Spin-orbit Interaction For Electrons in Periodic Solid

Jung Hoon Han

October 23, 2013

Today I want to talk about some aspects of spin-orbit interaction effects in condensed matter systems. The spin-orbit-driven phenomena came to dominate much of the activities in condensed matter physics that have taken place over the 15 years or so, and have led to a striking array of discoveries. To name a few, there is the re-interpretation of the anomalous Hall effect as a spin-orbit-driven Berry phase phenomenon, its cousin effect known as the spin Hall effect, its quantized version that we now call the quantized spin Hall effect. There is the famous generalization of it to three dimensions that led to the prediction and confirmation of three-dimensional topological insulators. A related phenomenon of quantized anomalous Hall effect was also predicted, and recently confirmed.

On the insulator side, a whole class of multiferroic materials exhibiting both electric and magnetic long range orders came into focus due to the interaction between the two orders that allowed (to some extent) the control of one degree of freedom through the manipulation of the other. Such coupling of electronic density (responsible for dipolar LRO) with electronic spin (responsible for spin LRO) would be unthinkable were it not for the underlying spin-orbit interaction coupling the two degrees of freedom.

What makes the spin-orbit-driven phenomena fascinating and yet accessible to most students without a lot of theoretical background is that, despite all its apparent complexity, it is inherently a one-body phenomenon. When we say one-body in condensed matter physics it means that one can neglect the influence of electron-electron interactions caused by their Coulombic charges. So there will be the kinetic energy of the electron plus the periodic potential arising from the lattice of ions, and those two constitute the usual one-body Hamiltonian. There is another term that should have been included in this one-body physics, which is the spin-orbit interaction, that so far has been almost completely neglected in textbook treatments of electron behavior inside the solid. Now, I am certain that situation will gradually change over the years as the forthcoming textbooks will inevitably include discussions of spin-orbit-induced physical phenomena. So what we are going to do, in a sense, is a preview of what the future textbooks on condensed matter physics might look like given what we already know about them. As people are ever more pushing the boundaries there are theories being proposed that include both spin-orbit interaction and the electron-electron interaction. These topics will be omitted in the present lecture. For one reason I do not know too much about them, and for another the level of preparation required to understand them are quite a bit higher than what I have in mind for this audience.

There are roughly two usages of spin-orbit interaction in condensed matter physics these days. I often experience that a person brought up on one perspective of this interaction does not feel comfortable in the language of the other usage, so let me try to lay down both of them here to allow an easy comparison.

Figure 1: Two views of spin-orbit interaction pertinent to condensed matter: (a) free electron-like motion in the plane. (b) Electrons orbiting around the nucleus.

The first view, which is probably more straightforward to grasp, arises from the relativistic physics of a charged particle moving in the plane. A particle of velocity $\mathbf{v}$ sees a part of the electric field as magnetic field and a part of the magnetic field becomes the electric field, according to the principles of relativity. (If you have forgotten about these please refer to the electrodynamics textbooks for a reminder.) Electrons move in two dimensions when they are confined to the surface of a three-dimensional solid, or at the interface of two different materials. If the confinement of the electrons were to take place, there must be a confining electric field prohibiting the electrons to leave the surface region. Therefore the electric field relevant for the electron motion at the surface is pointing in the perpendicular direction to the plane. We call this direction $\hat{z}$. A moving electron with velocity $\mathbf{v}$ then sees this electric field partly as a magnetic field, according to the formula

$$\mathbf{B}' = \mathbf{B} - \mathbf{v} \times \mathbf{E}/c^2 = -\mathbf{v} \times \mathbf{E}/c^2 = -(\hbar\mathbf{k}/mc^2) \times \mathbf{E}. \tag{1}$$

I am employing a CGS unit system in this equation. The electron velocity is replaced by the group velocity $\hbar\mathbf{k}/m$ in the solid with the effective mass m. The procedure is admittedly semi-classical and without a careful justification, but let's keep it at that since a more careful derivation would have to involve some elaborate wave-packet analysis. Remember that electrons carry spin, and the spin sees the $\mathbf{B}'$ as a source of Zeeman energy, splitting the electron energy according to the spin orientation

$$H_{\mathrm{Z}} = -\mu_{\mathrm{B}}\boldsymbol{\sigma} \cdot \mathbf{B}', \tag{2}$$

where $\mu_{\mathrm{B}} = e\hbar/2m$[1] represents the Bohr magneton of the electron. Combining

[1]We are assuming the gyromagnetic ratio $g = 2$ for the electron spin. The charge e is positive here.

both expressions (1) and (2) we find the energy splitting due to relativistic effects takes place as

$$H_{\rm Z} = \left(\frac{\hbar}{mc}\right)^2 \frac{e}{2}\boldsymbol{\sigma}\cdot\mathbf{k}\times\mathbf{E} = (a_{\rm B}\alpha_f)^2\frac{e}{2}\boldsymbol{\sigma}\cdot\mathbf{k}\times\mathbf{E}. \tag{3}$$

Bohr radius and the fine structure constant are introduced as $a_{\rm B}$ and α_f, satisfying $a_{\rm B}\alpha_f = \hbar/mc$. On a typical surface its strength may be estimated as $\sim \alpha_f^2(a_{\rm B}k)(e\mathcal{E}a_{\rm B})$, with $e\mathcal{E}a_B$ of order of the surface work function, eV. For electrons in a solid there is a wide distribution of momentum vector $\mathbf{k}$ with the magnitude reaching up to $\sim \pi/a_{\rm B}$. For "slow" electrons we have the inequality $a_{\rm B}k \lesssim 1$. The estimated Zeeman energy is therefore α_f^2 times the work function at the most, which is admittedly a tiny splitting. The said effect is therefore present, in principle, but hardly observable in practice for the slow-moving, free electrons in the solid.

The other view is more atomistic and involves the concept of electronic orbitals. And this is probably the view you were taught about in advanced quantum mechanics classes. Let's start with a heuristic derivation of the atomic spin-orbit interaction. Unlike with the electrons moving on the plane, the source of electric field for electrons in an atom is the atomic nucleus, that disperses electric field in the radial direction. We may write such a field as $\mathbf{E} = -\hat{r}(dV/dr)$, V being the electrostatic potential. The other difference from the planar motion of the electron is that the electron orbit inside the atom tends to be curved (again, in the semi-classical picture). So instead of introducing the wave vector $\mathbf{k}$ we start from the more generic form, and arrive at the Zeeman Hamiltonian

$$\begin{aligned} H_{\rm Z} = -\mu_B\boldsymbol{\sigma}\cdot\left(-\frac{\mathbf{v}}{c^2}\times\mathbf{E}\right) &= \frac{e\hbar}{2mc^2}\boldsymbol{\sigma}\cdot\mathbf{r}\times\mathbf{v}\left(\frac{1}{r}\frac{dV}{dr}\right) \\ &= \frac{e}{2}(a_{\rm B}\alpha_f)^2\left(\frac{1}{r}\frac{dV}{dr}\right)\boldsymbol{\sigma}\cdot\left(\frac{\mathbf{L}}{\hbar}\right). \end{aligned} \tag{4}$$

The audience must have noticed the rather hand-waving nature of this derivation. As with all such semi-classical derivations, the final form may be right, but not necessarily the factors in front. For one thing the Thomas correction by multiplying the above result by 1/2 must take place in order to get the right, relativistic expression, but that hardly matters because for most elements we don't really know what to write down for dV/dr anyway. Because of the inner product between the spin angular momentum $\boldsymbol{\sigma}$ and the orbital angular momentum $\mathbf{L}$ in the above Hamiltonian, this is commonly called the spin-orbit Hamiltonian, a terminology we shall stick to from now on. I have divided $\mathbf{L}$ by $\hbar$ to make it dimensionless, and in the further discussion below I will re-define $\mathbf{L}/\hbar$ as $\mathbf{L}$ and call it the orbital angular momentum operator.

In a given atom, electrons exist as certain angular momentum eigenstates. When the above Zeeman effect is introduced as a perturbation the true eigenstate becomes something of a mixture between different spin states and different orbital states due to the fact that one can re-write $\boldsymbol{\sigma}\cdot\mathbf{L}$ as

$$\boldsymbol{\sigma}\cdot\mathbf{L} = \sigma^z L^z + \frac{1}{2}(\sigma^+L^- + \sigma^-L^+). \tag{5}$$

The raising and lowering operators for each kind of operator is defined by $\sigma^{\pm} = \sigma^x \pm i\sigma^y$, $L^{\pm} = L^x \pm iL^y$. This operator acts within the finite-dimensional Hilbert space consisting of $2l+1$ kinds of orbital states, where l comes from $\mathbf{L}^2 = l(l+1)$, labeled by m ranging from $-l$ to $+l$, and two kinds of spin states, $\sigma = \uparrow, \downarrow$. This allows us to view the spin-orbit Hamiltonian as a kind of finite-dimensional matrix, provided we are allowed to replace the radial factors by some average. In practice this is almost always done in the condensed matter context, absorbing the average of $1/r(dV/dr)$ over some radial function as a simple constant $\lambda_{\rm so}$. With this gross approximation the spin-orbit interaction in a given atom, in a given orbital state of quantum number l, reads

$$H_{\rm so} = \lambda_{\rm so} \boldsymbol{\sigma} \cdot \mathbf{L}. \tag{6}$$

Keep in mind that the orbital operator $\mathbf{L}$ is to be defined within a particular orbital subspace, say s-, $p-$, or d-orbital spaces.

To have some feeling for what the spin-orbit Hamiltonian does to the physics, let's examine its matrix elements in various orbital spaces. Beginning with the s-orbital, one quickly sees that the s-orbital subspace, consisting of only one orbital state, has zero spin–orbit matrix elements because

$$L^z|s\rangle = 0, \; L^{\pm}|s\rangle = 0. \tag{7}$$

A lesson can be drawn, therefore, that spin-orbit interaction is only effective in a multi-orbital subspace, requiring at least a p-orbital manifold.

Going to the p-orbital subspace, I challenge the audience to work out the following matrix elements of $H_{\rm so}$ valid in the six-dimensional p-orbital basis $(p_{x\uparrow},\; p_{y\uparrow},\; p_{z\uparrow},\; p_{x\downarrow},\; p_{y\downarrow},\; p_{z\downarrow})$:

$$H_{\rm so} = \lambda_{\rm so} \begin{pmatrix} 0 & -i & 0 & 0 & 0 & 1 \\ i & 0 & 0 & 0 & 0 & -i \\ 0 & 0 & 0 & -1 & i & 0 \\ 0 & 0 & -1 & 0 & i & 0 \\ 0 & 0 & -i & -i & 0 & 0 \\ 1 & i & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{8}$$

The above matrix is straightforward to work out if one first writes the orbitals in the angular momentum basis,

$$\begin{aligned} |p_x\rangle &= \frac{1}{\sqrt{2}}(|-1\rangle - |1\rangle), \\ |p_y\rangle &= \frac{1}{\sqrt{2}}(|1\rangle + |-1\rangle), \\ |p_z\rangle &= |0\rangle, \end{aligned} \tag{9}$$

and applied raising and lowering operators to them.

Formerly all six states, three orbitals and two spins, were degenerate in energy. The spin-orbit interaction splits the degeneracy, in a well-known manner, into a group of four $J = 3/2$ states and another group of two $J = 1/2$ states.

The energies can be found without actually diagonalizing the above Hamiltonian, because $H_{\rm so}$ can be re-written as

$$H_{\rm so} = \lambda_{\rm so}\mathbf{J}^2, \quad \mathbf{J} = \mathbf{L} + \boldsymbol{\sigma}/2, \tag{10}$$

minus some constant, and $\mathbf{J}^2$ is either $3/2(3/2+1) = 15/4$ or $\mathbf{J}^2 = 1/2(1/2+1) = 3/4$. There is an observable splitting of the energy levels of $3|\lambda_{\rm so}|$ within the p-orbital subspace of the atomic states. By measuring the energy splitting one can infer the amount, as well as the sign, of the spin-orbit energy $\lambda_{\rm so}$ in a given atom.

This is probably all that has been said in classrooms regarding spin-orbit interaction - in single atoms - if you are a beginning graduate student. In solid state courses, where one learns about the behavior of electrons in a periodic arrangement of atoms, the spin-orbit interaction is still present but is rarely discussed. Among the usual rationale was the smallness of the spin-orbit energy in the majority of elements people in the condensed matter community were interested in, but an even more significant reason to have ignored spin-orbit effect is because of the belief that there is no spectacular effect associated with spin-orbit interaction. Now both excuses are being challenged in recent years, one because of a lot of exciting things being discovered in materials involving heavy elements (such as Bi) where spin-orbit energy becomes a good fraction of an eV that approaches the bandwidth itself, and secondly because some novel physics such as all kinds of exotic Hall effects and topological insulators are due to, and only due to, the spin-orbit interaction.

As you see from the above discussion, spin-orbit physics is inherently a multi-orbital phenomenon. What you need is p- or d-orbital states (or higher angular momentum states) in which different angular momentum orientations are mixed by the spin-orbit interaction. Even for p-orbitals, if the energy difference between the p_z-orbital and the rest of the p-orbitals were so large that electrons practically occupied the p_z-level only, the situation effectively reduces to the single-orbital problem wherein the spin-orbit interaction matrix elements vanish again. This is exactly what happens in graphene band structure, which is formed almost entirely out of electrons in the p_z-orbital state. So now you see why graphene has very little spin-orbit effect. To put it precisely, spin-orbit interactions become effective for multi-orbital bands whose orbital states have nearly degenerate energies.

At first it seems d-orbital spin-orbit problem is much more complicated simply by virtue of the ten-dimensional matrix elements for the spin-orbit interaction. As nature would have it, however, those five d-levels often split up into three nearly degenerate levels called t_{2g} and another two nearly degenerate levels called e_g. The t_{2g} levels refer to the trio of d_{xy}, d_{yz}, d_{zx} orbitals, while e_g consists of $d_{x^2-y^2}$ and $d_{3z^2-r^2}$ orbitals. There is often a substantial energy gap between t_{2g} and e_g levels, with electrons occupying the t_{2g} levels but not the e_g levels. This will then be an ideal situation in which spin-orbit effects can be examined within the subspace of t_{2g} levels. The mathematics of spin-orbit interaction in the t_{2g} is almost identical to that of the p-orbitals. To show this, first re-write the three t_{2g} orbitals in the angular momentum basis

$$
\begin{aligned}
|d_{xy}\rangle &= \frac{i}{\sqrt{2}}(|-2\rangle - |2\rangle), \\
|d_{yz}\rangle &= \frac{i}{\sqrt{2}}(|1\rangle + |-1\rangle), \\
|d_{zx}\rangle &= \frac{1}{\sqrt{2}}(|-1\rangle - |1\rangle),
\end{aligned} \tag{11}
$$

then apply the raising/lowering operators. If done properly you will find the spin-orbit interaction matrix in the t_{2g} subspace ($d_{xy,\uparrow}, d_{yz,\uparrow}, d_{zx,\uparrow}, d_{xy,\downarrow}, d_{yz,\downarrow}, d_{zx,\downarrow}$):

$$
H_{\rm so} = \lambda_{\rm so} \begin{pmatrix}
0 & 0 & 0 & 0 & 1 & -i \\
0 & 0 & i & -1 & 0 & 0 \\
0 & -i & 0 & i & 0 & 0 \\
0 & -1 & -i & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & -i \\
i & 0 & 0 & 0 & i & 0
\end{pmatrix}. \tag{12}
$$

It is useful to keep in mind therefore that an excellent platform to look for significant spin-orbit-derived effects in solids are the materials where the dominant energy bands consists of either p-orbitals, or of t_{2g}-orbitals, without a significant energy splitting among the constituent orbital states. Meanwhile the e_g-level subspace does not have any spin-orbit matrix elements. This is because spin-orbit interaction has finite matrix elements only if the two states are connected by the angular momentum difference of 0, or ± 1. The e_g orbitals are made up of $m = \pm 2$ and $m = 0$ orbitals as

$$
\begin{aligned}
|x^2 - y^2\rangle &= \frac{1}{\sqrt{2}}(|2\rangle + |-2\rangle), \\
|3z^2 - r^2\rangle &= |0\rangle.
\end{aligned} \tag{13}
$$

The matrix elements of $\mathbf{L}$ in the e_g subspace is strictly zero.

Figure 2: A new dogma. (a) Central questions in many-body physics and how they are addressed. (b) Central questions in spin-orbit physics and how they are addressed.

In the traditional many-body approaches that people have worked on over the years, the question of central importance was the fate of the single-particle

spectral function, or the one-particle Green's function, in the face of increasing interaction. To answer this one breaks the Hamiltonian into a non-interacting one H_0, and the interacting part H_1, and try to answer how the spectral function for H_0 alone gets modified or even destroyed as H_1 gradually increases in strength. In the spin-orbit physics, quasi-particles are, by definition, always well-defined because we choose not to deal with the interaction effects. Rather, the question of real significance now is how the wave functions are modified by the inclusion of spin-orbit term, and what it means for the changes in physical properties.

In my view the single most important change induced by the spin-orbit physics is the *complexification* of the wave function. Let me illustrate this by discussing the wave functions of an electron moving in the periodic potential of the solid. A simple model for such motion is afforded by the tight-binding equation

$$-t \sum_{\mathbf{a}} \Psi(\mathbf{r}+\mathbf{a}) = E\Psi_\sigma(\mathbf{r}). \tag{14}$$

for the wave function $\Psi_\sigma(\mathbf{r})$ of either spin orientation $\sigma =\uparrow,\downarrow$. You might wonder "this doesn't look like Schrödinger's equation!!" but the reality is that on the lattice the electrons hop from one atomic site to the next, located at $\mathbf{a}$ apart from one another, instead of moving continuously through space. On a square lattice with lattice spacing of a we should solve

$$-t\left(\Psi_\sigma(\mathbf{r}+a\hat{x}) + \Psi_\sigma(\mathbf{r}-a\hat{x}) + \Psi_\sigma(\mathbf{r}+a\hat{y}) + \Psi_\sigma(\mathbf{r}-a\hat{y})\right) = E\Psi_\sigma(\mathbf{r}). \tag{15}$$

By the Bloch's theorem the solution to any periodic problem like the one above can be found in the form $\Psi_{\mathbf{k},\sigma}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r})$. It is labeled by a quantum number $\mathbf{k}$ (which we know to be the crystal momentum of the electron) and a function $u_{\mathbf{k}}(\mathbf{r})$ which has the periodic property under the translation by any of the lattice translation vectors $\mathbf{a}$: $u_{\mathbf{k}}(\mathbf{r}+\mathbf{a}) = u_{\mathbf{k}}(\mathbf{r})$. By insertion one can easily find what $u_{\mathbf{k}}(\mathbf{r})$ is,

$$-t\left(\sum_{\mathbf{a}} e^{i\mathbf{k}\cdot\mathbf{a}}\right) u_{\mathbf{k}}(\mathbf{r}) = E_{\mathbf{k}}u_{\mathbf{k}}(\mathbf{r}), \tag{16}$$

which gives the energy $E_{\mathbf{k}} = -2t(\cos k_x a + \cos k_y a)$ in the case of square lattice. As to the wave function, it is simply given by a constant, $u_{\mathbf{k}}(\mathbf{r}) = 1$.

If you looked into a particular condensed matter question, the part that really governs the properties of the periodic solid is not the plane-wave part $e^{i\mathbf{k}\cdot\mathbf{r}}$, but the periodic function $u_{\mathbf{k}}(\mathbf{r})$. To motivate this statement physically, the plane-wave $e^{i\mathbf{k}\cdot\mathbf{r}}$ is merely a consequence of the periodic nature of the solid, regardless of the details of the particular solid in question. All the material-specific information is therefore encoded in the periodic part, $u_{\mathbf{k}}(\mathbf{r})$, which in the above example was found to be simplest function imaginable - a constant.

The situation changes dramatically if magnetic field is imposed on the electron motion. The magnetic field creates the well-known Aharonov-Bohm (AB) phase on the electron wave function, giving rise to phase-dependent hopping.

What this means is that as the electron hops from site $\mathbf{r}$ to site $\mathbf{r}+\mathbf{a}$, the phase of the electron wave function picks up an extra factor

$$\exp\left(i\frac{q}{\hbar}\int_{\mathbf{r}}^{\mathbf{r}+\mathbf{a}}\mathbf{A}\cdot d\mathbf{r}\right)\equiv e^{i\phi(\mathbf{r}+\mathbf{a},\mathbf{r})}. \tag{17}$$

The tight-binding equation of motion is modified to reflect this phase factor:

$$-t\sum_{\mathbf{a}}e^{i\phi(\mathbf{r}+\mathbf{a},\mathbf{r})}\Psi_\sigma(\mathbf{r}+\mathbf{a})=E\Psi_\sigma(\mathbf{r}). \tag{18}$$

Due to the position-dependent, complex-valued phase factor one can no longer obtain the solution by a simple guesswork based on Bloch's theorem. The problem can still be solved, by employing the idea of magnetic translational symmetry (instead of ordinary translational symmetry) inherent in the Eq. (18). The idea is that if the magnetic flux in a single plaquette of the lattice (a square of size a^2 if it were a square lattice) was equal to a rational fraction of 2π,

$$\sum_{plaquette}\phi(\mathbf{r}+\mathbf{a},\mathbf{r})=2\pi\frac{p}{q}, \tag{19}$$

then one can always treat a problem like Eq. (18) as the electron motion subject to the periodic potential of period q times the lattice constant. Hence one can again apply Bloch's theorem to write the general solution as $e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r})$ and find what $u_{\mathbf{k}}(\mathbf{r})$ is. This time, however, $u_{\mathbf{k}}(\mathbf{r})$ becomes both position-dependent and complex, due to the complex phase factors. With the magnetic field and the accompanying AB phases, the electron motion is only described with the complex-valued wave function.

Figure 3: Harper's model for electron hopping on a flux-threaded lattice. It is the first microscopic model in which the exact quantization of Hall conductance was demonstrated. The same model is romantically associated with the Hofstadter's butterfly.

On a historical footnote, Eq. (18) is called the Harper's equation. The energy spectrum of the Harper's equation on a two-dimensional square lattice was first obtained numerically by Douglas Hofstadter in his Ph. D. thesis work. In a Phys. Rev. B paper published in 1976 he demonstrated the energy spectrum plotted against the flux ϕ, which revealed one of the first examples of self-similar

structures in physical systems. The famous butterfly pattern of the energies he observed is now called the Hofstadter butterfly. He moved on and out of physics to compose such classic as *Gödel, Escher, Bach*, while David Thouless, during his visit to University of Oregon where Hofstadter was a Ph. D. student, heard about this from Hofstadter's thesis supervisor Gregory Wannier and took keen interest in it. When the discovery of integer quantum Hall effect was announced, Thouless looked for the origin of the conductance quantization in the context of Harper's equation, so he studied the Hall response of the Harper model and found the now-famous Thouless-Kohmoto-den Nijs-Nightingale (TKNN) formula for quantized Hall conductance. You might be surprised that a simple equation like (18) is responsible for two of the most famous jargons in modern condensed matter physics: Hofstadter's butterfly and TKNN number for integer quantum Hall effect.

Beginning around the new millennium people, in particular Qian Niu (a star pupil of David Thouless) and Naoto Nagaosa, began to realize that another systematic way to guarantee the complexity of the wave function was to introduce spin-orbit interaction $H_{\rm so}$ into the problem. This time, one doesn't need the external magnetic field, so the time-reversal symmetry remains unbroken, but the wave function still become complex-valued. The reason for the complexity can be understood by analyzing the tight-binding problem which includes $H_{\rm so}$. Recall that the spin-orbit effect mixes different spins and different orbitals, and therefore the wave function must become multi-component. In the case of p-orbital problem it would include six different wave functions covering three orbitals and two spins. In short, we write

$$\Psi(\mathbf{r}) = \big(\psi_{x,\uparrow}(\mathbf{r}), \psi_{y,\uparrow}(\mathbf{r}), \psi_{z,\uparrow}(\mathbf{r}), \psi_{x,\downarrow}(\mathbf{r}), \psi_{y,\downarrow}(\mathbf{r}), \psi_{z,\downarrow}(\mathbf{r})\big)\,. \tag{20}$$

The Schrödinger problem satisfied by this enlarged wave function is

$$-t\sum_{\mathbf{a}} \Psi(\mathbf{r}+\mathbf{a}) + \lambda_{\rm so}(\boldsymbol{\sigma}\cdot\mathbf{L})\Psi(\mathbf{r}) = E\Psi(\mathbf{r}). \tag{21}$$

Do the same division of the wave function $\Psi(\mathbf{r})$ into the plane-wave part $e^{i\mathbf{k}\cdot\mathbf{r}}$ and the periodic part $u_{\mathbf{k}}(\mathbf{r})$, except that now $u_{\mathbf{k}}(\mathbf{r})$ will itself carry six components, the problem is reduced to

$$\left(-t\sum_{\mathbf{a}} e^{i\mathbf{k}\cdot\mathbf{a}}\right) u_{\mathbf{k}} + \lambda_{\rm so}(\boldsymbol{\sigma}\cdot\mathbf{L})u_{\mathbf{k}} = E_{\mathbf{k}}u_{\mathbf{k}}. \tag{22}$$

The position dependence has dropped out from $u_{\mathbf{k}}$. To fully solve the problem there is the remaining step to diagonalize the spin-orbit matrix, as given previously in Eq. (8). The elements of the spin-orbit matrix contain some imaginary numbers, which forces some of the coefficients in the eigenstate $u_{\mathbf{k}}$ to also become complex. Final solution for the spin-orbit problem will read something like

$$\Psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}\left(\sum_{m=x,y,z}\sum_{\sigma=\uparrow,\downarrow}\alpha_{m,\sigma}\phi_m(\mathbf{r})|m,\sigma\rangle\right). \tag{23}$$

The basis $|m,\sigma\rangle$ refers to the state of particular orbital m and spin σ, $\phi_m(\mathbf{r})$ is the orbital wave function, and $\alpha_{m,\sigma}$ are the complex-valued coefficients diagonalizing Eq. (22). This way it is clear that the periodic part of the Bloch state in the spin-orbit-coupled bands are inherently complex in character.

So why are complex-valued wave functions any more interesting than real-valued ones? It turns out that complex-valued wave functions give rise to the concept of "connection" while the real-valued counterparts do not. The connection is a quantity defined by

$$A_\mu(\mathbf{k}) = -i\langle u_\mathbf{k}|\partial_\mu|u_\mathbf{k}\rangle = -iu_\mathbf{k}^\dagger \frac{\partial u_\mathbf{k}}{\partial \mathbf{k}}. \tag{24}$$

Instead acting on the space coordinate $\mathbf{r}$, the derivative $\partial_\mu \equiv \partial/\partial k_\mu$ acts on the momentum index $\mathbf{k}$ of the eigenstate. In some ways it captures how the wave functions defined at adjacent quantum numbers $\mathbf{k}$ and $\mathbf{k}+\Delta\mathbf{k}$ are "connected" to each other. It is straightforward to show that this connection vanishes identically if the wave function $u_\mathbf{k}$ was real-valued. Even though $u_\mathbf{k}$ is complex, the corresponding connection is always real-valued. The connection defined in this way behaves much like the vector potential of electromagnetism, except that this new kind of vector potential resides in the momentum space rather than in the real space. Nevertheless, since mathematics couldn't care less whether the variables being adopted refer to real or momentum space, one might as well push the analogy further to define the "magnetic field" $\mathbf{B}(\mathbf{k})$ in momentum space according to the usual rule of electromagnetism:

$$\mathbf{B}(\mathbf{k}) = \boldsymbol{\nabla}_\mathbf{k} \times \mathbf{A}(\mathbf{k}). \tag{25}$$

Note how I promoted the $\mathbf{k}$ vector from being an index, labeling different eigenstates in the solid, to being a variable. In momentum space and for the physics that arises within it, $\mathbf{k}$ is indeed the pertinent dynamical variable.

Finally, integrating the magnetic field over a certain two-dimensional area will give rise to the concept of magnetic flux. Except, in this case the area to be integrated over is the Brillouin zone (BZ). Take, for instance, the two-dimensional solid (with the corresponding two-dimensional Brillouin zone) and integrate $B_z(\mathbf{k}) = \partial_{k_x} A_y(\mathbf{k}) - \partial_{k_y} A_x(\mathbf{k})$ over the BZ:

$$\frac{1}{2\pi}\int_{\mathrm{BZ}} dk_x dk_y B_z(\mathbf{k}). \tag{26}$$

Up to this point, despite all its formal fascination, what I did was simply to develop an analogy to standard electromagnetism. It offers no clue for us to believe that this has anything to do with physics. Then comes the surprising discovery that dates back to Thouless and co-workers in 1982, that this integral, from Stokes' theorem and due to the uniqueness of the wave function, must always be an integer, and furthermore, is related to the Hall conductance σ_{xy} through the relation

$$\sigma_{xy} = \frac{e^2}{h}\left(\frac{1}{2\pi}\int dk_x dk_y B_z(\mathbf{k})\right). \tag{27}$$

Actually, the idea that the curl of the connection (called the curvature) integrated over a closed space must be an integer, a topological number, was well-known in the mathematics community from the work of mathematician S. S. Chern in 1940s, and it is even named the first Chern number. What was not known until 1982 was that this same number had a direct physical manifestation, as the Hall conductance of a two-dimensional solid. The integer quantization of Hall conductance discovered by Klaus von Klitzing is thus a consequence of the macroscopic solid object displaying the Chern quantization in differential geometry. When this formula and the beautiful connection of topology to quantized Hall effect were discovered in 1982, the needed complexity of the wave function was provided by the huge external magnetic field. About 20 years later people began to realize that the "emergent magnetic field" in momentum space need not come from external magnetic field at all. It can equally well come from the spin-orbit interaction!

Figure 4: A recipe for producing spontaneous Hall effect without an external magnetic field.

I have argued that spin-orbit effects generically lead to some non-zero connection in momentum space. Of course all that means is that $\mathbf{B}(\mathbf{k})$ need not be zero at any point in $\mathbf{k}$-space. The measured Hall conductance, on the other hand, is the sum of this number over all the occupied $\mathbf{k}$-states, so if there were some kind of symmetry in the ground state such that the sum vanished, we wouldn't get any Hall effect. For ground states that are invariant under the time-reversal operation one can easily prove that σ_{xy} must be zero. The way this comes about is, for example, if $B_z(-\mathbf{k}) = -B_z(\mathbf{k})$ and one always had an equal occupation of $+\mathbf{k}$ and $-\mathbf{k}$ states. This symmetry gets broken once the time-reversal symmetry is lifted. One way to do it is, again, by imposing external magnetic field. The second way, as noted independently by Nagaosa and Qian Niu among others around 2001, is by having spontaneous macroscopic magnetization, breaking the time-reversal symmetry by providing a net internal magnetic field. So you add on the spin-orbit effect in a typical ferromagnet, and find that a spontaneous Hall effect is predicted. Indeed, the first microscopic Hamiltonian shown to exhibit Hall effect without an external magnetic field was of the form

$$H = H_0^{(t_{2g})} + H_{\rm so}^{(t_{2g})} - J \sum_{m=xy,yz,zx} \mathbf{S} \cdot (\psi_m^\dagger \boldsymbol{\sigma} \psi_m). \tag{28}$$

The Hall effect taking place without the imposition of external magnetic field

is known as anomalous Hall effect since its first observation in 1950s. Historically there have been several proposals for explaining the anomalous Hall effect, until more recently the momentum-space flux idea gained recognition. In this new picture, the underlying band structure in its non-magnetic phase (say above Curie temperature) was already in a stage of high momentum flux density across the band; application of the symmetry-breaking field in the form of spontaneous magnetization then tips the subtle balance to produce nonzero average flux when integrated over the occupied states in the momentum space. This is the origin of anomalous Hall effect, along with other historically well-known sources such as side jump and skew scattering mechanisms. A poetic way of saying this is that a spin-orbit-coupled material is already on the verge of producing spontaneous momentum-space flux, and all one needs is some symmetry-breaking perturbation to swing the system into a state carrying non-zero Hall conductivity. This is also why the element Bismuth is such a heavily diamagnetic material. It is an extremely heavy element, with huge spin-orbit interaction (in fact has the largest spin-orbit interaction among the non-radioactive materials), and carrying huge momentum-space flux density in its band structure. Application of the external magnetic field will immediately induce substantial amount of Hall response σ_{xy} in the material, which induces large amount of diamagnetic current around the boundary, and makes the object made of Bismuth float.

Figure 5: (a) Schematic picture of the band structure in momentum space for periodic lattice. (b) Modified picture of the band structure with non-zero flux density. (c) Uneven distribution of the flux density brought by time-reversal symmetry breaking. (d) Now we know why Bismuth floats under the magnetic field.

In summary, the conventional picture of energy bands in periodic solid as curves representing the energy-momentum relation is to be supplemented by the additional quantity representing the distribution of flux densities at each

momentum point. The revised energy-momentum and flux distribution curves may look like the one shown in Fig. 5. Spontaneous magnetization induced on one of those bands will lead to topologically driven Hall effect.

One may wonder if anything interesting can be said about the case with finite distribution of flux densities $B_z(\mathbf{k}) \neq 0$ but with zero average $\int_{\rm BZ} B_z(\mathbf{k}) = 0$. Is there any interesting, observable phenomena that might take place? For one thing I strongly suspect that a local magnetic impurity, placed at random sites of the solid, may nucleate finite flux, in real space, around the site of the impurity. The flux thus induced may lead to Landau-level quantization of energy levels locally, as in the case of graphene bubble, which may be detected in an STM probe. It might even affect the way the local magnetic moment interacts with the surrounding conduction spins and modify the Kondo coupling behavior. These are just my guesses, but you may go ahead and work them out. More realistically, a striking phenomenon known as the spin Hall effect takes place when the average flux is zero in the band, but something else is non-zero.

To understand the origin of topologically driven spin Hall effect, we need to assume that two bands, for some reason, are completely degenerate. This is not as unreasonable as it seems. GaAs has a band structure captured by the Hamiltonian

$$H \sim -(\mathbf{k}\cdot\mathbf{J})^2 \tag{29}$$

where $\mathbf{J}$ is the total spin $J = 3/2$ operator. The three p-orbitals and two spin states of the Ga valence band are re-grouped as four $J = 3/2$ and two $J = 1/2$ states, and the four $J = 3/2$ states form a band whose dispersion is captured by the above effective Hamiltonian, valid around the $\mathbf{k} = 0$ (Γ) point. The energy levels are two-fold degenerate, at $-\mathbf{k}^2/4$ and $-9\mathbf{k}^2/4$, respectively. We can take a pair of eigenstates whose energy is, for example, $-\mathbf{k}^2/4$ and write them as $|1,\mathbf{k}\rangle$ and $|2,\mathbf{k}\rangle$. Because spin-orbit coupling is at work in producing the effective Hamiltonian (29), both states should have complex-valued wave functions. We may try to go ahead and write down the connections between adjacent $\mathbf{k}$-vectors, but there occurs a problem that didn't exist before with the non-degenerate band. For each $\mathbf{k}$ there are two states of identical energy, and any linear combination of them is also an eigenstate, so which state do we choose to compute the connection? Because of energy degeneracy, the only reasonable choice appears to be the most democratic one, written as a 2×2 matrix:

$$\mathcal{A}_\mu(\mathbf{k}) = -i \begin{pmatrix} \langle 1,\mathbf{k}|\partial_\mu|1,\mathbf{k}\rangle & \langle 1,\mathbf{k}|\partial_\mu|2,\mathbf{k}\rangle \\ \langle 2,\mathbf{k}|\partial_\mu|1,\mathbf{k}\rangle & \langle 1,\mathbf{k}|\partial_\mu|1,\mathbf{k}\rangle \end{pmatrix}. \tag{30}$$

This is what the particle physicists would call a non-Abelian connection, or non-Abelian gauge potential. Then there is the non-Abelian flux density, to be derived from the above by the usual rule of non-Abelian gauge theory:

$$\mathcal{B}_{\mu\nu}(\mathbf{k}) = \partial_\mu \mathcal{A}_\nu(\mathbf{k}) - \partial_\nu \mathcal{A}_\mu(\mathbf{k}) + i[A_\mu(\mathbf{k}), A_\nu(\mathbf{k})]. \tag{31}$$

(A lot of us in condensed matter community get confused at first by this expression. It takes some time to absorb the idea that for each space-time component $\mu = t, x, y, z$ there is an associated 2×2 matrix, rather than a number.)

The quantized Hall effect discussed earlier was a consequence of a miraculous coincidence between the measurable quantity, the Hall conductance, and a certain topological number defined long time ago in mathematics called the first Chern number. The coincidence was established by the linear response calculation. (Linear response theory tells us how to compute electrical, magnetic, and thermal susceptibilities in a condensed matter system.) With regard to the non-Abelian magnetic field just defined above, what physical quantity in the linear response theory would correspond to an integral of it over, say, the Brillouin zone? Unfortunately, the answer is not as rigorous here as it was in the theory of Hall conductance. The mathematical rigorous answer, it turns out, can only be found in four dimensions and is known as the second Chern number. For the condensed matter systems that have the spatial dimension at most equal to three, such ideal results seem, well, too ideal to have much real-life consequences.

Figure 6: Schematic plot of the degenerate band structure and the opposite flux density distributions that give rise to spin Hall effect.

What we know to be true, and it's still very useful, is that if we can somehow neglect the off-diagonal components of the non-Abelian vector potential, and further show that the two diagonal components have the opposite sign, then we can define $a_\mu(\mathbf{k}) = -i\langle 1, \mathbf{k}|\partial_\mu|1, \mathbf{k}\rangle$ and write the non-Abelian potential as

$$\mathcal{A}_\mu(\mathbf{k}) = a_\mu(\mathbf{k})\sigma^z. \tag{32}$$

The appearance of the Pauli matrix σ^z suggests, quite obviously, that the two overlapping bands have the opposite Hall conductances, $\sigma_{xy}^{(1)} = -\sigma_{xy}^{(2)}$. The sum is zero, but the difference is not, $\sigma_{xy}^{(1)} - \sigma_{xy}^{(2)} \neq 0$, and this gives rise to the notion of spin Hall conductance as opposed to the charge Hall conductance. What will happen is that under the application of the electric field along the x-direction, Hall current develops in the y-direction, but the flow direction is $+y$ for spin-up electrons and $-y$ for spin-down electrons. So the net charge flow is zero, but the net spin flow gets established.

The strict degeneracy of energy levels is not a pre-requisite for the spin Hall effect to be observable. One can imagine a material having lots of bands near the Fermi level all with complex-valued wave functions due to significant spin-orbit interaction. Some bands have more of a spin-up character and have one

sign of emergent magnetic field while other bands have opposite spin character and an opposite sign of the magnetic field. Averaging over all these bands will give rise to some non-zero spin Hall effect. The issue is that of finding the right material and being able to engineer it in the right way in order to maximize the spin Hall signal. The spintronics community has made a lot of progress in this direction recently. For the topologically driven spin Hall phenomena the basic picture is as given in Fig. 6 - pair of degenerate bands with opposite flux densities.

Basically the same picture applies to the quantized spin Hall systems. Whether some materials behave as topological spin Hall system, or a quantized spin Hall system, all depends on the how the various hopping processes are arranged on a lattice. In other words, going back to Eq. (22), the hopping part of the Hamiltonian can be carefully arranged in such a way to produce various spin Hall systems. Such developments culminated in the prediction, and verification, of the three-dimensional topological insulators. This is a beautiful and deep subject which I have no time to cover. In the next talk you will hear about a particularly beautiful consequence of having a topologically non-trivial bulk band structure - the emergence of novel quasiparticles called the Majorana fermions, at the edge of the solid.

Figure 7: (a) ISB for heterogeous arrangement of atoms on hexagonal lattice. (b) ISB of geometric origin for atoms on the top layers of the solid.

Everything I have said so far referred to changes in the electronic properties with the addition of spin-orbit interaction in metals and insulators. In classifying solids one looks into various crystal symmetries or lack thereof in a given solid. An important class of symmetries of solid is the inversion symmetry. It means, if you choose the origin wisely, any atom in a solid located at $\mathbf{r}$ is paired by the same atom at its inversion point, $-\mathbf{r}$. All Bravais lattices in two dimensions satisfy this symmetry, provided every site is occupied by the same atom species. Such would not be the case if, for example, the hexagonal lattice is alternatively occupied by A and B atoms on adjacent sites. Then no matter how one chooses the origin, an atom A at position $\mathbf{r}$ is met by an atom B at $-\mathbf{r}$, but not by the same atom. In this case the inversion symmetry is broken by the atomic arrangement in a solid. There is another, more geometric origin of inversion symmetry breaking taking place at all surfaces of the solid. An atom at the top layer of the solid sees a vacuum at $z > 0$ but other atoms for $z < 0$. The top

and the bottom half of the space are geometrically different for the atoms on the top few layers of any solid.

What happens if we imposed spin-orbit interaction $H_{\rm so}$ on a crystal that lacks the inversion symmetry? Such question is of general relevance to all of surface science, to a degree that will depend sensitively on the strength of spin-orbit interaction as well as the extent of inversion symmetry breaking (ISB). The amount of ISB is roughly measured by the confining electric field which acts on the surface electrons. You multiply this number by a rough estimate of the surface layer depth, and you will arrive at a number comparable to the work function. For typical surfaces of metals and semiconductors this is of order a few eV regardless of the exact types of atoms forming the surface layer. There is a very different story with the strength of spin-orbit interaction energy, which may vary from tens of meV for $3d$ transition elements to a good fraction of an electron volt for the heaviest elements like Pt ($Z = 78$) and Bi ($Z = 83$). The combined influence of ISB at the surface and the spin-orbit interaction is the well-known Rashba interaction coupling electron spins to its momentum. The Rashba interaction was already written down in Eq. (3) as an example of spin-orbit coupling phenomena arising from relativistic effects. Earlier I also showed that the conventional free-electron argument fails to capture the right order of magnitudes for the observed Rashba effects in real solids. For the sake of argument we ignore this inconsistency and write down a simple, phenomenological model for electrons with Rashba interaction:

$$H = \frac{\hbar^2 \mathbf{k}^2}{2m} + \alpha_{\rm R} \boldsymbol{\sigma} \cdot (\mathbf{k} \times \hat{z}). \tag{33}$$

The Rashba constant $\alpha_{\rm R}$ is to be treated as a phenomenological constant, to match the experimental values and so on. Since the first term in the Hamiltonian does not care about the spin orientation, one can choose the eigenstate of the combined Hamiltonian as the spin coherent state $|\pm \mathbf{n}_{\mathbf{k}}\rangle$ satisfying

$$(\hat{n}_{\mathbf{k}} \cdot \boldsymbol{\sigma})|\pm \hat{n}_{\mathbf{k}}\rangle = \pm|\pm \hat{n}_{\mathbf{k}}\rangle, \quad \hat{n}_{\mathbf{k}} = \hat{\mathbf{k}} \times \hat{z}. \tag{34}$$

Such states have their averaged spin orientation either parallel, or anti-parallel to the direction $\hat{n}_{\mathbf{k}} = \hat{\mathbf{k}} \times \hat{z}$ and have their energies split by

$$\frac{\hbar^2 \mathbf{k}^2}{2m} \pm \alpha_{\rm R} |\mathbf{k}|. \tag{35}$$

This gives rise to a modified band structure as depicted in Fig. 8. The split energy levels are, in principle, visible in the spectroscopic probe such as ARPES. In practice it took a while for the resolution of the ARPES to be high enough to resolve the relatively meager energy difference of the two Rashba-split bands. The first such observation was made in 1996 by La Shell *et al*, on the surface bands of gold. Since then, other materials with significantly larger Rashba splitting have been found, most of them involving the heavy element Bi.

The limit of large Rashba splitting raises an interesting possibility. Suppose the Rashba constant $\alpha_{\rm R}$ is so large that the two bands split by their respective spin chiralities are essentially an infinity apart. Then only one species

of electrons carrying one sign of chirality will dominate the electronic properties, forming what one may regard as a chiral electron liquid. In the world of elementary particles massless particles have this one-to-one correspondence of momentum to chirality. The behavior of massless elementary particles are nicely mimicked by the massive electrons in solids moving under the heavy influence of Rashba interaction.

Figure 8: Rashba-split electronic band structure.

This leads us to an inevitable question of practical importance. How do we understand the large value of the Rashba parameter $\alpha_{\rm R}$ found in real solids, and what can we do to beef up its strength to obtain genuinely chiral matter? A recent development showed that multi-orbital physics is as critical to the large energy scale of the Rashba splitting as the spin-orbit interaction and the inversion symmetry breaking electric field. I will talk more about this new and improved (and hopefully correct) view of the Rashba splitting tomorrow morning.

The other topic of much importance, which I must omit here for lack of time, is the issue of what happens to spin systems, as opposed to electronic systems, in possession of spin-orbit interaction and the inversion symmetry breaking. The resulting Dzyaloshinskii-Moriya interaction among the spin degrees of freedom gives rise to a fascinating array of new phenomena ranging from spiral spins to Skyrmion lattice. Those topics will be covered to some extent in several of the invited talks on Thursday afternoon.

There is little doubt in my mind that spin-orbit-dominated physics will continue to be an integral part of the condensed matter physics activities over the next few years. It is my hope that a lecture like this will motivate some of you to start thinking about the subject and come up with some original ideas of your own. Thank you.