \sum_{\alpha=1,2}\sum_{\beta=+,-} C_{\alpha\beta}\psi_{\alpha\beta}^\uparrow \mathrm{e}^{\beta\lambda_\alpha z}. \tag{29}$$

Applying the boundary conditions (16) to this general solution, we obtain two transcendental equations:

$$\frac{(C-M-E-D_+\lambda_1^2)\lambda_2}{(C-M-E-D_+\lambda_2^2)\lambda_1}=\frac{\tanh(\lambda_2 L/2)}{\tanh(\lambda_1 L/2)}, \tag{30}$$

and

$$\frac{(C-M-E-D_+\lambda_1^2)\lambda_2}{(C-M-E-D_+\lambda_2^2)\lambda_1}=\frac{\tanh(\lambda_1 L/2)}{\tanh(\lambda_2 L/2)}. \tag{31}$$

The solutions to equations (30) and (31) give two energies at the Γ point, designated as $E_+^0 \equiv E_+(k=0)$ and $E_-^0 \equiv E_-(k=0)$, respectively. The eigen-wavefunctions for E_+^0 and E_-^0 are, respectively,

$$\varphi(A_1) \equiv \Psi_\uparrow^+ = C_+ \begin{bmatrix} -D_+\eta_1^+ f_-^+ \\ \mathrm{i}A_1 f_+^+ \end{bmatrix}, \tag{32}$$

$$\chi(A_1) \equiv \Psi_\uparrow^- = C_- \begin{bmatrix} -D_+\eta_2^- f_+^- \\ \mathrm{i}A_1 f_-^- \end{bmatrix}, \tag{33}$$

where $C_\pm$ are the normalization factors. The superscripts of $f_\pm^\pm$ and $\eta_{1,2}^\pm$ stand for $E_\pm^0$ and the subscripts of $f_\pm^\pm$ for parity, respectively. The expressions for $f_\pm^\pm$ and $\eta_{1,2}^\pm$ are given by

$$f_+^\pm(z) = \left.\frac{\cosh(\lambda_1 z)}{\cosh\left(\frac{\lambda_1 L}{2}\right)} - \frac{\cosh(\lambda_2 z)}{\cosh\left(\frac{\lambda_2 L}{2}\right)}\right|_{E=E_\pm^0}, \tag{34}$$

$$f_-^\pm(z) = \left.\frac{\sinh(\lambda_1 z)}{\sinh\left(\frac{\lambda_1 L}{2}\right)} - \frac{\sinh(\lambda_2 z)}{\sinh\left(\frac{\lambda_2 L}{2}\right)}\right|_{E=E_\pm^0}, \tag{35}$$

$$\eta_1^\pm = \left.\frac{(\lambda_1)^2-(\lambda_2)^2}{\lambda_1 \coth\left(\frac{\lambda_1 L}{2}\right) - \lambda_2 \coth\left(\frac{\lambda_2 L}{2}\right)}\right|_{E=E_\pm^0}, \tag{36}$$

$$\eta_2^\pm = \left.\frac{(\lambda_1)^2-(\lambda_2)^2}{\lambda_1 \tanh\left(\frac{\lambda_1 L}{2}\right) - \lambda_2 \tanh\left(\frac{\lambda_2 L}{2}\right)}\right|_{E=E_\pm^0}. \tag{37}$$

The energy spectra and wavefunctions of the lower block $h(-A_1)$ of H_0 can be obtained directly by replacing A_1 by $-A_1$. Based on the above discussions, the four eigenstates of H_0 can be given by

$$\begin{aligned} \Phi_1 &= \begin{bmatrix} \varphi(A_1) \\ 0 \end{bmatrix}, \quad \Phi_2 = \begin{bmatrix} \chi(A_1) \\ 0 \end{bmatrix}, \\ \Phi_3 &= \begin{bmatrix} 0 \\ \varphi(-A_1) \end{bmatrix}, \quad \Phi_4 = \begin{bmatrix} 0 \\ \chi(-A_1) \end{bmatrix}, \end{aligned} \tag{38}$$

with $\Phi_1 \to \Phi_3$ and $\Phi_2 \to \Phi_4$ under the time-reversal operation. We should emphasize that these four solutions are for the surface states, and the solutions for the bulk states are not presented here. We use the four states as the basis states, and other states are discarded (except that in figure 3 where four extra bulk states are also included by the same approach), because of a large gap between the valence and conduction bands.

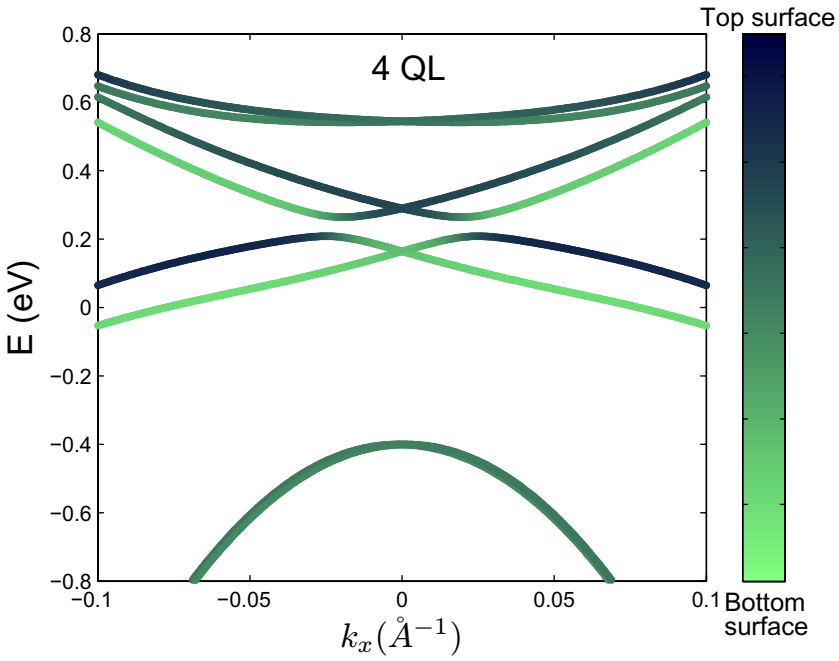

Figure 3. Energy spectra of surface states (four branches in the middle) and several branches of bulk states (those at the top and bottom) for a film with the thickness of 4 QL in the presence of SIA. The color of the lines corresponds to the spatial distribution of the wavefunctions in the z-direction. Dark blue (light green) shows that the wavefunctions are mainly distributed on the side of the top (substrate) surface. The model parameters are listed in the first row of table 1.

3.2. Effective model for 3DTI films

With the help of the four states, equation (38), at the Γ point, we can expand Hamiltonian equation (1) to obtain a new effective Hamiltonian

$$H_{\text{eff}} \equiv \int_{-L/2}^{L/2} \mathrm{d}z[\Phi_1, \Phi_4, \Phi_2, \Phi_3]^{\dagger} H[\Phi_1, \Phi_4, \Phi_2, \Phi_3], \tag{39}$$

where for convenience we organize the sequence of the basis states following $\{\Phi_1, \Phi_4, \Phi_2, \Phi_3\}$. Under the reorganized basis, the effective Hamiltonian is found to be

$$H_{\text{eff}} = \begin{bmatrix} h_+ & 0 \\ 0 & h_- \end{bmatrix}, \tag{40}$$

with

$$\begin{aligned} h_+ &= E_0 - Dk^2 + \begin{bmatrix} \frac{\Delta}{2} - Bk^2 & \tilde{A}_2 k_- \\ \tilde{A}_2^* k_+ & -\frac{\Delta}{2} + Bk^2 \end{bmatrix}, \\ h_- &= E_0 - Dk^2 + \begin{bmatrix} -\frac{\Delta}{2} + Bk^2 & -\tilde{A}_2^* k_- \\ -\tilde{A}_2 k_+ & \frac{\Delta}{2} - Bk^2 \end{bmatrix}, \end{aligned} \tag{41}$$

and

$$\begin{aligned} B &= \frac{\tilde{B}_1 - \tilde{B}_2}{2}, \quad D = \frac{\tilde{B}_1 + \tilde{B}_2}{2} - D_2, \\ E_0 &= (E_+ + E_-)/2, \quad \Delta = E_+ - E_-, \\ \tilde{B}_1 &= B_2 \langle \varphi(A_1) | \sigma_z | \varphi(A_1) \rangle, \\ \tilde{B}_2 &= B_2 \langle \chi(A_1) | \sigma_z | \chi(A_1) \rangle, \\ \tilde{A}_2 &= A_2 \langle \varphi(A_1) | \sigma_x | \chi(-A_1) \rangle. \end{aligned} \tag{42}$$

We find that $\tilde{A}_2$ here can be either real or purely imaginary (see the appendix for details), classifying the model into two cases:

Case I is for a real $\tilde{A}_2 \equiv \hbar v_F$, and the effective Hamiltonian is further written as

$$h_{\tau_z} = E_0 - Dk^2 + \hbar v_F \tau_z \vec{\sigma} \cdot \vec{k} + \tau_z \sigma_z \left(\frac{\Delta}{2} - Bk^2 \right) \quad \text{(case I)}, \tag{43}$$

and case II is for a purely imaginary $\tilde{A}_2 \equiv \mathrm{i}\hbar v_F$,

$$h_{\tau_z} = E_0 - Dk^2 + \hbar v_F (\vec{\sigma} \times \vec{k})_z + \tau_z \sigma_z \left(\frac{\Delta}{2} - Bk^2 \right) \quad \text{(case II)} \tag{44}$$

where $\tau_z = \pm 1$ corresponds to the upper (lower) 2×2 block in equation (40), v_F is the defined Fermi velocity and $\vec{\sigma}$ and $\vec{k}$ here refer only to the components in the x–y plane. In fact, these two effective Hamiltonians can consist of the invariants of the irreducible representation $D_{1/2}$ of SU(2) group [34].

Equation (41) can also be expressed in terms of the Pauli matrices

$$h_{\tau_z} = E_0 - Dk^2 + \mathbf{d} \cdot \sigma, \tag{45}$$

with

$$\mathbf{d} = \begin{cases} \tau_z \left(\hbar v_F k_x, \ \hbar v_F k_y, \ \dfrac{\Delta}{2} - Bk^2 \right) & \text{(case I)}, \\ \left(\hbar v_F k_y, \ -\hbar v_F k_x, \ \tau_z \left(\dfrac{\Delta}{2} - Bk^2 \right) \right) & \text{(case II)}. \end{cases} \tag{46}$$

The $\mathbf{d}(k)$ vectors in case I and case II, respectively, correspond to Dresselhaus- and Rashba-like textures. Note that case I is essentially the effective 4×4 model for the CdTe/HgTe quantum wells [5]. However, we find that case I only occurs for quite a small range of thickness. For most thicknesses of interest, $\tilde{A}_2$ is purely imaginary. Therefore, we only focus on case II in the following discussions. So far, we have reduced the anisotropic 3D Dirac model into a generalized effective model for 2D thin films, under the freestanding open boundary conditions.

3.3. Effective continuous model for surface states

Despite the simple explicit form, the parameters in Hamiltonian (40) need to be determined numerically. Before that, we can take two limits to see their behaviors. The first limit is $\lambda_\alpha L \gg 1$, for $\alpha = 1, 2$. In this case, $\tanh(\lambda_\alpha L/2) \simeq 1$, and both equations (30) and (31) reduce to

$$(C - M - E - D_+ \lambda_1^2)\lambda_2 = (C - M - E - D_+ \lambda_2^2)\lambda_1. \tag{47}$$

Solving this equation, we have an effective continuous model for the surface states (ss) of the 3DITI as

$$H_{\rm ss} = C + \frac{D_1 M}{B_1} + \left(D_2 - B_2\frac{D_1}{B_1}\right)k^2 + A_2\sqrt{1-\left(\frac{D_1}{B_1}\right)^2}(\sigma_x k_y - \sigma_y k_x), \tag{48}$$

which has the same dispersion as equation (12) and the same Fermi velocity $v_{\rm F} = \frac{A_2}{\hbar}\sqrt{1-(D_1/B_1)^2}$ as for the semi-infinite boundary conditions. In an isotropic case, $D_1 = D_2$ and $B_1 = B_2$, the quadratic term disappears and we have a linear dispersion for the Dirac cone. Finally, it is noticed that the models for the surface states at the top and bottom surfaces have the same form assuming $\lambda_\alpha L \gg 1$. We will see that these results work well even for films down to 5 QL of atoms in thickness (1 QL is about 1 nm).

3.4. The ultrathin limit

Another opposite limit is $L \to 0$, which is a little bit complicated since $\lambda_\alpha L$ does not approach zero when L is very small. In equation (30), the left side has an order of L^2 when $L \to 0$, so $\tanh(\lambda_1 L/2)$ must have the order of L^{-2}, which means

$$\tanh\left(\frac{\lambda_1 L}{2}\right) = 0 \Rightarrow \lambda_1 = i\frac{\pi}{L}. \tag{49}$$

Combining this result with equation (5), the model becomes

$$h_{\tau_z} = \frac{D_1\pi^2}{L^2} + D_2 k^2 + A_2(\vec{\sigma}\times\vec{k})_z + \tau_z\left(\frac{B_1\pi^2}{L^2} + B_2 k^2\right)\sigma_z. \tag{50}$$

It is found that a finite energy gap opens at $k = 0$, i.e. $\Delta = 2B_1\pi^2/L^2$ as shown in figure 2. Note that this result in the $L \to 0$ limit even provides a rough estimate of the gap for most thicknesses. Besides, the continuum limit generally assumed in this work also works well even for several QLs.

4. Structure inversion asymmetry

4.1. SIA

A recent experiment [28] revealed that the substrate on which the film is grown influences dramatically the electronic structure inside the film. Because the top surface of the film is usually exposed to the vacuum and the bottom surface is attached to a substrate, the inversion symmetry does not hold along the z-direction, leading to the Rashba-like energy spectra for the gapped surface states. In this case, an extra term that describes the SIA needs to be taken into account in the effective model.

We use the same method as that in section 3 to include the SIA term. Without loss of generality, we add a potential energy $V(z)$ to the Hamiltonian. Generally speaking, $V(z)$ can be expressed as $V(z) = V_{\rm s}(z) + V_{\rm a}(z)$, in which $V_{\rm s}(z) = V_{\rm s}(-z)$ and $V_{\rm a}(z) = -V_{\rm a}(-z)$. The symmetric term $V_{\rm s}$ could contribute to the mass term Δ in the effective model, which may lead to energy splitting of the Dirac cone at the Γ point. We do not discuss this in detail in this paper. Here, we focus on the case of the antisymmetric term, $V(z) = V_{\rm a}(z)$, which breaks the top–bottom inversion symmetry in the Hamiltonian. A detailed analysis demonstrates that

$V_a(z)$ couples Φ_1 (Φ_3) to Φ_2 (Φ_4), which can be readily seen according to their spin and parity natures. The modified effective Hamiltonian in the presence of $V(z)$ becomes

$$H_{\rm eff}^{\rm SIA} = H_{\rm eff} + \begin{bmatrix} 0 & 0 & \tilde{V} & 0 \\ 0 & 0 & 0 & \tilde{V}^* \\ \tilde{V}^* & 0 & 0 & 0 \\ 0 & \tilde{V} & 0 & 0 \end{bmatrix}, \tag{51}$$

where

$$\tilde{V} = \int_{-L/2}^{L/2} {\rm d}z \langle \varphi(A_1)|V_a(z)|\chi(A_1)\rangle. \tag{52}$$

Comparing this definition with that of $\tilde{A}_2$ in equation (42), we find that $\tilde{V}$ can be either real or purely imaginary. In the case of a purely imaginary (case II) $\tilde{A}_2$, $\tilde{V}$ must be real (see the appendix), and the effective Hamiltonian with SIA can be written as

$$H_{\rm eff}^{\rm SIA} = \begin{bmatrix} h_+(k) & \tilde{V}\sigma_0 \\ \tilde{V}\sigma_0 & h_-(k) \end{bmatrix}. \tag{53}$$

In the case of a real $\tilde{A}_2$, $\tilde{V}$ must be purely imaginary, and the effective Hamiltonian with SIA then has the form

$$H_{\rm eff}^{\rm SIA} = \begin{bmatrix} h_+(k) & \tilde{V}\sigma_z \\ -\tilde{V}\sigma_z & h_-(k) \end{bmatrix}. \tag{54}$$

Without the SIA term, the effective Hamiltonian (44) gives the energy spectra of the gapped surface states as

$$E_\pm = E_0 - Dk^2 \pm \sqrt{\left(\frac{\Delta}{2} - Bk^2\right)^2 + (\hbar v_{\rm F})^2 k^2}, \tag{55}$$

where the + (−) sign stands for the conduction (valence) band, each of which has double spin degeneracy due to time-reversal symmetry. When the SIA term is included, the Hamiltonian (51) gives

$$E_{1,\pm} = E_0 - Dk^2 \pm \sqrt{\left(\frac{\Delta}{2} - Bk^2\right)^2 + (|\tilde{V}| + \hbar v_{\rm F} k)^2},$$

$$E_{2,\pm} = E_0 - Dk^2 \pm \sqrt{\left(\frac{\Delta}{2} - Bk^2\right)^2 + (|\tilde{V}| - \hbar v_{\rm F} k)^2}, \tag{56}$$

where the extra index 1 (2) stands for the inner (outer) branches of the conduction or valence bands. The energy spectra in the presence of $\tilde{V}$ is shown in figure 3. Each spin-degenerate dispersion in equation (55) shifts away from the other along the k axis. Both the conduction and valence bands show Rashba-like splitting. An intuitive understanding of the energy spectra in figure 3 can be given with the help of figure 4. The part on the left is for a thicker freestanding symmetric TI film, and it has a single gapless Dirac cone on each of its two surfaces, with solid and dashed lines for the top and bottom surfaces, respectively. The two Dirac cones are degenerate. The top of figure 4 indicates that the inter-surface coupling across an ultrathin film

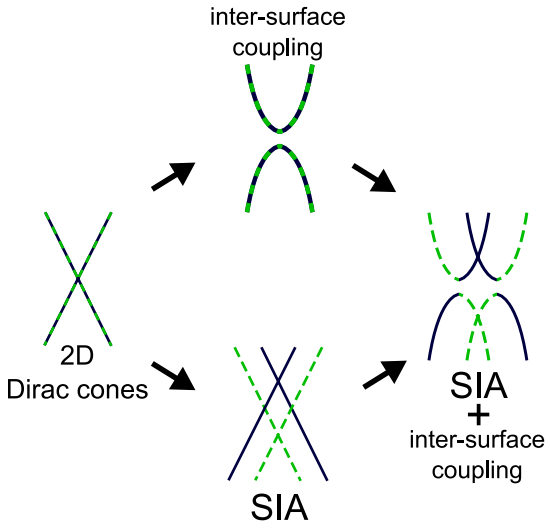

Figure 4. The evolution of the doubly degenerate gapless Dirac cones for the 2D surface states, in the presence of both inter-surface coupling and SIA, into gapped hyperbolas that also split in the k-direction. The blue solid and green dashed lines correspond to the states residing near the top and bottom surfaces, respectively.

will turn the Dirac cones into gapped Dirac hyperbolas. At the bottom of figure 4, SIA lifts the Dirac cone at the top surface while lowering the Dirac cone at the bottom surface. The potential difference at the top and bottom surfaces removes the degeneracy of the Dirac cones. On the right of figure 4, the coexistence of both inter-surface coupling and SIA leads to two gapped Dirac hyperbolas which also split in the k-direction, as shown in figure 3.

4.2. Location of the surface states

The location of the surface states can be revealed by evaluating the expectation of position z of these states. The spatial distributions along the z-direction of a state ψ_α can be evaluated by the expectation of position in the z-direction $\langle z\rangle$,

$$\langle z\rangle_\alpha = \int_{-L/2}^{L/2} z|\psi_\alpha|^2\,\mathrm{d}z. \tag{57}$$

By this definition, $\langle z\rangle_\alpha \in [-L/2, L/2]$ and $\langle z\rangle_\alpha$ becomes 0 for a symmetric spatial distribution.

With the SIA, the eigen-wavefunctions are found to be

$$\psi_{1\pm} = \frac{1}{2\sqrt{E_\pm^{\mathrm{in}}(E_\pm^{\mathrm{in}}+t)k}} \begin{bmatrix} \mathrm{i}(t+E_\pm^{\mathrm{in}})k \\ (|\tilde{V}|+\hbar v_{\mathrm{F}}k)k_+ \\ \mathrm{i}(|\tilde{V}|+\hbar v_{\mathrm{F}}k)k \\ (t+E_\pm^{\mathrm{in}})k_+ \end{bmatrix}, \tag{58}$$

with $E_{\pm}^{\rm in} = E_{1\pm} - E_0 + Dk^2$, $t = \frac{\Delta}{2} - Bk^2$ and

$$\psi_{2\pm} = \frac{1}{2\sqrt{E_{\pm}^{\rm out}(E_{\pm}^{\rm out}+t)}k} \begin{bmatrix} -\mathrm{i}(t+E_{\pm}^{\rm out})k \\ (|\tilde{V}| - \hbar v_{\rm F} k)k_+ \\ \mathrm{i}(-|\tilde{V}| + \hbar v_{\rm F} k)k \\ (t+E_{\pm}^{\rm out})k_+ \end{bmatrix} \tag{59}$$

with $E_{\pm}^{\rm out} = E_{2\pm} - E_0 + Dk^2$. Figure 3 demonstrates $\langle z \rangle$ by the brightness of lines, with dark blue for $\langle z \rangle = L/2$ (the top surface) and light green for $\langle z \rangle = -L/2$ (the substrate or bottom surface).

For a thin film of 4 QL, $L = 3.8$ nm, it is found that the two surface states are well separated and dominantly distributed near the two surfaces. The averaged $\langle z \rangle \simeq \pm L/3$, which is about 2/3 of a QL ($\approx L/6$) deviating from the surface. In this case, the top and bottom surface states are well defined even without the SIA ($\tilde{V} = 0$). The average value remains almost unchanged in a large range of k. However, at the crossing point of the spectra of the top and bottom surface states, the averaged $\langle z \rangle$ changes from $+L/3$ to 0, and then goes to the value of $-L/3$. This demonstrates that the finite thickness makes the two states couple with each other as their wavefunctions along the z-direction have a finite overlap. As a result, the two states open an energy gap as in the case of edge states in the QSH system [33]. The value of the gap is a function of L as shown in figures 2(a) and (b). Near this region, $\langle z \rangle$ varies from $\langle z \rangle \simeq L/3$ to $-L/3$, and becomes zero exactly when two states are mixed completely. For a large L, we find that the averaged distance of the surface states deviating from the surface remains about 1 QL.

Simply speaking, the states close to the top surface are easier to probe by light than those close to the bottom surface. This provides a hint to understand why there are branches in energy spectra with much more faint ARPES signals [28].

5. Bi_2Se_3 thin films and QSH states

In this section, we will investigate the realization of the QSH effect in thin films and apply the effective model to Bi_2Se_3 thin films. When the system does not break the inversion symmetry, the effective Hamiltonian is block-diagonalized by $\tau_z = \pm 1$. This is in good agreement with the theory of Murakami *et al* [35]. In this case we can define a τ_z-dependent Chern number (Hall conductance) for each block like the spin Chern number [36], from which the nontrivial QSH phase can be identified. After introducing the SIA term, the τ_z-dependent Chern number loses its meaning as the two blocks are mixed together. However, we can still employ the Z_2 topological classification [4], which requires no inversion symmetry, to identify possible QSH thin films in experiment.

5.1. QSH effect without SIA

Considering the block-diagonal form of the effective model without SIA (40), we can derive the Hall conductance for each block, separately. For the 2×2 Hamiltonian in terms of the $\mathbf{d}(k)$ vectors and Pauli matrices in equation (45), the Kubo formula for the Hall conductance can be generally expressed as [37, 38]

$$\sigma_{xy} = -\frac{e^2}{2\Omega\hbar} \sum_k \frac{(f_{k,-} - f_{k,+})}{d^3} \epsilon_{\alpha\beta\gamma} \frac{\partial d_a}{\partial k_x} \frac{\partial d_\beta}{\partial k_y} d_\gamma \tag{60}$$

where Ω is the volume of the system; d is the norm of (d_x, d_y, d_z); and $f_{k,\pm} = 1/\{\exp[(E_\pm(k) - \mu)/k_B T] + 1\}$ is the Fermi distribution function of electron (+) and hole (−) bands, with μ being the chemical potential, k_B the Boltzmann constant and T the temperature.

At zero temperature and when the chemical potential μ lies between the bandgap $(-\frac{|\Delta|}{2}, \frac{|\Delta|}{2})$, the Fermi functions reduce to $f_{k,+} = 0$ and $f_{k,-} = 1$. In this case, we have [31]

$$\sigma_{xy}^{\tau_z} = -\tau_z \frac{e^2}{2h}[\mathrm{sgn}(\Delta) + \mathrm{sgn}(B)]. \tag{61}$$

This result intuitively shows that only when B and Δ have the same sign is the Chern number equal to +1 or −1, which is topologically nontrivial, and the Hall conductance is quantized to be $\pm e^2/h$. In other words, the QSH depends not only on the sign of Δ at the Γ point but also on that of B for large enough k. Experimentally, the τ_z-dependent Hall conductance can be probed by a nonlocal measurement, just like that for 2D CdTe/HgTe quantum wells [7].

5.2. *QSH effect with SIA: Z_2 invariant*

In the presence of SIA, $\tilde{V}$ couples the blocks h_+ and h_-, so the τ_z-dependent Hall conductance becomes nonsense. Following Kane and Mele [4], we can employ the Z_2 topological classification to give a criterion of the QSH phase, because it does not require inversion symmetry as a necessary condition. The Z_2 index can be obtained by counting the number of pairs of complex zeros of $P(\mathbf{k}) \equiv \mathrm{Pf}[A(\mathbf{k})]$, where the *Pfaffian* is defined as

$$\mathrm{Pf}[A(\mathbf{k})] = \frac{1}{2^n n!} \sum_{\text{Permutations of } \{i_1,\ldots,i_{2n}\}} (-1)^N A_{i_1 i_2} \ldots A_{i_{2n-1} i_{2n}}, \tag{62}$$

in which N counts the number of times of permutations, and $A(\mathbf{k})$ is a $2n$ order anti-symmetric matrix defined by the overlaps of time reversal

$$A_{ij}(\mathbf{k}) = \langle u_i(\mathbf{k})|\Theta|u_j(\mathbf{k})\rangle \tag{63}$$

with i, j run over all the bands below the Fermi surface, i.e. ψ_{1-} and ψ_{2-} in the present case according to equations (58) and (59). Based on the spin nature of the basis states $\{\Phi_1, \Phi_4, \Phi_2, \Phi_3\}$ in our effective model, the time-reversal operator here is defined as $\Theta \equiv i\sigma_x \otimes \sigma_y \mathcal{K}$, where σ_x and σ_y are the x- and y-components of Pauli matrices, respectively, and $\mathcal{K}$ the complex conjugate operator. The number of pairs of zeros can be counted by evaluating the winding of the phase of $P(\mathbf{k})$ around a contour C enclosing half of the complex plane of $\mathbf{k} = k_x + ik_y$,

$$I = \frac{1}{2\pi i} \oint_C d\mathbf{k} \cdot \nabla_{\mathbf{k}} \log[P(\mathbf{k}) + i\delta]. \tag{64}$$

Because the model is isotropic, we can choose $\mathcal{C}$ to enclose the upper half-plane; the integral then reduces to only the path along the k_x-axis, while the part of the half-circle integral vanishes for $\delta > 0$ and $|\mathbf{k}| \to +\infty$.

In the absence of the SIA term, $P(\mathbf{k})$ is found, for the Hamiltonian (44), to be

$$P(k) = \frac{-\frac{\Delta}{2} + Bk^2}{\sqrt{\left(\frac{\Delta}{2} - Bk^2\right)^2 + (\hbar v_F)^2 k^2}}, \tag{65}$$

Table 2. The fitting parameters for the Bi_2Se_3 thin films, using the energy spectra equation (56) from our effective model. Adapted from [28].

Layers (QL)	E_0 (eV)	D (eV Å ^{2})	Δ (eV)	B (eV Å ^{2})	v_F (10^5 m s^{-1})	$\lvert\tilde{V}\rvert$ (eV)
2	−0.470	−14.4	0.252	21.8	4.47	0
3	−0.407	−9.7	0.138	18.0	4.58	0.038
4	−0.363	−8.0	0.070	10.0	4.25	0.053
5	−0.345	−15.3	0.041	5.0	4.30	0.057
6	−0.324	−13.0	0	0	4.28	0.068

in which one can check that the zero points exist only when $k^2 = \Delta/2B > 0$, and form a circular ring. Along the k_x-axis, only one of a pair of zeros in the ring is enclosed in the contour $\mathcal{C}$, which gives a Z_2 index $I = 1$. This defines the nontrivial QSH phase and is consistent with the conclusion from the Hall conductance in equation (61).

In the presence of a small SIA term $\tilde{V} < \hbar v_F\sqrt{|\Delta/2B|}$, with the help of the eigen-wavefunctions (58) and (59), real $P(\mathbf{k})$ can be found (after a $U(1)$ rotation) to be

$$P(k) = \frac{(t+E_-^{in})(t+E_-^{out}) + |\tilde{V}|^2 - (\hbar v_F k)^2}{2\sqrt{E_-^{in}E_-^{out}(t+E_-^{in})(t+E_-^{out})}} \begin{cases} \mathrm{sgn}(\hbar v_F k - |\tilde{V}|), & \Delta > 0, \\ 1, & \Delta \leqslant 0, \end{cases} \tag{66}$$

where the sgn is to secure the continuity of $P(\mathbf{k})$. One can check that $P(0) = -\mathrm{sgn}(\Delta)$ and $P(\infty) = \mathrm{sgn}(B)$. Besides, for a small $\tilde{V}$, the behavior of $P(\mathbf{k})$ between $P(0)$ and $P(\infty)$ will not change qualitatively (see figure 5). Therefore, for $\Delta B > 0$, $P(k_x, 0)$ should still have odd pairs of zeros. For a large $\tilde{V} \geqslant \hbar v_F\sqrt{|\Delta/2B|}$,

$$P(k) = \frac{(t+E_-^{in})(t+E_-^{out}) + |\tilde{V}|^2 - (\hbar v_F k)^2}{2\sqrt{E_-^{in}E_-^{out}(t+E_-^{in})(t+E_-^{out})}} \begin{cases} \mathrm{sgn}(\hbar v_F k - |\tilde{V}|), & B < 0. \\ 1, & B \geqslant 0. \end{cases} \tag{67}$$

One can check that for this case, $P(0)P(\infty)$ is always positive; thus $P(k)$ has even pairs of zeros, regardless of the signs and values of Δ and B. In other words, a large SIA will always destroy the QSH phase.

5.3. Bi_2Se_3 *thin films and the QSH effect*

Recently, the thickness-dependent band structure of molecular beam epitaxy grown ultrathin films Bi_2Se_3 was investigated by *in situ* angle-resolved photoemission spectroscopy [28]. An energy gap was first observed experimentally in the surface states of Bi_2Se_3 below the thickness of 6 QL, which confirms the theoretical prediction as a finite size effect [30]–[33] .

Table 2 shows the fitting parameters for the ARPES data of Bi_2Se_3 thin films [28] using the energy spectrum formula (equation (56)). For the films with thickness ranging from 2 QL to 5 QL, all of them satisfy $\mathrm{sgn}(\Delta B) > 0$ and $\tilde{V} < \hbar v_F\sqrt{|\Delta/2B|}$, hence the films are possibly in the QSH regime. We identify that only 2 QL, 3 QL, and 4 QL belong to the nontrivial case for the potential QSH effect. 5 QL is an exceptional case where the fitted parameters B and D do not satisfy the existence condition of an edge state solution [33]. The condition of $B^2 < D^2$ will lead to the bandgap closing for a large k. However, it is understood that the model is only valid

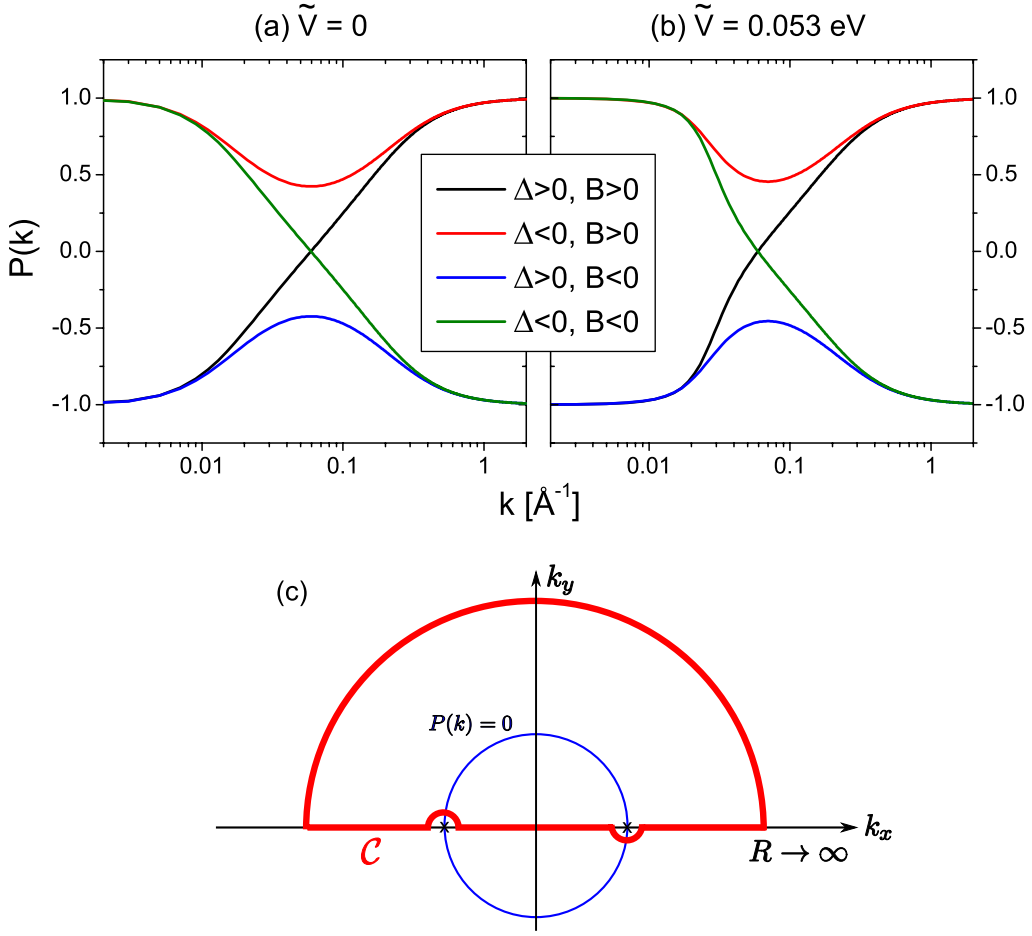

Figure 5. (a, b) $P(k)$ for four combinations of Δ and B, in the absence (a) and presence (b) of a small SIA $\tilde{V}$. Note that $|\Delta| = 0.07\,\text{eV}$ and $|B| = 10\,\text{eV\AA}^2$. (c) The contour $\mathcal{C}$ is used to count the number of pairs of the zeros of $P(k)$, which form a circular ring when $\Delta B > 0$.

near the Γ point, and the fitting parameters are limited to the case of small k. The band gap was measured clearly for the film of 5 QL.

It was previously predicted, using the parameters from first-principles calculations [19], that the gap Δ should oscillate as a function of the film thickness [30]–[32]. However, this oscillation is not reflected in the measured results.

5.4. QSH effect of SIA and the edge states

In the quantum Hall effect, the Chern number of the bulk states has an explicit correspondence to the number of edge states in an open boundary condition [39]. In the TI or QSH system, the Z_2 topological invariant has also a relation to the number of helical edge states [40]. As supplementary support to the above conclusion, we demonstrated the presence of edge states in a periodic boundary condition along the x-direction and an open boundary condition (say along the y-direction) imposed in a geometry of a strip of the thin film by means of numerical

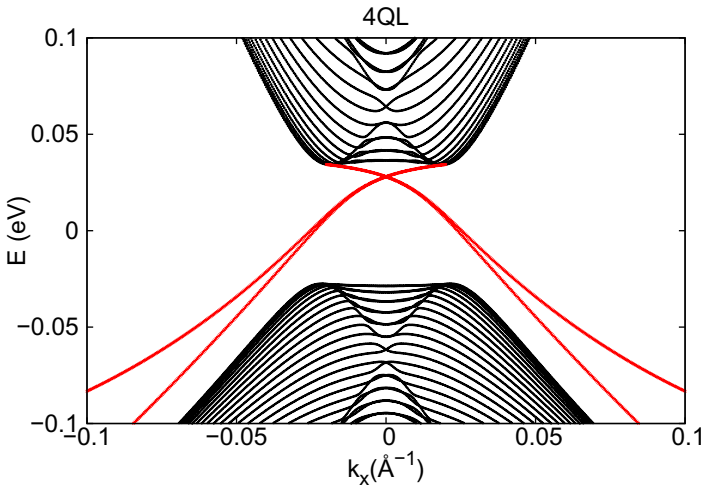

Figure 6. The energy spectra obtained by the tight-binding calculation for 4 QL of Bi_2Se_3 thin films, with the width along the y-direction of 200 nm. The parameters are given in table 2.

calculation. Using the parameters in table 1, we have concluded that a strip of 2–4 QL will exhibit helical edge states. More specifically, we present the energy dispersion for 4 QL in figure 6. There is a doubly degenerate Dirac point inside the gap of the 2D surface states for 4 QL consistent with the results obtained in the above sections.

6. Conclusions

We derived 2D effective continuous models for the surface states and thin films of 3DTI. A gapless Dirac cone was confirmed for the surface states of a 3DTI. For a thin film, the coupling between opposite topological surface states in space opens an energy gap, and the Dirac cone evolves into a gapped Dirac hyperbola. The thin film may break the top–bottom symmetry. For example, the thin film grows on a substrate, and possesses SIA. This SIA leads to Rashba-like coupling and energy splitting in the momentum space. It also leads to asymmetric distributions of states along the film growth direction.

The ARPES measurements on Bi_2Se_3 films have demonstrated that the surface spectra open a visible energy gap when the thickness is below 6 QLs. The energy gap was observed to be a function of the thickness of the thin film, and in good agreement with theoretical prediction as a finite size effect of the thickness of thin film. The Rashba-like splitting was measured clearly in the thin film of 2–6 QLs. This can be explained very well from the inclusion of the SIA. Since the thin film was grown on an SiC substrate and the other surface is exposed to the vacuum, this fact results in the SIA in the thin film. Another piece of direct evidence to support the SIA is the signal intensity pattern of the energy spectra of ARPES. Usually the surface states are located dominantly near the top and bottom surfaces. The signal intensity for these two branches of energy spectra of ARPES are different. The SIA will cause the coupling between two surface states near their crossing point. That is why the Rashba-like splitting of the ARPES spectra has a bright crossing point near the Γ point, with one branch bright and the other almost invisible.

Table A.1. Four possible combinations of λ_1 and λ_2, according to equation (5), and the resulting $f_\pm$ and $\eta_{1,2}$ according to equation (38). According to equation (5), $\lambda_1^2 < \lambda_2^2$, so there does not exist a case when λ_1 is real and λ_2 is purely imaginary.

	λ_1	λ_2	$f_\pm$	$\eta_{1,2}$
Case A	λ_2^*	λ_1^*	Imaginary	Real
Case B	Real	Real	Real	Real
	Imaginary	Real	Real	Real
	Imaginary	Imaginary	Real	Real

Table A.2. Four possible groups of E_+ and E_- and resulting values of $\tilde{A}_2$ and $\tilde{V}$.

E_+	E_-	$\widetilde{A}_2$	$\widetilde{V}$
Case A	Case A	Imaginary	Real
Case A	Case B	Real	Imaginary
Case B	Case A	Real	Imaginary
Case B	Case B	Imaginary	Real

Thus the SIA term can be used to describe the ARPES measurements on Bi_2Se_3 thin films very well.

Our effective model demonstrates that the 3DTI can be reduced to a 2D QSH due to spatial confinement. Strictly speaking, the system is no longer a 3DTI in the original sense once the energy gap opens in the surface bands, since the Z_2 invariant for the bulk states becomes zero. However, the surface bands themselves may contribute a nontrivial one in the Z_2 invariant even when the SIA term is included. Our calculation demonstrates that a strong SIA always tends to destroy the QSH effect. A critical value for SIA exists at the point where there is a transition from a topological trivial to nontrivial phase. Based on the model parameters fitted from the experimental data of ARPES, we conclude that the Bi_2Se_3 thin films should exhibit the QSH effect once the energy gap opens in the surface spectra due to the spatial confinement of the thin film.

Acknowledgments

We thank Ke He and Qi-Kun Xue for providing experimental data prior to publication, and Qian Niu for helpful discussions. This work was supported by the Research Grant Council of Hong Kong under grant numbers HKU 7037/08P and HKU 10/CRF/08.

Appendix. Model parameters $\tilde{A}_2$ and $\tilde{V}$

In this appendix, we demonstrate that both parameters $\tilde{A}_2$ and $\tilde{V}$ in the effective model (40) can be either real or purely imaginary, and the product of $\tilde{A}_2\tilde{V}$ must be purely imaginary. By putting the wavefunctions equations (32) and (33) into the definitions in equations (42) and (52),

we have

$$\tilde{A}_2 = \mathrm{i}A_1A_2D_+C_+C_-\int_{-L/2}^{L/2}\mathrm{d}z[\eta_2^-f_+^{+*}f_+^- + \eta_1^{+*}f_-^{+*}f_-^-],$$
$$\tilde{V} = C_+C_-\int_{-L/2}^{L/2}\mathrm{d}zV(z)[D_+^2\eta_1^{+*}\eta_2^-f_-^{+*}f_+^- + A_1^2f_+^{+*}f_-^-]. \tag{A.1}$$

For arbitrary energy, equation (5) requires that the values of λ_1 and λ_2 be only one of the combinations shown in table A.1. By putting these combinations into equations (34)–(37), one can show that for the first entry, η_1 and η_2 are real, while f_+ and f_- are purely imaginary, referred to as case A; while for entries 2–4, all of η_1, η_2, f_+ and f_- are real, referred to as case B. For an arbitrary group of E_+ and E_-, each of them belongs to either case A or B, leading to four possibilities, as shown in table A.2. In particular, according to equation (15), when $A_1^2/(-D_+D_-) > 4M/B_1$, there is no complex $\lambda_{1,2}$, corresponding to the last row of table A.2, i.e. $\tilde{A}_2$ is purely imaginary, while $\tilde{V}$ is real.

References

[1] Kane C L and Mele E J 2006 *Science* **314** 1692
Zhang S C 2008 *Physics* **1** 6
Buttiker M 2009 *Science* **325** 278
Moore J 2009 *Nat. Phys.* **5** 378
[2] Sarma S D and Pinczuk A 1997 *Perspectives in Quantum Hall Effects* (New York: Wiley)
[3] Kane C L and Mele E J 2005 *Phys. Rev. Lett.* **95** 226801
[4] Kane C L and Mele E J 2005 *Phys. Rev. Lett.* **95** 146802
[5] Bernevig B A, Hughes T L and Zhang S 2006 *Science* **314** 1757
[6] Konig M, Wiedmann S, Brune C, Roth A, Buhmann H, Molenkamp L W, Qi X L and Zhang S C 2007 *Science* **318** 766
[7] Roth A, Brune C, Buhmann H, Molenkamp L W, Maciejko J, Qi X L and Zhang S C 2009 *Science* **325** 294
[8] Li J, Chu R L, Jain J K and Shen S Q 2009 *Phys. Rev. Lett.* **102** 136806
[9] Jiang H, Wang L, Sun Q F and Xie X C 2009 *Phys. Rev.* B **80** 165316
[10] Groth C W, Wimmer M, Akhmerov A R, Tworzydlo J and Beenakker C W J 2009 *Phys. Rev. Lett.* **103** 196805
[11] Fu L, Kane C L and Mele E J 2007 *Phys. Rev. Lett.* **98** 106803
[12] Moore J E and Balents L 2007 *Phys. Rev.* B **75** 121306
[13] Murakami S 2007 *New J. Phys.* **9** 356
[14] Teo J C Y, Fu L and Kane C L 2008 *Phys. Rev.* B **78** 045426
[15] Hsieh D, Qian D, Wray L, Xia Y, Hor Y S, Cava R J and Hasan M Z 2008 *Nature* **452** 970
[16] Hsieh D *et al* 2009 *Science* **323** 919
[17] Xia Y *et al* 2009 *Nat. Phys.* **5** 398
[18] Chen Y L *et al* 2009 *Science* **325** 178
[19] Zhang H, Liu C X, Qi X L, Dai X, Fang Z and Zhang S C 2009 *Nat. Phys.* **5** 438
[20] Qi X L, Li R, Zang F and Zhang S C 2009 *Science* **323** 1184
[21] Fu L and Kane C L 2008 *Phys. Rev. Lett.* **100** 096407
[22] Nilsson J, Akhmerov A R and Beenakker C W J 2008 *Phys. Rev. Lett.* **101** 120403
[23] Fu L and Kane C L 2009 *Phys. Rev. Lett.* **102** 216403
[24] Akhmerov A R, Nilsson J and Beenakker C W J 2009 *Phys. Rev. Lett.* **102** 216404
[25] Tanaka Y, Yokoyama T and Nagaosa N 2009 *Phys. Rev. Lett.* **103** 107002
[26] Law K T, Lee P A and Ng T K 2009 *Phys. Rev. Lett.* **103** 237001
[27] Zhang G, Qin H, Teng J, Guo J, Guo Q, Dai X, Fang Z and Wu K 2009 *Appl. Phys. Lett.* **95** 053114

[28] Zhang Y *et al* 2009 arXiv:0911.3706
[29] Peng H, Lai K, Kong D, Meister S, Chen Y, Qi X L, Zhang S C, Shen Z X and Cui Y 2009 *Nat. Mater.* **9** 225
[30] Linder J, Yokoyama T and Sudbø A 2009 *Phys. Rev.* B **80** 205401
[31] Lu H Z, Shan W Y, Yao W, Niu Q and Shen S Q 2010 *Phys. Rev.* B **81** 115407
[32] Liu C X, Zhang H J, Yan B H, Qi X L, Frauenheim T, Dai X, Fang Z and Zhang S C 2009 *Phys. Rev.* B **81** 041307
[33] Zhou B, Lu H Z, Chu R L, Shen S Q and Niu Q 2008 *Phys. Rev. Lett.* **101** 246807
[34] Winkler R 2003 *Spin–Orbit Coupling Effect in Two-Dimensional Electron and Hole System* (Berlin: Springer)
[35] Murakami S, Iso S, Avishai Y, Onoda M and Nagaosa N 2007 *Phys. Rev.* B **76** 205304
[36] Sheng D N, Weng Z Y, Sheng L and Haldane F D M 2006 *Phys. Rev. Lett.* B **97** 036808
[37] Qi X L, Wu Y S and Zhang S C 2006 *Phys. Rev.* B **74** 085308
[38] Zhou B, Ren L and Shen S Q 2006 *Phys. Rev.* B **73** 165303
[39] Hatsugai Y 1993 *Phys. Rev. Lett.* **71** 3697
[40] Qi X L, Wu Y S and Zhang S C 2006 *Phys. Rev.* B **74** 045125