A high-order procedure for computing globally optimal Wannier functions in one-dimensional crystalline insulators

Abinand Gopal and Hanwen Zhang

A standard task in solid state physics and quantum chemistry is the computation of localized molecular orbits known as Wannier functions. In this manuscript, we propose a new procedure for computing Wannier functions in one-dimensional crystalline materials. Our approach proceeds by first performing parallel transport of the Bloch functions using numerical integration. Then, using novel analysis, we show that a simple correction can be analytically computed that yields the optimally localized Wannier function. The resulting scheme is robust and capable of achieving high-order accuracy. We illustrate this in a number of numerical experiments.

**A high-order procedure for computing globally optimal
Wannier functions in one-dimensional crystalline insulators**

Abinand Gopal[†], Hanwen Zhang [‡]

September 19, 2024

[†] Department of Mathematics, UC Davis, Davis, CA 95616
[‡] Department of Mathematics, Yale University, New Haven, CT 06511

1   Introduction

Wannier functions are localized molecular orbitals in crystalline materials. They provide a convenient basis for many tasks in electronic structure calculations, such as when using the tight-binding approximation and understanding the conductivity and polarization properties of certain materials (cf. [20]). As such, the computation of Wannier functions is of interest in the development of new materials, such as semiconductors, solar cells, and topological insulators.

From the point of view of mathematics, the problem amounts to solving a series of parameter-dependent eigenvalue problems, such that the eigenvectors are smooth as a function of the parameter. The most common approach is to minimize the Marzari–Vanderbilt functional using some form of gradient descent with high quality initial guesses computed by density-matrix based methods [6, 7, 16]. In this manuscript, we propose an alternative approach that requires no iterative optimization. Instead, we solve the parallel transport problem via numerical integration to determine the Bloch functions and band functions. We then show that a phase correction yields the maximally-localized Wannier function. The analysis in this paper for exponential localization is based on Kato's analytic perturbation theory and is related to the work [5]. The identification of the global optimal solution is a simple consequence of the analysis.

The algorithm in this paper is closely related to the one dimensional case in [4, 15]. However, this approach differs from these existing works in three crucial aspects. First, we provide a complete numerical procedure that starts with the Schrödinger equation and ends with the evaluation of the Wannier function. Second, this approach is high-order convergent and amendable to further improvement by switching to arbitrarily higher-order ODE solvers. Finally and most significantly, we are able to directly compute a provably maximally localized Wannier function without any type of iterative optimization. We refer the reader to Remark 4.1 for further details.

This is the first in a series of papers; in subsequent papers we will consider the multiband case in one dimension, and single and multiband cases in higher dimensions.

We now outline the remainder of this manuscript. In Section 2, we review some background material on Wannier functions. This is followed by Section 3 where we develop the analytic apparatus for constructing globally optimal Wannier functions. After this are Sections 4 and 5 where we describe our numerical procedure. Our numerical procedure is illustrated in Section 6 which contains numerical experiments. We provide our conclusions and outline future work in Section 7.

2   Physical and Mathematical Preliminaries

2.1   Notation

We let $\mathbb{N}$ denote the positive integers. We let δ_{mn} denote the Kronecker delta function, defined by

$$\delta_{mn} = \begin{cases} 1 & m = n \\ 0 & m \neq n \end{cases}.$$

(1)

We let $\ell^2(\mathbb{Z})$ denote the Hilbert space for complex sequences $\mathbf{x} = \{x_i\}_{i=-\infty}^{\infty}$ such that $\sum_{i=-\infty}^{\infty} |x_i|^2 < \infty$ with the standard inner product. We denote the right shift operator in $\ell^2(\mathbb{Z})$ by R, where

$$R(\mathbf{x}) = \{x_{i-1}\}_{i=-\infty}^{\infty}.$$

(2)

For a vector $\mathbf{x} \in \ell^2(\mathbb{Z})$ or $\mathbb{C}^N$, we denote by $\mathbf{x}^*$ the conjugate transpose and by $\|\mathbf{x}\|$ the standard 2-norm, $\|\mathbf{x}\| = \sqrt{\mathbf{x}^*\mathbf{x}}$.

Given sets $A \subset \mathbb{C}$ and $B \subset \mathbb{C}$, we write $f \in C(A; B)$ if $f : A \to B$ is a continuous function. Similarly, we write $f \in C^n(A; B)$ if f is a n-times continuously differentiable. For an $\mathbf{A}$ as an operator in $\ell^2(\mathbb{Z})$ or a matrix in $\mathbb{C}^{M \times M}$, we let $\mathbf{A}^\dagger$ denote its Moore–Penrose pseudoinverse.

We choose the principal branch of the log function so that we assume all real phases ϕ in $e^{\mathrm{i}\phi}$ take value in $(-\pi, \pi]$.

2.2 Bloch functions

In this section, we introduce Bloch functions in one dimension. These are standard facts in the theory of bands and can all be found in [14], for example.

We first define a periodic lattice with a period (lattice constant) a

$$\Lambda = \{na : n \in \mathbb{Z}\}. \tag{3}$$

We refer to the interval $[-a/2, a/2)$ as the primitive unit cell:

$$I_{\mathrm{uc}} = [-a/2, a/2). \tag{4}$$

It is also convenient to define

$$\Omega = \frac{2\pi}{a} \tag{5}$$

as the periodic for the reciprocal lattice

$$\Lambda^* = \{n\Omega : n \in \mathbb{Z}\}. \tag{6}$$

The interval $[-\Omega/2, \Omega/2)$ is usually called the first Brillouin zone (BZ), and we denote the interval by I_{bz} :

$$I_{\mathrm{bz}} = [-\Omega/2, \Omega/2). \tag{7}$$

We assume that $V : \mathbb{R} \to \mathbb{R}$ is a piecewise continuous function that has a period a:

$$V(x + y) = V(x), \quad x \in \mathbb{R}, \ y \in \Lambda. \tag{8}$$

For any $k \in \mathbb{R}$, the relevant Schrödinger equation is given by the eigenvalue problem

$$-\frac{\mathrm{d}^2\psi_k^{(j)}}{\mathrm{d}x^2} + V(x)\psi_k^{(j)} = E_k^{(j)}\psi_k^{(j)}, \tag{9}$$

subject to the boundary condition

$$\psi_k^{(j)}(x + y) = e^{iky}\psi_k^{(j)}(x), \quad y \in \Lambda. \tag{10}$$

The value k is called the quasimomentum, $\{\psi_k^{(j)}\}_{j=1}^\infty$ in (9) are called the Bloch functions, and $\{E_k^{(j)}\}_{j=1}^\infty$ are called the energy of the states represented by $\{\psi_k^{(j)}\}_{j=1}^\infty$.

Bloch's theorem states two simplifications of the above problem. First, it is sufficient to only consider $k \in I_{\mathrm{bz}}$, the first BZ. Second, the Bloch functions are of the form

$$\psi_k^{(j)}(x) = e^{\mathrm{i}kx} u_k^{(j)}(x), \tag{11}$$

for some function $u_k^{(j)}$ where

$$u_k^{(j)}(x+y) = u_k^{(j)}(x), \quad y \in \Lambda. \tag{12}$$

Substituting (11) into (9), shows $u_k^{(j)}$ satisfies

$$\left[-\left(\frac{\mathrm{d}}{\mathrm{d}x} + \mathrm{i}k \right)^2 + V \right] u_k^{(j)} = E_k^{(j)} u_k^{(j)}, \tag{13}$$

subject to the boundary condition in (12). It is convenient to denote the operator in (13) densely defined in $L^2(I_{\mathrm{uc}})$ by

$$H(k) = -\left(\frac{\mathrm{d}}{\mathrm{d}x} + \mathrm{i}k \right)^2 + V. \tag{14}$$

For any real piecewise continuous function V, $H(k)$ is self-adjoint for real k and the operator has the symmetry

$$H(k) = \overline{H}(-k) \tag{15}$$

for real k. This symmetry is referred to as time-reversal symmetry in physics literature.

The following lemma, stating that the operator defining (13) is analytic in k and has a discrete spectrum for piecewise continuous potential V in I_{uc}, can be found in [17].

Lemma 2.1. *Suppose the potential V in (13) is piecewise continuous in I_{uc}. Then the following holds for any $k \in I_{\mathrm{bz}}$:*
1. *The operator $H(k)$ in (14) is analytic in k.*
2. *The operator $H(k)$ in (14) has purely discrete spectrum.*

This justifies the labeling of energies and Bloch functions in (9) and (13).

Remark 2.2. *The eigenfunctions $\{\psi_k^{(j)}\}_{j=1}^{\infty}$ are only defined up to a nonzero constant scaling. We can assume that these modulus of these constants are chosen such that*

$$\int_{-a/2}^{a/2} |\psi_k^{(j)}(x)|^2 \, \mathrm{d}x = \int_{-a/2}^{a/2} |u_k^{(j)}(x)|^2 \, \mathrm{d}x = 1, \quad k \in \mathbb{R}, \, j = 1, 2, \ldots. \tag{16}$$

There are different conventions for extending k outside of the first BZ. In the context of Wannier functions, it is natural to use the so-called periodic zone scheme, where the Bloch functions and eigenvalues are copies of those in the first BZ [20]. In other words, we can extend the Bloch functions and eigenvalues to be periodic functions in k with period Ω. More explicitly, we have

$$E_{k+g}^{(j)} = E_k^{(j)} \tag{17}$$

and

$$\psi_{k+g}^{(j)} = \psi_k^{(j)} \tag{18}$$

for any $k \in I_{\mathrm{bz}}, g \in \Lambda^*$ and $j = 1, 2, \ldots$. Combining (18) and (11) shows that

$$u_{k+g}^{(j)}(x) = e^{-\mathrm{i}gx} u_k^{(j)}(x). \tag{19}$$

4

2.3 Wannier functions

Due to the periodic condition (18), the Wannier functions are defined as the Fourier coefficients of $\{\psi_k^{(j)}(x)\}_{j=1}^{\infty}$:

$$W_n^{(j)}(x) = \frac{1}{\Omega} \int_{-\Omega/2}^{\Omega/2} e^{-inak} \psi_k^{(j)}(x)\, dk, \quad j = 1, 2, \ldots. \tag{20}$$

The Wannier functions with index $n \neq 0$ are copies of $W_0^{(j)}$ shifted to be centered around the nth lattice site [20]. More explicitly, we have

$$W_n^{(j)}(x) = W_0^{(j)}(x - na), \quad j = 1, 2, \ldots. \tag{21}$$

Hence, we only need to consider the case where $n = 0$:

$$W_0^{(j)}(x) = \frac{1}{\Omega} \int_{-\Omega/2}^{\Omega/2} \psi_k^{(j)}(x)\, dk = \frac{1}{\Omega} \int_{-\Omega/2}^{\Omega/2} e^{ixk} u_k^{(j)}(x)\, dk, \quad j = 1, 2, \ldots. \tag{22}$$

It was shown in [14], in the case of a symmetric potential, and in [9], more generally, that $\{\psi_k^{(j)}\}_{j=1}^{\infty}$ can be scaled such that the $\{W_0^{(j)}\}_{j=1}^{\infty}$ are exponentially localized. In other words, there exist constants $C > 0$ and $D > 0$ such that

$$|W_0^{(j)}(x)| \leq Ce^{-D|x|}, \quad x \in \mathbb{R}. \tag{23}$$

We also define the moment functions

$$\langle x \rangle = \int_{-\infty}^{\infty} x |W_0^{(j)}(x)|^2\, dx, \tag{24}$$

and

$$\langle x^2 \rangle = \int_{-\infty}^{\infty} x^2 |W_0^{(j)}(x)|^2\, dx. \tag{25}$$

2.4 Analyticity of eigenvalues and eigenvectors

We next introduce standard results in analytic perturbation theory for linear operators in [13]. This includes smoothness results for eigenvalues/vectors of an analytic family of operators, and formulas for their derivatives.

Suppose we have a family of matrices $T(z) \in \mathbb{C}^{n \times n}$, where $T(z)$ is analytic for z in a domain $D \subset \mathbb{C}$ intersecting the real axis. We call an analytic family of matrices $T(z)$ self-adjoint if it satisfies $T^*(z) = T(\bar{z})$ for $z \in D$. The following theorem is Theorem 6.1 in [13].

Theorem 2.3. *If an analytic family $T(z) \in \mathbb{C}^{n \times n}$ is self-adjoint, its eigenvalues $\{\lambda_i(z)\}_{i=1}^{n}$ and eigenprojectors $\{P_i(z)\}_{i=1}^{n}$ are analytic for real $z \in D$.*

Remark 2.4. *The above theorem is still applicable when the eigenvalues become degenerate. In this case, the eigenprojectors at degenerate points must be viewed as limits of those approaching the degenerate points. We refer the reader to Theorem 1.10 and Remark 1.11 in [13] for details.*

Remark 2.5. *Theorem 2.3 concerns only the eigenprojectors, not the eigenvectors; this is because for any an eigenvector $\mathbf{v}_i(z)$ of $T(z)$, one can multiply the vector by a z-dependent phase $e^{-i\phi(z)}$ ($\phi(z)$ is real for real z) while the new vector $e^{-i\phi(z)}\mathbf{v}_i(z)$ is still an eigenvector. The phase could change the smoothness of $\mathbf{v}_i(z)$ while the projector $P_i(z) = \mathbf{v}_i(z)\mathbf{v}_i^*(z)$ is unchanged for real z.*

The following theorem shows that it is possible to find a family of eigenvectors $\mathbf{v}_i(z)$ that are analytic for real $z \in D$. It is a special case of the construction in Section 2.6.2 in [13].

Theorem 2.6. *Suppose that the analytic matrix family $T(z)$ is self-adjoint in D, and that $\mathbf{v}_0$ is an eigenvector of $T(z_0)$ with a nondegenerate eigenvalue λ_0 for some real z_0 in D. Let $\lambda(z)$ and $P(z)$ be analytic families of eigenvalues and eigenprojectors in D, the existence of which follows from Theorem 2.3. Consider the the initial value problem*

$$\frac{\mathrm{d}}{\mathrm{d}z}\mathbf{v}(z) = Q(z)\mathbf{v}(z), \quad \text{with the initial condition} \quad \mathbf{v}(z_0) = \mathbf{v}_0, \tag{26}$$

where $Q = \frac{\mathrm{d}P}{\mathrm{d}z}P - P\frac{\mathrm{d}P}{\mathrm{d}z}$. The solution $\mathbf{v}(z)$ to (26) is analytic and satisfies $T(z)\mathbf{v}(z) = \lambda(z)\mathbf{v}(z)$ for real z in D.

While the above results only apply to self-adjoint analytic family of finite-dimensional matrices $T(z)$ in D, they straightforwardly generalize to a class of infinite-dimensional operators in Hilbert space. The definition for the analytic family $T(z)$ being self-adjoint in a domain D intersecting the real axes is unchanged. To extend Theorems 2.3 and 2.6 to the operator case Kato defined a class of operators where a subset of their spectra can be separated by a simple close curve in the complex plane from the rest.

Kato showed that, for a self-adjoint analytic family $T(z)$ possessing a finite set of isolated spectrum that can be separated for $z \in D$, the finite-dimensional version Theorems 2.3 and 2.6 can be applied to the finite-dimensional subspace corresponding to the separated eigenvalues. We summarize this extension in the following theorem and refer the reader to Sections 7.1.3 and 7.3.1 in [13] for details.

Theorem 2.7. *Suppose that $T(z)$ is an analytic family of operators in Hilbert space for $z \in D$. Suppose further that $T(z)$ is self-adjoint, defined analogously as the matrix case in Theorem 2.3. Theorems 2.3 and 2.6 also apply to any finite systems of eigenvalues, eigenprojectors and eigenvectors of $T(z)$, where the eigenvalues can be separated from the rest of the spectrum of $T(z)$.*

In this paper, we deal with operators with discrete spectra, where any eigenvalues of interests can be separated from the rest. Theorem 2.7 is applicable to the eigensubspaces of any such operator.

2.5 Derivatives of eigenvalues and eigenvectors

Next, we provide expressions for the derivatives of eigenvalues and eigenvectors. For simplicity, we only state the result for the Hilbert space $\ell^2(\mathbb{Z})$ and operators with a purely discrete spectrum, which will suffice for our purpose. These formulas can be found in Section 8.2.3 in [13]. The formula (28) is a consequence of the derivative of the projection $\frac{\mathrm{d}P}{\mathrm{d}z}$ substituted into (26).

Theorem 2.8. *Suppose that $T(z)$ is a self-adjoint analytic family of operators in $\ell^2(\mathbb{Z})$ for $z \in D$ and the operators have a purely discrete spectrum. Suppose further that $z_0 \in D$ is real and $\lambda(z_0)$ is a*

nondegenerate eigenvalue of $T(z_0)$ with the eigenvector $\mathbf{v}$, i.e. $T(z_0)\mathbf{v}(z_0) = \lambda(z_0)\mathbf{v}(z_0)$. The derivative of λ at z_0 is given by

$$\frac{\mathrm{d}\lambda(z_0)}{\mathrm{d}z} = \mathbf{v}(z_0)^* \frac{\mathrm{d}T(z_0)}{\mathrm{d}z} \mathbf{v}(z_0), \tag{27}$$

and the derivative of $\mathbf{v}$ at z_0 is given by

$$\frac{\mathrm{d}\mathbf{v}(z_0)}{\mathrm{d}z} = -(T(z_0) - \lambda(z_0))^\dagger \frac{\mathrm{d}T(z_0)}{\mathrm{d}z} \mathbf{v}(z_0), \tag{28}$$

where $(T(z_0) - \lambda(z_0))^\dagger$ is the pseudoinverse of $T(z_0) - \lambda(z_0)$ ignoring the eigenspace of $\lambda(z_0)$.

We note that in obtaining (28) from (26) we also used $P\frac{\mathrm{d}P}{\mathrm{d}z}P = 0$, $P\mathbf{v} = \mathbf{v}$ for $\mathbf{v}$ in the range of P. We observe that, by the definition of pseudoinverse, the range of $(T - \lambda)^\dagger$ is orthogonal to that of P, so we have the following formula

$$(T - \lambda)^\dagger P = P(T - \lambda)^\dagger = 0. \tag{29}$$

Remark 2.9. *In the physics literature, the formula (28) is usually written using the spectral decomposition of T in the following form*

$$\frac{\mathrm{d}\mathbf{v}(z_0)}{\mathrm{d}z} = \sum_{\substack{j \\ \lambda_j \neq \lambda}} \mathbf{v}_j(z_0) \frac{\mathbf{v}_j(z_0)^* \frac{\mathrm{d}T(z_0)}{\mathrm{d}z} \mathbf{v}(z_0)}{\lambda(z_0) - \lambda_j(z_0)}, \tag{30}$$

where $\lambda_j(z_0)$ and $\mathbf{v}_j(z_0)$ are the rest of the eigenvalues and eigenvectors of $T(z_0)$. This is the formula one would use to compute $\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}z}$ using the singular value decomposition of the operator $T(z_0) - \lambda(z_0)$.

Remark 2.10. *We observe that (28) holds true if $\frac{\mathrm{d}T(z_0)}{\mathrm{d}z}$ is replaced with $\frac{\mathrm{d}T(z_0)}{\mathrm{d}z} - \frac{\mathrm{d}\lambda(z_0)}{\mathrm{d}z}$. Since $(\frac{\mathrm{d}T(z_0)}{\mathrm{d}z} - \frac{\mathrm{d}\lambda(z_0)}{\mathrm{d}z})\mathbf{v}(z_0)$ is orthogonal to $\mathbf{v}(z_0)$ (see (29)), we can compute $\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}z}$ by*

$$\frac{\mathrm{d}\mathbf{v}(z_0)}{\mathrm{d}z} = -(T(z_0) - \lambda(z_0) + \gamma P(z_0))^{-1} \left(\frac{\mathrm{d}T(z_0)}{\mathrm{d}z} - \frac{\mathrm{d}\lambda(z_0)}{\mathrm{d}z} \right) \mathbf{v}(z_0), \tag{31}$$

where $P(z_0) = \mathbf{v}(z_0)\mathbf{v}^*(z_0)$ is the projector and γ is any reasonable constant such that the addition $\gamma P(z_0)$ makes the original operator invertible. As a result, the pseudoinverse is replaced by the inverse. Although (31), (28) and (30) are all equivalent, (31) is more attractive numerically in cases where the discretization of $T(z_0)$ is large.

3 Analytic apparatus

3.1 Analysis of Bloch functions

In this section, we introduce results about analyticity and the Fourier analysis of the Bloch functions in Section 2.2. In what follows, it will be convenient to view $\{\psi_k^{(j)}\}_{j=1}^\infty$ as defined in (9) and $\{u_k^{(j)}\}_{j=1}^\infty$ defined in (13) as functions of both x and k. We will also view E_k in (9) as a function of k.

In this paper, we will restrict ourselves to a single band whose energy E_k does not become degenerate for any k (i.e., a single value of j), so we will omit the superscript. We thus rewrite (13), (12) and (19) as

$$H(k)u = -(\partial_{xx} + ik)^2 u(x,k) + V(x)u(x,k) = E(k)u(x,k), \tag{32}$$

and

$$u(x,k) = u(x+y,k), \quad y \in \Lambda,\ k \in I_{\mathrm{bz}} \tag{33}$$

$$u(x,k+g) = e^{-igx}u(x,k), \quad g \in \Lambda^*,\ x \in I_{\mathrm{uc}}. \tag{34}$$

Since $H(k)$ is self-adjoint analytic family of operators and the eigenvalue $E(k)$ considered is isolated, Theorem 2.7 is applicable and we have the following lemma about the analyticity of $E(k)$ and u in k.

Lemma 3.1. *The energy $E(k)$ in (32) is analytic in $k \in I_{\mathrm{bz}}$. The function $u(x,k)$ can be chosen to be analytic in $k \in I_{\mathrm{bz}}$ for $x \in I_{\mathrm{uc}}$.*

Remark 3.2. *We observe that Theorem 2.7 does not guarantee the function $E(k)$ or $u(x,k)$ to be periodically analytic in $k \in I_{\mathrm{bz}}$, which will be shown to be necessary for exponentially localized Wannier functions. If $E(k)$ were to become degenerate for some $k_0 \in I_{\mathrm{bz}}$, it could go to another branch so that it does not return to $E(-\Omega/2)$ when it reaches $E(\Omega/2)$. When E is not degenerate, $E(k)$ will be periodically analytic in $k \in I_{\mathrm{bz}}$. For the eigenvector $u(x,k)$, as discussed in Remark 2.5, whenever $u(x,k)$ is analytic in $k \in I_{\mathrm{bz}}$, so is any $e^{-i\phi(k)}u(x,k)$ with a real analytic ϕ, which need not be periodic in I_{bz}.*

Exploiting (33), we take the Fourier expansion of u with respect to the x variable to get

$$u(x,k) = \sum_{m=-\infty}^{\infty} a_m(k)e^{im\Omega x}, \tag{35}$$

where

$$\frac{1}{a}\int_{-a/2}^{a/2} u(x,k)e^{-im\Omega x}\mathrm{d}x = a_m(k), \quad m \in \mathbb{Z}. \tag{36}$$

By Parseval's theorem, the normalization condition for $u(x,k)$ in (16) shows that

$$\sum_{m=-\infty}^{\infty} |a_m(k)|^2 = \frac{1}{a}. \tag{37}$$

For simplicity, we assume for the rest of the paper that $u(x,k)$ as a function of x is a periodic function in $C^1(I_{\mathrm{uc}};\mathbb{C})$ and $\partial_x u(x,k)$ has bounded variation for all $k \in I_{\mathrm{bz}}$, so that the Fourier coefficients $a_m(k)$ decays as $|m|^{-2}$ for large m [23, Section 2.6]. Thus, we have a positive constant D independent of k, such that

$$|a_m(k)| \le \frac{D}{|m|^2}, \quad m \in \mathbb{Z}. \tag{38}$$

Hence the Fourier series in (35) converges absolutely independent of k. By (36), if $u(x,k)$ is chosen to be analytic for $k \in I_{\mathrm{bz}}$ by Lemma 3.1, we observe that all the Fourier coefficients $a_m(k)$ are analytic for $k \in I_{\mathrm{bz}}$.

Lemma 3.3. *Suppose $u(x,k)$ is chosen to be analytic in $k \in I_{\mathrm{bz}}$. Its Fourier coefficients $a_m(k)$ in (35) are analytic for $k \in I_{\mathrm{bz}}$.*

The following lemma provides a simple relation for the dependence of k and m in the Fourier coefficients in (35). It is a well-known relation in the context of periodic zone scheme (see (17)-(19)). It is a simple consequence of (34).

Lemma 3.4. *There exists a function $\alpha : \mathbb{R} \to \mathbb{C}$ such that $a_m(k)$ in (35) can be expressed as*

$$a_m(k) = \alpha(k + m\Omega), \quad k \in I_{\mathrm{bz}}, \, m \in \mathbb{Z}. \tag{39}$$

Proof. We let $n \in \mathbb{Z}$ and $g = n\Omega$. From (34) and (35) that

$$u(x, k+g) = \sum_{m=-\infty}^{\infty} a_m(k+g)e^{im\Omega x} \tag{40}$$

$$= \sum_{m=-\infty}^{\infty} a_m(k)e^{i(m-n)\Omega x} \tag{41}$$

$$= \sum_{l=-\infty}^{\infty} a_{l+n}(k)e^{il\Omega x}. \tag{42}$$

It follows from (40) and (42) that a_m satisfies the functional relation

$$a_m(k + n\Omega) = a_{m+n}(k), \quad m, n \in \mathbb{Z}, \, k \in \mathbb{R}. \tag{43}$$

Setting

$$\alpha(k + m\Omega) = a_m(k), \quad k \in I_{\mathrm{bz}}, \, m \in \mathbb{Z} \tag{44}$$

yields the result. $\qquad\square$

Lemma 3.4 shows that we can obtain the function α defined in $\mathbb{R}$ by "unfolding" the Fourier coefficients $a_m(k)$ defined in I_{bz} for $m \in \mathbb{Z}$ such that $a_m(k)$ takes up the interval $[-\Omega/2 + m\Omega, \Omega/2 + m\Omega)$. We summarize this fact in the following lemma, together with Lemma 3.3 and the decay condition (38) in terms of α.

Lemma 3.5. *Suppose that $u(\cdot, k) \in C^1(I_{\mathrm{uc}}; \mathbb{C})$ is periodic and $\partial_x u(\cdot, k)$ is of bounded variation for $k \in I_{\mathrm{bz}}$. The Fourier series (35) can be expressed as*

$$u(x, k) = \sum_{m=-\infty}^{\infty} \alpha(k + m\Omega)e^{im\Omega x}, \tag{45}$$

where α is defined in (39). The function $\alpha(k)$ can be chosen to be analytic in the intervals $[-\Omega/2 + m\Omega, \Omega/2 + m\Omega)$ for $m \in \mathbb{Z}$. Furthermore, $|\alpha(k)|$ decays no slower than $|k|^{-2}$ as $|k| \to \infty$.

Remark 3.6. *Since $u(x, k)$ may not be periodically analytic in $k \in I_{\mathrm{bz}}$ (see Remark 3.2), the function α can be a discontinuous function in $\mathbb{R}$. Lemma 3.3 only ensures α is analytic in each interval $[-\Omega/2 + m\Omega, \Omega/2 + m\Omega)$ for $m \in \mathbb{Z}$. Moreover, the function α is formed from a very special choice of coefficients a_m since the corresponding $u(x, k)$ is chosen to be analytic in k; this does not happen if one computes $u(x, k)$ at each point k independently (see Remark 2.5) so that $u(x, k)$ is an arbitrarily irregular function in k.*

9

3.2 Properties of Wannier functions

In this section, we introduce the Fourier transform of the Wannier functions introduced in Section 2.3, and their variance in terms of the α function defined in (45). Following the discussion above, we suppress the band index and we rewrite (22) as

$$W_0(x) = \frac{1}{\Omega} \int_{-\Omega/2}^{\Omega/2} \psi_k(x) \, \mathrm{d}k = \frac{1}{\Omega} \int_{-\Omega/2}^{\Omega/2} e^{\mathrm{i}xk} u_k(x) \, \mathrm{d}k. \tag{46}$$

First, we introduce the Fourier transform of $W_0(x)$. By substituting (45) into (46), we obtain

$$W_0(x) = \frac{1}{\Omega} \int_{-\Omega/2}^{\Omega/2} \sum_{m=-\infty}^{\infty} \alpha(k+m\Omega) e^{\mathrm{i}(k+m\Omega)x} \mathrm{d}k = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{2\pi}{\Omega} \alpha(\xi) e^{\mathrm{i}\xi x} \mathrm{d}\xi. \tag{47}$$

We observe that (47) only hold formally since α may not even be integrable (see Remark 3.6). However, whenever α is chosen to be integrable, (47) shows that $\frac{2\pi}{\Omega}\alpha$ is the Fourier transform of W_0 in (46) and we have the following theorem.

Theorem 3.7. *Suppose u in Lemma 3.5 is chosen so that the function α is integrable. The corresponding Wannier function W_0 in (46) is well defined and is given by the formula*

$$W_0(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{2\pi}{\Omega} \alpha(\xi) e^{\mathrm{i}\xi x} \mathrm{d}\xi. \tag{48}$$

Its Fourier transform denoted by $\widehat{W}_0$ is given by the formula

$$\widehat{W}_0(\xi) = \int_{-\infty}^{\infty} W_0(x) e^{-\mathrm{i}\xi x} \, \mathrm{d}x = \frac{2\pi}{\Omega} \alpha(\xi). \tag{49}$$

By the Riemann–Lebesgue lemma, $|W_0(x)|$ goes to zero as $|x| \to \infty$.

Thus, if α is chosen to have poor regularity, the Wannier function W_0 is poorly localized. To define its variance α needs to be smooth so that $|W_0(x)|$ goes to zero faster than $1/|x|$ as $|x| \to \infty$. Next, we assume that α is chosen so that its derivative is integrable. The following lemma contains the formulas for the first and second moments of W_0 in terms of α for defining its variance. They are consequences of the one-dimensional case of (7) and (8) in [15], first derived in [2].

Lemma 3.8. *Suppose that u in Lemma 3.5 is chosen so that the function α has an integrable derivative α' in $\mathbb{R}$. We have the following formulas for the first and second moments of W_0:*

$$\langle x \rangle = \int_{-\infty}^{\infty} x \overline{W}_0(x) W_0(x) \, \mathrm{d}x = \frac{2\pi}{\Omega^2} \mathrm{i} \int_{-\infty}^{\infty} \overline{\alpha}(\xi) \alpha'(\xi) \, \mathrm{d}\xi, \tag{50}$$

and

$$\langle x^2 \rangle = \int_{-\infty}^{\infty} x^2 \overline{W}_0(x) W_0(x) \, \mathrm{d}x = \frac{2\pi}{\Omega^2} \int_{-\infty}^{\infty} |\alpha'(\xi)|^2 \mathrm{d}\xi. \tag{51}$$

Hence the variance $\langle x^2 \rangle - \langle x \rangle^2$ is given by the formula

$$\langle x^2 \rangle - \langle x \rangle^2 = \frac{2\pi}{\Omega^2} \int_{-\infty}^{\infty} |\alpha'(\xi)|^2 \mathrm{d}\xi + \frac{4\pi^2}{\Omega^4} \left(\int_{-\infty}^{\infty} \overline{\alpha}(\xi) \alpha'(\xi) \, \mathrm{d}\xi \right)^2. \tag{52}$$

3.3 Perturbation analysis in the Fourier domain

The function u in Lemma 3.5 can be obtained by applying Kato's construction in Theorem 2.6. This involves the ODE defined by (27) for the eigenvalue problem defined by (32). In order to carry out this procedure computationally, we transform (32) into the Fourier domain. In this section, we introduced the Fourier domain version of (32) and apply (27) to the result.

First, we introduce Fourier series of the potential V given by the formula

$$V(x) = \sum_{l=-\infty}^{\infty} \widehat{V}_l e^{il\Omega x}, \tag{53}$$

where the vector with elements $\{\widehat{V}_l\}_{l=-\infty}^{\infty}$ is in $\ell^2(\mathbb{Z})$ for piecewise continuous V in I_{uc}. Moreover, we define a vector $\mathbf{y}$ containing the Fourier coefficients of u in (35) and (45):

$$\mathbf{y}_m(k) = a_m(k) = \alpha(k + m\Omega), \quad m \in \mathbb{Z}, \, k \in I_{\mathrm{bz}}. \tag{54}$$

It follows that $\mathbf{y} \in \ell^2(\mathbb{Z})$ for any u that satisfies the assumptions of Lemma 3.5. Moreover, it follows from (37) that

$$\|\mathbf{y}(k)\|^2 = \frac{1}{a}, \quad k \in I_{\mathrm{bz}}. \tag{55}$$

Inserting (35) into (32) yields the infinite linear system

$$(\mathbf{D}(k) + \mathbf{V})\mathbf{y}(k) = E(k)\mathbf{y}(k), \quad k \in I_{\mathrm{bz}}, \tag{56}$$

where for all $m, n \in \mathbb{Z}$ we have

$$\mathbf{D}_{m,n}(k) = (k + m\Omega)^2 \delta_{mn}, \tag{57}$$

$$\mathbf{V}_{m,n} = \widehat{V}_{m-n}. \tag{58}$$

It is convenient to define

$$\mathbf{\Theta}(k) = \mathbf{D}(k) + \mathbf{V} - E(k)\mathbf{I}, \tag{59}$$

where $\mathbf{I}_{m,n} = \delta_{mn}$ is the identity. Applying (27) and (28) to (56), we obtain the following formulas

$$\frac{\mathrm{d}}{\mathrm{d}k} E(k) = \frac{\mathbf{y}(k)^* \mathbf{S}(k)\mathbf{y}(k)}{\mathbf{y}(k)^* \mathbf{y}(k)}, \tag{60}$$

$$\frac{\mathrm{d}}{\mathrm{d}k}\mathbf{y}(k) = -\mathbf{\Theta}^\dagger(k)\mathbf{S}(k)\mathbf{y}(k), \quad k \in I_{\mathrm{bz}}, \tag{61}$$

where

$$\mathbf{S}_{m,n}(k) = 2(k + m\Omega)\delta_{mn}. \tag{62}$$

The significance of (60) and (61) is that they provide an infinite system of ODEs for the energy E and the vector containing the Fourier coefficients of u. If E_0 and $\mathbf{y}_0$ satisfy (56) at $k_0 = -\Omega/2$, solving (60) and (61) for $k \in I_{\mathrm{bz}}$ subject to the initial conditions

$$E(k_0) = E_0, \quad \mathbf{y}(k_0) = \mathbf{y}_0 \tag{63}$$

where $k_0 = -\Omega/2$, produces E and u that are analytic functions of k.

It is worth mentioning that (61) is sometimes referred to as the parallel-transport equation due to the relation

$$\mathbf{y}(k)^* \frac{\mathrm{d}}{\mathrm{d}k}\mathbf{y}(k) = 0, \tag{64}$$

which holds due to the relation

$$\boldsymbol{\Theta}^\dagger(k)\widetilde{\mathbf{y}}(k) = 0, \tag{65}$$

which in turn is a consequence of (29). In other words, $\frac{\mathrm{d}}{\mathrm{d}k}\mathbf{y}(k)$ is always orthogonal to $\mathbf{y}(k)$. The relation (64) also implies that $\|\mathbf{y}(k)\|$ is constant for all k.

Remark 3.9. *We observe that the operator* $-\boldsymbol{\Theta}^\dagger(k)\mathbf{S}(k)$ *in (61) has a spectrum that decays as $1/m$ asymptotically for any $k \in I_{\mathrm{bz}}$, so it is a compact operator on $\ell^2(\mathbb{Z})$. As a result, despite the fact that (61) is infinite-dimensional, (61) is very benign both mathematically and numerically.*

3.4 Analysis of the discontinuities at zone boundaries

In this section, we show that the discontinuities of the Fourier transform $\alpha(k)$ (and its derivatives of all orders) at $k = \Omega/2 + m\Omega$ for $m \in \mathbb{Z}$ are identical.

Suppose that $\mathbf{y}(k)$ for $k \in I_{\mathrm{bz}}$ is a solution to the ODE (61) with initial condition (63) at $k_0 = -\Omega/2$. The vector $\mathbf{y}(k)$ defines a function $\alpha(k)$ (see (54)) by

$$\mathbf{y}_m(k) = a_m(k) = \alpha(k + m\Omega), \quad m \in \mathbb{Z}, \ k \in I_{\mathrm{bz}}. \tag{66}$$

We can extend $\mathbf{y}(k)$ to $k = \Omega/2$, so that we have

$$\mathbf{y}_m(\Omega/2) = a_m(\Omega/2), \quad m \in \mathbb{Z}. \tag{67}$$

As discussed in Remark 3.6, $a_m(\Omega/2)$ is not guaranteed to be equal to $a_{m+1}(-\Omega/2)$. Since $a_m(\Omega/2)$ and $a_{m+1}(-\Omega/2)$ may be viewed as limits of $\alpha(k)$ at the boundary points $k = \Omega/2 + m\Omega$ for $m \in \mathbb{Z}$ from both sides, their differences characterize the jump of $\alpha(k)$ at these boundary points. The following lemma shows that $a_m(\Omega/2)$ and $a_{m+1}(-\Omega/2)$ only differ by a phase independent of m.

Lemma 3.10. *For the vector $\mathbf{y}(k)$ with $k \in [-\Omega/2, \Omega/2]$ defined above, we have the following relation*

$$a_{m+1}(-\Omega/2) = e^{\mathrm{i}\phi_{\mathrm{zak}}} a_m(\Omega/2), \quad m \in \mathbb{Z}, \tag{68}$$

where ϕ_{zak} is a real number independent of m.

Proof. We observe that both vectors $\mathbf{y}(-\Omega/2)$ and $\mathbf{y}(\Omega/2)$ satisfy (56):

$$(\mathbf{D}(-\Omega/2) + \mathbf{V})\mathbf{y}(-\Omega/2) = E\mathbf{y}(-\Omega/2), \tag{69}$$

$$(\mathbf{D}(\Omega/2) + \mathbf{V})\mathbf{y}(\Omega/2) = E\mathbf{y}(\Omega/2), \tag{70}$$

where $E(-\Omega/2) = E(\Omega/2) = E$ (see Remark 3.2). Exploiting the structure in (57) and (58), we can reindex (70) to yield

$$(\mathbf{D}(-\Omega/2) + \mathbf{V})R(\mathbf{y}(\Omega/2)) = ER(\mathbf{y}(\Omega/2)), \tag{71}$$

12

where R is the right shift operator defined in (2). Since the eigenvalue E is not degenerate, $\mathbf{y}(-\Omega/2)$ and $R(\mathbf{y}(\Omega/2))$ can only differ by a constant of unity modulus $e^{i\phi_{\text{zak}}}$ for some real ϕ_{zak}. The component form of $\mathbf{y}(-\Omega/2) = e^{i\phi_{\text{zak}}}R(\mathbf{y}(\Omega/2))$ yields (68). $\qquad\square$

A similar argument applied to (61) shows that

$$\frac{\mathrm{d}}{\mathrm{d}k}\mathbf{y}(-\Omega/2) = -\boldsymbol{\Theta}^{\dagger}(-\Omega/2)\mathbf{S}(-\Omega/2)\mathbf{y}(-\Omega/2), \tag{72}$$

$$\frac{\mathrm{d}}{\mathrm{d}k}R(\mathbf{y}(\Omega/2)) = -\boldsymbol{\Theta}^{\dagger}(-\Omega/2)\mathbf{S}(-\Omega/2)R(\mathbf{y}(\Omega/2)) \tag{73}$$

$$= -e^{-i\phi_{\text{zak}}}\boldsymbol{\Theta}^{\dagger}(-\Omega/2)\mathbf{S}(-\Omega/2)\mathbf{y}(-\Omega/2) \tag{74}$$

$$= e^{-i\phi_{\text{zak}}}\frac{\mathrm{d}}{\mathrm{d}k}\mathbf{y}(-\Omega/2). \tag{75}$$

This shows that the jump between $a'_m(\Omega/2)$ and $a'_{m+1}(-\Omega/2)$ is also a factor of $e^{i\phi_{\text{zak}}}$. Repeated differentiation of (72) and (73) shows the jump between the nth derivatives $a_m^{(n)}(\Omega/2)$ and $a_{m+1}^{(n)}(-\Omega/2)$ are all identical.

Remark 3.11. *The quantity ϕ_{zak} was introduced by Zak in [22] in a different context and is known as the Zak phase. It was shown in [22] that the Zak phase depends only on the Schrödinger equation (9) and determines the center of the Wannier function.*

3.5   Gauge transformation

In this section, we introduce gauge transformation and show that a simple class of gauge choices fixes all discontinuities of the Fourier transform α discussed in Section 3.4, thus obtaining exponentially localized Wannier functions.

Given $\phi \in C^1(I_{\text{bz}};\mathbb{R})$ (not necessarily periodic), we can modify any solution $\mathbf{y}$ of (61) via

$$\widetilde{\mathbf{y}}(k) = e^{-i\phi(k)}\mathbf{y}(k), \quad \text{for } k \in I_{\text{bz}}, \tag{76}$$

where $\widetilde{\mathbf{y}}(k)$ still satisfies (56) with the same $E(k)$ as $\mathbf{y}(k)$ and the normalization (55). Such a transform turns α defined by $\mathbf{y}$ into $\widetilde{\alpha}$ given by the formula

$$\widetilde{\alpha}(k + m\Omega) = e^{-i\phi(k)}\mathbf{y}_m(k) = e^{-i\phi(k)}a_m(k), \quad m \in \mathbb{Z}, \, k \in I_{\text{bz}}. \tag{77}$$

This process is known as gauge transformation in the physics literature. From (61), it follows that $\widetilde{\mathbf{y}}$ satisfies

$$\widetilde{\mathbf{y}}'(k) = -\boldsymbol{\Theta}^{\dagger}(k)\mathbf{S}(k)\widetilde{\mathbf{y}}(k) - i\phi'(k)\widetilde{\mathbf{y}}(k), \quad k \in I_{\text{bz}}. \tag{78}$$

In the physics literature, ϕ' is known as the Berry connection and the phase ϕ is known as the Berry phase. The Berry connection is usually denoted by A and defined as $A(k) = i\langle u_k | u'_k\rangle$ in Bra-Ket notation for some function u_k satisfying (13). In this paper, this corresponds to the identity:

$$i\widetilde{\mathbf{y}}^*(k)\widetilde{\mathbf{y}}'(k) = \phi'(k)/a, \tag{79}$$

13

where we used (65) and the factor $1/a$ comes from the normalization (55).

We choose ϕ such that $\phi(k) = -\phi(-k)$ and $\phi'(k) = \phi_{\mathrm{zak}}/\Omega$, where ϕ_{zak} is defined in (68):

$$\phi(k) = \int_{-\Omega/2}^{k} \phi'(k')\mathrm{d}k' - \frac{\Omega}{2} = \frac{\phi_{\mathrm{zak}}}{\Omega}k, \quad k \in I_{\mathrm{bz}}. \tag{80}$$

By Lemma 3.10, this choice of $\phi(k)$ fixes the jump so that

$$\widetilde{\mathbf{y}}(-\Omega/2) = R(\widetilde{\mathbf{y}}(\Omega/2)). \tag{81}$$

Furthermore, analogously to (72) and (73), we have

$$\widetilde{\mathbf{y}}'(-\Omega/2) = -\boldsymbol{\Theta}^{\dagger}(-\Omega/2)\mathbf{S}(-\Omega/2)\widetilde{\mathbf{y}}(-\Omega/2) - \mathrm{i}\phi'(-\Omega/2)\widetilde{\mathbf{y}}(-\Omega/2), \tag{82}$$

$$R(\widetilde{\mathbf{y}}'(\Omega/2)) = -\boldsymbol{\Theta}^{\dagger}(-\Omega/2)\mathbf{S}(-\Omega/2)R(\widetilde{\mathbf{y}}(\Omega/2)) - \mathrm{i}\phi'(\Omega/2)R(\widetilde{\mathbf{y}}(\Omega/2)). \tag{83}$$

Since ϕ' in (80) is a constant, we have $\phi'(-\Omega/2) = \phi'(\Omega/2)$. Together with (81), we conclude that

$$\widetilde{\mathbf{y}}'(-\Omega/2) = R(\widetilde{\mathbf{y}}'(\Omega/2)). \tag{84}$$

By repeated differentiation of (82) and (83), we conclude that

$$\widetilde{\mathbf{y}}^{(n)}(-\Omega/2) = R(\widetilde{\mathbf{y}}^{(n)}(\Omega/2)), \tag{85}$$

holds for all derivatives of order $n \geq 1$. From (81), (85) and Lemma 3.5, we conclude that the transformed function $\widetilde{\alpha}$ in (77) corresponding to $\widetilde{\mathbf{y}}(k)$ is analytic on $\mathbb{R}$. Thus, its corresponding Wannier function defined by (48) will be exponentially localized.

We also observe that the choice $\phi'(k) = \phi_{\mathrm{zak}}/\Omega$ is not the only possibility that yields exponential localization. The above will still hold if $\phi'(k)$ is chosen to be a real, periodically analytic function in I_{bz} with the zeroth Fourier coefficient $(\phi_{\mathrm{zak}} + 2\pi n)/\Omega$ for any $n \in \mathbb{Z}$. Such functions can be written as

$$\phi'(k) = \frac{\phi_{\mathrm{zak}} + 2\pi n}{\Omega} + \sum_{\substack{m=-\infty \\ m\neq 0}}^{\infty} c_m e^{\mathrm{i}mak}, \tag{86}$$

where the Fourier coefficients c_m decay exponentially as $|m| \to \infty$, and the Berry phase ϕ is then given by

$$\phi(k) = \int_{-\Omega/2}^{k} \phi'(k')\mathrm{d}k' - \frac{\Omega}{2}, \quad k \in I_{\mathrm{bz}}. \tag{87}$$

We summarize this fact in the following theorem. It will be the key tool in constructing exponentially localized Wannier functions.

Theorem 3.12. *Suppose that $\mathbf{y}(k)$ with $k \in I_{\mathrm{bz}}$ is a solution to (61) with initial conditions (63). Then, $\mathbf{y}(k)$ defines a function α in $\mathbb{R}$ by (54). Suppose further that we choose $\phi'(k)$ to be of the form in (86) so that a new function $\widetilde{\mathbf{y}}$ is defined by the gauge transform (76). The function $\widetilde{\alpha}$ defined by $\widetilde{\mathbf{y}}$ in (77) is analytic in $\mathbb{R}$. As a result, the Wannier function as the Fourier transform of $\widetilde{\alpha}$ in (48) is exponentially localized.*

3.6 Gauge choice and Wannier localization

In the previous section, we have seen that many choices of the Berry phase ϕ (see Theorem 3.12 and (86)) are able to produce exponentially localized Wannier functions. In this section, we derive the optimal choice in terms of minimizing the variance defined by (52). It is a standard measure of the localization of Wannier functions [15]. First, we modify the formulas in Lemma 3.8 to account for the inclusion of ϕ. The following is a result of (77), (78), (79) with ϕ' given in (86) and the formulas in Lemma 3.8.

Lemma 3.13. *Suppose that $\widetilde{\mathbf{y}}$ is defined in Theorem 3.12. We have the following formulas*

$$\langle x \rangle = \frac{a}{2\pi} \left[\phi\left(\frac{\Omega}{2}\right) - \phi\left(-\frac{\Omega}{2}\right) \right] = \frac{\phi_{\mathrm{zak}}}{2\pi} a + na, \tag{88}$$

$$\langle x^2 \rangle = \frac{a^2}{2\pi} \int_{-\Omega/2}^{\Omega/2} \|\mathbf{y}'(k)\|^2 \, \mathrm{d}k + \frac{a}{2\pi} \int_{-\Omega/2}^{\Omega/2} \phi'(k)^2 \, \mathrm{d}k \tag{89}$$

$$= \frac{a^2}{2\pi} \int_{-\Omega/2}^{\Omega/2} \|\mathbf{y}'(k)\|^2 \, \mathrm{d}k + \left(\frac{\phi_{\mathrm{zak}}}{2\pi} a + na \right)^2 + \sum_{\substack{m=-\infty \\ m \neq 0}}^{\infty} |c_m|^2, \tag{90}$$

and

$$\langle x^2 \rangle - \langle x \rangle^2 = \frac{a^2}{2\pi} \int_{-\Omega/2}^{\Omega/2} \|\mathbf{y}'(k)\|^2 \, \mathrm{d}k + \sum_{\substack{m=-\infty \\ m \neq 0}}^{\infty} |c_m|^2. \tag{91}$$

As a consequence, we see that the optimally localized Wannier function for a single band is constructed based on the gauge choice

$$\phi'(k) = \frac{\phi_{\mathrm{zak}} + 2\pi n}{\Omega}, \tag{92}$$

i.e., the choice of (86) with all of the coefficients c_m set to zero. For such a choice, we observe that the variance in (91) is gauge-independent since $\mathbf{y}$ is constructed by solving (61), which only depends on the original Schrödinger equation (9).

We observe that the integer n in (92) only results in a shift of the center (88). Since every lattice point has an identical copy of the Wannier function (see (21)), any nonzero choice of n amounts to shifting the index of the Wannier functions. As a result, we simply set $n = 0$, which is the choice in (80). We thus have the following corollary.

Corollary 3.14. *An optimal choice of gauge in term of the variance of Wannier functions is given by*

$$\phi(k) = \int_{-\Omega/2}^{k} \phi'(k') \, \mathrm{d}k' - \frac{\Omega}{2} = \frac{\phi_{\mathrm{zak}}}{\Omega} k, \quad k \in I_{\mathrm{bz}}. \tag{93}$$

The center $\langle x \rangle$ and the variance $\langle x^2 \rangle - \langle x \rangle^2$ of the optimally localized Wannier function are given by the formulas

$$\langle x \rangle = \frac{\phi_{\mathrm{zak}}}{2\pi} a. \tag{94}$$

15

$$\langle x^2 \rangle - \langle x \rangle^2 = \frac{a^2}{2\pi} \int_{-\Omega/2}^{\Omega/2} \|\mathbf{y}'(k)\|^2 \, \mathrm{d}k. \tag{95}$$

Both quantities only depend on the given Schrödinger equation.

Theorem 3.12 and Corollary 3.14 complete the construction of optimally localized Wannier functions.

Remark 3.15. *The condition in (92) can also be obtained by applying the calculus of variations on the variance in (91). Since the center in (91) must equal $\frac{\phi_{\mathrm{zak}}}{2\pi}a + na$ for (81) to hold, it can be excluded from the objective function. Therefore, to minimize the variance it is sufficient for ϕ to minimize $\langle x^2 \rangle$, where the only gauge-dependent term is*

$$\int_{-\Omega/2}^{\Omega/2} \phi'(k)^2 \, \mathrm{d}k. \tag{96}$$

The optimal solution is characterized by the solution to the Laplace equation

$$\phi'' = 0, \quad \text{subject to } \phi(\Omega/2) - \phi(-\Omega/2) = \frac{\phi_{\mathrm{zak}}}{2\pi}a + na. \tag{97}$$

The derivative of the solution is precisely (92). Results of this form are known in physics literature [15, 20, 21], but they are usually stated in a nonconstructive form and are only meant to be theoretical tools characterizing the minimum.

3.7 Realty of Wannier functions

While we have completed the construction of optimally localized Wannier functions, there is one remaining degree of freedom, the phase choice of the vector $\mathbf{y}_0$ in (63); we could replace $\mathbf{y}_0$ by $e^{-\mathrm{i}\phi_0}\mathbf{y}_0$ for any real ϕ_0 and all above formulas would still hold, including the expressions in Corollary 3.14. However, the phase $e^{-\mathrm{i}\phi_0}$ does change the resulting W_0. In this section, we prove that the constructed Wannier function W_0 can always be chosen to be real if $\mathbf{y}_0$ in (63) satisfies a simple condition.

First, we observe that, since the potential V is real, we have

$$\widehat{V}_m = \overline{\widehat{V}}_{-m}, \quad m \in \mathbb{Z}. \tag{98}$$

Moreover, the symmetry in (15) implies that the energy E satisfies

$$E(k) = E(-k), \quad k \in I_{\mathrm{bz}}. \tag{99}$$

In the following, we show that these two symmetries allow us to run the construction backward in k to show the realty of W_0; it is a consequence of the uniqueness of infinite-dimensional, linear initial value problems with compact coefficients (cf. e.g., [8]).

Theorem 3.16. *Suppose the vector $\widetilde{\mathbf{y}}(k)$ for $k \in I_{\mathrm{bz}}$ and its corresponding function $\widetilde{\alpha}$ are constructed based on Theorem 3.12 with ϕ given in Corollary 3.14. Suppose further that the vector $\mathbf{y}(k) = \{a_m(k)\}_{m=-\infty}^{\infty}$ is related to $\widetilde{\mathbf{y}}(k)$ by (76). It is always possible to choose the elements in the initial vector $\mathbf{y}(-\Omega/2) = \mathbf{y}_0$ in (63) to satisfy*

$$a_m(-\Omega/2) = \bar{a}_{-m}(\Omega/2), \quad m \in \mathbb{Z}, \tag{100}$$

so that the function $\widetilde{\alpha}$ *satisfies*

$$\widetilde{\alpha}(k + m\Omega) = \overline{\widetilde{\alpha}}(-k - m\Omega), \quad k \in I_{\mathrm{bz}}, \, m \in \mathbb{Z}. \tag{101}$$

As a result, the Wannier function W_0 *corresponding to* $\widetilde{\alpha}$ *can always be chosen to be real.*

Proof. Suppose that $\mathbf{y}(k) = \{a_m(k)\}_{m=-\infty}^{\infty}$ satisfies the ODE (61). We first show that there is a choice of $\mathbf{y}_0$ such that if $\mathbf{y}$ also satisfies (63), then

$$a_m(-\Omega/2) = \overline{a}_{-m}(\Omega/2), \quad m \in \mathbb{Z}. \tag{102}$$

For now, we let $\mathbf{y}_0$ be an arbitrary solution to (56) with $k = -\Omega/2$.

By (57), (58) and (98), we observe that the operator $\mathbf{D}(k) + \mathbf{V}$ in (56) is invariant under the operation consisting of the substitutions $m \to -m$, $n \to -n$, $k \to -k$ and complex conjugation. Applying such a transformation to (56), together with (99), gives

$$(\mathbf{D}(k) + \mathbf{V})\mathbf{x}(k) = E(k)\mathbf{x}(k), \tag{103}$$

where $\mathbf{x}(k) = \{\overline{a}_{-m}(-k)\}_{m=-\infty}^{\infty}$. Since $E(k)$ is non-degenerate and $\|\mathbf{y}(k)\|$ is constant, there must exist a real θ such that

$$a_m(-\Omega/2) = e^{\mathrm{i}\theta}\overline{a}_{-m}(\Omega/2), \quad m \in \mathbb{Z}. \tag{104}$$

Replacing the original $\mathbf{y}_0$ with $e^{-\mathrm{i}\phi_0}\mathbf{y}_0$ with $\phi_0 = \theta/2$ produces $\mathbf{y}$ that satisfies (102).

Next, suppose that $\mathbf{y}(k) = \{a_m(k)\}_{m=-\infty}^{\infty}$ is a solution to the ODE (61) satisfying (102). Applying the aforementioned transformation to (61) yields

$$\frac{\mathrm{d}}{\mathrm{d}k}\mathbf{x}(k) = -\mathbf{\Theta}^{\dagger}(k)\mathbf{S}(k)\mathbf{x}(k), \tag{105}$$

where again we have $\mathbf{x}(k) = \{\overline{a}_{-m}(-k)\}_{m=-\infty}^{\infty}$. Noting that (102) is equivalent to the condition that $\mathbf{y}(-\Omega/2) = \mathbf{x}(-\Omega/2)$, it follows from the uniqueness theorem for initial value problems that $\mathbf{x}(k) = \mathbf{y}(k)$ for $k \in [-\Omega/2, \Omega/2]$. Consequently, we have

$$a_m(k) = \overline{a}_{-m}(-k), \quad k \in [-\Omega/2, \Omega/2], \, m \in \mathbb{Z}. \tag{106}$$

Together with (77) and $\phi(k) = -\phi(-k)$ in (93), the above relation yields the desired result:

$$\widetilde{\alpha}(k + m\Omega) = e^{-\mathrm{i}\phi(k)}a_m(k) = e^{\mathrm{i}\phi(-k)}\overline{a}_{-m}(-k) = \overline{\widetilde{\alpha}}(-k - m\Omega) \quad m \in \mathbb{Z}. \tag{107}$$

$$\square$$

We observe that the condition in (100) is applicable after solving (61) for determining $\mathbf{y}$ in $[-\Omega/2, \Omega/2]$; we can obtain a constant phase $e^{-\mathrm{i}\phi_0}$ to enforce (100) so the the resulting W_0 is real.

4 Constructing optimally localized Wannier functions

In this section, we use the apparatus in Section 3 to construct optimally localized Wannier function. Suppose we are given a real, piecewise continuous potential V with period given by the lattice constant a. Let Ω be defined by (5) and I_{bz} by (7). We also select the band for which the Wannier function will be constructed.

The construction can be divided into three stages. First, we compute the Fourier series of V as in (53) and transform (32) into its Fourier version in (56). Then, we compute the energy E_0 and the eigenvector $\mathbf{y}_0$ at $k = -\Omega/2$ for the selected band. Second, we solve the ODE defined by (60) and (61) in $[-\Omega/2, \Omega/2]$ with initial conditions given by E_0 and $\mathbf{y}_0$ at $k = -\Omega/2$. From $\mathbf{y}(-\Omega/2)$ and $\mathbf{y}(\Omega/2)$, we determine a constant ϕ_0 to enforce (100) and the Zak phase ϕ_{zak} in (68). This is followed by carrying out a gauge transform to obtain the new $\widetilde{\mathbf{y}}$ in (76) with the Berry phase ϕ given by (80). Using the definition of $\widetilde{\mathbf{y}}$ in (77), we obtain a function $\widetilde{\alpha}$ in $\mathbb{R}$ that defines the Fourier transform of the Wannier function by (49). Finally, we Fourier transform $\widetilde{\alpha}$ to obtain the Wannier function W_0 by (48). This Wannier function is guaranteed to be optimal by Theorem 3.12 and Corollary 3.14, and real by Theorem 3.16.

We now describe a high-order numerical procedure for the construction described above. Sections 4.1.1 and 4.1.2 contain the computation for the first stage, the Sections 4.1.3 and 4.1.4 for the second stage, and Sections 4.1.5 and 4.1.6 for the last stage.

Remark 4.1. *There are various schemes resembling the one above. For example, [20] and Section IV.C in [15] describe a conceptually similar algorithm. Instead of solving the ODE in the second stage above, u_k, for varying values of k on a grid in I_{bz}, are computed independently and an approximate SVD-based approach is used to fix the broken phase. For a large number of grid points, the SVD-based approach should yield a decent approximation to the ODE solution above (see Remark 4.8 for further discussion). However, the authors in [15, 20] did not prove the solution's optimality as stated in Corollary 3.14. Another similar approach can be found in Section III in [4], where an approximate parallel transport scheme was used to construct u_k. Different matching conditions are imposed to obtain assignments of u_k that are only continuous in k, producing Wannier functions that decay as $1/x^2$, which are then used as inputs for the general local optimization approach in [15].*

4.1 Numerical procedure

The input to our method are the potential function V and its period a (so $\Omega = \frac{2\pi}{a}$), $M \in \mathbb{N}$ specifying the number of degrees of freedom in the real space, $K \in \mathbb{N}$ specifying the number of degrees of freedom in the momentum space, and $\ell \in \mathbb{N}$ indicating the band index.

4.1.1 Computing the Fourier interpolant of the potential

We first let $\{t_j\}_{j=1}^{2M+1}$ be given by

$$t_j = -\frac{\pi}{\Omega} + \frac{2\pi(j-1)}{\Omega(2M+1)}, \quad j = 1, 2, \ldots, 2M+1. \tag{108}$$

We then evaluate $V(t_j)$ for all $j = 1, 2, \ldots, 2M+1$. The $2M+1$ Fourier coefficients $\{\widehat{V}_j\}_{j=-M}^{M}$ in (53) are then computed via a discrete Fourier transform. We then form the potential matrix $\mathbf{V} \in \mathbb{C}^{(2M+1)\times(2M+1)}$

with entries given by

$$\mathbf{V}_{m,n} = \begin{cases} \widehat{V}_{m-n} & |m-n| \leq M \\ 0 & |m-n| > M \end{cases}, \quad m, n = 1, 2, \ldots, 2M+1. \tag{109}$$

Remark 4.2. *Computing $\{\widehat{V}_j\}_{j=-M}^M$ can be done in $O(M \log M)$ floating point operations using the fast Fourier transform. Forming the matrix $\mathbf{V}$ takes $O(M^2)$ floating point operations.*

4.1.2 Obtaining an initial condition

We define $\mathbf{D}(k) \in \mathbb{C}^{(2M+1) \times (2M+1)}$ to have entries

$$\mathbf{D}(k)_{m,n} = (k + (m - M - 1)\Omega)^2 \delta_{mn}, \quad m, n = 1, 2, \ldots, 2M+1. \tag{110}$$

We then solve the eigenvalue problem

$$\left(\mathbf{D}\left(-\frac{\Omega}{2} \right) + \mathbf{V} \right) \mathbf{y}^{(0)} = E_0 \mathbf{y}^{(0)}, \tag{111}$$

where E_0 is the ℓth smallest eigenvalue and $\mathbf{y}^{(0)}$ is the corresponding eigenvector, normalized as in (55) such that

$$\|\mathbf{y}^{(0)}\| = 1/a. \tag{112}$$

This can be done using the standard QR algorithm (cf. e.g., [11, Section 8.3]).

Remark 4.3. *The convergence of the discretization depends on the smoothness of V. If $V \in C^m(\mathbb{R}; \mathbb{R})$, then the error will decrease at a rate of $O(M^{-m})$. In particular, if $V \in C^\infty(\mathbb{R}; \mathbb{R})$, then the error will decrease superalgebraically (i.e., at a rate faster than $O(M^{-m})$ for any $m \in \mathbb{N}$). We refer the reader to [3, Chapter 2] for more details.*

Remark 4.4. *If the standard QR algorithm is used, this procedure takes $O(M^3)$ floating point operations. It should be noted that the coefficient matrix in (111) is highly structured. In particular, it consists of a diagonal matrix plus a Toeplitz matrix, which can likely be exploited to accelerate this step.*

4.1.3 Solving the initial value problem

We now define $\mathbf{S} \in \mathbb{C}^{(2M+1) \times (2M+1)}$ to have entries

$$\mathbf{S}(k)_{m,n} = 2(k + (m - M - 1)\Omega))\delta_{mn}, \quad m, n = 1, 2, \ldots, 2M+1. \tag{113}$$

We let $\mathbf{I}$ denote the $(2M+1) \times (2M+1)$ identity matrix. We now solve the initial value problem

$$\mathbf{y}'(k) = -(\mathbf{D}(k) - E(k)\mathbf{I} + \mathbf{V})^\dagger \mathbf{S}(k)\mathbf{y}(k) \tag{114}$$

$$E'(k) = \frac{\mathbf{y}(k)^* \mathbf{S}(k) \mathbf{y}(k)}{\mathbf{y}(k)^* \mathbf{y}(k)} \tag{115}$$

$$\mathbf{y}\left(-\frac{\Omega}{2} \right) = \mathbf{y}^{(0)} \tag{116}$$

$$E\left(-\frac{\Omega}{2} \right) = E_0, \tag{117}$$

where E_0 and $\mathbf{y}^{(0)}$ were obtained in (111). We would like to solve this initial value problem from $-\Omega/2$ to $\Omega/2$. We note that by (112) $\|\mathbf{y}(k)\| = 1/a$ for $k \in [-\Omega/2, \Omega/2]$.

We use a time-stepping scheme to determine the values

$$\mathbf{y}^{(j)} \approx \mathbf{y}(k_j), \quad j = 1, 2, \ldots, K, \tag{118}$$

where

$$k_j = -\frac{\Omega}{2} + \frac{\Omega(j-1)}{K-1}, \quad j = 1, 2, \ldots, K. \tag{119}$$

At each time-step the coefficient matrix

$$\boldsymbol{\Theta}(k) = \mathbf{D}(k) + \mathbf{V} - E(k)\mathbf{I} \tag{120}$$

must be inverted. In this paper, we compute $\boldsymbol{\Theta}(k)^\dagger$ by first computing a singular value decomposition of $\boldsymbol{\Theta}(k)$, discarding the mode corresponding to the smallest singular value (see Remark 2.9), and applying the inverse of the decomposition directly to the right-hand side (cf. e.g., [1, Section 2.6.1]). It is not as efficient as the approach in Remark 2.10, but this is not significant for one dimension problems.

Remark 4.5. *In our numerical experiments, we use the standard 4th-order Runge–Kutta method to solve the initial value problem for simplicity (cf. e.g., [18, Section 12.5]); the resulting solution produces a solution that converges at rate $O(K^{-4})$. Even higher order, for example $O(K^{-12})$, can be achieved by switching to a spectral deferred scheme (cf. [10]).*

Remark 4.6. *At each time-step, the dominant cost is that of computing the singular value decomposition of $\boldsymbol{\Theta}(k)$, which consists of $O(M^3)$ floating point operations. Since K time steps are taken, this stage of our procedure requires $O(KM^3)$ floating point operations.*

Remark 4.7. *The explicit dependence of (114) on $E(k)$ and the ODE for the energy (115) can be eliminated altogether by replacing $E(k)$ with the Rayleigh quotient*

$$\frac{\mathbf{y}(k)^*(\mathbf{D}(k) + \mathbf{V})\mathbf{y}(k)}{\mathbf{y}(k)^*\mathbf{y}(k)} \tag{121}$$

when $\mathbf{y}(k)$ is known. This also has the advantage that the error the computed $E(k)$ is the squared error in the eigenvectors from the ODE solve.

Remark 4.8. *If two vectors $\mathbf{v}_1 = \mathbf{y}(k)$ and $\mathbf{v}_2 = \mathbf{y}(k + \Delta k)$ for small Δk are obtained by direct eigensolves, the phases of $\mathbf{v}_1$ and $\mathbf{v}_2$ are independent. In [15] and [20], it is pointed out that the (approximate) parallel transported vector $\widetilde{\mathbf{v}}_2$ from $\mathbf{v}_1$ given by $\widetilde{\mathbf{v}}_2 = e^{-i\beta}\mathbf{v}_2$, where β is the phase given by $\beta = -\operatorname{Im}\log(\mathbf{v}_1^*\mathbf{v}_2)$. When used as a time-stepping algorithm, we empirically observed that the resulting solutions have accuracy of order $O(K^{-2})$ for $k \in I_{\mathrm{bz}}$. Thus, when only low accuracy is required, this is a viable algorithm, provided the eigenvalues and eigenvectors for the band are computed first at points $\{k_j\}_{j=1}^K$.*

4.1.4 Correcting the phase

According to Lemma 3.10, we can determine ϕ_{zak} from

$$\mathbf{y}_{m+1}^{(0)} = e^{i\phi_{\mathrm{zak}}}\mathbf{y}_m^{(K)}, \tag{122}$$

for any $m \in \{1, 2, \ldots, 2M\}$ to fix the jumps. Moreover, the condition in (100) allows us to determine ϕ_0 by

$$e^{-i\phi_0}\mathbf{y}_m^{(0)} = e^{i\phi_0}\overline{\mathbf{y}}_{2M+2-m}^{(K)} \tag{123}$$

for any $m \in \{M+1, M+2, \ldots, 2M+1\}$ for the realty of Wannier functions. By (93), we define the Berry phase

$$\phi(k_j) = \frac{\phi_{\text{zak}}}{\Omega}k_j, \quad j = 1, 2, \ldots, K. \tag{124}$$

We form the new vector

$$\widetilde{\mathbf{y}}^{(j)} = e^{-i\phi_0}e^{-i\phi(k_j)}\mathbf{y}^{(j)}, \quad j = 1, 2, \ldots, K, \tag{125}$$

which will lead to the real optimally localized Wannier function.

4.1.5 Computing $\widehat{W}_0$

Given $x \in \mathbb{R}$, we then evaluate $W_0(x)$ using a trapezoidal rule discretization of (48). We set

$$N = (2M+1)K \tag{126}$$

and define $\{\widetilde{\alpha}_l\}_{l=1}^N$ by

$$\widetilde{\alpha}_{(m-1)K+j} = \widetilde{\mathbf{y}}_m^{(j)}, \quad j = 1, 2, \ldots, K, \, m = 1, 2, \ldots, 2M+1. \tag{127}$$

4.1.6 Evaluate the Wannier function

Finally, we compute

$$W_0(x) = \frac{1}{K}\sum_{j=1}^{K}\sum_{m=1}^{2M+1}\widetilde{\alpha}_{(m-1)K+j}e^{ix(k_j+(m-M-1)\Omega)} \tag{128}$$

Remark 4.9. *Since the integrand in (48) is analytic and exponentially decaying as $\xi \to \pm\infty$, the error in the trapezoidal rule approximation decays with respect to K as $O(e^{-cK})$ for some $c > 0$ (cf. e.g., [19]).*

Remark 4.10. *The cost of direct evaluation of (128) for each x is $O(KM)$. The evaluation of (128) for a collection of points can be accelerated using the fast Fourier transform or non-uniform fast Fourier transform (cf. e.g., [12]).*

5 Detailed description of the procedure

The input of our numerical procedure is the lattice spacing $a > 0$, potential function V, $M \in \mathbb{N}$ specifying the number of spatial degrees of freedom, $K \in \mathbb{N}$ specifying the number of degrees of freedom in momentum space, $\ell \in \mathbb{N}$ indicating the band index, and a collection of points $\{x_j\}_{j=1}^P$ on which the Wannier function is to be evaluated. The output of Algorithm 1 is $\{W_0(x_j)\}_{j=1}^P$.

For clarity, we have isolated the evaluation of the right-hand side of the ODEs (114) and (115) into Algorithm 1A. Algorithm 1A takes as input $\mathbf{y}(k) \in \mathbb{C}^{2M+1}$, $E(k) \in \mathbb{R}$, and $k \in \mathbb{R}$ and outputs the right-hand side of the ODE in (114) and (115).

Algorithm 1

Input: $a > 0$, $M \in \mathbb{N}$, $K \in \mathbb{N}$, $V : \mathbb{R} \to \mathbb{R}$, $\ell \in \mathbb{N}$, $\{x_j\}_{j=1}^{P} \subset \mathbb{R}$

Output: $\{W_0(x_j)\}_{j=1}^{P}$

Step 1 [Forming the potential matrix $\mathbf{V}$].

1. Evaluate $\{t_j\}_{j=1}^{2M+1}$ by (108).

2. Evaluate $\{V(t_j)\}_{j=1}^{2M+1}$.

3. Apply the discrete Fourier transform to $\{V(t_j)\}_{j=1}^{2M+1}$ to produce the coefficients of the Fourier interpolant $\{\widehat{V}_j\}_{j=-M}^{M}$.

4. Form $\mathbf{V} \in \mathbb{C}^{(2M+1) \times (2M+1)}$ using the formula (109).

Step 2 [Obtaining the initial condition].

1. Compute the matrix $\mathbf{D}(-\Omega/2)$ using the formula (110).

2. Solve the eigenvalue problem (111) for the ℓth smallest eigenvalue $E^{(0)}$ and corresponding eigenvector $\mathbf{y}^{(0)}$ using the QR algorithm and ensuring $\|\mathbf{y}^{(0)}\| = 1$.

Step 3 [Solving the initial value problem using RK4].

1. Set k_j for $j = 1, 2, \ldots, K$ via (119). Set $\Delta k = \Omega/K$.

2. Do $j = 1, 2, \ldots, K$

   (a) Use Algorithm 1A with inputs $\mathbf{V}$, $\mathbf{y}^{(j-1)}$, and k_j to obtain $\mathbf{w}^{(1)}$ and $F^{(1)}$.

   (b) Use Algorithm 1A with inputs $\mathbf{V}$, $\mathbf{y}^{(j-1)} + (\Delta k/2)\mathbf{w}^{(1)}$, and $k_j + \Delta k/2$ to obtain $\mathbf{w}^{(2)}$ and $F^{(2)}$.

   (c) Use Algorithm 1A with inputs $\mathbf{V}$, $\mathbf{y}^{(j-1)} + (\Delta k/2)\mathbf{w}^{(2)}$, and $k_j + \Delta k/2$ to obtain $\mathbf{w}^{(3)}$ and $F^{(3)}$.

   (d) Use Algorithm 1A with inputs $\mathbf{V}$, $\mathbf{y}^{(j-1)} + (\Delta k)\mathbf{w}^{(3)}$, and $k_j + \Delta k$ to obtain $\mathbf{w}^{(4)}$ and $F^{(4)}$.

   (e) Set

$$\mathbf{y}^{(j)} = \mathbf{y}^{(j-1)} + \frac{\Delta k}{6} \left(\mathbf{w}^{(1)} + 2\mathbf{w}^{(2)} + 2\mathbf{w}^{(3)} + \mathbf{w}^{(4)} \right) \tag{129}$$

$$E^{(j)} = E^{(j-1)} + \frac{\Delta k}{6} \left(F^{(1)} + 2F^{(2)} + 2F^{(3)} + F^{(4)} \right). \tag{130}$$

   End Do

Step 4 [Phase correction]

1. Set ϕ_0 by (125).

2. Set $\phi_{\text{zak}} = \arg(\mathbf{y}_m^{(K)}/\mathbf{y}_{m+1}^{(0)})$.

3. Do $j = 1, 2, \ldots, K$

    (a) Set $\widetilde{\mathbf{y}}^{(j)} = \exp(-\mathrm{i}\phi_0)\exp\left(-\mathrm{i}\phi_{\text{zak}}k_j/\Omega\right)\mathbf{y}^{(j)}$.

    End Do

Step 5 [Tabulate Fourier transform of W_0]

1. Set $N = (2M + 1)K$.

2. Compute $\{\widetilde{\alpha}_l\}_{l=1}^{N}$ by (127).

Step 6 [Evaluate Wannier function]

1. Compute $\{W(x_j)\}_{j=1}^{P}$ using (128).

Algorithm 1A

Input: $\mathbf{V} \in \mathbb{C}^{(2M+1)\times(2M+1)}$, $\mathbf{y}(k) \in \mathbb{C}^{2M+1}$, $E(k) \in \mathbb{R}$, $k \in \mathbb{R}$
Output: $\mathbf{y}'(k) \in \mathbb{C}^{2M+1}$, $E'(k) \in \mathbb{R}$

1. Evaluate $\mathbf{S}(k)$ using (113).

2. Compute $E'(k)$ using (115).

3. Compute $\mathbf{z} = \mathbf{S}(k)\mathbf{y}(k)$.

4. Evaluate $\mathbf{\Theta}(k)$ defined in (120) using (57).

5. Compute the singular value decomposition of $\mathbf{\Theta}(k) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ using the QR algorithm.

6. Set $\mathbf{y}'(k) = \mathbf{V}_{1:2M+1,1:2M}(\mathbf{\Sigma}_{1:2M,1:2M})^{-1}(\mathbf{U}_{1:2M+1,1:2M})^*\mathbf{z}$.

6 Numerical experiments

We implemented the procedure described in Section 4 in Fortran. We fix V, a, M, and ℓ and increase K until 10 digits of accuracy is obtained. The timings reported are the CPU times required to obtain the Fourier transform of W_0 evaluated on a uniform grid (i.e., Steps 1–5 in Section 5). To obtain an estimate of the error we compute

$$E_{\text{RK4}}^{(\ell)} = \left\|\mathbf{y}^{(K)} - \mathbf{y}\left(\frac{\Omega}{2}\right)\right\|, \tag{131}$$

where $\mathbf{y}^{(K)}$ is defined in (129) and $\mathbf{y}(\Omega/2)$ is computed by directly solving the eigenvalue problem via the QR algorithm. We also report

$$E_{\text{imag}}^{(\ell)} = \frac{\max_{j=1,2,\ldots,1000}\left|\text{Im}\left[W_0\left(-\frac{\pi}{\Omega} + \frac{2\pi j}{1000\Omega}\right)\right]\right|}{\max_{j=1,2,\ldots,1000}\left|W_0\left(-\frac{\pi}{\Omega} + \frac{2\pi j}{1000\Omega}\right)\right|}, \tag{132}$$

where we evaluate W_0 using (128). All experiments were conducted in double precision on a single core of an Apple M3 PRO CPU with the GNU Fortran compiler.
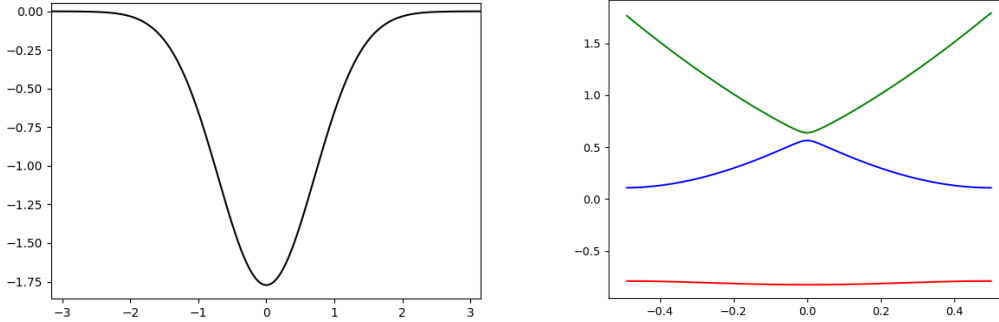
Figure 1: Plot of potential (left) and $E^{(j)}(k)$ (right) for $j = 1, 2, 3$ (red, blue, green, resp.)

6.1 Gaussian potential

In this experiment, we set $a = 2\pi$ (so $\Omega = 1$), $M = 10$, and $\ell = 1, 2, 3$. The potential is taken to be

$$V(x) = -\frac{1}{2} - \sum_{j=1}^{5} e^{-j^2/4} \cos(j\Omega x) \tag{133}$$

and plotted in Figure 1 (left). The corresponding energies are plotted in Figure 1 (right). The Wannier functions are plotted in Figure 2.

The relevant quantities are:

K	time (s)	$E_{\mathrm{RK4}}^{(1)}$	$E_{\mathrm{imag}}^{(1)}$	$E_{\mathrm{RK4}}^{(2)}$	$E_{\mathrm{imag}}^{(2)}$	$E_{\mathrm{RK4}}^{(3)}$	$E_{\mathrm{imag}}^{(3)}$
51	2.11E–01	2.04E–09	5.28E–10	1.66E–04	3.81E–04	7.17E+00	3.53E–01
101	3.42E–01	1.33E–10	3.26E–11	8.18E–06	1.36E–05	6.53E+00	1.20E–01
201	6.66E–01	8.51E–12	2.03E–12	5.58E–07	9.56E–07	2.66E+00	1.56E–02
401	1.25E+00			3.60E–08	6.30E–08	4.35E–01	7.57E–04
801	2.42E+00			2.28E–09	4.02E–09	3.51E–02	2.36E–05
1601	4.82E+00			1.43E–10	2.54E–10	1.56E–03	1.62E–06
3201	9.39E+00			8.94E–12	1.59E–11	9.17E–05	1.27E–07
6401	1.84E+01					5.59E–06	8.67E–09
12801	3.62E+01					3.44E–07	5.64E–10
25601	7.06E+01					2.22E–08	3.78E–11
51201	1.38E+02					2.70E–11	2.57E–12

We observe that the expected fourth-order convergence with respect to K down to an accuracy of about 10 digits. Moreover, the $O(K)$ complexity can be observed.

6.2 Asymmetric potential

In this experiment, we set $a = 2\pi$ (so that $\Omega = 1$), $M = 15$, and $\ell = 1, 2, 3$. The potential is given by

$$V(x) = \frac{1}{4}\left(1 + 2\sin(2\Omega x) + 3e^{\cos(\Omega x)}\right). \tag{134}$$
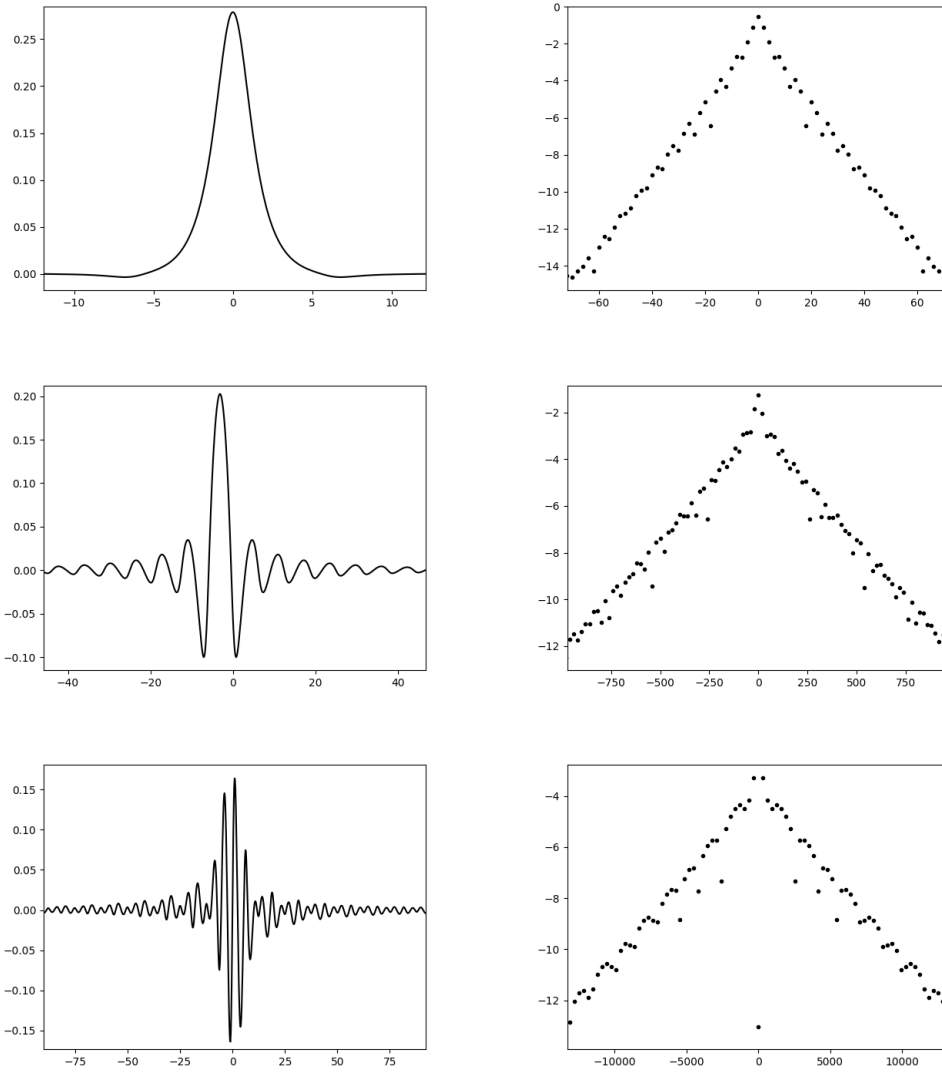
24

Figure 2: $W_0^{(j)}(x)$ (left) and $\log_{10}(|W_0^{(j)}(x)|)$ (right) for $j = 1, 2, 3$ (top, middle, bottom, resp.)

The potential is plotted in Figure 3 (left) and the energies are plotted in Figure 3 (right). The Wannier functions are plotted in Figure 4. The relevant quantities are:
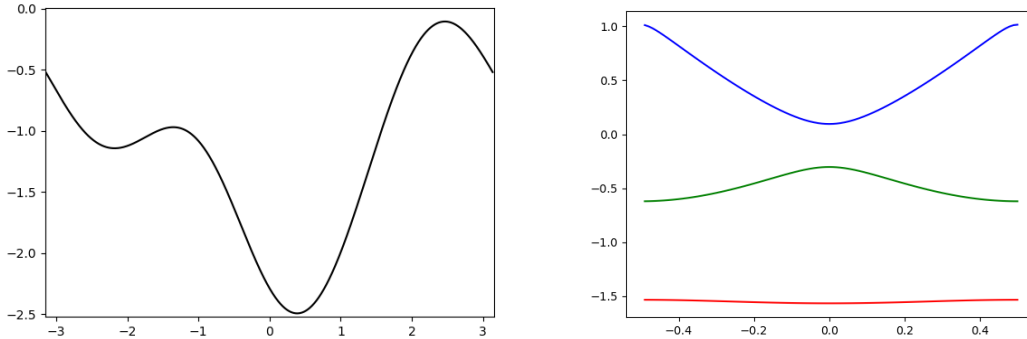
Figure 3: $W_0^{(j)}(x)$ (left) and $\log_{10}(|W_0^{(j)}(x)|)$ (right) for $j = 1, 2, 3$ (top, middle, bottom, resp.)

K	time (s)	$E_{\text{RK4}}^{(1)}$	$E_{\text{imag}}^{(1)}$	$E_{\text{RK4}}^{(2)}$	$E_{\text{imag}}^{(2)}$	$E_{\text{RK4}}^{(3)}$	$E_{\text{imag}}^{(3)}$
51	1.00E–01	2.79E–09	7.18E–10	9.63E–07	5.08E–07	1.16E–02	2.68E–05
101	1.68E–01	1.78E–10	4.45E–11	6.20E–08	3.22E–08	6.03E–04	2.00E–06
201	2.90E–01	1.12E–11	2.77E–12	3.92E–09	2.02E–09	3.57E–05	1.43E–07
401	5.36E–01			2.47E–10	1.27E–10	2.16E–06	9.46E–09
801	1.02E+00			1.57E–11	8.07E–12	1.33E–07	6.08E–10
1601	2.01E+00					8.24E–09	3.86E–11
3201	4.00E+01					5.07E–10	2.47E–12
6401	7.89E+01					6.49E–11	4.07E–13

We are again able to rapidly compute a solution down to 10 digits of accuracy.

7 Conclusions

In this work, we developed a robust, high-order numerical method for computing Wannier functions for one-dimensional crystalline insulators. We presented complete analysis that shows the constructed Wannier functions are exponentially localized and globally optimal in terms of their variance. Furthermore, the Wannier functions can always be chosen to be real. The analysis and algorithm in this paper can also be viewed as a constructive proof of the existence of exponentially localized Wannier functions for an isolated single band in one dimension. We note that the construction can be applied easily to many other self-adjoint analytic families of operators beyond the Schrödinger operators in (9).

Since Kato's theorem (Theorem 2.3) guarantees analyticity even when eigenvalues become degenerate in one dimension, the approach in this paper has been easily extended to isolated multiband cases in one dimension. Similarly, globally optimal solutions can be computed analytically. For higher dimensions, topological phenomena show up in the construction of Wannier functions and new ideas beyond this paper are required. Nonetheless, the algorithms in one dimension are the basic building block for higher-dimensional cases, and the extension of the approach in this paper has been worked out and will be reported at a later date.
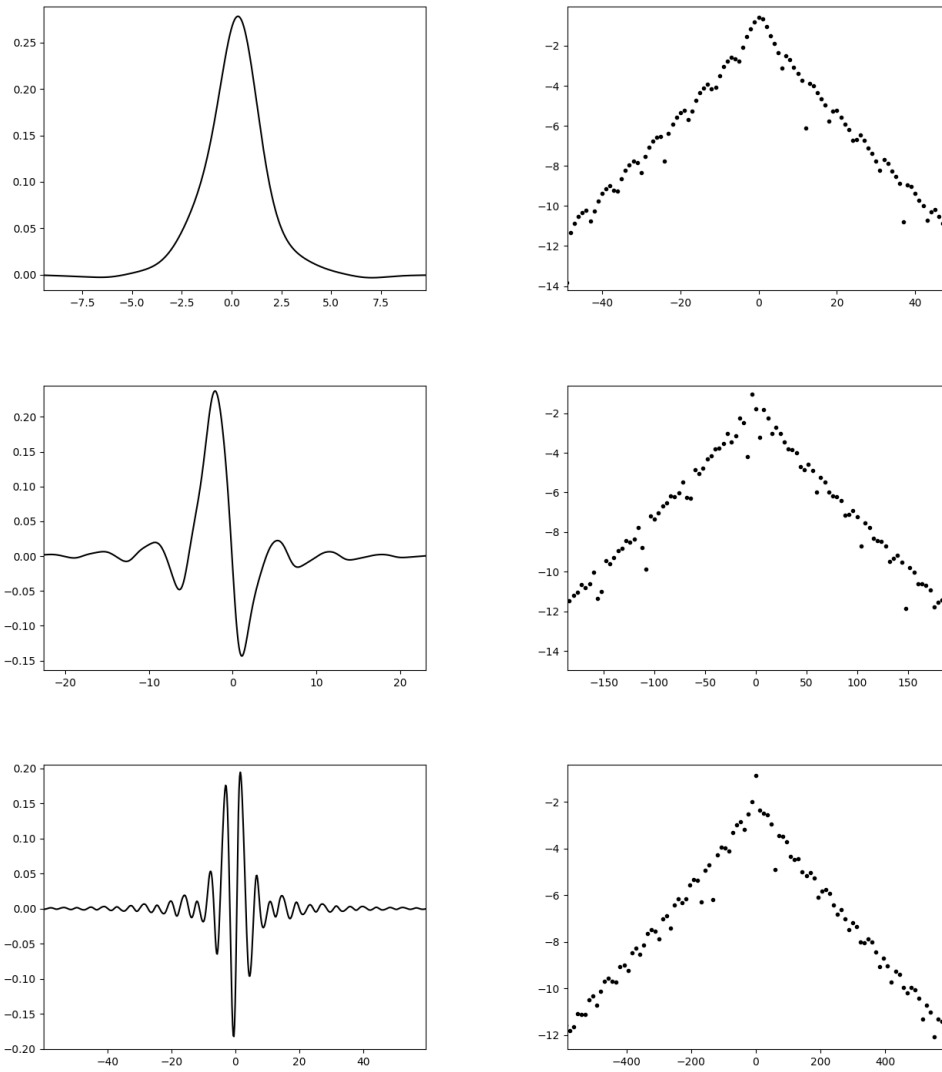
Figure 4: $W_0^{(j)}(x)$ (left) and $\log_{10}(|W_0^{(j)}(x)|)$ (right) for $j = 1, 2, 3$ (top, middle, bottom, resp.)

References

[1] Åke Björck. *Numerical methods in matrix computations*, volume 59. Springer, 2015.

[2] EI Blount. Formalisms of band theory. In *Solid state physics*, volume 13, pages 305–373. Elsevier, 1962.

[3] John P. Boyd. *Chebyshev and Fourier spectral methods*. Courier Corporation, 2001.

[4] Éric Cancès, Antoine Levitt, Gianluca Panati, and Gabriel Stoltz. Robust determination of maximally localized Wannier functions. *Physical Review B*, 95(7):075114, 2017.

[5] Horia D Cornean, Ira Herbst, and Gheorghe Nenciu. On the construction of composite wannier functions. In *Annales Henri Poincaré*, volume 17, pages 3361–3398. Springer, 2016.

[6] Anil Damle, Antoine Levitt, and Lin Lin. Variational formulation for Wannier functions with entangled band structure. *Multiscale Modeling & Simulation*, 17(1):167–191, 2019.

[7] Anil Damle and Lin Lin. Disentanglement via entanglement: a unified method for wannier localization. *Multiscale Modeling & Simulation*, 16(3):1392–1410, 2018.

[8] Klaus Deimling. *Ordinary differential equations in Banach spaces*, volume 596. Springer, 2006.

[9] Jacques Des Cloizeaux. Analytical properties of n-dimensional energy bands and Wannier functions. *Physical Review*, 135(3A):A698, 1964.

[10] Alok Dutt, Leslie Greengard, and Vladimir Rokhlin. Spectral deferred correction methods for ordinary differential equations. *BIT Numerical Mathematics*, 40:241–266, 2000.

[11] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Univeristy Press, 4 edition, 2013.

[12] Leslie Greengard and June-Yub Lee. Accelerating the nonuniform fast Fourier transform. *SIAM review*, 46(3):443–454, 2004.

[13] Tosio Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 1980.

[14] Walter Kohn. Analytic properties of Bloch waves and Wannier functions. *Physical review*, 115(4):809, 1959.

[15] Nicola Marzari and David Vanderbilt. Maximally localized generalized Wannier functions for composite energy bands. *Physical review B*, 56(20):12847, 1997.

[16] Giovanni Pizzi, Valerio Vitale, Ryotaro Arita, Stefan Blügel, Frank Freimuth, Guillaume Géranton, Marco Gibertini, Dominik Gresch, Charles Johnson, Takashi Koretsune, et al. Wannier90 as a community code: new features and applications. *Journal of Physics: Condensed Matter*, 32(16):165902, 2020.

[17] M Reed and B Simon. Analysis of operators (methods of modern mathematical physics iv), acad. *Press, Orlando*, 1978.

[18] Endre Süli and David F. Mayers. *An introduction to numerical analysis.* Cambridge University Press, 2003.

[19] Lloyd N. Trefethen and J. A. C. Weideman. The exponentially convergent trapezoidal rule. *SIAM review*, 56(3):385–458, 2014.

[20] David Vanderbilt. *Berry phases in electronic structure theory: electric polarization, orbital magnetization and topological insulators.* Cambridge University Press, 2018.

[21] Gabriel Weinreich. Solids: elementary theory for advanced students. 1965.

[22] Joshua Zak. Berry's phase for energy bands in solids. *Physical review letters*, 62(23):2747, 1989.

[23] Antoni Zygmund. *Trigonometric series*, volume 1. Cambridge university press, 2002.