Topological insulators: a simple example

Consider first a simple model of a 2D crystal lattice (fig. 1) with the electrons described by a nearest-neighbor tight-binding Hamiltonian:

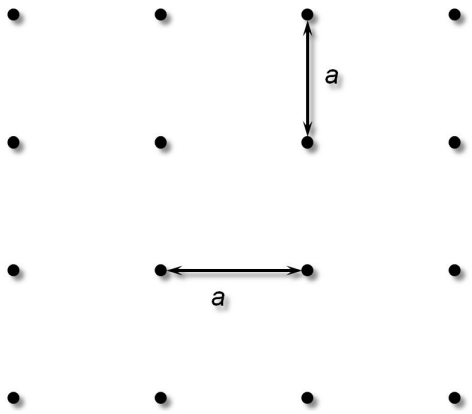

Fig. 1

$$\hat{H} = E_{at}\sum_i a_i^+ a_i - \frac{1}{2}t\sum_{ij\in n.n.}(a_i^+ a_j + \text{H.c.})$$

$$\equiv E_o\sum_i a_i^+ a_i + \frac{1}{2}t\sum_{ij\in n.n.}(a_i^+ - a_j^+)(a_i - a_j) \quad (1)$$

"$n.n.$" $\equiv$ nearest neighbors, $E_o \equiv E_{at} - 4t$
The single-particle eigenstates are 2D Bloch waves:

$$|\boldsymbol{k}\rangle = N^{-1/2}\sum_i \exp\,(i\boldsymbol{k}\cdot\boldsymbol{R}_i)a_i^+|\text{vac}\rangle \equiv a_{\mathbf{k}}^+|\text{vac}\rangle \quad (2)$$

with $\boldsymbol{k} \in$ FBZ (i.e. $-\pi/a \le k_x, k_y \le \pi/a$). The corresponding energy is

$$E(\boldsymbol{k}) = E_o + t\sum_{\hat{R}=a\hat{x},a\hat{y}}\left|e^{i\boldsymbol{k}\cdot\boldsymbol{R}} - 1\right|^2 = E_o + 4t\left\{\sin^2 k_x a/2 + \sin^2 k_y a/2\right\} \quad (3)$$

so that the total bandwidth is $8t$

Do surface modes automatically exist? From the symmetry of the problem, any surface mode would have to be, in bulk, of the form, as regards its n-dependence.

$$|\kappa\rangle = \text{const.}\ \textstyle\sum_n e^{-n\kappa}a_n^+|vac\rangle$$

But to the extent that there is no "extra" potential associated with the surface, the above form (4) must extend also to *negative* values of n, which is not physically allowable. If there is an appropriate potential associated with the surface as such, surface modes may exist (problem) but do not have any specially interesting properties.

Next consider the generalization to the case of a 2D lattice of atoms with two atomic states, say "s" and "p", described by a nearest-neighbor tight-binding Hamiltonian: Actually we have to be rather careful here: in 3D there are actually 3 atomic p-states which are degenerate in the absence of spin-orbit interaction, and while the 2-dimensionality of the lattice may split off the "p_z" state, the "p_x" and "p_y" states will still be degenerate by symmetry considerations. Moreover, we would expect that the Bloch waves derived from those atomic states in the TB limit would break the degeneracy, so that for given (general) $\boldsymbol{k}$ the p_x- and p_y- derived states would differ in energy; thus one has to deal not with two but with three bands, and the problem gets quite messy. To get around this complication let us assume that the intra-atomic spin-orbit coupling is large compared not only to the p-state hopping matrix element t but to the p-state bandwidth 8t; although this assumption is likely to be unrealistic in most experimentally viable situations, it permits a simple conceptual picture. In fact, from the 2D symmetry we may reasonably assume in this limit that the z component of (real) spin, which we denote as s_z to avoid confusion with the pseudospin, is a good quantum number, so that the intra-atomic spin-orbit interaction (SOI) is of the form const. $s_z\hat{L}_z$, where $\hat{L}_z$ is the z component of orbital angular momentum. Now, the "chemists' eigenstates" $|p_x\rangle$ and $|p_y\rangle$ are not even approximately eigenstates of $\hat{L}_z$; however, we can form states which diagonalize this operator within the p-state manifold by taking the "physicists' eigenstates", namely $|p_x\rangle \pm i|p_y\rangle$. Schematically, the state $|\hat{L}_z = +1\rangle = (|p_x\rangle + i|p_y\rangle)$ is of the form shown in fig. 2; the state $|\hat{L}_z = -1\rangle = (|p_x\rangle - i|p_y\rangle)$ is of the same "shape" but has the amplitudes $\pm i$ exchanged. Since spin is conserved, the 4

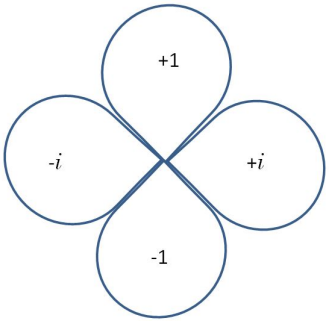

Fig. 2

states in the p-manifold are split into doublets $J_z = \pm 3/2, \pm 1/2$ and from our assumption above we may assume that the splitting is large compared to the bandwidth and we may keep only the lower doublet (let us say for definiteness $J_z = \pm 3/2$). Then for the subsequent

analysis we may fix s_z to be (say) $+1/2$ and thus concentrate in the orbital state $|L_z = +1\rangle$; since all the terms in the Hamiltonian which we will consider conserve spin, we may neglect the $|L_z = -1\rangle$ state for the moment (although, we shall see, its existence is important to the eventual argument).

With this simplifying assumption and neglecting for the moment inter-atomic SOI, the structure of the single p-band derived from $|L_z = +1\rangle$ state in the TB limit is quite similar to that of the s-band: despite the sign variation on the atomic states, it is intuitively clear (cf. fig. 3) that the matrix element t_p for hopping between any pair of nearest-neighbor atoms is simply a constant (which is actually likely to be negative, though we shall not need this result): thus, the total TB Hamiltonian needs

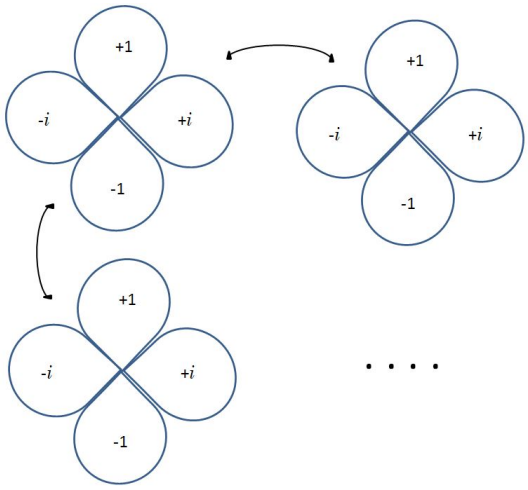

Fig. 3

$$\hat{H} = E_{at}^{(s)} \sum_i a_i^+ a_i - \frac{1}{2} t_s \sum_{ij \in nn} (a_i^+ a_j + \text{H.c.}) + E_{at}^{(p)} \sum_i b_i^+ b_i - \frac{1}{2} t_p \sum_{ij \in nn} (b_i^+ b_j + \text{H.c.}) \quad (4)$$

where b_i^+ creates the $|L_z = +1\rangle$ state on atom i. By the same analysis as in the simple s-band case we transform eqn. (4) to the form

$$\hat{H} = \sum_k \{E^{(s)}(\boldsymbol{k}) a_k^+ a_k + E^{(p)}(\boldsymbol{k}) b_k^+ b_k\} \quad (5)$$

with

$$E^{(\alpha)}(k) = E_o^{(\alpha)} + 4t_\alpha \{\sin^2(k_x a/2) + \sin^2(k_y a/2)\} \qquad (\alpha \equiv s, p) \quad (6)$$

At this point it is convenient to subtract out a term which is proportional to the unit vector in the 2D Hilbert space spaniel by $a_k^+|\text{vac}\rangle$ and $b_k^+|\text{vac}\rangle$ by writing

$$E_{av}(\boldsymbol{k}) \equiv \frac{1}{2}(E^{(p)}(k) + E^{(s)}(k), \delta E(k) \equiv \frac{1}{2}(E^{(p)}(k) - E^{(s)}(k))) \tag{7}$$

so that

$$\hat{H} = \sum_k E_{(av)}(k)(a_k^+ a_k + b_k^+ b_k) + \sum_k \delta E(k)(b_k^+ b_k - a_k^+ a_k) \tag{8}$$

For most of the subsequent discussion the term in $E_{av}(\boldsymbol{k})$ is irrelevant and may be neglected. Introducing the pseudospin language, with $\hat{\sigma}_z(k) \equiv a_k^+ a_k - b_k^+ b_k$, the rest of the Hamiltonian (8) can be rewritten in the form

$$\hat{H} = -\sum_{\boldsymbol{k} \in FBZ} \sigma_{z\boldsymbol{k}} \mathcal{H}_{\boldsymbol{k}} \tag{9}$$

with $\mathcal{H}_{\boldsymbol{k}}$ a "field" in the z-direction in the pseudospin space with a value given by

$$\mathcal{H}_{\boldsymbol{k}} = M - B\{\sin^2(k_x a/2) + \sin^2(k_y a/2)\} \tag{10}$$

with the coefficients M and B given by

$$M \equiv \frac{1}{2}(E_0^{(p)} - E_0^{(s)}) \quad B \equiv \frac{1}{2}(t_s - t_p) \tag{11}$$

The model given by eqns. (9) and (10) is a slight generalization of the model for CdTe-HgTe alloys used in the classic paper of Bernevig et al[1]; it has the advantage of explicitly respecting the periodicity of $\mathcal{H}_k$ with respect to the FBZ. We will usually be interested in the case where not only $M > 0$ but also $B > 0$ (recall that t_p is likely to be negative).

The crucial question which arises in connection with eqn. (10) is: what is the locus in the FBZ (if any) of $\mathcal{H}_{\|} = 0$? We will call this locus the "inversion contour". Evidently we can distinguish three cases, depending on the relative magnitude of B and M (which we will take to be both positive):

Case I ($B < M/2$):

There is no point in the FBZ at which $\mathcal{H}_k = 0$ (fig 4a). We will see below that (when the inter-atomic SOI is taken into account) this is the case of a simple textbook band insulator.

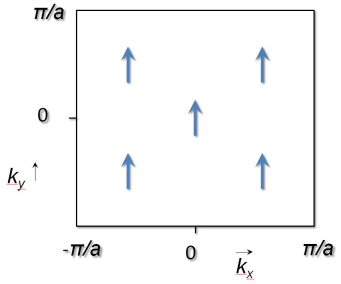

Fig. 4. a

Case II ($B > M$):

The locus of $\mathcal{H}_k = 0$ lies entirely in the interior of the FBZ (fig. 4b). This is the simplest case of a topological insulator.

[1] Science **314**, 1757 (2006)

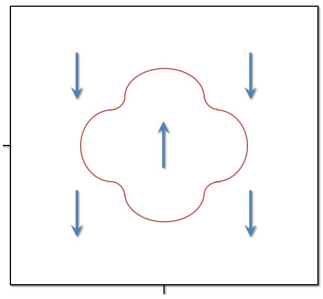

Fig. 4b

Case III ($M/2 < B < M$):

In this case the locus of $\mathcal{H}_k = 0$ intersects the FBZ boundary (fig. 4c). However, it is clear that by an appropriate redefinition of the FBZ accompanied by an inversion of the pseudospin axis we can reduce this case to case II, so again it will turn out to correspond to a topological insulator.

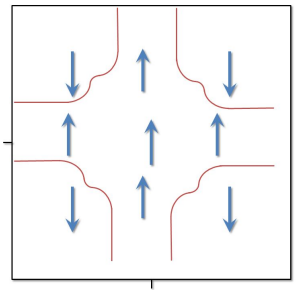

Fig. 4c

Were eqn. (10) in fact the complete Hamiltonian, then as argued in L. 22 not only would case I be a simple textbook insulator but cases II and cases III would correspond to textbook semi-metals. What makes things much more interesting is the presence of a term in the Hamiltonian not included in (10), namely the inter-atomic spin-orbit interaction. To discuss this, we shall concentrate as above on the band derived from the atomic state $J_z = +3/2, (|s_z = +1/2, L_z = +1\rangle)$ and drop explicit reference to the spin index.

The general form of the SOI (whether intra-atomic or inter-atomic) is

$$\hat{H}_{so} = \text{const. } \hat{\boldsymbol{s}} \cdot \hat{\boldsymbol{p}} \times \boldsymbol{\nabla} \hat{V} \tag{12}$$

where $\boldsymbol{p}$ is the momentum operator and $V(r)$ the electronic potential. Thus, the matrix elements between a state $\mu(= s, p)$ on atom i and a state $\nu(= s, p)$ on atom j is

$$\left\langle \mu, i|\hat{H}_{So}|\nu, j \right\rangle = \langle \mu, i|\hat{\boldsymbol{s}} \cdot \boldsymbol{p} \times \boldsymbol{\nabla} V|p, j\rangle \tag{13}$$

It is clear from the symmetry of the problem (the quantity $\boldsymbol{p} \times \boldsymbol{\nabla} V$ transforms as a pseudovector) that there are no nonzero s-s matrix elements, and a little thought shows that there are no p-p ones either. Hence, if we restrict ourselves to interactions, the only nonvanishing nearest-neighbor

matrix elements are of the form

$$\langle s,i|\hat{\boldsymbol{s}}\cdot\boldsymbol{p}\times\boldsymbol{\nabla}V|p,j\rangle_{ij\in nn} \tag{14}$$

(and the Hermitian conjugate element). From inspection of fig. 5 we see that this should be proportional to the overlap

$$\langle s,i|p,j\rangle$$

which in turn should be proportional to exp $i\theta_{ij}$, where θ_{ij} is the angle of the vector $\boldsymbol{r}_{ij}$ connecting atoms i and j; note in particular that the phase changes by π for opposite nearest neighbors (i.e.$\langle s,i|p,j\rangle = -\langle s,j|p,i\rangle$).

Thus, the general form of the inter-atomic SOI is

$$\hat{H}_{\text{so, int}} = \Delta' \sum_{ij} (\exp\,(i\theta_{ij}) a_i^+ b_j + \text{H. c.}) \tag{15}$$

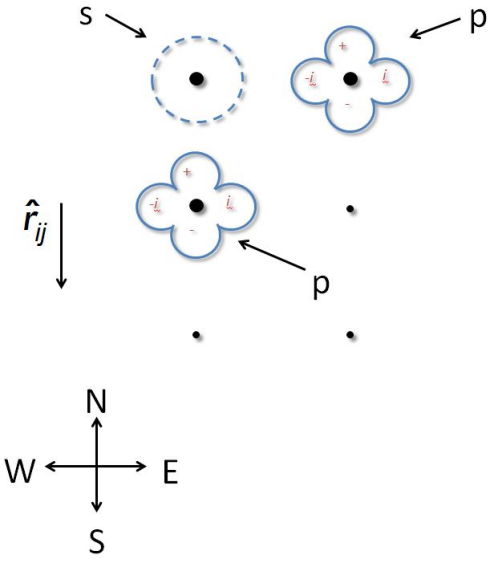

Fig. 5

where $\theta_{ij} = 0, \pi/2, \pi, 3\pi/2$ for nearest neighbors to the N, E, S, W respectively. The effect of $\hat{H}_{\text{so, int}}$ is to *hybridize* the s- and p-bands, so that the energy eigenstates (Bloch waves) are now spinors:

$$\psi_k = N^{-1/2} \sum_i e^{i\boldsymbol{k}\cdot\boldsymbol{R}i}(u_k a_i^+ + v_k b_i^+)|\text{vac}\rangle \tag{16}$$

or in explicit (pseudo-) spinor notation,

$$\psi_k = \begin{pmatrix} u_k \\ v_k \end{pmatrix} \cdot N^{-1/2} \sum_i e^{i\boldsymbol{k}\cdot\boldsymbol{R}_i} \tag{17}$$

When projected on to the (spinor) Bloch state $\psi_k, \hat{H}_{so}$ is equivalent[2] to a "field" in the pseudospin xy-plane:

$$(\mathcal{H}_k)_{so} = A(\hat{\boldsymbol{x}}\sin k_x a + \hat{\boldsymbol{y}}\sin k_y a) \quad (A = 4\Delta') \tag{18}$$

Combining eqn. (18) with eqn. (10), we finally obtain the complete Hamiltonian for our problem:

$$\hat{H} = -\sum_k \hat{\boldsymbol{\sigma}}_k \cdot \boldsymbol{\mathcal{H}}_k \tag{19}$$

where the (pseudo) field $\boldsymbol{\mathcal{H}}_k$ is given by the expression [3]

$$\boldsymbol{\mathcal{H}}_k \equiv \left(M - B\left\{\sin^2(k_x a/2) + \sin^2(k_y a/2)\right\}\right)\hat{\boldsymbol{z}} + A\sin(k_x a)\hat{\boldsymbol{x}} + A\sin(k_y a)\hat{\boldsymbol{y}} \tag{20}$$

[2] This is not quite obvious and needs a little calculation.

[3] I introduce the notation M, A, B, to make contact with that of Bernevig et al., but note that A and B have different dimension from their quantities of the same name

We note that for the band which evolves from the $J_z = -3/2 (s_z = -1/2)$ atomic state the only difference is that the sign of the y-component of $\mathcal{H}_k$ is reversed (this may be seen from fig. 5 by changing i to $-i$)

It follows immediately from (20) that the energy eigenvalues for the Bloch state $\boldsymbol{k}$ are $\pm E_k$, where

$$E_k \equiv |\boldsymbol{\mathcal{H}}_k| = \left\{(M - B[\sin^2(k_x a/2) + \sin^2(k_y a/2)])^2 + A^2(\sin^2 k_x a + \sin^2 k_y a)\right\}^{1/2} \quad (21)$$

(and those for the $J_z = -3/2$ band are of course identical). It is immediately clear from (21) that barring pathology the energy gap $2E_k$ is nonzero everywhere in the FBZ, so the system is an insulator. However, we do not yet know what kind of insulator ... we do however know that in the GS $\boldsymbol{\sigma}_k$ is parallel to $\boldsymbol{\mathcal{H}}_k$ for all $\boldsymbol{k}$

Some notes on the Hamiltonian (19), with $\boldsymbol{\mathcal{H}}_k$ given by expression (20):

1. For $|k|a \ll 1$, $\boldsymbol{\mathcal{H}}_k$ reduces to

$$\boldsymbol{\mathcal{H}}_k = \left(M - \frac{a^2}{4} B k^2\right) \hat{\boldsymbol{z}} + Aa\hat{\boldsymbol{k}} \quad (22)$$

   which is the form used by Bernevig et al. However, the full expression (20) is properly periodic in the reciprocal lattice vectors $(2\pi/a)(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$.

2. In the limit $\boldsymbol{k} \to 0$, the "transverse" term has the "Dirac" form ($\boldsymbol{p} \equiv \hbar \boldsymbol{k}$)

$$\hat{H}_D = A\boldsymbol{\sigma} \cdot \boldsymbol{p} \qquad \text{(which is invariant under } T \text{ but not under } P)$$

   and the z-component is sometimes regarded as a "mass" term.

3. By contrast, in *real* space the Hamiltonian (with the $J_z = -3/2$ contribution included) is invariant under both P and T, as it must be since it was derived from the SOI which has both these properties.

4. Consider case II, ($B > M$) and imagine sweeping k_y up from 0 to π/a and looking, for each k_y, for the minimum of E_k as a function of k_x. (see fig. 6). From fig. 6 we see that at the "critical" value of $k_y, k_c = 2(a^{-1} \sin^{-1} \sqrt{M/B})$, we have (trivially)

$$E_{\min}(k_x) \leq E(k_x = 0) = A \sin k_c a. \quad (23)$$

   - a result which will be useful in the discussion of surface states. (Actually, for the form (21) eqn. (23) is an equality, but it is useful to keep the $\leq$).

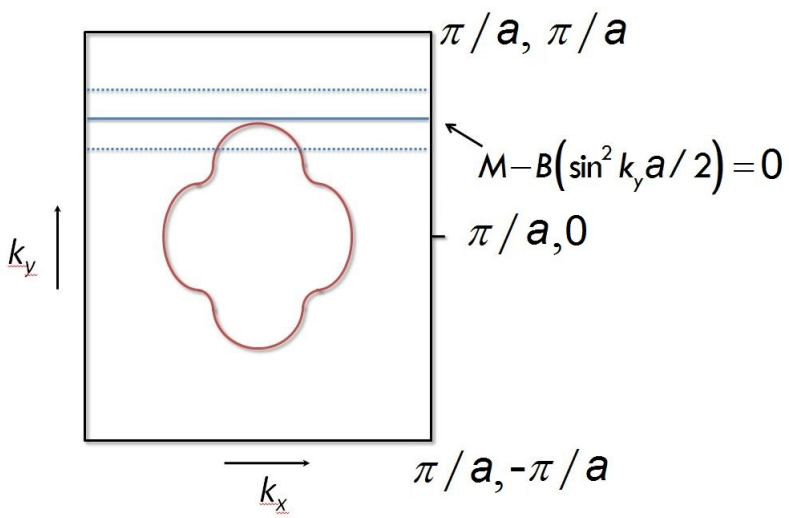

Fig. 6

5. Consider the corners of the FBZ ($k_x, k_y = \pm\pi/a$). Since the transverse field components both vanish at these points, the field, and therefore the pseudospin of the electron occupying this state in the groundstate, must be along the z-axis ("$\mathcal{H}_\perp = 0$"), and we see that it is parallel to the spin direction at the origin for case I (fig. 4a) but opposite for case II (fig. 4b). This suggests that case II (the TI) might be somewhat analogous to the "boojum" texture we discussed for superfluid $^3He - A$. Now, at first sight this analogy does not seem to be good, because in the "boojum" case we could find a contour at the edge of the disk on which $\boldsymbol{l}$ was everywhere down, while in the TI case it is clear from eqn. (20) that the locus of $\mathcal{H}_\perp = 0$ does not constitute a complete curve but is only a set of distinct points. Thus at first sight it appears impossible to reproduce the considerations regarding the MH theorem, etc.

What saves us is the theorem derived in the appendix to lecture 22. If we consider any one edge of the FBZ, say the right-hand one in fig. 6, then along this edge according to eqn (20), the effective field $\mathcal{H}_k$, and hence the pseudospin $\boldsymbol{\sigma}_k$, lies entirely in the yz-plane, and hence the phase χ can be globally defined; in particular the difference in phase between its value at the corners $(\pi/a, \pi/a)$ and $(\pi/a, -\pi/a)$ (when $\boldsymbol{l}$ lies along the z-axis) is well-defined. Applying a similar argument to each of the four edges, we see the **χ is well-defined everywhere along the circumference of the FBZ** and hence, since it

must be single-valued mod 2π, we have for a contour C exactly enclosing the FBZ[4]

$$\oint \boldsymbol{\nabla}\chi \cdot \boldsymbol{dl} = 2\pi n \qquad , n \text{ integral} \tag{24}$$

Moreover, application of the Stokes and MH theorems (of course in $\boldsymbol{k}$-space) leads to the conclusion that this n is just the Chern number of the filled "pseudo-valence" band:

$$\frac{1}{2\pi}\int_{FBZ} dS_k \epsilon_{\alpha\beta\gamma}\sigma_\alpha(\boldsymbol{\nabla}_k\sigma_\beta \times \boldsymbol{\nabla}_k\sigma_\gamma) = n \tag{25}$$

It is exceedingly plausible that the case $n = 0$ corresponds to a band insulator and $n = 1$ to a topological insulator, and we can check that this is correct by carrying out the explicit rotation of $\boldsymbol{\sigma}_k$ along the lines connecting the center of the zone to the four corners; since for case II these operations are just special cases of the one already carried out for the cylindrically symmetric case, we can say right away that the resulting phases at the corners must be just what they would be if the square FBZ were deformed into a circle, namely (const. +) $0, \pi/2, \pi, 3\pi/2$. Thus from eqn. (24), we conclude that for case II (the TI) $n = 1$. On the other hand, it is clear that for case I, where $\boldsymbol{\sigma}_k$ is rotated away from the $\uparrow$ direction but returns to it at the corners, n must be zero. (For the $J_z = -3/2 (S_2 = -1)$ pseudo-valence band, since the helicity of the orbital configuration is reversed, the Chern number n is -1).

Thus, we obtain the conclusion that a topological insulator is distinguished from a band insulator by a *topological* property, namely the Chern number n, defined by eqn (25). This is the central qualitative output of theory of the TI state.

The above argument implicitly assumes that the value of $\boldsymbol{\sigma}_k$ in the GS is well-defined for all $\boldsymbol{k} \in$ FBZ, which in turn implies that the value of $\boldsymbol{\mathcal{H}}_k$ is non-zero for all k, i.e. the energy gap is non-zero throughout the whole of the FBZ. To the extent that this is true, it is clear that since the Chern number can take only discrete values, it is impossible to deform the GS of a TI continuously into that of a band insulator, or vice versa.

It is clear that the above remarks should apply much more generally than for the specific Hamiltonian represented by eqns. (19) and (20).

Although it is the nontrivial Chern number which is in some sense the most fundamental characteristic of a TI as opposed to a band insulator, this is not an easy quantity to measure experimentally; indeed, it is rather difficult to think of ways of distinguishing the two kinds of system by looking at their bulk behavior. The recent interest in TI's stems largely from the realization that the properties of their *surfaces* are very different from those of a band insulator. In the latter case, while surface modes may or may not exist, their properties

[4]The difference from eqn. (50) of L. 22 (where the more general form of the RNS is $4\pi n$) is due to the fact that we are discussing (pseudo-)spin 1/2 rather than (orbital) angular momentum 1, cf. p. 12 of L 22.

are not particularly exotic, and in particular there is no requirement for them to span the bulk energy gap.

What happens in the case of a TI? The following qualitative argument is somewhat persuasive: Let's go back to the expression (21) for the energy eigenvalues E_k, and set $B \gg M > 0$; this corresponds to case (2) of lecture 21, and it is clear that the minimum value of $2E_k$ (the "energy gap") will occur for small values of k. For such values eqn. (21) can be simplified to the form

$$E_k = \left\{(M - B'k^2)^2 + A'^2k^2\right\}^{1/2} \qquad (B' \equiv Ba^2/4, A' \equiv Aa) \tag{26}$$

(which is in fact the form used by Bernevig et al. in their 2006 paper).

The locus of reversal of $\mathcal{H}_z$ (inversion contour) occurs on a circle of radius $(M/B)^{1/2}$ (fig. 7), but for $M < A^{'2}$ the minimum value of $2E_k$ (the "energy gap") actually occurs at the origin and is simply equal to M. If now we consider the case $M < 0$, this corresponds to case (1) of lecture 22, i.e. the system is a band insulator and the gap occurs (for $|M| \ll B$) at the origin independently of the value of $A^{'2}$.

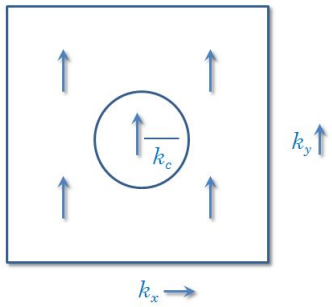

Fig. 7

Thus, if we sweep M from positive to negative values, the transition from a TI to a BI occurs, as anticipated, precisely at the point $M = 0$ at which the gap closes. At this point in the sweep $\mathcal{H}_z = 0$, so the Hamiltonian has the simple "Dirac" form

$$\hat{H}_{Dirac} = A'\boldsymbol{\sigma} \cdot \boldsymbol{k} \equiv A'(k_x\hat{\sigma}_x + k_y\hat{\sigma}_y) \tag{27}$$

with eigenvalues $\pm A|k|$.

Now imagine a spatially extended system, with the parameter M allowed to be a function of the spatial coordinater $\boldsymbol{r}$ and to cross zero on some surface [5] If we imagine that $M(\boldsymbol{r})$ varies slowly enough to allow a semiclassical treatment, we would conclude that close to this surface the Hamiltonian will have approximately the Dirac form (27),

[5] This problem was actually analyzed in some detail by Pankratov and Volkov (Sov. Phys. Uspekhi **29** 579 (1986)) many years before the recent surge of interest in TI's.

and thus does not possess an energy gap: $E(\boldsymbol{k}=0)=0$! Thus, qualitatively, we expect that close to this surface the system will behave as a (semi)-metal. Note also that the surface modes described by (27) have a definite helicity (handedness): if we go back to the original derivation of the Hamiltonian (19) we see that this handedness is specific to the band derived from the state $J_z=+3/2 (S_z=+1/2)$, so that the corresponding state for $S_z=-1/2$ should have $\boldsymbol{k}$ in (27) replaced by $\mathbf{k}^* \equiv (k_x,-k_y)$, i.e. $\hat{H}=A'(k_x\hat{\sigma}_x - k_y\hat{\sigma}_y)$, so opposite helicity.

Finally, we might guess that for the more experimentally interesting case of a real surface of the system with vacuum, qualitatively similar conclusions would apply. However, it is clear that in this case the concept of a component of the wave vector normal to the surface becomes somewhat dubious, so an explicit treatment is needed.

For this purpose, we need to go back to the original real-space Hamiltonian described by the sum of eqns (4) and (15), which it is convenient to write explicitly in terms of the a_i^+'s and b_i^+'s rather than the pseudospin rotation:

$$\hat{H} = M'\sum_i (b_i^+ b_i - a_i^+ a_i) + B\sum_{i,j\in nn}\{(a_i^+ - a_j^+)(a_i - a_j) - (b_i^+ - b_j^+)(b_i - b_j)\} + A\sum_{i,j\in nn}\{(\exp i\theta_{ij})a_i^+ b_j + \text{ H. c.}\} \qquad M' \equiv M - 2B \quad (28)$$

where θ_{ij} indicates the direction of atomic site j relative to site i (cf. fig. 5). Consider a surface (edge) of the 2D atomic array parallel to the y-direction in real space (fig. 8). It is

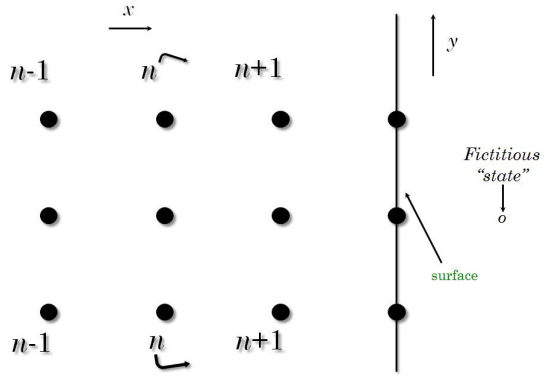

Fig. 8

clear that k_y will be a constant of the motion; for the moment let us set it equal to zero, i.e. $\psi_{n,m}$ is independent of m, the index in the y-direction. Then sitting the (c-number) amplitude a_n equal to $\langle\psi_n|a_n^+|vac\rangle$ (etc.) we obtain by commuting $\hat{H}$ with the operator which creates the excitation ψ_n the pair of equations, for any "n" not on the surface,

$$M'a_n - B(a_{n+1}+a_{n-1}) + A(b_{n+1}+b_{n-1}) = Ea_n \quad (29a)$$

$$-M'b_n - B(b_{n+1} + b_{n-1}) + A(a_{n+1} + a_{n-1}) = Eb_n \tag{29b}$$

Note that the sign of the term in A in the second equation is reversed relative to that of the other terms because of the factor exp $i\theta_{ij}$; this reversal is essential for subsequent conclusions.

In the case where n corresponds to an atom on the surface, the terms in a_{n+1}, etc., are of course absent. Equivalently, we may solve the equations (32) as they stand but require that the physical solution corresponds to $a_{n+1}, b_{n+1} = 0$ for the "fictitious" state $n+1$.

In general, the 2×2 matrix eqn. (32) will have two different solutions corresponding to different eigenvalues E, which cannot be superposed to give a (time-independent) solution which cancels the amplitude on the fictitious state. However, for the special case $E = 0$ there may be two different solutions which can be superposed. To see this, set

$$a_n = a\lambda^n, b_n = b\lambda^n \tag{30}$$

where for normalizability we must require $|\lambda| > 1$ Then eqns. (32) yield

$$(M' - B(\lambda + \lambda^{-1}))a + A(\lambda - \lambda^{-1})b = 0 \tag{31a}$$

$$(M' - B(\lambda + \lambda^{-1}))b + A(\lambda - \lambda^{-1})a = 0 \tag{31b}$$

It is clear that these equations tolerate a solution $a = \pm b$, i.e. with equal amplitudes for the s and p bands. Which sign we choose depends on the choice of sign of A, which is a matter of convention; if for definiteness we choose A to be positive, then b=+a. Then we find two roots for λ

$$\lambda = \frac{1}{2(A+B)} \left\{ -M' \pm \sqrt{M'^2 - 4(B^2 - A^2)} \right\} \tag{32}$$

and by superposing the eigenfunctions with appropriate amplitudes we may be able to cancel the total amplitude of the "fictitious" state $n+1$. However, it is clear that we cannot do this in the case where the expression under the square root is negative, so we may confine ourselves to the case $M'^2 \geqslant 4(B^2 - A^2)$. Then some rather tedious algebra shows that the condition for the two eigenvalues to be both > 1 is

$$M > B > 0 \tag{33}$$

i.e. the inversion contour must lie wholly within the FBZ.

Now consider the case of nonzero k_y. This is equivalent to the case $k_y = 0$ with two modifications:

1. M' is replaced by $M'_{\text{eff}} \equiv M' - B \sin^2(k_y a/2)$
2. There is an additional term in the energy corresponding to the "y-axis nearest-neighbor SO interaction;" it is straightforward to show that this produces a nonzero energy which (for $s_z = 1/2$) is of the "bulk" form

$$E(k_y) = A\sin(k_y a) \tag{34}$$

(Intuitively, we have at the surface an "imaginary" field in the $\hat{\boldsymbol{z}}$-direction whose magnitude is equal to $\mathcal{H}_x$, so $|\boldsymbol{\mathcal{H}}| = \sqrt{\mathcal{H}_z^2 + \mathcal{H}_k^2 + \mathcal{H}_y^2}$ is just $\mathcal{H}_y$). Note that $E(k_y)$ is odd in k_y and so apparently breaks time-reversal invariance. However, the latter is restored by the observation that for the band corresponding to $s_2 = -1/2$ the sign of $E(k_y)$ is reversed, so the pattern looks like fig. 9.

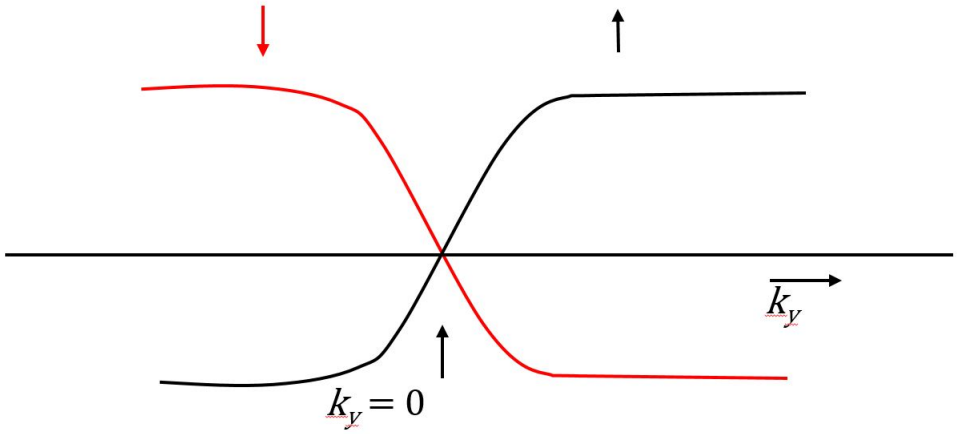

Fig. 9

How far does the surface state extend in k_y? To see this, note that the equations which follow for λ for $k_y \neq 0$ are identical to eqns. (32) except for the replacement of M by $M_{\text{eff}} \equiv M - B\sin^2(k_y a/2)$. For the original ($k_y = 0$) problem the surface state existed provided $B > M > 0$); thus the range of k_y for which the state continues to exist is $M_{\text{eff}} > 0$, i.e.

$$k_y < \frac{2}{a}\sin^{-1}(M/B)^{1/2} \equiv k_c \tag{35}$$

and its maximum energy is therefore A sin $k_c a$. However, we saw earlier (lecture 22) that for $k_y = k_c$ the minimum value of the gap as a function of k_x was $\leqslant A$ sin $(k_c a)$! Hence we conclude that *the surface state must completely cross the band gap* see fig. 10.
An important point is that because the states $(k_y \uparrow, -k_y \downarrow)$ are related by time reversal, *no time-reversal-invariant perturbation* (such as scattering by spinless static impurities) *can mix them*; the situation is somewhat similar to that occurring at the edge of a QH state. To the extent that this is true, one expects, a "surface metal"; since there is only one "transverse" channel (for a given spin) and it moreover has definite helicity, one expects according to the Landauer formula a "spin Hall conductance" of exactly e^2/h.

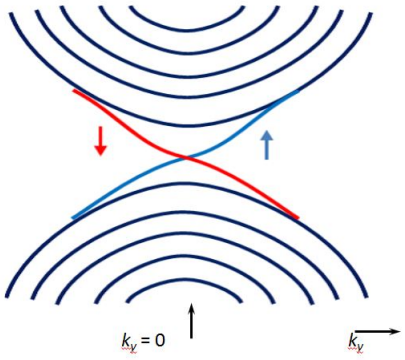

Fig. 10