



# Numerical quadrature in the Brillouin zone for periodic Schrödinger operators

Éric Cancès<sup>1</sup> · Virginie Ehrlicher<sup>1</sup> · David Gontier<sup>2</sup> · Antoine Levitt<sup>3</sup> · Damiano Lombardi<sup>4</sup>

Received: 22 May 2018 / Revised: 18 October 2019 / Published online: 7 January 2020  
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

## Abstract

As a consequence of Bloch's theorem, the numerical computation of the fermionic ground state density matrices and energies of periodic Schrödinger operators involves integrals over the Brillouin zone. These integrals are difficult to compute numerically in metals due to discontinuities in the integrand. We perform an error analysis of several widely-used quadrature rules and smearing methods for Brillouin zone integration. We precisely identify the assumptions implicit in these methods and rigorously prove error bounds. Numerical results for two-dimensional periodic systems are also provided. Our results shed light on the properties of these numerical schemes, and provide guidance as to the appropriate choice of numerical parameters.

**Mathematics Subject Classification** 65D30 · 65L20 · 65Z05

---

✉ David Gontier  
gontier@ceremade.dauphine.fr

Éric Cancès  
eric.cances@enpc.fr

Virginie Ehrlicher  
virginie.ehrlicher@enpc.fr

Antoine Levitt  
antoine.levitt@inria.fr

Damiano Lombardi  
damiano.lombardi@inria.fr

<sup>1</sup> CERMICS, Ecole des Ponts ParisTech and Inria Paris, Université Paris-Est, 6-8 avenue Blaise Pascal, 77455 Marne-la-Vallée, France

<sup>2</sup> CEREMADE, Université Paris-Dauphine, PSL University, 75016 Paris, France

<sup>3</sup> Inria Paris, CERMICS (ENPC), Université Paris-Est, 75589 Paris Cedex 12, France

<sup>4</sup> Inria Paris, COMMEDIA, 2 rue Simone Iff, 75012 Paris, France

## 1 Introduction

The computation of the electronic properties of a  $d$ -dimensional perfect crystal in a mean-field setting (e.g. Kohn–Sham density functional theory) formally requires to solve a periodic problem with infinitely many electrons. In practice, a truncation to a finite supercell composed of  $L^d$  crystal unit cells with periodic boundary conditions is necessary for the actual computation, and  $L$  is increased until an acceptable accuracy is achieved. Bloch’s theorem allows for a tremendous reduction of computational costs by an explicit block-diagonalization of the Hamiltonian operator, transforming an electronic problem for  $L^d N$  one-body wave functions, where  $N$  is the number of electron pairs per unit cell, to  $L^d$  electronic problems for  $N$  one-body wave functions. In the infinite- $L$  limit, the theorem states that properties of the perfect crystal can be obtained as an integral over the Brillouin zone (a  $d$ -dimensional torus) of properties of a parametrized system of  $N$  electron pairs. The truncation to a supercell of  $L^d$  unit cells can then be seen as a numerical quadrature of this integral. This leads to the famous Monkhorst–Pack numerical scheme [20].

Mathematically, the natural question is that of the speed of convergence of a given electronic property as  $L \rightarrow \infty$ . There appears a distinction between *insulators*, characterized by a band gap, and *metals*, with no band gap. In a sense that will be made precise later, electrons are localized in insulators, but delocalized in metals. Accordingly, the convergence of electronic properties is much faster for insulators than for metals. This translates to quantities of interest being very smooth across the Brillouin zone in insulators, so that the quadrature is very efficient: see for instance [11], which proves the exponential convergence with  $L$  of a number of properties of interest for insulators in the reduced Hartree–Fock model. In this paper, we aim to extend these results to the case of metals, under natural genericity assumptions on the band structure at the Fermi level (see Assumptions 1 and 2 in Sect. 3.1). To reduce the technical content of the paper, we limit ourselves to the case of independent electrons modeled by a single-particle Hamiltonian  $H = -\frac{1}{2}\Delta + V$  on  $L^2(\mathbb{R}^d)$ , where  $V$  is a given (non self-consistent) periodic potential.

For metals, because of the absence of a band gap, quantities of interest are discontinuous when the electronic bands  $\varepsilon_{n\mathbf{k}}$  cross the Fermi level  $\varepsilon_F$ , and specific quadrature rules have to be used to locate this singular set (the *Fermi surface*) and recover an acceptable convergence speed. In the simple case when the Fermi level intersects a single isolated band  $\mathcal{B} \ni \mathbf{k} \mapsto f(\mathbf{k}) := \varepsilon_{n_0, \mathbf{k}} \in \mathbb{R}$  (which can be the case for a metal with  $2n_0 + 1$  electrons per unit cell), the problem of computing the ground-state energy boils down to evaluating

$$E := E(\varepsilon_F)$$

where the function  $\mathbb{R} \ni \varepsilon \mapsto E(\varepsilon) \in \mathbb{R}$  is defined by

$$E(\varepsilon) = \int_{\mathcal{B}} f(\mathbf{k}) \mathbb{1}(f(\mathbf{k}) \leq \varepsilon) d\mathbf{k} \quad (1.1)$$

and the Fermi level  $\varepsilon_F$  by the constraint

$$\frac{1}{|\mathcal{B}|} \int_{\mathcal{B}} \mathbb{1}(f(\mathbf{k}) \leq \varepsilon_F) d\mathbf{k} = \frac{1}{2}. \quad (1.2)$$

The Fermi surface then is the level set  $\{\mathbf{k} \in \mathcal{B} \mid f(\mathbf{k}) = \varepsilon_F\}$  of the function  $f$ , and the Fermi level is chosen such that the volume of the set  $\Omega := \{\mathbf{k} \in \mathcal{B} \mid f(\mathbf{k}) < \varepsilon_F\}$  is half the one of the Brillouin zone. Similar quadrature problems are encountered in the level set method introduced by Osher and Sethian [23]. However, the Brillouin zone integration problem encountered in electronic structure calculation has some specificities. First, for one of the most important quantities of interest, namely the ground-state energy, the function to be integrated on  $\Omega$  is precisely the level set function  $f$  (see (1.1)), which requires a specific analysis. Second, the shape of the Fermi surface can be very complicated for real materials, and the required accuracy is much higher than in standard applications of the level set methods, where linear approximations of the boundary of  $\Omega$  from a fixed uniform grid are usually sufficient [22, 26]. Additional technical difficulties appear when the Fermi level intersects several bands, and when the quantity of interest is not the ground-state energy, but some observable involving the Bloch eigenfunctions of  $H$  and non only the Bloch eigenvalues  $\varepsilon_{n\mathbf{k}}$ , such as e.g. the ground-state density.

The most famous Brillouin zone integration method is the linear tetrahedron method and its improved version by Blöchl [2] (the Blöchl scheme is not covered by the results in this paper, and we plan to investigate it in a forthcoming paper). Other numerical quadratures have been proposed [19, 24] (see also [12] for an adaptive numerical scheme). In this paper, we study these quadrature rules, and prove that an interpolation of quantities of interest to order  $p$  coupled to a reconstruction of the Fermi surface with a method of order  $q$  leads, in general, to a total error of order  $L^{-(\min(p+1, q+1))}$ : this is the content of Theorem 4.5. On the other hand, the error made on the ground state energy is, to leading order, proportional to the error made on the number of electrons, which is kept fixed by varying the Fermi level: therefore, the energy is less sensitive to the location of the Fermi surface, and the leading order contribution to the error vanishes, leaving a total error of order  $L^{-(\min(p+1, 2q+2))}$ .

Another way to improve the convergence, and the most widely used method to compute properties of metals, is the *smearing* method [21]. This amounts to regularizing the discontinuities of the occupation numbers, restoring smoothness across the Brillouin zone. The smearing parameter  $\sigma > 0$ , which has the dimension of an energy, should be chosen small enough so that it does not change the properties of interest too much, but large enough so that the quadrature is efficient. In numerical codes, this choice is left to the users, who must use their expertise to select an appropriate value for the parameter  $\sigma$ . This is a complex task, and rules of thumb provide suboptimal choices of  $\sigma$ .

We show in this paper that, up to sub-exponential factors, the total error for a given smearing parameter  $\sigma$  and supercell of size  $L$  is bounded by  $C(\sigma^{p+1} + e^{-\eta\sigma L})$  for some  $C \in \mathbb{R}_+$  and  $\eta > 0$ , where  $p \geq 0$  is the order of the smearing method used (Theorem 5.11). This leads to the conclusion that  $\sigma$  should in practice be varied at the same time as  $L$  to balance the two sources of error. We also investigate the precise

convergence with respect to  $L$  at  $\sigma > 0$  fixed, and find the surprising result that, while the convergence is exponential when the Fermi–Dirac smearing is used, it is super-exponential (bounded by  $Ce^{-\eta L^{4/3}}$ ) when Gaussian-based smearing methods are used, due to the different complex-analytic properties of these functions. Such a phenomenon has already been observed in [27], in the context of the locality of the density matrix of metals with Gaussian smearing.

The structure of the paper is as follows. We introduce our notation and recall the basic properties of the periodic Schrödinger operator  $H = -\frac{1}{2}\Delta + V$  in Sect. 2. We carefully study the band structure of this operator in the vicinity of the Fermi level in Sect. 3. We analyze interpolation methods in Sect. 4 and smearing methods in Sect. 5. Technical results on the complex-analytic properties of the integrand in smearing methods are proved in “Appendix”. Two-dimensional numerical results illustrating our theoretical results are presented in Sect. 6. Some tests are also given where our assumptions on the band structure are violated (in the presence of a van Hove singularity or an eigenvalue crossing at the Fermi level).

## 2 Notation and model

In this section, we set our notation, and define the different quantities of interest.

Let  $d \in \{1, 2, 3\}$  denote the dimension of the crystal, and  $\mathcal{R} \subset \mathbb{R}^d$  the crystalline lattice. We denote by  $\mathcal{R}^*$  the dual (or reciprocal) lattice, by  $\Gamma$  the fundamental unit cell of  $\mathcal{R}$ , and we let  $\mathcal{B}$  be either the first Brillouin zone, or the fundamental unit cell of  $\mathcal{R}^*$ . Our results being independent of this choice, we will call  $\mathcal{B}$  “the Brillouin zone” for simplicity. The periodicities in  $\mathcal{R}$  and  $\mathcal{R}^*$  equip the sets  $\Gamma$  and  $\mathcal{B}$  with the topology of a  $d$ -dimensional torus.

Throughout this paper,  $\mathcal{C}^k(E, F)$  denotes the usual class of  $k$  times continuously differentiable functions from  $E$  to  $F$ , and  $L^p(\mathbb{R}^d)$  (resp.  $H^s(\mathbb{R}^d)$ ) denotes the usual Lebesgue (resp. Sobolev) space on  $\mathbb{R}^d$ , while  $L^p_{\text{per}}$  (resp.  $H^s_{\text{per}}$ ) denote spaces of  $\mathcal{R}$ -periodic complex-valued functions on the torus  $\Gamma$ . In particular, the space  $L^2_{\text{per}}$  is a Hilbert space when endowed with its natural inner product

$$\forall f, g \in L^2_{\text{per}}, \quad \langle f, g \rangle := \int_{\Gamma} \overline{f}(\mathbf{r})g(\mathbf{r})d\mathbf{r}.$$

We use the notation  $\mathcal{B} := \mathcal{B}(L^2_{\text{per}})$  to denote the space of bounded operators on  $L^2_{\text{per}}$ , and  $\mathfrak{S}_p := \mathfrak{S}_p(L^2_{\text{per}})$  to denote the Schatten class of compact operators on  $L^2_{\text{per}}$  with finite norms  $\|A\|_{\mathfrak{S}_p} := \left(\text{Tr}_{L^2_{\text{per}}} |A|^p\right)^{1/p}$ . In particular,  $\mathfrak{S}_1$  is the space of trace-class operators on  $L^2_{\text{per}}$ , while  $\mathfrak{S}_2$  is the space of Hilbert–Schmidt operators on  $L^2_{\text{per}}$ . We finally introduce, for  $s \geq 0$ , the space

$$\mathfrak{S}_{1,s} := \left\{ \gamma \in \mathfrak{S}_1 \mid \gamma^* = \gamma, (1 - \Delta)^{s/2} \gamma (1 - \Delta)^{s/2} \in \mathfrak{S}_1 \right\},$$

which we endow with the norm

$$\|\gamma\|_{\mathfrak{S}_{1,s}} := \|(1 - \Delta)^{s/2} \gamma (1 - \Delta)^{s/2}\|_{\mathfrak{S}_1}.$$

Let  $\mathcal{O}$  be an operator acting on  $L^2(\mathbb{R}^d)$  which commutes with  $\mathcal{R}$ -translations. Thanks to the Bloch–Floquet transform  $\mathcal{Z}$  [25, Chapter XIII], we have the decomposition

$$\mathcal{Z}^* \mathcal{O} \mathcal{Z} = \int_B^{\oplus} \mathcal{O}_{\mathbf{k}} d\mathbf{k},$$

where we denote by  $\int_B := \frac{1}{|\mathcal{B}|} \int_{\mathcal{B}}$ , and each of the operators  $\mathcal{O}_{\mathbf{k}}$ , the Bloch fibers of  $\mathcal{O}$ , acts on  $L^2_{\text{per}}$ . If  $\mathcal{O}$  is locally trace-class, its *trace per unit cell* is then given by

$$\underline{\text{Tr}}(\mathcal{O}) := \int_B \text{Tr}_{L^2_{\text{per}}}(\mathcal{O}_{\mathbf{k}}) d\mathbf{k}.$$

We consider the one-body electronic Hamiltonian

$$H := -\frac{1}{2} \Delta + V \quad \text{acting on } L^2(\mathbb{R}^d),$$

where  $V \in L^\infty_{\text{per}}(\mathbb{R}^d)$  is a real-valued  $\mathcal{R}$ -periodic potential. This hypothesis could be relaxed and a more general class of potentials could be considered; this choice simplifies some technical proofs in the “Appendix” by ensuring that  $V$  is bounded as an operator on  $L^2_{\text{per}}$ . It is well-known that  $H$  is a bounded from below self-adjoint operator on  $L^2(\mathbb{R}^d)$  with domain  $H^2(\mathbb{R}^d)$ , whose spectrum is purely absolutely continuous (see e.g. [25, Theorem XIII.100]). Since the operator  $H$  commutes with  $\mathcal{R}$ -translations, we can consider its Bloch–Floquet transform. For  $\mathbf{k} \in \mathcal{B}$ , the fiber  $H_{\mathbf{k}}$  is given by

$$H_{\mathbf{k}} = \frac{1}{2} (-i\nabla + \mathbf{k})^2 + V = -\frac{1}{2} \Delta - i\mathbf{k} \cdot \nabla + \frac{\mathbf{k}^2}{2} + V, \quad (2.1)$$

which is a self-adjoint operator on  $L^2_{\text{per}}$  with domain  $H^2_{\text{per}}$ . It is bounded from below, and with compact resolvent. We denote by  $\varepsilon_{1\mathbf{k}} \leq \varepsilon_{2\mathbf{k}} \leq \dots$  its eigenvalues ranked in increasing order, counting multiplicities, and by  $(u_{n\mathbf{k}})_{n \in \mathbb{N}^*} \in \left(H^2_{\text{per}}\right)^{\mathbb{N}^*}$  a corresponding  $L^2_{\text{per}}$ -orthonormal basis of eigenvectors, so that

$$H_{\mathbf{k}} u_{n\mathbf{k}} = \varepsilon_{n\mathbf{k}} u_{n\mathbf{k}}, \quad \langle u_{n\mathbf{k}}, u_{m\mathbf{k}} \rangle = \delta_{nm}.$$

Seeing  $H_{\mathbf{k}}$  as a bounded perturbation of the operator  $\frac{1}{2} (-i\nabla + \mathbf{k})^2$ , standard min–max arguments show that there exist  $\underline{C}_1, \overline{C}_1 \in \mathbb{R}$ ,  $\underline{C}_2, \overline{C}_2 > 0$  such that

$$\underline{C}_1 + \underline{C}_2 n^{2/d} \leq \varepsilon_{n\mathbf{k}} \leq \overline{C}_1 + \overline{C}_2 n^{2/d}. \quad (2.2)$$

Let us now introduce several physical observables. A fundamental quantity in our study is the *one-body density matrix* at level  $\varepsilon \in \mathbb{R}$ , which is the bounded non-negative self-adjoint operator acting on  $L^2(\mathbb{R}^d)$  and defined by

$$\gamma(\varepsilon) := \mathbb{1}(H \leq \varepsilon).$$

Its Bloch–Floquet decomposition is simply

$$\gamma_{\mathbf{k}}(\varepsilon) := \mathbb{1}(H_{\mathbf{k}} \leq \varepsilon) = \sum_{n \in \mathbb{N}^*} \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon) |u_{n\mathbf{k}}\rangle \langle u_{n\mathbf{k}}|.$$

The *integrated density of states* is the function  $\mathcal{N}$  from  $\mathbb{R}$  to  $\mathbb{R}_+$  defined by

$$\forall \varepsilon \in \mathbb{R}, \quad \mathcal{N}(\varepsilon) := \underline{\text{Tr}}(\gamma(\varepsilon)) = \sum_{n \in \mathbb{N}^*} \int_{\mathcal{B}} \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon) d\mathbf{k}. \quad (2.3)$$

The function  $\mathcal{N}$  is a non-decreasing continuous function, with  $\mathcal{N}(-\infty) = 0$  and  $\mathcal{N}(+\infty) = +\infty$ . In particular, if  $N$  denotes the number of electron pairs per unit cell, then  $\mathcal{N}^{-1}(\{N\})$  is a non-empty interval, of the form  $[\varepsilon_-, \varepsilon_+]$ . If  $\varepsilon_- < \varepsilon_+$ , the system is an *insulator*. In this case, supercell methods are very efficient to compute numerically the properties of the crystals (see for instance [11,20]). In this article, we focus on the *metallic* case  $\varepsilon_- = \varepsilon_+$ . In this case, the *Fermi level* of the system is the unique number  $\varepsilon_F := \varepsilon_- = \varepsilon_+$ .

We then introduce the *integrated density of energy*  $E : \mathbb{R} \rightarrow \mathbb{R}$  defined by

$$E(\varepsilon) := \underline{\text{Tr}}(H\gamma(\varepsilon)) = \sum_{n \in \mathbb{N}^*} \int_{\mathcal{B}} \varepsilon_{n\mathbf{k}} \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon) d\mathbf{k}, \quad (2.4)$$

and the (zero-temperature) *ground state energy* (per unit cell)  $E := E(\varepsilon_F)$ . Finally, the *electronic density* up to level  $\varepsilon$  is defined as the density of the locally trace-class  $\mathcal{R}$ -periodic operator  $\gamma(\varepsilon)$ , that is the real-valued function  $\rho_\varepsilon \in L^1_{\text{per}}$  characterized by

$$\forall v \in L^\infty_{\text{per}}, \quad \int_{\Gamma} \rho_\varepsilon(\mathbf{r}) v(\mathbf{r}) d\mathbf{r} = \underline{\text{Tr}}(v\gamma(\varepsilon)) = \sum_{n \in \mathbb{N}^*} \int_{\mathcal{B}} \langle u_{n\mathbf{k}} | v | u_{n\mathbf{k}} \rangle \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon) d\mathbf{k}.$$

We therefore have

$$\forall \mathbf{r} \in \Gamma, \quad \rho_\varepsilon(\mathbf{r}) := \sum_{n \in \mathbb{N}^*} \int_{\mathcal{B}} |u_{n\mathbf{k}}|^2(\mathbf{r}) \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon) d\mathbf{k}. \quad (2.5)$$

The (zero-temperature) *ground state electronic density* then is  $\rho := \rho_{\varepsilon_F}$ .

**Remark 2.1** (Observables) In this article, we focus on the numerical calculation of the integrated density of states  $\mathcal{N}$ , the Fermi level  $\varepsilon_F$ , the ground state energy per unit cell  $E$  and the ground state electronic density  $\rho$  of the system. It is possible to extend our

results to a broader class of observables, but precisising the complete set of assumptions needed to formulate our results is cumbersome and we will not proceed further in this direction.

**Remark 2.2** (Discretization errors) The goal of this paper is to study various numerical schemes to compute Brillouin zone integrals of the form above. In particular, we assume that the eigenvalues  $\varepsilon_{n\mathbf{k}}$  and eigenvectors  $u_{n\mathbf{k}}$  are perfectly known on some mesh of the Brillouin zone  $\mathcal{B}$ , and we study the numerical errors coming from the discretization of the Brillouin zone in (2.3), (2.4) and (2.5). We do not study the effects of numerical errors in the computation of the  $\varepsilon_{n\mathbf{k}}$  and  $u_{n\mathbf{k}}$  themselves. We also do not study more complicated nonlinear models such as the periodic Kohn–Sham model.

It is however interesting to note that the discretization of the eigenvalue problem  $H_{\mathbf{k}}u_{n\mathbf{k}} = \varepsilon_{n\mathbf{k}}u_{n\mathbf{k}}$  in a  $\mathbf{k}$ -consistent manner is not trivial: since  $H_{\mathbf{k}}$  is not equal but unitarily equivalent to  $H_{\mathbf{k}+\mathbf{K}}$  for  $\mathbf{K} \in \mathcal{R}^*$ , a fixed Galerkin space will yield eigenvalues  $\varepsilon_{n\mathbf{k}}$  that are not  $\mathcal{R}^*$ -periodic. Conversely, popular choices such as a  $\mathbf{k}$ -dependent Galerkin space  $V_{\mathbf{k}}$  consisting of all the plane waves  $e^{i\mathbf{K}\cdot\mathbf{r}}$  such that  $\frac{1}{2}|\mathbf{k} + \mathbf{K}|^2 \leq E_{\text{cut}}$  will yield eigenvalues that are periodic, but not continuous as a function of  $\mathbf{k}$ . It is possible to restore continuity by using a smooth cutoff; we plan to explore this possibility from a numerical analysis viewpoint in a forthcoming paper.

### 3 Properties of the band structure at the Fermi level

We first partition the Brillouin zone  $\mathcal{B}$  into several sets, whose definitions are gathered here for the sake of clarity. For a given energy level  $\varepsilon \in \mathbb{R}$ , we introduce, for  $n \in \mathbb{N}$  (and with the convention that  $\varepsilon_{0\mathbf{k}} = -\infty$ ), the sets

$$\begin{aligned} \mathcal{B}_n(\varepsilon) &:= \{\mathbf{k} \in \mathcal{B}, \varepsilon_{n\mathbf{k}} < \varepsilon < \varepsilon_{n+1,\mathbf{k}}\} \quad n\text{-th component of the Brillouin zone,} \\ \mathcal{S}_n(\varepsilon) &:= \{\mathbf{k} \in \mathcal{B}, \varepsilon_{n\mathbf{k}} = \varepsilon\} \quad n\text{-th sheet of the level set,} \\ \mathcal{S}(\varepsilon) &:= \bigcup_{n \in \mathbb{N}} \mathcal{S}_n(\varepsilon) = \{\mathbf{k} \in \mathcal{B}, \exists n \in \mathbb{N}, \varepsilon_{n\mathbf{k}} = \varepsilon\} \quad \text{level set.} \end{aligned}$$

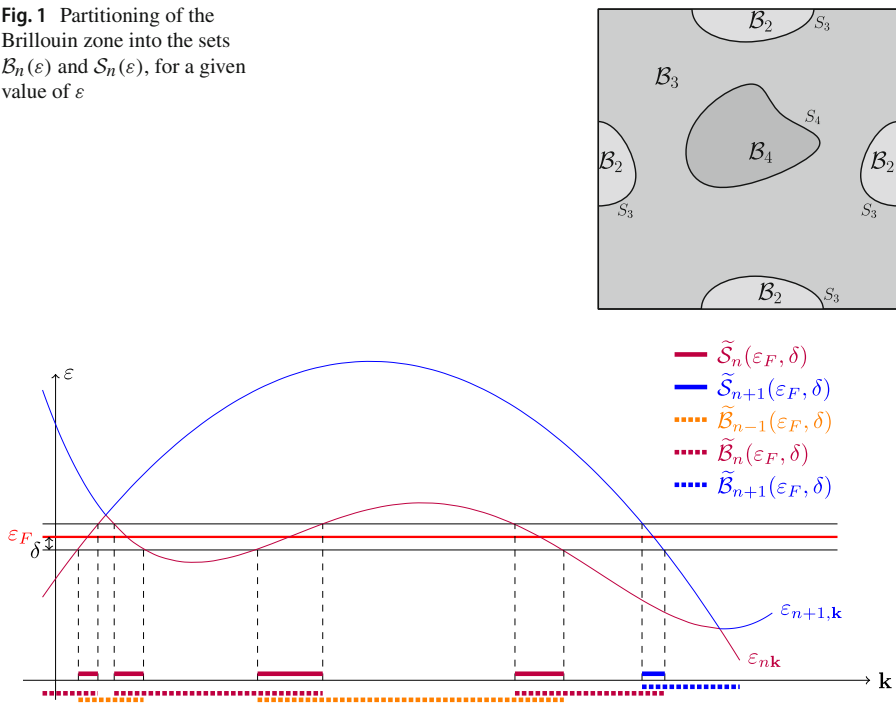
The set  $\mathcal{S}(\varepsilon_F)$  is called the Fermi surface. For all  $\varepsilon \in \mathbb{R}$ , it holds that

$$\mathcal{B} = \mathcal{S}(\varepsilon) \cup \left( \bigcup_{n \in \mathbb{N}} \mathcal{B}_n(\varepsilon) \right). \quad (3.1)$$

From (2.2), the unions in the above definition of  $\mathcal{S}(\varepsilon)$  and in (3.1) are finite. The sets  $\mathcal{S}_1(\varepsilon)$ ,  $\mathcal{S}_2(\varepsilon)$ ,  $\dots$  are pairwise disjoint outside of *band crossings* where  $\varepsilon_{n\mathbf{k}} = \varepsilon_{n+1,\mathbf{k}} = \varepsilon$  for some  $n$ . The boundary of  $\mathcal{B}_n(\varepsilon)$  is  $\partial\mathcal{B}_n(\varepsilon) = \mathcal{S}_n(\varepsilon) \cup \mathcal{S}_{n+1}(\varepsilon)$ . A typical example of the sets  $\mathcal{B}_n(\varepsilon)$  and  $\mathcal{S}_n(\varepsilon)$  is represented in Fig. 1.

As we shall see in Lemma 3.2, various spectral quantities are smooth in  $\mathcal{B}_n(\varepsilon)$  and on  $\mathcal{S}_n(\varepsilon)$ . It will be useful in our analysis to extend this smoothness to the following

**Fig. 1** Partitioning of the Brillouin zone into the sets  $\mathcal{B}_n(\varepsilon)$  and  $\mathcal{S}_n(\varepsilon)$ , for a given value of  $\varepsilon$



**Fig. 2** A schematic view of the sets  $\tilde{\mathcal{B}}_n$  and  $\tilde{\mathcal{S}}_n$

neighborhoods of these sets (see Fig. 2): for  $\delta > 0$ , we set

$$\tilde{\mathcal{B}}_n(\varepsilon, \delta) := \bigcup_{\varepsilon' \in (\varepsilon - \delta, \varepsilon + \delta)} \mathcal{B}_n(\varepsilon') = \{\mathbf{k} \in \mathcal{B}, \exists \varepsilon' \in (\varepsilon - \delta, \varepsilon + \delta), \varepsilon_{n\mathbf{k}} < \varepsilon' < \varepsilon_{n+1,\mathbf{k}}\},$$

$$\tilde{\mathcal{S}}_n(\varepsilon, \delta) := \bigcup_{\varepsilon' \in (\varepsilon - \delta, \varepsilon + \delta)} \mathcal{S}_n(\varepsilon') = \{\mathbf{k} \in \mathcal{B}, \varepsilon_{n\mathbf{k}} \in (\varepsilon - \delta, \varepsilon + \delta)\}.$$

Here and thereafter, smooth means infinitely differentiable.

### 3.1 Assumptions on the Fermi level

Recall that we are studying metallic systems, so that the Fermi level  $\varepsilon_F$  is uniquely defined. In particular, the Fermi surface  $\mathcal{S}(\varepsilon_F)$  is non empty. We make the following two assumptions to ensure a good mathematical structure of the Fermi surface:

**Assumption 1 (no band crossings at  $\varepsilon_F$ ):**  $\forall n \neq m, \mathcal{S}_n(\varepsilon_F) \cap \mathcal{S}_m(\varepsilon_F) = \emptyset$ ;

**Assumption 2 (no flat bands at  $\varepsilon_F$ ):**  $\forall n \in \mathbb{N}^*, \forall \mathbf{k} \in \mathcal{S}_n(\varepsilon_F), \nabla_{\mathbf{k}} \varepsilon_{n\mathbf{k}} \neq \mathbf{0}$ .

From Assumption 2, we see that, for all  $n \in \mathbb{N}^*$ , the map  $\varepsilon_n : \mathbf{k} \mapsto \varepsilon_{n\mathbf{k}}$  is a submersion near the Fermi surface, so that  $\mathcal{S}_n(\varepsilon_F) = \varepsilon_n^{-1}(\varepsilon_F)$  is either empty or a smooth compact co-dimension 1 submanifold of the torus  $\mathcal{B}$ . From Assumption 1,  $\mathcal{S}(\varepsilon_F)$  is itself a smooth compact manifold, as the finite disjoint union of the  $\mathcal{S}_n(\varepsilon_F)$ .

**Remark 3.1** (Genericity of hypotheses) It is an interesting question to know whether such assumptions hold generically, i.e. for almost every potential  $V$ . For a generic smooth family  $H_{\mathbf{k}}$  of self-adjoint operators on  $L^2_{\text{per}}$  with compact resolvents, eigenvalue crossings happen on a set of codimension 3 [29], and flat bands on isolated points. Such singularities, called van Hove singularities, thus do not appear in general at the Fermi level in the physical cases  $d \leq 3$ , and we would naturally expect both these assumptions to be generically true. There are however two important caveats: first, many natural conjectures on the genericity of properties of the band structure still remain open in general (see [14] for an overview), and second, symmetries may force van Hove singularities. For instance, Assumption 1 is violated in the case of the free electron gas, or in graphene [10], due to the high symmetries of these systems. We will treat the case of the graphene in future work. In the sequel, the quality of Assumption 1 and Assumption 2 are measured by quantities  $\delta_0 > 0$  and  $C_{\nabla} > 0$  respectively (see Lemma 3.2 below). For instance, for systems with Fermi level close to van-Hove singularity,  $C_{\nabla}$  will be small.

Let us define by

$$\underline{M} := \min\{n \in \mathbb{N}^*, \mathcal{S}_n(\varepsilon_F) \neq \emptyset\} \quad \text{and} \quad \overline{M} := \max\{n \in \mathbb{N}^*, \mathcal{S}_n(\varepsilon_F) \neq \emptyset\}.$$

The existence of  $\underline{M}$  and  $\overline{M}$  comes from (2.2), and it naturally holds that  $\underline{M} \leq \overline{M}$ .

In the next lemma, we collect a number of properties of the Fermi surface and of spectral quantities on the sets  $\tilde{\mathcal{B}}_n(\varepsilon_F, \delta)$  and  $\tilde{\mathcal{S}}_n(\varepsilon_F, \delta)$ . In order to state these results, we introduce the density matrices

$$\gamma_{n\mathbf{k}} := \sum_{m=1}^n |u_{m\mathbf{k}}\rangle \langle u_{m\mathbf{k}}| \quad \text{acting on} \quad L^2_{\text{per}}, \quad (3.2)$$

which are well-defined operators whenever  $\varepsilon_{n\mathbf{k}} < \varepsilon_{n+1,\mathbf{k}}$ , and the associated densities

$$\rho_{n\mathbf{k}} = \sum_{m=1}^n |u_{m\mathbf{k}}|^2 \in L^1_{\text{per}}. \quad (3.3)$$

We recall that a smooth map  $F : \mathbb{R}^d \rightarrow E$  where  $E$  is a Banach space is real-analytic if it is locally equal to its Taylor series.

**Lemma 3.2** *Under Assumptions 1 and 2, there exists  $\delta_0 > 0$  and  $C_{\nabla} > 0$  such that*

- (i) *For any  $n \in \mathbb{N}^*$  and for all  $0 < \delta \leq \delta_0$ ,  $\tilde{\mathcal{S}}_n(\varepsilon_F, \delta) \neq \emptyset$  if and only if  $\underline{M} \leq n \leq \overline{M}$ ;*
- (ii) *for all  $\underline{M} \leq m < n \leq \overline{M}$ ,  $\tilde{\mathcal{S}}_m(\varepsilon_F, \delta_0) \cap \tilde{\mathcal{S}}_n(\varepsilon_F, \delta_0) = \emptyset$ ;*
- (iii) *for all  $\underline{M} \leq n \leq \overline{M}$  and all  $\mathbf{k} \in \tilde{\mathcal{S}}_n(\varepsilon_F, \delta_0)$ ,  $|\nabla_{\mathbf{k}} \varepsilon_{n\mathbf{k}}| \geq C_{\nabla}$ ;*
- (iv) *for all  $\underline{M} \leq n \leq \overline{M}$  and all  $\varepsilon \in (\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)$ ,  $\mathcal{S}_n(\varepsilon)$  is a non-empty smooth compact manifold of co-dimension 1, with non-zero Hausdorff measure  $|\mathcal{S}_n(\varepsilon)|_{\text{Haus}} > 0$ . The same properties hold for  $\mathcal{S}(\varepsilon)$ ;*
- (v) *assume in addition that  $V \in H^s_{\text{per}}$  for some  $s \geq 0$ . Then, for all  $\underline{M} \leq n \leq \overline{M}$ ,*

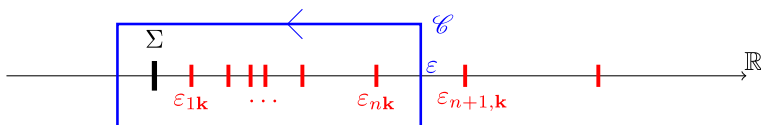


Fig. 3 The contour  $\mathcal{C}$

- the map  $\mathbf{k} \mapsto \gamma_{n\mathbf{k}}$  is real-analytic from  $\tilde{\mathcal{B}}_n(\varepsilon_F, \delta_0)$  to  $\mathfrak{S}_{1,s+2}$ ;
- the map  $\mathbf{k} \mapsto \rho_{n\mathbf{k}}$  is real-analytic from  $\tilde{\mathcal{B}}_n(\varepsilon_F, \delta_0)$  to  $H_{\text{per}}^{s+2}$ ;
- the map  $\mathbf{k} \mapsto \varepsilon_{n\mathbf{k}}$  is real-analytic from  $\tilde{\mathcal{S}}_n(\varepsilon_F, \delta_0)$  to  $\mathbb{R}$ .

**Proof** Assertions (i) and (ii) come from Assumption 1 and the continuity of the maps  $\mathbf{k} \mapsto \varepsilon_{n\mathbf{k}}$ . We now prove (v). Using (v) and Assumption (ii) we will deduce (iii), which implies (iv).

Let  $\underline{M} \leq n \leq \overline{M}$  and  $\delta_0 > 0$  small enough so that (i) and (ii) hold true.

The map  $\mathbf{k} \mapsto \varepsilon_{n\mathbf{k}}$  is continuous on  $\tilde{\mathcal{B}}_n(\varepsilon_F, \delta_0)$  and for all  $\mathbf{k} \in \tilde{\mathcal{B}}_n(\varepsilon_F, \delta_0)$ , we have  $\varepsilon_{n+1,\mathbf{k}} - \varepsilon_{n\mathbf{k}} > 0$ . Therefore, there exists  $0 < \delta_1 < \delta_0$  and  $g > 0$  such that

$$\forall \mathbf{k} \in \tilde{\mathcal{B}}_n(\varepsilon_F, \delta_1), \quad \varepsilon_{n+1,\mathbf{k}} - \varepsilon_{n\mathbf{k}} \geq g.$$

We first prove the analyticity of  $\mathbf{k} \mapsto \gamma_{n\mathbf{k}}$  on  $\tilde{\mathcal{B}}_n(\varepsilon_F, \delta_1)$ . Let  $\mathbf{k}_0 \in \tilde{\mathcal{B}}_n(\varepsilon_F, \delta_1)$ . We then set  $\varepsilon = \varepsilon_{n,\mathbf{k}_0} + g/2$  and  $\Sigma := \min \sigma(H)$ , and consider the positively oriented loop (see Fig. 3)

$$\mathcal{C} := [\Sigma - 1 - i, \varepsilon - i] \cup [\varepsilon - i, \varepsilon + i] \cup [\varepsilon + i, \Sigma - 1 + i] \cup [\Sigma - 1 + i, \Sigma - 1 - i].$$

From the definition of  $\mathcal{C}$ , we see that there exists  $0 < \delta_2 < \delta_1$  such that

$$\forall \mathbf{k} \in \mathcal{B} \text{ s.t. } |\mathbf{k} - \mathbf{k}_0| \leq \delta_2, \quad \forall \lambda \in \mathcal{C}, \quad |\lambda - H_{\mathbf{k}}| \geq g/4.$$

In particular, we see that Cauchy's residual formula

$$\gamma_{n\mathbf{k}} = \frac{1}{2\pi i} \oint_{\mathcal{C}} \frac{d\lambda}{\lambda - H_{\mathbf{k}}}, \quad (3.4)$$

holds for  $\mathbf{k}$  in a neighborhood of  $\mathbf{k}_0$ . In addition,

$$H_{\mathbf{k}} = H_{\mathbf{k}_0} + (\mathbf{k} - \mathbf{k}_0) \cdot (-i\nabla + \mathbf{k}_0) + \frac{|\mathbf{k} - \mathbf{k}_0|^2}{2},$$

so that, for  $\mathbf{k} - \mathbf{k}_0$  small enough,

$$(\lambda - H_{\mathbf{k}})^{-1} = (\lambda - H_{\mathbf{k}_0})^{-1} \left( 1 - \left[ \left( (\mathbf{k} - \mathbf{k}_0) \cdot (-i\nabla + \mathbf{k}_0) + \frac{|\mathbf{k} - \mathbf{k}_0|^2}{2} \right) (\lambda - H_{\mathbf{k}_0})^{-1} \right] \right)^{-1}.$$

For all  $0 \leq s' \leq s$ , the linear operator  $(\lambda - H_{\mathbf{k}_0})^{-1}$  is continuous from  $H_{\text{per}}^{s'}$  to  $H_{\text{per}}^{s'+2}$  by classical elliptic regularity results, and the operator in brackets is bounded on  $H^{s'}$ .

Therefore we obtain, by expanding in Neumann series, that the map  $\mathbf{k} \mapsto \gamma_{n\mathbf{k}}$  is real-analytic from a neighborhood of  $\mathbf{k}_0$  to  $\mathcal{B}(H_{\text{per}}^{s'}, H_{\text{per}}^{s'+2})$ . Using the fact that  $\gamma_{n\mathbf{k}} = \gamma_{n\mathbf{k}}^2$  and a bootstrap argument, this implies that the map  $\mathbf{k} \mapsto \gamma_{n\mathbf{k}}$  is real-analytic from a neighborhood of  $\mathbf{k}_0$  to  $\mathcal{B}(L_{\text{per}}^2, H_{\text{per}}^{s+2})$ .

Let  $(v_{1\mathbf{k}_0}, \dots, v_{n\mathbf{k}_0})$  be an  $L_{\text{per}}^2$ -orthonormal basis of  $\text{Ran}(\gamma_{n\mathbf{k}_0})$ ,  $\tilde{v}_{j\mathbf{k}} = \gamma_{n\mathbf{k}} v_{j\mathbf{k}_0}$  and

$$v_{i\mathbf{k}} = \sum_{j=1}^n \tilde{v}_{j\mathbf{k}} [S_{\mathbf{k}}^{-1/2}]_{ji},$$

where  $S_{\mathbf{k}}$  is the overlap matrix defined by  $[S_{\mathbf{k}}]_{ji} = \langle \tilde{v}_{j\mathbf{k}}, \tilde{v}_{i\mathbf{k}} \rangle$ , so that  $(v_{1\mathbf{k}}, \dots, v_{n\mathbf{k}})$  is an  $L_{\text{per}}^2$ -orthonormal basis of  $\text{Ran } \gamma_{n\mathbf{k}}$ . Of course, it holds that  $\tilde{v}_{j\mathbf{k}_0} = v_{j\mathbf{k}_0}$  and  $S_{\mathbf{k}_0} = \text{Id}_n$ . It is easily checked that the map  $\mathbf{k} \mapsto (v_{1\mathbf{k}}, \dots, v_{n\mathbf{k}}) \in (H_{\text{per}}^{s+2})^n$  is well-defined and real-analytic in a neighborhood of  $\mathbf{k}_0$ . In particular, we have, in a neighborhood of  $\mathbf{k}_0$ , that

$$(1 - \Delta)^{(s+2)/2} \gamma_{n\mathbf{k}} (1 - \Delta)^{(s+2)/2} = \sum_{j=1}^n |w_{j\mathbf{k}}\rangle \langle w_{j\mathbf{k}}|,$$

where the maps  $\mathbf{k} \mapsto w_{j\mathbf{k}} := (1 - \Delta)^{(s+2)/2} v_{j\mathbf{k}} \in L_{\text{per}}^2$  are real-analytic in a neighborhood of  $\mathbf{k}_0$ . It follows that the map  $\mathbf{k} \mapsto \gamma_{n\mathbf{k}}$  is analytic from a neighborhood of  $\mathbf{k}_0$  to  $\mathfrak{S}_{1,s+2}$ , hence from  $\tilde{\mathcal{B}}_n(\varepsilon_F, \delta_1)$  to  $\mathfrak{S}_{1,s+2}$ .

On the other hand, whenever  $s + 2 > \frac{d}{2}$  (which is the case whenever  $s \geq 0$  and  $d \leq 3$ ), it holds that  $H_{\text{per}}^{s+2}$  is an algebra. As a result, from the analyticity of the map  $\mathbf{k} \mapsto (v_{1\mathbf{k}}, \dots, v_{n\mathbf{k}}) \in (H_{\text{per}}^{s+2})^n$ , we deduce the analyticity of the maps  $\mathbf{k} \mapsto \rho_{n\mathbf{k}} := \sum_{i=1}^n |u_{i\mathbf{k}}|^2 = \sum_{i=1}^n |v_{i\mathbf{k}}|^2 \in H_{\text{per}}^{s+2}$ .

To prove the analyticity of  $\mathbf{k} \mapsto \varepsilon_{n\mathbf{k}}$  on  $\tilde{\mathcal{S}}_n(\varepsilon_F, \delta_1)$ , we notice that  $\tilde{\mathcal{S}}_n(\varepsilon_F, \delta_1) = \tilde{\mathcal{B}}_{n-1}(\varepsilon_F, \delta_1) \cap \tilde{\mathcal{B}}_n(\varepsilon_F, \delta_1)$ . As a result, both  $\mathbf{k} \mapsto \gamma_{n-1,\mathbf{k}}$  and  $\mathbf{k} \mapsto \gamma_{n\mathbf{k}}$  are analytic from  $\tilde{\mathcal{S}}_n(\varepsilon_F, \delta_1)$  to  $\mathfrak{S}_{1,s+2}$ . Therefore, on  $\tilde{\mathcal{S}}_n(\varepsilon_F, \delta_1)$ , it holds that

$$\begin{aligned} \varepsilon_{n\mathbf{k}} &= \text{Tr}_{L_{\text{per}}^2} [H_{\mathbf{k}}(\gamma_{n\mathbf{k}} - \gamma_{n-1,\mathbf{k}})] \\ &= \text{Tr}_{L_{\text{per}}^2} \left[ (1 - \Delta)^{-1/2} H_{\mathbf{k}} (1 - \Delta)^{-1/2} (1 - \Delta)^{1/2} (\gamma_{n\mathbf{k}} - \gamma_{n-1,\mathbf{k}}) (1 - \Delta)^{1/2} \right]. \end{aligned}$$

Since the maps  $\tilde{\mathcal{S}}_n(\varepsilon_F, \delta_1) \ni \mathbf{k} \mapsto (1 - \Delta)^{-1/2} H_{\mathbf{k}} (1 - \Delta)^{-1/2} \in \mathcal{B}$  and  $\tilde{\mathcal{S}}_n(\varepsilon_F, \delta_1) \ni \mathbf{k} \mapsto (1 - \Delta)^{1/2} (\gamma_{n\mathbf{k}} - \gamma_{n-1,\mathbf{k}}) (1 - \Delta)^{1/2} \in \mathfrak{S}_1$  are real-analytic, this proves the real-analyticity of the map  $\mathbf{k} \mapsto \varepsilon_{n\mathbf{k}}$  on  $\tilde{\mathcal{S}}_n(\varepsilon_F, \delta_1)$ .  $\square$

### 3.2 Density of states

We are now concerned with the properties of the density of states  $\mathcal{D}(\varepsilon)$ , defined as the derivative of the integrated density of states  $\mathcal{N}(\varepsilon)$  defined in (2.3). Since  $\mathcal{N}$  is a non-decreasing continuous function which is increasing on  $\sigma(H)$  and constant outside  $\sigma(H)$ ,  $\mathcal{D}$  is a positive measure on  $\mathbb{R}$  whose support is exactly  $\sigma(H)$ . Our goal in this

section is to establish that, under Assumptions 1 and 2, both  $\mathcal{N}$  and  $\mathcal{D}$  are smooth around  $\varepsilon_F$ .

We recall some tools of differential geometry. In the sequel, if  $S$  is a (smooth) hypersurface of  $\mathcal{B}$ , we denote by  $d\sigma_S$  the Hausdorff measure on  $S$ . We write  $d\sigma$  instead of  $d\sigma_S$  if there is no risk of confusion. We first recall the co-area formula, which allows the integration of a function  $g : \mathcal{B} \rightarrow \mathbb{R}$  along the level sets of another function  $\mathcal{E} : \mathcal{B} \rightarrow \mathbb{R}$ .

**Lemma 3.3** (Co-area formula [8]) *Let  $\mathcal{E} : \mathcal{B} \rightarrow \mathbb{R}$  be a Lipschitz function and  $g \in L^1(\mathcal{B})$ , then*

$$\int_{\mathcal{B}} g(\mathbf{k}) |\nabla \mathcal{E}(\mathbf{k})| d\mathbf{k} = \int_{\mathbb{R}} \left( \int_{\mathcal{E}^{-1}\{\varepsilon\}} g(\mathbf{k}) d\sigma(\mathbf{k}) \right) d\varepsilon.$$

Setting  $f = g|\nabla \mathcal{E}|$ , we deduce that for all  $f : \mathcal{B} \rightarrow \mathbb{R}$  such that  $\mathbf{k} \mapsto \frac{f(\mathbf{k})}{|\nabla \mathcal{E}(\mathbf{k})|} \in L^1(\mathcal{B})$ , then

$$\int_{\mathcal{B}} f(\mathbf{k}) d\mathbf{k} = \int_{\mathbb{R}} \left( \int_{\mathcal{E}^{-1}\{\varepsilon\}} \frac{f(\mathbf{k})}{|\nabla \mathcal{E}(\mathbf{k})|} d\sigma(\mathbf{k}) \right) d\varepsilon. \quad (3.5)$$

This allows us to differentiate functions defined as integrals on the sets  $\mathcal{S}_n(\varepsilon)$  and  $\mathcal{B}_n(\varepsilon)$  with respect to the energy level  $\varepsilon$ .

**Lemma 3.4** *Let  $f \in \mathcal{C}^p(\mathcal{B}, \mathbb{R})$ . Under Assumptions 1 and 2 and with the notation of Lemma 3.2, the maps  $F_n : (\varepsilon_F - \delta_0, \varepsilon_F + \delta_0) \rightarrow \mathbb{R}$ ,  $\underline{M} \leq n \leq \overline{M}$ , defined by*

$$\forall \varepsilon \in (\varepsilon_F - \delta_0, \varepsilon_F + \delta_0), \quad F_n(\varepsilon) := \int_{\mathcal{S}_n(\varepsilon)} f(\mathbf{k}) d\sigma(\mathbf{k}),$$

*are of class  $\mathcal{C}^p$ , and we have*

$$\forall \varepsilon \in (\varepsilon_F - \delta_0, \varepsilon_F + \delta_0), \quad F'_n(\varepsilon) = \int_{\mathcal{S}_n(\varepsilon)} \frac{\operatorname{div} \left( f(\mathbf{k}) \frac{\nabla''_{n\mathbf{k}}}{|\nabla''_{n\mathbf{k}}|} \right)}{|\nabla \varepsilon_{n\mathbf{k}}|} d\sigma(\mathbf{k}). \quad (3.6)$$

**Proof** Using suitable cut-off functions, it is sufficient to prove the result for  $f$  compactly supported in  $\tilde{\mathcal{S}}_n(\varepsilon_F, \delta_0)$ . The outgoing normal unit vector of  $\mathcal{S}_n(\varepsilon)$  at  $\mathbf{k}$  (oriented so that its interior is  $\{\mathbf{k} \in \mathcal{B}, \varepsilon_{n\mathbf{k}} < \varepsilon\}$ ) is  $\nu_{\mathbf{k}} := \nabla \varepsilon_{n\mathbf{k}} / |\nabla \varepsilon_{n\mathbf{k}}|$ . Using the divergence theorem, we get

$$\begin{aligned} F_n(\varepsilon) &= \int_{\mathcal{S}_n(\varepsilon)} f(\mathbf{k}) d\sigma(\mathbf{k}) = \int_{\mathcal{S}_n(\varepsilon)} f(\mathbf{k}) \frac{\nabla \varepsilon_{n\mathbf{k}}}{|\nabla \varepsilon_{n\mathbf{k}}|} \cdot \nu_{\mathbf{k}} d\sigma(\mathbf{k}) \\ &= \int_{\{\mathbf{k} \in \mathcal{B}, \varepsilon_{n\mathbf{k}} < \varepsilon\}} \operatorname{div} \left( f(\mathbf{k}) \frac{\nabla''_{n\mathbf{k}}}{|\nabla''_{n\mathbf{k}}|} \right) d\mathbf{k}. \end{aligned}$$

We now use the co-area formula (3.5) with  $\tilde{f}(\mathbf{k}) := \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon) \operatorname{div} \left( f(\mathbf{k}) \frac{\nabla''_{n\mathbf{k}}}{|\nabla''_{n\mathbf{k}}|} \right)$  and  $\mathcal{E}(\mathbf{k}) = \varepsilon_{n\mathbf{k}}$ , so that

$$F_n(\varepsilon) = \int_{-\infty}^{\varepsilon} \left( \int_{S_n(\varepsilon')} \frac{\operatorname{div} \left( f(\mathbf{k}) \frac{\nabla''_{n\mathbf{k}}}{|\nabla''_{n\mathbf{k}}|} \right)}{|\nabla \varepsilon_{n\mathbf{k}}|} d\sigma(\mathbf{k}) \right) d\varepsilon'.$$

Differentiating this expression leads to (3.6), and iterating  $p$  times leads to the result.  $\square$

Formally, using the co-area formula on the integrated density of states  $\mathcal{N}(\varepsilon)$  would yield

$$\mathcal{N}(\varepsilon) = \frac{1}{|\mathcal{B}|} \sum_{n \in \mathbb{N}^*} \int_{-\infty}^{\varepsilon} \left( \int_{S_n(\varepsilon')} \frac{1}{|\nabla \varepsilon_{n\mathbf{k}}|} d\sigma(\mathbf{k}) \right) d\varepsilon',$$

but the integrand may not be well-defined for all  $\varepsilon' < \varepsilon$ . This argument is however valid close to the Fermi surface, and we therefore have the following result.

**Lemma 3.5** *Under Assumptions 1 and 2 and with the notation of Lemma 3.2, the integrated density of states  $\mathcal{N}$  is smooth on  $(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)$ . Moreover, we have*

$$\forall \varepsilon \in (\varepsilon_F - \delta_0, \varepsilon_F + \delta_0), \quad \mathcal{D}(\varepsilon) := \mathcal{N}'(\varepsilon) = \frac{1}{|\mathcal{B}|} \sum_{n=\underline{M}}^{\overline{M}} \int_{S_n(\varepsilon)} \frac{1}{|\nabla \varepsilon_{n\mathbf{k}}|} d\sigma(\mathbf{k}) > 0.$$

**Proof** Applying the co-area formula with  $f(\mathbf{k}) = \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon_F)$ , we have

$$\mathcal{N}(\varepsilon) := \sum_{n=\underline{M}}^{\overline{M}} \int_{\mathcal{B}} \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon) d\mathbf{k} = \mathcal{N}(\varepsilon_F - \delta_0) + \frac{1}{|\mathcal{B}|} \sum_{n=\underline{M}}^{\overline{M}} \int_{\varepsilon_F - \delta_0}^{\varepsilon} d\varepsilon' \int_{S_n(\varepsilon')} \frac{1}{|\nabla \varepsilon_{n\mathbf{k}}|} d\sigma(\mathbf{k}),$$

from which we get

$$\mathcal{D}(\varepsilon) := \mathcal{N}'(\varepsilon) = \frac{1}{|\mathcal{B}|} \sum_{n=\underline{M}}^{\overline{M}} \int_{S_n(\varepsilon)} \frac{1}{|\nabla \varepsilon_{n\mathbf{k}}|} d\sigma(\mathbf{k}). \quad (3.7)$$

By Lemmas 3.2 and 3.4, this function is smooth and positive.  $\square$

The function  $\mathcal{D}$  appearing in (3.7) is called the *density of states*. This lemma justifies our Assumptions 1 and 2 as natural assumptions to ensure a smooth density of states at the Fermi level. The presence of crossings at the Fermi level may indeed yield singularities of the density of states, as is well-known for instance in graphene. Similarly, a zero of the band gradient (leading to so-called “flat bands”) produces *van Hove singularities* in the density of states.

Following the same steps as in Lemma 3.5, we obtain that the integrated density of energy defined in (2.4) is also smooth on  $(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)$ , and that its derivative (the *density of energy*) satisfies

$$\forall \varepsilon \in (\varepsilon_F - \delta_0, \varepsilon_F + \delta_0), \quad E'(\varepsilon) = \varepsilon \mathcal{D}(\varepsilon). \quad (3.8)$$

We also record here the following technical lemma on the volume of the sets  $\tilde{\mathcal{S}}_n(\varepsilon, \delta)$ , which will be used in our analysis. It is an easy consequence of the co-area formula.

**Lemma 3.6** *Under Assumptions 1 and 2 and with the notation of Lemma 3.2, there exists  $C \in \mathbb{R}_+$  such that, for all  $\varepsilon \in (\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)$  and all  $0 \leq \delta < \delta_0$  such that  $(\varepsilon - \delta, \varepsilon + \delta) \subset (\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)$ , it holds that  $|\tilde{\mathcal{S}}_n(\varepsilon, \delta)| \leq C\delta$  for all  $\underline{M} \leq n \leq \overline{M}$ .*

**Proof** We apply the co-area formula (3.5) with  $f(\mathbf{k}) := \mathbb{1}(\varepsilon - \delta \leq \varepsilon_{n\mathbf{k}} \leq \varepsilon + \delta)$  and  $\mathcal{E}(\mathbf{k}) = \varepsilon_{n\mathbf{k}}$ , and get that

$$|\tilde{\mathcal{S}}_n(\varepsilon, \delta)| = \int_{\varepsilon - \delta}^{\varepsilon + \delta} \left( \int_{\mathcal{S}_n(\varepsilon')} \frac{1}{|\nabla \varepsilon_{n\mathbf{k}}|} d\sigma(\mathbf{k}) \right) d\varepsilon'.$$

From Lemma 3.2, the map  $\varepsilon' \mapsto \int_{\mathcal{S}_n(\varepsilon')} |\nabla \varepsilon_{n\mathbf{k}}|^{-1} d\sigma(\mathbf{k})$  is continuous and bounded on  $(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)$ . The proof follows.  $\square$

## 4 Interpolation methods

In this section, we investigate methods based on the local interpolation of the functions  $\mathbf{k} \mapsto \varepsilon_{n\mathbf{k}}$ . We consider families of linear interpolation operators  $\Pi^L : C^0(\mathcal{B}, \mathbb{R}) \rightarrow C^0(\mathcal{B}, \mathbb{R})$  indexed by  $L \in \mathbb{N}^*$ , which strongly converge to the identity operator when  $L$  goes to infinity, i.e.

$$\forall f \in C^0(\mathcal{B}, \mathbb{R}), \quad \Pi^L f \xrightarrow{L \rightarrow \infty} f \text{ in } C^0(\mathcal{B}, \mathbb{R}).$$

We say that  $(\Pi^L)_{L \in \mathbb{N}^*}$  is of order  $(p+1) \in \mathbb{N}$  if, for all  $\eta > 0$ , there exists  $C_\Pi^\eta \in \mathbb{R}_+$  such that, for all  $p' \in \mathbb{N}$ , all open sets  $\Omega \subset \mathcal{B}$ , and all  $f \in C^{p'+1}(\Omega_\eta, \mathbb{R})$ , where  $\Omega_\eta := \{\mathbf{k} \in \mathcal{B}, d(\mathbf{k}, \Omega) \leq \eta\}$  is the  $\eta$ -neighborhood of  $\Omega$ , it holds that

$$\sup_{\mathbf{k} \in \Omega} |f(\mathbf{k}) - \Pi^L f(\mathbf{k})| \leq \frac{C_\Pi^\eta}{L^{\min(p, p')+1}} \sup_{\mathbf{k} \in \Omega_\eta} |f^{(p'+1)}(\mathbf{k})|. \quad (4.1)$$

One of the most used interpolation operator is the *linear tetrahedron method* (and its *improved* version [2]—see also [13]—that will be studied in a future work). In this case, we choose a sequence of uniform tetrahedral meshes  $(\mathcal{T}^L)_{L \in \mathbb{N}}$ , and we define  $\Pi^L f$  as the piecewise linear function (linear on each tetrahedron  $T \in \mathcal{T}^L$ ) interpolating  $f$  at the vertices of  $\mathcal{T}^L$ . In three dimensions, a linear tetrahedron method constructed

from a regular  $L \times L \times L$  mesh of the torus  $\mathcal{B}$  is of order 2. Similarly, the quadratic method described in [3, 17] and the cubic tetrahedron method described in [30] are of order 3 and 4 respectively. We chose this convention so that  $p$  denotes the degree of the polynomial in a usual polynomial interpolation (of order  $p + 1$ ).

**Remark 4.1** (Local interpolation) The approximation property above is local, in the sense that the quality of the approximation at a point depends only on the smoothness of the function near this point. This is necessary to interpolate efficiently the functions  $\varepsilon_{n\mathbf{k}}$ , which are not smooth across the whole Brillouin zone. By contrast, a Fourier interpolation on the whole Brillouin zone would not satisfy this condition: discontinuities in the interpolated function produce Gibbs oscillations, which slow down the convergence of the Fourier series even far from the point of discontinuity.

#### 4.1 Error on the Integrated density of states and on the Fermi level

Recall that the integrated density of states  $\mathcal{N}(\varepsilon)$  is defined in (2.3). In practice, we cannot compute  $\mathcal{N}(\varepsilon)$ , but only an approximation of it. We therefore introduce

$$\forall \varepsilon \in \mathbb{R}, \quad \mathcal{N}^{L,q}(\varepsilon) := \sum_{n \in \mathbb{N}^*} \int_{\mathcal{B}} \mathbb{1}(\varepsilon_{n\mathbf{k}}^{L,q} \leq \varepsilon) d\mathbf{k} \quad \text{with} \quad \varepsilon_{n\mathbf{k}}^{L,q} := \Pi^{L,q}[\varepsilon_{n\mathbf{k}}], \quad (4.2)$$

where  $(\Pi^{L,q})_{L \in \mathbb{N}^*}$  is a family of interpolation operators of order  $q + 1$ . In practice, the integral in (4.2) is performed analytically (hence at low computational cost). Because of the smoothness of  $\varepsilon_{n\mathbf{k}}$  near  $\mathcal{S}_n(\varepsilon_F)$ , we are able to control the error on this function.

**Lemma 4.2** (Error on the integrated density of states) *Under Assumptions 1 and 2, there exist  $C \in \mathbb{R}_+$  and  $\delta > 0$  such that*

$$\forall L \in \mathbb{N}^*, \quad \max_{\varepsilon \in [\varepsilon_F - \delta, \varepsilon_F + \delta]} \left| \mathcal{N}(\varepsilon) - \mathcal{N}^{L,q}(\varepsilon) \right| \leq \frac{C}{L^{q+1}}.$$

**Proof** Let  $\delta_B^{L,q}$  be the maximum error between  $\varepsilon_{n\mathbf{k}}$  and  $\varepsilon_{n\mathbf{k}}^{L,q}$  on the whole Brillouin zone, i.e.

$$\delta_B^{L,q} := \max_{n \leq M} \max_{\mathbf{k} \in \mathcal{B}} |\varepsilon_{n\mathbf{k}}^{L,q} - \varepsilon_{n\mathbf{k}}|.$$

From the fact that  $\mathbf{k} \mapsto \varepsilon_{n\mathbf{k}}$  is Lipschitz and (4.1) in the case  $p' = 0$ , we deduce that  $\lim_{L \rightarrow \infty} \delta_B^{L,q} = 0$ . For  $L$  large enough,  $\delta_B^{L,q} < \delta_0/2$ , where  $\delta_0$  was defined in Lemma 3.2. Let  $\varepsilon \in [\varepsilon_F - \delta_0/2, \varepsilon_F + \delta_0/2]$ . We have

$$\mathcal{N}(\varepsilon) - \mathcal{N}^{L,q}(\varepsilon) = \sum_{n \in \mathbb{N}^*} \int_{\mathcal{B}} (\mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon) - \mathbb{1}(\varepsilon_{n\mathbf{k}}^{L,q} \leq \varepsilon)) d\mathbf{k}. \quad (4.3)$$

The integrand in (4.3) can only be nonzero in the *discrepancy regions* where  $\varepsilon_{n\mathbf{k}} \leq \varepsilon < \varepsilon_{n\mathbf{k}}^{L,q}$  or  $\varepsilon_{n\mathbf{k}}^{L,q} \leq \varepsilon < \varepsilon_{n\mathbf{k}}$ . In these regions, it holds that  $|\varepsilon_{n\mathbf{k}} - \varepsilon| \leq |\varepsilon_{n\mathbf{k}} - \varepsilon_{n\mathbf{k}}^{L,q}|$ ,

so that they are included in  $\tilde{\mathcal{S}}_n(\varepsilon, \delta_B^{L,q})$ . We can therefore rewrite (4.3) as

$$\mathcal{N}(\varepsilon) - \mathcal{N}^{L,q}(\varepsilon) = \frac{1}{|\mathcal{B}|} \sum_{n \in \mathbb{N}^*} \int_{\tilde{\mathcal{S}}_n(\varepsilon, \delta_B^{L,q})} (\mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon) - \mathbb{1}(\varepsilon_{n\mathbf{k}}^{L,q} \leq \varepsilon)) d\mathbf{k}. \quad (4.4)$$

From Lemma 3.6, we easily deduce that  $|\mathcal{N}(\varepsilon) - \mathcal{N}^{L,q}(\varepsilon)| \leq C\delta_B^{L,q}$ . This is however a very crude approximation, since  $\delta_B^{L,q}$  only decays as  $L^{-1}$  (and not as  $L^{-(q+1)}$  as wanted). This comes from the fact that  $\varepsilon_{n\mathbf{k}}$  is Lipschitz but not  $C^1$  on  $\mathcal{B}$ . However, according to Lemma 3.2,  $\varepsilon_{n\mathbf{k}}$  is analytic on  $\tilde{\mathcal{S}}_n(\varepsilon, \delta_B^{L,q})$ . Hence, by setting

$$\delta_S^{L,q} := \max_{n \leq M} \max_{\mathbf{k} \in \tilde{\mathcal{S}}_n(\varepsilon_F, \delta_0)} |\varepsilon_{n\mathbf{k}}^{L,q} - \varepsilon_{n\mathbf{k}}|,$$

we first deduce from (4.1) that  $\delta_S^{L,q} = O(L^{-(q+1)})$ , then that the integrand in (4.4) is non-zero only on  $\tilde{\mathcal{S}}_n(\varepsilon, \delta_S^{L,q})$ , which again from Lemma 3.6 is of Lebesgue measure  $O(L^{-(q+1)})$ .  $\square$

For all  $L \in \mathbb{N}$ , the function  $\mathcal{N}^{L,q}$  is continuous and non-decreasing from  $\mathcal{N}^{L,q}(-\infty) = 0$  to  $\mathcal{N}^{L,q}(+\infty) = +\infty$ . However, it is not necessarily increasing, and the non-empty set  $(\mathcal{N}^{L,q})^{-1}(\{N\})$  may contain more than one point. Our results however will be independent of the choice of the approximated Fermi level  $\varepsilon_F^{L,q} \in (\mathcal{N}^{L,q})^{-1}(\{N\})$ . We state the next lemma in a very general setting.

**Lemma 4.3** (Error on the Fermi level) *Under Assumptions 1 and 2, there is  $\delta_1 > 0$  such that, for all  $0 < \delta \leq \delta_1$  and all continuous function  $\tilde{\mathcal{N}}_\delta : \mathbb{R} \rightarrow \mathbb{R}$  satisfying*

$$\max_{\varepsilon \in [\varepsilon_F - \delta, \varepsilon_F + \delta]} |\mathcal{N}(\varepsilon) - \tilde{\mathcal{N}}_\delta(\varepsilon)| \leq \frac{\mathcal{D}(\varepsilon_F)}{2} \delta, \quad (4.5)$$

*the equation  $\tilde{\mathcal{N}}_\delta(\varepsilon) = N$  has at least one solution  $\tilde{\varepsilon}_F$  in the range  $[\varepsilon_F - \delta, \varepsilon_F + \delta]$ , and any such solution satisfies*

$$|\varepsilon_F - \tilde{\varepsilon}_F| \leq \frac{2}{\mathcal{D}(\varepsilon_F)} \max_{\varepsilon \in [\varepsilon_F - \delta, \varepsilon_F + \delta]} |\mathcal{N}(\varepsilon) - \tilde{\mathcal{N}}_\delta(\varepsilon)| \quad (\leq \delta). \quad (4.6)$$

*Together with Lemma 4.2, we deduce that there exists  $C \in \mathbb{R}^+$  such that*

$$\forall L \in \mathbb{N}^*, \quad |\varepsilon_F - \varepsilon_F^{L,q}| \leq \frac{C}{L^{q+1}}.$$

**Remark 4.4** (Failure of hypotheses) Lemma 4.3 fails when  $\mathcal{D}(\varepsilon_F) = 0$ , i.e. when the density of states is zero at the Fermi level (i.e. in semimetals such as graphene). In that case, the bound depends on the local behavior of  $\mathcal{D}$  around  $\varepsilon_F$ .

**Proof of Lemma 4.3** Thanks to Lemma 3.5, there exists  $\delta_1 > 0$  such that  $\inf_{\varepsilon \in [\varepsilon_F - \delta_1, \varepsilon_F + \delta_1]} \mathcal{D}(\varepsilon) > \mathcal{D}(\varepsilon_F)/2$ . In particular, for all  $\varepsilon \in (\varepsilon_F - \delta_1, \varepsilon_F + \delta_1)$  and  $0 \leq \delta < \delta_1$  such that  $[\varepsilon - \delta, \varepsilon + \delta] \subset [\varepsilon_F - \delta_1, \varepsilon_F + \delta_1]$ , we have

$$\mathcal{N}(\varepsilon + \delta) = \mathcal{N}(\varepsilon) + \int_{\varepsilon}^{\varepsilon + \delta} \mathcal{D}(\varepsilon) d\varepsilon \geq \mathcal{N}(\varepsilon) + \frac{\mathcal{D}(\varepsilon_F)}{2} \delta, \quad (4.7)$$

and similarly,  $\mathcal{N}(\varepsilon - \delta) \leq \mathcal{N}(\varepsilon) - \frac{\mathcal{D}(\varepsilon_F)}{2} \delta$ . Let now  $\tilde{\mathcal{N}}_\delta : \mathbb{R} \rightarrow \mathbb{R}$  be a continuous function satisfying (4.5). Then, it holds that

$$\tilde{\mathcal{N}}_\delta(\varepsilon_F - \delta) \leq N \leq \tilde{\mathcal{N}}_\delta(\varepsilon_F + \delta).$$

Hence, by continuity of  $\tilde{\mathcal{N}}_\delta$ , the equation  $\tilde{\mathcal{N}}_\delta(\varepsilon) = N$  has at least one solution in  $[\varepsilon_F - \delta, \varepsilon_F + \delta]$ . Let  $\tilde{\varepsilon}_F$  be such a solution. We denote by  $\kappa_{\mathcal{N}, \tilde{\mathcal{N}}_\delta} := \max_{\varepsilon \in [\varepsilon_F - \delta, \varepsilon_F + \delta]} |\mathcal{N}(\varepsilon) - \tilde{\mathcal{N}}_\delta(\varepsilon)|$ . Since  $\kappa_{\mathcal{N}, \tilde{\mathcal{N}}_\delta} \leq \frac{\mathcal{D}(\varepsilon_F)}{2} \delta$ , we can again use (4.7) and get

$$N = \mathcal{N}(\varepsilon_F) = \tilde{\mathcal{N}}_\delta(\tilde{\varepsilon}_F) \leq \mathcal{N}(\tilde{\varepsilon}_F) + \kappa_{\mathcal{N}, \tilde{\mathcal{N}}_\delta} \leq \mathcal{N}\left(\tilde{\varepsilon}_F + \frac{2}{\mathcal{D}(\varepsilon_F)} \kappa_{\mathcal{N}, \tilde{\mathcal{N}}_\delta}\right),$$

and, similarly,  $N = \mathcal{N}(\varepsilon_F) \geq \mathcal{N}\left(\tilde{\varepsilon}_F - \frac{2}{\mathcal{D}(\varepsilon_F)} \kappa_{\mathcal{N}, \tilde{\mathcal{N}}_\delta}\right)$ . From the fact that  $\mathcal{N}$  is non-decreasing, and the inequality

$$\mathcal{N}\left(\tilde{\varepsilon}_F - \frac{2}{\mathcal{D}(\varepsilon_F)} \kappa_{\mathcal{N}, \tilde{\mathcal{N}}_\delta}\right) \leq \mathcal{N}(\varepsilon_F) \leq \mathcal{N}\left(\tilde{\varepsilon}_F + \frac{2}{\mathcal{D}(\varepsilon_F)} \kappa_{\mathcal{N}, \tilde{\mathcal{N}}_\delta}\right),$$

we obtain (4.6).  $\square$

## 4.2 Error on the ground state energy and density

We now focus on the calculations of the ground state energy (2.4) and density (2.5). Let  $\Pi^{L,p}$  and  $\Pi^{L,q}$  be interpolation operators of order  $(p+1)$  and  $(q+1)$  respectively. For the total energy, we introduce

$$\varepsilon_{n\mathbf{k}}^{L,p} := \Pi^{L,p}(\varepsilon_{n\mathbf{k}}) \quad \text{and} \quad \varepsilon_{n\mathbf{k}}^{L,q} := \Pi^{L,q}(\varepsilon_{n\mathbf{k}}).$$

Using two different interpolation operators allows us to identify the error coming from the inexact approximation of  $\varepsilon_{n\mathbf{k}}$  everywhere in the Brillouin zone (*bulk error*) and the one coming from the inexact calculation of the Fermi energy (*surface error*). We assume that the Fermi level is approximated by  $\varepsilon_F^{L,q}$  as in the previous section, using the same interpolation operator  $\Pi^{L,q}$ . Altogether, we compare the ground state energy

$E = E(\varepsilon_F)$  defined in (2.4) with the approximate ground state energy

$$E^{L,p,q} := \sum_{n \in \mathbb{N}^*} \int_{\mathcal{B}} \varepsilon_{n\mathbf{k}}^{L,p} \mathbb{1}(\varepsilon_{n\mathbf{k}}^{L,q} \leq \varepsilon_F^{L,q}) d\mathbf{k}, \quad (4.8)$$

and the ground state electronic density  $\rho = \rho_{\varepsilon_F}$ , where  $\rho_{\varepsilon}$  is defined in (2.5), with the approximate ground state density

$$\rho^{L,p,q}(\mathbf{r}) := \sum_{n \in \mathbb{N}^*} \int_{\mathcal{B}} \Pi^{L,p}(|u_{n\mathbf{k}}(\mathbf{r})|^2) \mathbb{1}(\varepsilon_{n\mathbf{k}}^{L,q} \leq \varepsilon_F^{L,q}) d\mathbf{k}.$$

The main theorem of this section is the following.

**Theorem 4.5** Assume  $V \in H_{\text{per}}^s$  for some  $s \geq 0$ . Under Assumptions 1 and 2, there exists  $C \in \mathbb{R}_+$  such that, for all  $L \in \mathbb{N}$ ,

$$\begin{aligned} \|\rho - \rho^{L,p,q}\|_{H_{\text{per}}^{s+2}} &\leq C \left( \frac{1}{L^{p+1}} + \frac{1}{L^{q+1}} \right), \\ |E - E^{L,p,q}| &\leq C \left( \frac{1}{L^{p+1}} + \frac{1}{L^{2q+2}} \right). \end{aligned}$$

**Proof** We start with the density. Let  $W \in H_{\text{per}}^{-(s+2)}$  and introduce

$$W_{n\mathbf{k}} := \left\langle W, |u_{n\mathbf{k}}|^2 \right\rangle_{H_{\text{per}}^{-(s+2)}, H_{\text{per}}^{s+2}} \quad \text{and} \quad W_{n\mathbf{k}}^{L,p} := \Pi^{L,p}(W_{n\mathbf{k}}),$$

so that the error is  $\|\rho - \rho^{L,p,q}\|_{H_{\text{per}}^{s+2}} = \sup_{W \in H_{\text{per}}^{-(s+2)}, \|W\|_{H_{\text{per}}^{-(s+2)}}=1} e^{L,p,q}(W)$ , where we

set

$$\begin{aligned} e^{L,p,q}(W) &:= \left\langle W, \rho - \rho^{L,p,q} \right\rangle_{H_{\text{per}}^{-(s+2)}, H_{\text{per}}^{s+2}} \\ &= \sum_{n \in \mathbb{N}^*} \int_{\mathcal{B}} \left( W_{n\mathbf{k}} \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon_F) - W_{n\mathbf{k}}^{L,p} \mathbb{1}(\varepsilon_{n\mathbf{k}}^{L,q} \leq \varepsilon_F^{L,q}) \right) d\mathbf{k}. \end{aligned}$$

We decompose the error into two contributions: the bulk error and the surface error. We write  $e^{L,p,q}(W) = e_{\text{bulk}}^{L,p,q}(W) + e_{\text{surf}}^{L,p,q}(W)$  with

$$e_{\text{bulk}}^L(W) := \sum_{n \leq M} \int_{\mathcal{B}} \left( W_{n\mathbf{k}} - W_{n\mathbf{k}}^{L,p} \right) \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon_F) d\mathbf{k} \quad (4.9)$$

and

$$e_{\text{surf}}^{L,p,q}(W) := \sum_{n \leq M} \int_{\mathcal{B}} W_{n\mathbf{k}}^{L,p} \left[ \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon_F) - \mathbb{1}(\varepsilon_{n\mathbf{k}}^{L,q} \leq \varepsilon_F^{L,q}) \right] d\mathbf{k}. \quad (4.10)$$

The bulk error (4.9) is spread over the whole Brillouin zone, while the surface error (4.10) is localized near the Fermi surface  $\mathcal{S}$ . In order to control these two terms, we use Lemma 3.2 which shows that  $\mathbf{k} \mapsto \sum_{m=1}^n W_{m\mathbf{k}} = \langle W, \rho_{n\mathbf{k}} \rangle_{H_{\text{per}}^{-(s+2)}, H_{\text{per}}^{s+2}}$  is smooth on  $\mathbf{k} \mapsto \tilde{\mathcal{B}}_n(\varepsilon_F, \delta_0)$ , while the map  $\mathbf{k} \mapsto \varepsilon_{n\mathbf{k}}$  is smooth on  $\tilde{\mathcal{S}}_n(\varepsilon_F, \delta_0)$ .

**Bulk error** We have

$$\begin{aligned} e_{\text{bulk}}^{L,p}(W) &= \frac{1}{|\mathcal{B}|} \sum_{n \leq \overline{M}} \int_{\mathcal{B}_n(\varepsilon_F)} \sum_{m=1}^n (W_{m\mathbf{k}} - W_{m\mathbf{k}}^{L,p}) d\mathbf{k} \\ &= \frac{1}{|\mathcal{B}|} \sum_{n \leq \overline{M}} \int_{\mathcal{B}_n(\varepsilon_F)} \left( \langle W, \rho_{n\mathbf{k}} \rangle_{H_{\text{per}}^{-(s+2)}, H_{\text{per}}^{s+2}} - \Pi^{L,p} \left[ \langle W, \rho_{n\mathbf{k}} \rangle_{H_{\text{per}}^{-(s+2)}, H_{\text{per}}^{s+2}} \right] \right) d\mathbf{k}. \end{aligned}$$

Let us introduce the maps

$$F_{n,W} : \begin{cases} \mathcal{B} \rightarrow \mathbb{R} \\ \mathbf{k} \mapsto \langle W, \rho_{n\mathbf{k}} \rangle_{H_{\text{per}}^{-(s+2)}, H_{\text{per}}^{s+2}} \end{cases}.$$

According to Lemma 3.2, the maps  $F_{n,W}$  are analytic on  $\tilde{\mathcal{B}}_n(\varepsilon_F, \delta_0/2)$ , and it holds that

$$\forall n \leq \overline{M}, \forall \mathbf{k} \in \tilde{\mathcal{B}}_n(\varepsilon_F, \delta_0/2), \quad \left| F_{n,W}^{(p+1)} \right| \leq \|W\|_{H_{\text{per}}^{-(s+2)}} \sup_{\mathbf{k} \in \tilde{\mathcal{B}}_n(\varepsilon_F, \delta_0/2)} \left\| \partial_{\mathbf{k}}^{(p+1)} \rho_{n\mathbf{k}} \right\|_{H_{\text{per}}^{s+2}}.$$

Together with (4.1), we deduce that there exists  $C \in \mathbb{R}_+$  such that

$$\left| e_{\text{bulk}}^{L,p}(W) \right| \leq \frac{C}{L^{p+1}} \|W\|_{H_{\text{per}}^{-(s+2)}}.$$

**Surface error** For the integrand in (4.10) to be non-zero, it must hold that

$$\varepsilon_{n\mathbf{k}} \leq \varepsilon_F \text{ and } \varepsilon_{n\mathbf{k}}^{L,q} > \varepsilon_F^{L,q} \quad \text{or} \quad \varepsilon_{n\mathbf{k}} > \varepsilon_F \text{ and } \varepsilon_{n\mathbf{k}}^{L,q} \leq \varepsilon_F^{L,q}.$$

In the former case for instance, we have

$$0 \leq \varepsilon_F - \varepsilon_{n\mathbf{k}} = \left( \varepsilon_F - \varepsilon_F^{L,q} \right) + \left( \varepsilon_F^{L,q} - \varepsilon_{n\mathbf{k}}^{L,q} \right) + \left( \varepsilon_{n\mathbf{k}}^{L,q} - \varepsilon_{n\mathbf{k}} \right).$$

The middle term being negative, we deduce that  $|\varepsilon_F - \varepsilon_{n\mathbf{k}}| \leq |\varepsilon_F - \varepsilon_F^{L,q}| + |\varepsilon_{n\mathbf{k}}^{L,q} - \varepsilon_{n\mathbf{k}}|$ . The other case is similar. In particular, as in the proof of Lemma 4.3, for  $L$  large enough, we can first restrict the integral in (4.10) to  $\tilde{\mathcal{S}}_n(\varepsilon_F, \delta_0/2)$ , then to some  $\tilde{\mathcal{S}}_n(\varepsilon_F, CL^{-(q+1)})$ . Finally, since the maps  $\mathbf{k} \mapsto W_{n\mathbf{k}} = \langle W, \rho_{n+1,\mathbf{k}} - \rho_{n\mathbf{k}} \rangle_{H_{\text{per}}^{-(s+2)}, H_{\text{per}}^{s+2}}$  are smooth on  $\tilde{\mathcal{S}}_n(\varepsilon_F, \delta_0)$ , we deduce that there exists  $C_q \in \mathbb{R}_+$  such that

$$\left| e_{\text{surf}}^{L,p,q}(W) \right| \leq \frac{C_q}{L^{(q+1)}} \|W\|_{H_{\text{per}}^{-(s+2)}},$$

and the result follows.

**Remark 4.6** As we see from the proof, the surface error behaves as  $L^{-q-1}$ , while the bulk error behaves as  $L^{-p-1}$ . For the case of insulators (no Fermi surface), the surface error vanishes, and the bulk error can be exponentially small with good choices of interpolants. One can ask whether the bulk error could also be much smaller in the metallic case. We believe that our estimates are optimal, and that errors are much larger than in the insulating case, due to Gibbs oscillations. This is illustrated in our numerical simulations in Sect. 6.

**Case of the energy** We now focus on the energy. We follow the same lines as above, and decompose the error into a bulk error and a surface error. The bulk error is bounded as above. We focus on the surface error, which reads

$$\begin{aligned} e_{\text{surf}}^{L,p,q} &:= \sum_{n \leq \bar{M}} \int_{\mathcal{B}} \varepsilon_{n\mathbf{k}}^{L,p} \left[ \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon_F) - \mathbb{1}(\varepsilon_{n\mathbf{k}}^{L,q} \leq \varepsilon_F^{L,q}) \right] d\mathbf{k} \\ &= \sum_{n \leq \bar{M}} \int_{\mathcal{B}} (\varepsilon_{n\mathbf{k}}^{L,p} - \varepsilon_F) \left[ \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon_F) - \mathbb{1}(\varepsilon_{n\mathbf{k}}^{L,q} \leq \varepsilon_F^{L,q}) \right] d\mathbf{k} \\ &\quad + \underbrace{\varepsilon_F (\mathcal{N}(\varepsilon_F) - \mathcal{N}^{L,q}(\varepsilon_F^{L,q}))}_{=0}. \end{aligned}$$

Again, the integrand of  $e_{\text{surf}}^{L,p,q}$  is supported on sets of the form  $\tilde{\mathcal{S}}_n(\varepsilon_F, CL^{-(q+1)})$ . On these sets, it holds that

$$|\varepsilon_{n\mathbf{k}}^{L,p} - \varepsilon_F| \leq |\varepsilon_{n\mathbf{k}}^{L,p} - \varepsilon_{n\mathbf{k}}| + |\varepsilon_{n\mathbf{k}} - \varepsilon_F| = O(L^{-(p+1)} + L^{-(q+1)}).$$

We easily deduce that

$$\begin{aligned} |e_{\text{surf}}^{L,p,q}| &\leq \frac{1}{|\mathcal{B}|} \sum_{n \leq \bar{M}} \int_{\tilde{\mathcal{S}}_n(\varepsilon_F, CL^{-(q+1)})} C \left( L^{-(p+1)} + L^{-(q+1)} \right) d\mathbf{k} \\ &\leq C \left( \frac{1}{L^{p+q+2}} + \frac{1}{L^{2q+2}} \right), \end{aligned}$$

and the proof follows.  $\square$

**Remark 4.7** (Order gain on the energy) The interest of choosing two different interpolation operators  $\Pi^{L,p}$  and  $\Pi^{L,q}$  for the calculation of  $E^{L,p,q}$  is now clear: the integration zone  $\mathbb{1}(\varepsilon_{n\mathbf{k}}^{L,q} \leq \varepsilon_F^{L,q})$  can be approximated with a lower order term (i.e.  $q = \lceil \frac{p-1}{2} \rceil$ ) with no loss of order. For instance, when using a cubic method ( $p = 3$ ), it is enough to evaluate the integral of cubic functions on tetrahedra ( $q = 1$ ), and not on complicated intersections of cubic surfaces ( $q = 3$ ).

This is surprising at first: since the computation of the approximate energy  $E^{L,p,q}$  involves the approximate Fermi level  $\varepsilon_F^{L,q}$ , how can the energy be more accurate than

the Fermi level? The answer, as shown above, is that, to leading order, the variations of the energy caused by an error in the determination of the Fermi surface is proportional to  $\varepsilon_F$  times the error on the number of the particles. Since this number is kept fixed to  $N$  for all  $L$ , this leading order error vanishes. This means that, even if the exact Fermi level is known, it is still numerically advantageous to keep it determined implicitly through the equation  $\mathcal{N}^{L,q}(\varepsilon_F^{L,q}) = N$ .

## 5 Smearing methods

We now focus on smearing methods. Let  $A$  denote either the integrated density of states  $\mathcal{N}$ , the ground state density  $\rho$  or the ground state energy  $E$ . We want to approximate  $A$  by  $A^L$ , where  $A^L$  is obtained by replacing the integral in (2.3)–(2.5) by a corresponding Riemann sum on a regular grid with  $L^d$  points. However, since the step function  $f(x) := \mathbb{1}(x \leq 0)$  appearing in the integrand is discontinuous, we expect the convergence to be slow. The idea of smearing methods is to replace this step function by a *smear*ed function  $f^\sigma$  that is smooth: we define

$$A^\sigma(\varepsilon) = \int_{\mathcal{B}} \sum_{n \in \mathbb{N}^*} A_{n\mathbf{k}} f^\sigma(\varepsilon_{n\mathbf{k}} - \varepsilon) d\mathbf{k}, \quad (5.1)$$

where  $f^\sigma$  is a smooth approximation to  $f$ , as we will discuss below. This approximate quantity  $A^\sigma$  can then be efficiently computed by a Riemann sum. We introduce, for  $L \in \mathbb{N}^*$ , the uniform grid

$$\mathcal{B}_L := \mathcal{B} \cap L^{-1}\mathcal{R}^*,$$

where we see here  $\mathcal{B}$  as a torus, so that there are  $L^d$  points in  $\mathcal{B}_L$ . We then define

$$A^{\sigma,L}(\varepsilon) := \frac{1}{L^d} \sum_{\mathbf{k} \in \mathcal{B}_L} \sum_{n \in \mathbb{N}^*} A_{n\mathbf{k}} f^\sigma(\varepsilon_{n\mathbf{k}} - \varepsilon). \quad (5.2)$$

We define  $\varepsilon_F^\sigma$  and  $\varepsilon_F^{\sigma,L}$  to be the (a priori non-unique) solutions of the equations

$$\mathcal{N}^\sigma(\varepsilon_F^\sigma) = N, \quad \mathcal{N}^{\sigma,L}(\varepsilon_F^{\sigma,L}) = N, \quad (5.3)$$

and we finally set

$$A^\sigma := A^\sigma(\varepsilon_F^\sigma), \quad A^{\sigma,L} := A^{\sigma,L}(\varepsilon_F^{\sigma,L}). \quad (5.4)$$

The quantities  $A^{\sigma,L}$  are the ones that we can compute numerically. Our goal is to compute the error between  $A^{\sigma,L}$  and  $A$ . We first estimate the error between  $A^\sigma$  and  $A$  in Sect. 5.2. Then, we provide in Sect. 5.3 error estimates for the discretization error  $A^{\sigma,L} - A^\sigma$ . The combination of the two provides the total error estimates for smearing methods.

## 5.1 Smearing functions

In this section, we explain how smearing functions are constructed.

**Definition 5.1** (*Smearing mollifier*) We say that a function  $\delta^1 : \mathbb{R} \rightarrow \mathbb{R}$  is a smearing mollifier if it satisfies the following two properties:

- (P1)  $\delta^1 \in \mathcal{S}(\mathbb{R})$ , where  $\mathcal{S}(\mathbb{R})$  denotes the Schwartz space of fast decaying functions;  
 (P2)  $\int_{\mathbb{R}} \delta^1 = 1$ .

Such a smearing mollifier is of order at least  $p \in \mathbb{N}$  if

$$\int_{\mathbb{R}} P(x) \delta^1(x) dx = P(0) \quad \text{for all polynomials } P \text{ with } \deg(P) \leq p,$$

that is  $M_0(\delta^1) = 1$  and  $M_n(\delta^1) = 0$  for  $1 \leq n \leq p$ , where  $M_n(\phi)$  is the  $n$ -th momentum of the function  $\phi \in \mathcal{S}(\mathbb{R})$ :

$$\forall n \in \mathbb{N}, \quad M_n(\phi) = \int_{\mathbb{R}} x^n \phi(x) dx.$$

The *order* of a smearing method is the largest  $p$  for which the above property holds. We say that  $f^1 : \mathbb{R} \rightarrow \mathbb{R}$  is a smearing function if there exists a smearing mollifier  $\delta^1$  such that

$$f^1(x) = \int_{-\infty}^x \delta^1(y) dy.$$

For any smearing mollifier  $\delta^1$ , we set  $\delta^\sigma(x) := \sigma^{-1} \delta^1(\sigma^{-1}x)$  and

$$f^\sigma(x) = 1 - \int_{-\infty}^x \delta^\sigma(y) dy, \quad \text{so that} \quad f^\sigma(x) = f^1(\sigma^{-1}x). \quad (5.5)$$

Note that  $\delta^\sigma = -(f^\sigma)'$ , and that we have in a distributional sense  $\delta^\sigma \rightarrow \delta$  and  $f^\sigma \rightarrow f$  as  $\sigma \rightarrow 0$ .

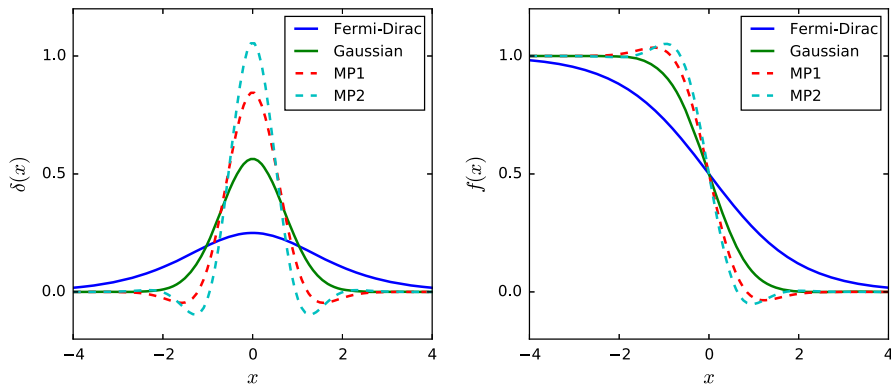
The true step function  $f$  is non-increasing (which implies that the set of possible Fermi levels  $\mathcal{N}^{-1}(\{N\})$  is an interval), and has values equal to either 0 or 1. In particular,  $f_{n\mathbf{k}} = f(\varepsilon_{n\mathbf{k}} - \varepsilon_F)$  is interpreted as the occupation number of the Bloch modes with energy  $\varepsilon_{n\mathbf{k}}$ . By contrast, this interpretation is not valid for a smearing method of order  $p \geq 2$ , since smearing functions of order  $p \geq 2$  necessarily have values outside the range  $[0, 1]$  (otherwise,  $\int (f^1 - f)x$  would be positive).

Let us mention some possible choices encountered in the literature and used in practice (see Fig. 4):

- the Fermi–Dirac smearing [7,9,18]:

$$f_{\text{FD}}^1(x) := \frac{1}{1 + e^x}, \quad \delta_{\text{FD}}^1(x) := \frac{1}{2 + e^x + e^{-x}}. \quad (5.6)$$

This method is of order 1 and  $f_{\text{FD}}^1$  is decreasing from 1 to 0;



**Fig. 4** Some smearing functions. Approximation to the Dirac function  $f^1$  (left), and occupation numbers  $\delta^1$  (right)

- the Gaussian smearing [6]:

$$f_G^1(x) := \frac{1}{2} (1 - \operatorname{erf}(x)), \quad \delta_G^1(x) := \frac{1}{\sqrt{\pi}} e^{-x^2}. \quad (5.6')$$

This method is of order 1 and  $f_G^1$  is decreasing from 1 to 0;

- the Methfessel–Paxton smearing [21]: this method is defined by the sequence of functions  $(f_{\text{MP},N}^1)_{N \in \mathbb{N}}$  given by

$$f_{\text{MP},N}^1(x) := f_G^1(x) + \sum_{n=1}^N A_n H_{2n-1}(x) e^{-x^2}, \quad \delta_{\text{MP},N}^1(x) := \sum_{n=0}^N A_n H_{2n}(x) e^{-x^2}. \quad (5.6'')$$

Here, the functions  $(H_n)_{n \in \mathbb{N}}$  are the Hermite polynomials (defined as  $H_0(x) = 1$ ,  $H_{n+1}(x) = 2xH_n(x) - H_n'(x)$ ), and the coefficients  $A_n := \frac{(-1)^n}{n!4^n\sqrt{\pi}}$  are chosen such that the method is of order  $2N + 1$ . For  $N \geq 1$ ,  $f_{\text{MP},N}^1$  is not monotone, and has negative occupation numbers;

- the Marzari–Vanderbilt cold smearing [15]:

$$f_{\text{cs}}^1(x) := f_G^1(x) + \frac{1}{4\sqrt{\pi}} (-aH_2(x) + H_1(x)) e^{-x^2}, \quad (5.6''')$$

corresponding to

$$\delta_{\text{cs}}^1(x) = \frac{1}{\sqrt{\pi}} \left( ax^3 - x^2 - \frac{3}{2}ax + \frac{3}{2} \right) e^{-x^2},$$

where  $a$  is a free parameter, usually chosen so that  $f_{\text{cs}}^1$  is always non-negative (avoiding negative occupation numbers). This method, like the  $N = 1$  case of the Methfessel–Paxton scheme above, is of order 3 (Fig. 4).

**Remark 5.2** (Temperature) The Fermi–Dirac distribution is used to model electronic systems at a finite temperature. In this case  $\sigma = k_B T$ , where  $k_B$  is the Boltzmann constant. The other smearing functions do not have such a physical interpretation and are only chosen for their mathematical properties.

The Fermi–Dirac function is meromorphic, but has poles at the imaginary energies  $(2\mathbb{Z} + 1)i\pi$  (called *Matsubara frequencies* in the context of field theory). By contrast, the other smearing functions introduced above (called *Gaussian-type* in the following) are entire. This will have an impact on our estimates.

## 5.2 Error between exact and smeared quantities

In the remaining of this section, we fix a smearing mollifier  $\delta^1$  of order  $p$ , and we are interested in the behavior of  $A^\sigma - A$  as  $\sigma$  goes to 0.

Consider a quantity of interest of the form

$$A(\varepsilon) = \sum_{n \in \mathbb{N}^*} \int_B A_{n\mathbf{k}} \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon) \, d\mathbf{k}. \quad (5.7)$$

Then, formally (we will justify this computation case by case for various  $A$  later)

$$\begin{aligned} \sum_{n \in \mathbb{N}^*} \int_B A_{n\mathbf{k}} f^1\left(\frac{\varepsilon_{n\mathbf{k}} - \varepsilon}{\sigma}\right) \, d\mathbf{k} &= \sum_{n \in \mathbb{N}^*} \int_B A_{n\mathbf{k}} \left( \int_{\frac{\varepsilon_{n\mathbf{k}} - \varepsilon}{\sigma}}^{\infty} \delta^1(x) \, dx \right) \, d\mathbf{k} \\ &= \frac{1}{\sigma} \sum_{n \in \mathbb{N}^*} \int_B A_{n\mathbf{k}} \left( \int_{\varepsilon_{n\mathbf{k}}}^{\infty} \delta^1\left(\frac{\varepsilon' - \varepsilon}{\sigma}\right) \, d\varepsilon' \right) \, d\mathbf{k} \\ &= \frac{1}{\sigma} \int_{\mathbb{R}} \left( \sum_{n \in \mathbb{N}^*} \int_B A_{n\mathbf{k}} \mathbb{1}(\varepsilon_{n\mathbf{k}} \leq \varepsilon) \, d\mathbf{k} \right) \delta^1\left(\frac{\varepsilon' - \varepsilon}{\sigma}\right) \, d\varepsilon' \\ &= \frac{1}{\sigma} \int_{\mathbb{R}} A(\varepsilon) \delta^1\left(\frac{\varepsilon' - \varepsilon}{\sigma}\right) \, d\varepsilon' = (A * \delta^\sigma)(\varepsilon). \end{aligned} \quad (5.8)$$

In other words, the effect of smearing is to smooth the function  $A(\varepsilon)$  by a convolution with  $\delta^\sigma$ . In order to understand the properties of the smearing method, we therefore have to study the asymptotic behavior of integrals of the form (5.8) for  $\sigma \rightarrow 0$ .

To make this precise, we introduce the mollification operator. Let us denote by  $\mathcal{S}'(\mathbb{R})$  the set of tempered distributions on  $\mathbb{R}$ . For  $g \in \mathcal{S}'(\mathbb{R})$  and  $\phi \in \mathcal{S}(\mathbb{R})$ , we define (in the sequel,  $g \sim A$ , and  $\phi \sim \delta^1$  is a mollifier)

$$\mathcal{M}_{g,\phi}(\varepsilon, \sigma) = \begin{cases} \langle g, \frac{1}{\sigma} \phi\left(\frac{\cdot - \varepsilon}{\sigma}\right) \rangle_{\mathcal{S}', \mathcal{S}} & \text{if } \sigma \neq 0 \\ g(\varepsilon) M_0(\phi) & \text{if } \sigma = 0 \end{cases} \quad (5.9)$$

Note that we extended  $\mathcal{M}$  to  $\sigma$  negative. Due to the change of variables in (5.8), this does not correspond to taking a negative smearing parameter in the original definition (5.1) of  $A^\sigma$ . The main idea of this section is that if  $g$  is smooth, then we can write that

$$\mathcal{M}_{g,\phi}(\varepsilon, \sigma) = \frac{1}{\sigma} \int_{\mathbb{R}} g(\varepsilon) \phi\left(\frac{\varepsilon' - \varepsilon}{\sigma}\right) d\varepsilon' = \int_{\mathbb{R}} g(\varepsilon + \sigma x) \phi(x) dx.$$

In particular, if  $\phi$  is a smearing mollifier of order  $p$ , then by Taylor-expanding  $g$  around  $\varepsilon$ , this quantity is also  $g(\varepsilon) + O(\sigma^{p+1})$ . We make this statement rigorous in the next Lemma, whose proof is postponed until the end of the section.

**Lemma 5.3** *Let  $g \in \mathcal{S}'(\mathbb{R})$  be such that  $g|_{(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)}$  is a function of class  $C^k$ , and let  $\phi \in \mathcal{S}(\mathbb{R})$ . Then, the function  $\mathcal{M}_{g,\phi}(\varepsilon, \sigma)$  is of class  $C^k$  on  $(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0) \times \mathbb{R}$ , and we have, for all  $(m, n) \in \mathbb{N} \times \mathbb{N}$  such that  $m + n \leq k$ ,*

$$\frac{\partial^{m+n} \mathcal{M}_{g,\phi}}{\partial \varepsilon^m \partial \sigma^n}(\varepsilon, \sigma) = \begin{cases} \frac{1}{\sigma^{n+1}} \left\langle g, \phi_{m,n}\left(\frac{\cdot - \varepsilon}{\sigma}\right) \right\rangle_{\mathcal{S}', \mathcal{S}} & \text{if } \sigma \neq 0, \\ g^{(m+n)}(\varepsilon) M_n(\phi) & \text{if } \sigma = 0, \end{cases}$$

where  $\phi_{m,n} \in \mathcal{S}(\mathbb{R})$  is defined by

$$\forall t \in \mathbb{R}, \quad \phi_{m,n}(t) := (-1)^{m+n} \frac{d^{m+n}}{dt^{m+n}} (t^n \phi(t)).$$

### 5.2.1 Error on the integrated density of states and on the Fermi level

By choosing  $A_{n\mathbf{k}} = 1$  in (5.7) and using the decay at infinity of  $\delta^1$  to justify the exchange of integrals in (5.8), we have  $\mathcal{N}^\sigma(\varepsilon) = \tilde{\mathcal{N}}(\varepsilon, \sigma)$ , where we set for clarity  $\tilde{\mathcal{N}} := \mathcal{M}_{\mathcal{N}, \delta^1}$ .

**Lemma 5.4** *For any smearing mollifier  $\delta^1$  of order  $p \geq 1$ , the function  $\tilde{\mathcal{N}}$  is smooth on  $(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0) \times \mathbb{R}$ , and satisfies*

$$\begin{aligned} \frac{\partial \tilde{\mathcal{N}}}{\partial \varepsilon}(\varepsilon, 0) &= \mathcal{D}(\varepsilon) > 0, \quad \forall 1 \leq n \leq p, \quad \frac{\partial^n \tilde{\mathcal{N}}}{\partial \sigma^n}(\varepsilon, 0) = 0, \\ \frac{\partial^{p+1} \tilde{\mathcal{N}}}{\partial \sigma^{p+1}}(\varepsilon, 0) &= \mathcal{D}^{(p)}(\varepsilon) M_{p+1}(\delta^1). \end{aligned}$$

In particular, there exists  $C \in \mathbb{R}^+$  such that

$$\max_{\varepsilon \in (\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)} |\mathcal{N}(\varepsilon) - \mathcal{N}^\sigma(\varepsilon)| \leq C \sigma^{p+1}. \quad (5.10)$$

**Proof** Applying Lemma 5.3 with  $g = \mathcal{N}$ , which is smooth on  $(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)$ , and  $\phi = \delta^1$ , we obtain that  $\tilde{\mathcal{N}}$  is smooth on  $(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0) \times \mathbb{R}$ . The proof is then a consequence of the fact that  $M_0(\delta^1) = 1$  and  $M_n(\delta^1) = 0$  for  $1 \leq n \leq p$ , together with the fact that  $\mathcal{N}'(\varepsilon) = \mathcal{D}(\varepsilon)$  by definition (see (3.7)).  $\square$

From (5.3),  $\varepsilon_F^\sigma$  is solution to  $\tilde{\mathcal{N}}(\varepsilon_F^\sigma, \sigma) = N$ . From the previous proposition together with the implicit function theorem we directly get the following result.

**Lemma 5.5** *For any smearing mollifier  $\delta^1$  of order  $p \geq 1$ , there exists  $\sigma_1, \delta_1 > 0$ , such that for  $|\sigma| < \sigma_1$ , the equation  $\tilde{\mathcal{N}}(\cdot, \sigma) = N$  has a unique solution  $\varepsilon_F^\sigma$  in  $(\varepsilon_F - \delta_1, \varepsilon_F + \delta_1)$ . In addition, the function  $(-\sigma_1, \sigma_1) \ni \sigma \mapsto \varepsilon_F^\sigma \in \mathbb{R}$  is smooth, and it holds that*

$$\varepsilon_F^{\sigma=0} = \varepsilon_F, \quad \forall 1 \leq n \leq p, \quad \left. \frac{d^n \varepsilon_F^\sigma}{d\sigma^n} \right|_{\sigma=0} = 0, \quad \left. \frac{d^{p+1} \varepsilon_F^\sigma}{d\sigma^{p+1}} \right|_{\sigma=0} = -\frac{\mathcal{D}^{(p)}(\varepsilon_F)}{\mathcal{D}(\varepsilon_F)} M_{p+1}(\delta^1).$$

In particular, it holds that

$$\varepsilon_F^\sigma = \varepsilon_F + \frac{1}{(p+1)!} \left( -\frac{\mathcal{D}^{(p)}(\varepsilon_F)}{\mathcal{D}(\varepsilon_F)} M_{p+1}(\delta^1) \right) \sigma^{p+1} + O(\sigma^{p+2}).$$

## 5.2.2 Error on the ground-state energy

By choosing  $A_{n\mathbf{k}} = \varepsilon_{n\mathbf{k}}$  in (5.8), we have that  $E^\sigma(\varepsilon) = \tilde{E}(\varepsilon, \sigma)$ , where we set for clarity  $\tilde{E} := \mathcal{M}_{E, \delta^1}$ . Numerically, the true ground state energy  $E$  is approximated by  $E^\sigma := \tilde{E}(\varepsilon_F^\sigma, \sigma)$ .

**Lemma 5.6** *For any smearing mollifier  $\delta^1$  of order  $p \geq 1$ , there exists  $\sigma_1 > 0$  such that the function  $\sigma \mapsto \tilde{E}$  is well-defined and smooth on  $(-\sigma_1, \sigma_1) \times \mathbb{R}$ , and satisfies*

$$\begin{aligned} \frac{\partial \tilde{E}}{\partial \varepsilon}(\varepsilon, 0) &= \varepsilon \mathcal{D}(\varepsilon), \quad \forall 1 \leq n \leq p, \quad \frac{\partial^n \tilde{E}}{\partial \sigma^n}(\varepsilon, 0) = 0, \\ \frac{\partial^{p+1} \tilde{E}}{\partial \sigma^{p+1}}(\varepsilon, 0) &= \left( \varepsilon \mathcal{D}^{(p)}(\varepsilon) + p \mathcal{D}^{(p-1)}(\varepsilon) \right) M_{p+1}(\delta^1) \end{aligned}$$

In particular, it holds that

$$E^{\sigma=0} = E_0, \quad \forall 1 \leq n \leq p, \quad \left. \frac{d^n E^\sigma}{d\sigma^n} \right|_{\sigma=0} = 0, \quad \left. \frac{d^{p+1} E^\sigma}{d\sigma^{p+1}} \right|_{\sigma=0} = p \mathcal{D}^{(p-1)}(\varepsilon_F) M_{p+1}(\delta^1).$$

Finally, we have

$$E^\sigma = E + \frac{1}{(p+1)!} \left( p \mathcal{D}^{(p-1)}(\varepsilon_F) M_{p+1}(\delta^1) \right) \sigma^{p+1} + O(\sigma^{p+2}). \quad (5.11)$$

**Proof** The first part is a consequence of Lemma 5.3 applied to  $g = E(\cdot)$  (which is smooth on  $(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)$ ) and  $\phi = \delta^1$ , together with (3.8). The second part comes from Lemma 5.5, and the chain rule.  $\square$

**Extrapolation of the energy** Following Marzari [16], we can introduce the entropy defined as

$$S^\sigma = \sum_{n \in \mathbb{N}^*} \int_{\mathcal{B}} s \left( \frac{\varepsilon_{n\mathbf{k}} - \varepsilon_F^\sigma}{\sigma} \right) d\mathbf{k}, \quad (5.12)$$

where  $s : \mathbb{R} \rightarrow \mathbb{R}$  is the unique function of  $\mathcal{S}(\mathbb{R})$  such that  $s'(t) = t\delta^1(t)$ . Following the same computation as in (5.8) with  $f^1 = s$ , we see that  $S^\sigma = \tilde{S}(\varepsilon_F^\sigma, \sigma)$ , where we set  $\tilde{S} := \mathcal{M}_{\mathcal{N}, -t\delta^1}$ . Note also that  $M_n(t\delta^1) = M_{n+1}(\delta^1)$  for all  $0 \leq n \leq p-1$ . As in the proof of Lemma 5.6, we deduce that if  $\delta^1$  is of order  $p \geq 1$ , then  $\tilde{S}$  is smooth on  $(-\sigma_1, \sigma_1) \times \mathbb{R}$ , with

$$\forall 1 \leq n \leq p-1, \quad \frac{\partial^n \tilde{S}}{\partial \varepsilon^n}(\varepsilon, 0) = 0, \quad \frac{\partial^n \tilde{S}}{\partial \sigma^n}(\varepsilon, 0) = 0, \quad \frac{\partial^p \tilde{S}}{\partial \sigma^p}(\varepsilon, 0) = -\mathcal{D}^{(p-1)} M_{p+1}(\delta^1).$$

Together with Lemma 5.5, and the chain rule, this lead to

$$\forall 1 \leq n \leq p-1, \quad \left. \frac{d^n S^\sigma}{d\sigma^n} \right|_{\sigma=0} = 0, \quad \left. \frac{d^p S^\sigma}{d\sigma^p} \right|_{\sigma=0} = -\mathcal{D}^{(p-1)}(\varepsilon_F) M_{p+1}(\delta^1).$$

Thus,

$$S^\sigma = \frac{1}{p!} \left( -\mathcal{D}^{(p-1)}(\varepsilon_F) M_{p+1}(\delta^1) \right) \sigma^p + O(\sigma^{p+1}). \quad (5.13)$$

From (5.11) and (5.13), we finally obtain that

$$E^\sigma - \sigma \frac{p}{p+1} S^\sigma = E + O(\sigma^{p+2})$$

As pointed out in [16], the right-hand side provides an approximation of  $E$  which is consistent of order  $p+2$ , and therefore outperforms the estimator  $E^\sigma$  in the asymptotic regime when  $\sigma$  goes to zero. This is numerically useful, as, from the definition of  $S^\sigma$  in (5.12),  $S^\sigma$  can be easily computed numerically.

### 5.2.3 Error on the ground-state density

**Lemma 5.7** Consider a smearing mollifier  $\delta^1$  of order  $p \geq 1$ . Under Assumptions 1 and 2 and with the notation of Lemma 3.2, there exists  $\sigma_1 > 0$  and  $C \in \mathbb{R}_+$  such that

$$\forall \sigma \in (-\sigma_1, \sigma_1), \quad \|\rho - \rho^\sigma\|_{H_{\text{per}}^{s+2}} \leq C\sigma^{p+1}. \quad (5.14)$$

**Proof** Let  $W \in H_{\text{per}}^{-(s+2)}$ , and set  $W_{n\mathbf{k}} := \langle W, |u_{n\mathbf{k}}|^2 \rangle_{H_{\text{per}}^{-(s+2)}, H_{\text{per}}^{s+2}}$ . We also set

$$A_W(\varepsilon) := \sum_{n \in \mathbb{N}^*} \int_{\mathcal{B}} W_{n\mathbf{k}} \mathbb{1}_{(\varepsilon_{n\mathbf{k}} \leq \varepsilon)} d\mathbf{k} \quad \text{and} \quad A_W^\sigma(\varepsilon) := \sum_{n \in \mathbb{N}^*} \int_{\mathcal{B}} W_{n\mathbf{k}} f^\sigma(\varepsilon_{n\mathbf{k}} - \varepsilon) d\mathbf{k}.$$

In this proof,  $C$  denotes a positive constant independent of  $W$ ,  $n$ ,  $\mathbf{k}$  and  $\sigma$ , for  $\sigma$  small enough, but whose value can vary from line to line. Since  $H_{\text{per}}^{s+2}$  is an algebra

for  $s \geq 0$  and  $d \leq 3$ , we have that

$$|W_{n\mathbf{k}}| = \left| \left\langle W, |u_{n\mathbf{k}}|^2 \right\rangle_{H_{\text{per}}^{-(s+2)}, H_{\text{per}}^{s+2}} \right| \leq C \|W\|_{H_{\text{per}}^{-(s+2)}} \|u_{n\mathbf{k}}\|_{H_{\text{per}}^{s+2}}^2.$$

Besides, using the equality  $H_{\mathbf{k}} u_{n\mathbf{k}} = \varepsilon_{n\mathbf{k}} u_{n\mathbf{k}}$ , (2.2), the continuity of the pointwise product  $H_{\text{per}}^{s+2} \times H_{\text{per}}^s$  to  $H_{\text{per}}^s$ , and a bootstrap argument, we see that

$$\forall n \geq 1, \quad \forall \mathbf{k} \in \mathcal{B}, \quad \|u_{n\mathbf{k}}\|_{H_{\text{per}}^{s+2}} \leq C \left(1 + \varepsilon_{n\mathbf{k}}^{s+2}\right) \leq C n^{\frac{2}{d}(s+2)}. \quad (5.15)$$

Therefore, we have

$$|W_{n\mathbf{k}}| \leq C \|W\|_{H_{\text{per}}^{-(s+2)}} n^{\frac{4}{d}(s+2)}.$$

This estimate shows that  $A_W$  is a tempered distribution (as a continuous function of polynomial growth), and that the computation in (5.8) is justified. In particular, we have

$$A_W^\sigma(\varepsilon) = \mathcal{M}_{A_W, \delta^1}(\varepsilon, \sigma).$$

From similar considerations as in Lemma 3.5,  $A_W$  is smooth on  $[\varepsilon_F - \delta_0, \varepsilon_F + \delta_0]$ , and

$$A'_W(\varepsilon) = \frac{1}{|\mathcal{B}|} \sum_{n \leq M} \int_{\mathcal{S}_n(\varepsilon)} \frac{W_{n\mathbf{k}}}{|\nabla \varepsilon_{n\mathbf{k}}|} d\sigma(\mathbf{k}).$$

It follows that

$$\begin{aligned} |\langle W, \rho - \rho^\sigma \rangle_{H_{\text{per}}^{-(s+2)}, H_{\text{per}}^{s+2}}| &= |A_W(\varepsilon_F) - A_W^\sigma(\varepsilon_F^\sigma)| \leq |A_W(\varepsilon_F) - A_W(\varepsilon_F^\sigma)| \\ &\quad + |A_W(\varepsilon_F^\sigma) - A_W^\sigma(\varepsilon_F^\sigma)| \\ &\leq \left( \max_{\varepsilon \in [\varepsilon_F - \delta_0, \varepsilon_F + \delta_0]} |A'_W(\varepsilon)| \right) |\varepsilon_F - \varepsilon_F^\sigma| + |A_W(\varepsilon_F^\sigma) - A_W^\sigma(\varepsilon_F^\sigma)| \\ &\leq \left( C \|W\|_{H_{\text{per}}^{-(s+2)}} \right) \sigma^{p+1}, \end{aligned}$$

where we have used Lemmas 5.3 and 5.5. The result follows.  $\square$

### 5.2.4 Proof of Lemma 5.3

We finally prove Lemma 5.3. Let  $g \in \mathcal{S}'(\mathbb{R})$ ,  $\varepsilon \in \mathbb{R}$  and  $\sigma \in \mathbb{R}^*$ . We define the shifted and scaled tempered distribution  $g(\varepsilon + \sigma \cdot)$  by duality:

$$\forall \phi \in \mathcal{S}(\mathbb{R}), \quad \langle g(\varepsilon + \sigma \cdot), \phi \rangle_{\mathcal{S}', \mathcal{S}} := \langle g, A_{\varepsilon, \sigma} \phi \rangle_{\mathcal{S}', \mathcal{S}},$$

where  $A_{\varepsilon, \sigma}$  is the linear map on  $\mathcal{S}(\mathbb{R})$  defined by

$$\forall \phi \in \mathcal{S}(\mathbb{R}), \quad \forall \varepsilon' \in \mathbb{R}, \quad (A_{\varepsilon, \sigma} \phi)(\varepsilon') := \frac{1}{|\sigma|} \phi \left( \frac{\varepsilon' - \varepsilon}{\sigma} \right).$$

This is consistent with the usual shift and scale operation for functions. If  $g$  is a tempered distribution that is continuous at  $\varepsilon$ , we define  $g(\varepsilon + 0\cdot)$  to be the constant tempered distribution with value  $g(\varepsilon)$ .

It is easy to check that the family  $(A_{\varepsilon,\sigma})_{(\varepsilon,\sigma) \in \mathbb{R} \times \mathbb{R}^*}$  forms a group of continuous linear operators on  $\mathcal{S}(\mathbb{R})$  satisfying for all  $(\varepsilon, \sigma)$  and  $(\varepsilon', \sigma')$  in  $\mathbb{R} \times \mathbb{R}^*$ ,

$$A_{\varepsilon,\sigma} A_{\varepsilon',\sigma'} = A_{\varepsilon+\sigma\varepsilon',\sigma\sigma'}.$$

In addition, we have the following properties on the derivatives of  $A_{\varepsilon,\sigma}$ : for all  $\phi \in \mathcal{S}(\mathbb{R})$ ,

$$\begin{aligned} A_{\varepsilon',\sigma'}\phi &\xrightarrow{(\varepsilon',\sigma') \rightarrow (\varepsilon,\sigma)} A_{\varepsilon,\sigma}\phi, \\ \frac{A_{\varepsilon',\sigma} - A_{\varepsilon,\sigma}}{\varepsilon' - \varepsilon}\phi &\xrightarrow{\varepsilon' \rightarrow \varepsilon} -\sigma^{-1} A_{\varepsilon,\sigma}\phi', \\ \frac{A_{\varepsilon,\sigma'} - A_{\varepsilon,\sigma}}{\sigma' - \sigma}\phi &\xrightarrow{\sigma' \rightarrow \sigma} -\sigma^{-1} A_{\varepsilon,\sigma}L\phi, \end{aligned}$$

the convergences holding in  $\mathcal{S}(\mathbb{R})$ , where  $L$  is the continuous operator on  $\mathcal{S}(\mathbb{R})$  defined by

$$\forall \phi \in \mathcal{S}(\mathbb{R}), \quad \forall t \in \mathbb{R}, \quad (L\phi)(t) = \frac{d}{dt}(t\phi(t)) = t\phi'(t) + \phi(t).$$

It immediately follows that  $\mathcal{M}_{g,\phi}$  is of class  $C^1$  on  $(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0) \times \mathbb{R}^*$  and that

$$\begin{aligned} \forall (\varepsilon, \sigma) \in \mathbb{R} \times \mathbb{R}^*, \quad \frac{\partial \mathcal{M}_{g,\phi}}{\partial \varepsilon}(\varepsilon, \sigma) &= -\sigma^{-1} \langle g, A_{\varepsilon,\sigma}\phi' \rangle_{\mathcal{S}',\mathcal{S}}, \\ \frac{\partial \mathcal{M}_{g,\phi}}{\partial \sigma}(\varepsilon, \sigma) &= -\sigma^{-1} \langle g, A_{\varepsilon,\sigma}L\phi \rangle_{\mathcal{S}',\mathcal{S}}. \end{aligned}$$

Let us prove that  $\mathcal{M}_{g,\phi}$  is also  $C^1$  at  $\sigma = 0$ . Let  $\varepsilon \in (\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)$ , and let  $\chi \in C_c^\infty(\mathbb{R})$  be a cut-off function supported in a compact interval  $K \subset (\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)$  and equal to 1 in a neighborhood  $\mathcal{V}$  of  $\varepsilon$ . For  $\phi \in \mathcal{S}(\mathbb{R})$ ,  $\varepsilon' \in \mathcal{V}$  and  $\sigma \in \mathbb{R}^*$ , we have

$$\begin{aligned} \langle g, A_{\varepsilon',\sigma}\phi \rangle_{\mathcal{S}',\mathcal{S}} &= \langle \chi g, A_{\varepsilon',\sigma}\phi \rangle_{\mathcal{S}',\mathcal{S}} + \langle (1 - \chi)g, A_{\varepsilon',\sigma}\phi \rangle_{\mathcal{S}',\mathcal{S}} \\ &= \langle A_{\varepsilon',\sigma}\phi, \chi g \rangle_{C^0(K)', C^0(K)} + \langle g, (1 - \chi)A_{\varepsilon',\sigma}\phi \rangle_{\mathcal{S}',\mathcal{S}} \xrightarrow{(\varepsilon',\sigma) \rightarrow (\varepsilon,0)} g(\varepsilon)M_0(\phi), \end{aligned}$$

since when  $(\varepsilon', \sigma) \rightarrow (\varepsilon, 0)$ ,  $A_{\varepsilon',\sigma}\phi \rightarrow M_0(\phi)\delta_\varepsilon$  in the space  $C^0(K)'$  of bounded Borel measures on  $K$ , while  $(1 - \chi)A_{\varepsilon',\sigma}\phi$  goes to zero in  $\mathcal{S}(\mathbb{R})$ . This proves that  $\mathcal{M}_{g,\phi}$  is continuous on  $(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0) \times \mathbb{R}$ .

If in addition,  $g$  is of class  $C^1$  on  $(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)$ , then for  $\sigma \neq 0$ ,

$$\begin{aligned} \frac{\mathcal{M}_{g,\phi}(\varepsilon, \sigma) - \mathcal{M}_{g,\phi}(\varepsilon, 0)}{\sigma} &= \frac{\langle \chi g, A_{\varepsilon,\sigma}\phi \rangle_{\mathcal{S}',\mathcal{S}} - M_0(\phi)(\chi g)(\varepsilon)}{\sigma} \\ &\quad + \frac{\langle (1 - \chi)g, A_{\varepsilon,\sigma}\phi \rangle_{\mathcal{S}',\mathcal{S}}}{\sigma} \end{aligned}$$

$$\begin{aligned}
&= \langle \sigma^{-1}(A_{\varepsilon,\sigma}\phi - M_0(\phi)\delta_\varepsilon), \chi g \rangle_{(C^1(K))', C^1(K)} \\
&\quad + \langle g, (1 - \chi)\sigma^{-1}A_{\varepsilon,\sigma}\phi \rangle_{S', S} \\
&\xrightarrow{\sigma \rightarrow 0} (\chi g)'(\varepsilon)M_1(\phi) = g'(\varepsilon)M_1(\phi),
\end{aligned}$$

since  $\sigma^{-1}(A_{\varepsilon,\sigma}\phi - M_0(\phi)\delta_\varepsilon)$  converges to  $M_1(\phi)\delta'_\varepsilon$  in  $C^1(K)'$ , while  $(1 - \chi)\sigma^{-1}A_{\varepsilon,\sigma}\phi$  converges to 0 in  $S(\mathbb{R})$ . Hence, for  $\varepsilon \in (\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)$ , we have

$$\frac{\partial \mathcal{M}_{g,\phi}}{\partial \varepsilon}(\varepsilon, 0) = g'(\varepsilon)M_0(\phi), \quad \frac{\partial \mathcal{M}_{g,\phi}}{\partial \sigma}(\varepsilon, 0) = g'(\varepsilon)M_1(\phi).$$

Observing that  $M_0(\phi') = 0$  and  $M_0(L\phi) = 0$ , so that  $M_1(\phi') = -M_0(\phi)$ , and that,  $M_1(L\phi) = -M_1(\phi)$  with an integration by part, and reasoning as above, we obtain that for  $\varepsilon' \in \mathcal{V}$  and  $\sigma \in \mathbb{R}^*$ ,

$$\begin{aligned}
\frac{\partial \mathcal{M}_{g,\phi}}{\partial \varepsilon}(\varepsilon, \sigma) &= -\sigma^{-1} \langle g, A_{\varepsilon,\sigma}\phi' \rangle_{S', S} \xrightarrow{\sigma \rightarrow 0} -g'(\varepsilon)M_1(\phi') = g'(\varepsilon)M_0(\phi), \\
\frac{\partial \mathcal{M}_{g,\phi}}{\partial \sigma}(\varepsilon, \sigma) &= -\sigma^{-1} \langle g, A_{\varepsilon,\sigma}L\phi \rangle_{S', S} \xrightarrow{\sigma \rightarrow 0} -g'(\varepsilon)M_1(L\phi) = g'(\varepsilon)M_1(\phi).
\end{aligned}$$

It follows that  $\mathcal{M}_{g,\varepsilon}$  is of class  $C^1$  in  $(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0) \times \mathbb{R}$ . Similar arguments allow one to show that  $\mathcal{M}_{g,\varepsilon}$  is of class  $C^k$  in  $(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0) \times \mathbb{R}$  whenever  $g$  is of class  $C^k$  in  $(\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)$ . The proof follows by a straightforward induction.

### 5.3 Error between smeared quantities and corresponding Riemann sums

We now investigate the convergence of  $A^{\sigma,L}$  to  $A^\sigma$ , for  $\sigma$  fixed. As we already mentioned, the advantage of using smearing functions is that the quantities  $A^\sigma$  are defined as the integral of smooth periodic functions. It is therefore natural to approximate numerically the integral by a regular Riemann sum, for which we can expect exponential convergence, depending on the analytic properties of the integrand (see for instance [28] for a review). For  $Y > 0$ , we introduce the (closed) complex strip

$$S_Y := \mathbb{R}^d + i[-Y, Y]^d = \left\{ \mathbf{z} \in \mathbb{C}^d, \quad |\operatorname{Im}(\mathbf{z})|_\infty \leq Y \right\}.$$

We recall the following classical result, which is proved as in [11, Lemma 5.1].

**Lemma 5.8** *There exists  $C \in \mathbb{R}_+$  and  $\eta > 0$  such that, for all  $Y > 0$  and all  $F : \mathbb{C}^d \rightarrow \mathbb{C}$  that is analytic on  $S_Y$  and  $\mathcal{R}^*$ -periodic on  $\mathbb{R}^d$ , we have*

$$\forall L \in \mathbb{N}^*, \quad \left| \int_{\mathcal{B}} F(\mathbf{k}) d\mathbf{k} - \frac{1}{L^d} \sum_{\mathbf{k} \in \mathcal{B}_L} F(\mathbf{k}) \right| \leq C \left( \max_{\mathbf{z} \in S_Y} |F(\mathbf{z})| \right) \frac{e^{-\eta Y L}}{Y^d}.$$

We see, for a fixed value of  $\sigma$ , the greater the region of analyticity of the integrand, the faster the convergence as  $L \rightarrow \infty$ . We distinguish here the different smearing

functions: the Fermi–Dirac function  $f_{\text{FD}}^\sigma$  has poles at  $(2\mathbb{Z} + 1)\pi\sigma i$  and displays exponential convergence, while the Gaussian-type smearing functions are entire, leading to super-exponential convergence. We collect in Lemmas A.1–A.2 estimates on the analytic behavior of the integrands of interest, from which the results in this section proceed.

We would like to emphasize at this point that, if the value of  $\sigma$  is not fixed but depends on  $L$ , obtaining rates of convergence becomes a more subtle and intricate task. We address this issue in more details in Remark 5.12.

### 5.3.1 Error for the integrated density of states

**Lemma 5.9** (Convergence of the integrated density of states) *It holds that*

- If  $f^1$  is any of the smearing functions (5.6–5.6'''), there exists  $C \in \mathbb{R}_+$  and  $\eta > 0$  such that for all  $0 < \sigma \leq \sigma_0$  and all  $L \in \mathbb{N}^*$ ,

$$\max_{\varepsilon \in [\varepsilon_F - \delta_0, \varepsilon_F + \delta_0]} \left| \mathcal{N}^{\sigma, L}(\varepsilon) - \mathcal{N}^\sigma(\varepsilon) \right| \leq C \sigma^{-(d+1)} e^{-\eta \sigma L}. \quad (5.16)$$

- If  $f^1$  is a Gaussian-type smearing function (5.6'–5.6'''), then there exists  $C' \in \mathbb{R}_+$  and  $\eta' > 0$  such that, for all  $0 < \sigma \leq \sigma_0$  and all  $L \in \mathbb{N}^*$ , such that  $\sigma^2 L \geq 4$ , it holds that

$$\max_{\varepsilon \in [\varepsilon_F - \delta_0, \varepsilon_F + \delta_0]} \left| \mathcal{N}^{\sigma, L}(\varepsilon) - \mathcal{N}^\sigma(\varepsilon) \right| \leq C' \sigma^{-\frac{5d}{3}} L^{-\frac{d}{3}} e^{-\eta' \sigma^{2/3} L^{4/3}}. \quad (5.17)$$

**Proof** We want to compare  $\mathcal{N}^{\sigma, L}(\varepsilon)$  with  $\mathcal{N}^\sigma(\varepsilon)$ . This is exactly the framework of Lemma 5.8, with integrand

$$F_{\varepsilon, \sigma}^{\mathcal{N}}(\mathbf{k}) := \sum_{n \in \mathbb{N}^*} f^\sigma(\varepsilon_{n\mathbf{k}} - \varepsilon) = \text{Tr}_{L_{\text{per}}^2} [f^\sigma(H_{\mathbf{k}} - \varepsilon)].$$

In “Appendix”, we study the analytic property of such functions. From Lemma A.1, there exists  $C \in \mathbb{R}_+$  and  $Y > 0$  such that, for all  $\varepsilon \in [\varepsilon_F - \delta_0, \varepsilon_F + \delta_0]$  and all  $0 < \sigma \leq \sigma_0$ , the map  $F_{\varepsilon, \sigma}^{\mathcal{N}}$  admits an analytic continuation on  $S_{\sigma Y}$ , and it holds that

$$\sup_{\mathbf{z} \in S_{\sigma Y}} \left| F_{\varepsilon, \sigma}^{\mathcal{N}}(\mathbf{z}) \right| \leq C \sigma^{-1}.$$

Together with Lemma 5.8, we deduce that there exists  $C, C' \in \mathbb{R}^+$  and  $\eta' > 0$  such that

$$\left| \mathcal{N}^{\sigma, L}(\varepsilon) - \mathcal{N}^\sigma(\varepsilon) \right| \leq C \sigma^{-1} (\sigma Y)^{-d} e^{-\eta \sigma Y L} \leq C' \sigma^{-(d+1)} e^{-\eta' \sigma L}.$$

This proves the first part (5.16). For the second part, we use Lemma A.2, and get in a similar way that if  $f^1$  is a Gaussian-type smearing function, then there exists  $C \in \mathbb{R}^+$

and  $\eta > 0$  such that for all  $Y \geq 1$ , all  $\varepsilon \in [\varepsilon_F - \delta_0, \varepsilon_F + \delta_0]$  and all  $0 < \sigma \leq \sigma_0$ ,

$$\left| \mathcal{N}^{\sigma, L}(\varepsilon) - \mathcal{N}^{\sigma}(\varepsilon) \right| \leq C(\sigma Y)^{-d} e^{\eta \left( \frac{Y^4}{\sigma^2} - YL \right)}.$$

Taking  $Y = Y(\sigma, L) = \left( \frac{1}{4} \sigma^2 L \right)^{1/3}$  leads to the result. The condition  $\sigma^2 L \geq 4$  comes from the fact that we need  $Y \geq 1$  for this result to be valid.  $\square$

### 5.3.2 Error for the Fermi level, the total energy, and the density

We now turn to the Fermi energy, the total energy, and the density. As above,  $\mathcal{N}^{\sigma, L}$  is a continuous function that satisfies  $\mathcal{N}^{\sigma, L}(-\infty) = 0$  and  $\mathcal{N}^{\sigma, L}(+\infty) = +\infty$ , hence there exists  $\varepsilon_F^{\sigma, L} \in \mathbb{R}$  so that  $\mathcal{N}^{\sigma, L}(\varepsilon_F^{\sigma, L}) = N$ . As  $\mathcal{N}^{\sigma, L}$  is not necessarily increasing,  $\varepsilon_F^{\sigma, L}$  may be non unique. However, since  $\mathcal{N}^{\sigma, L}$  is continuous, close to  $\mathcal{N}^{\sigma}$  (in the sense of the lemma above), and  $\mathcal{N}^{\sigma}(\varepsilon_F^{\sigma}) = N$  with  $\partial_{\varepsilon} \mathcal{N}^{\sigma}(\varepsilon_F^{\sigma}) > 0$ , it follows from the intermediary value theorem that there exists an  $\varepsilon_F^{\sigma, L}$  close to  $\varepsilon_F^{\sigma}$  so that  $\mathcal{N}^{\sigma, L}(\varepsilon_F^{\sigma, L}) = N$ . It is this  $\varepsilon_F^{\sigma, L}$  that we assume to be chosen for the rest of this paper.

**Lemma 5.10** (Convergence of the Fermi energy, the total energy, and the density) *It holds that*

- If  $f^1$  is any of the smearing functions in (5.6–5.6'''), there exists  $C \in \mathbb{R}_+$  and  $\eta > 0$  such that for all  $0 < \sigma \leq \sigma_0$  and all  $L \in \mathbb{N}^*$ ,

$$\left| \varepsilon_F^{\sigma} - \varepsilon_F^{\sigma, L} \right| + \left| E^{\sigma} - E^{\sigma, L} \right| + \left\| \rho^{\sigma} - \rho^{\sigma, L} \right\|_{H_{\text{per}}^{s+2}} \leq C \sigma^{-(d+1)} e^{-\eta \sigma L}. \quad (5.18)$$

- If  $f^1$  is a Gaussian-type smearing function (5.6'–5.6'''), there exists  $C \in \mathbb{R}_+$  and  $\eta > 0$  such that, for all  $0 < \sigma \leq \sigma_0$  and all  $L \in \mathbb{N}^*$  with  $\sigma^2 L \geq 4$ , it holds that

$$\left| \varepsilon_F^{\sigma} - \varepsilon_F^{\sigma, L} \right| + \left| E^{\sigma} - E^{\sigma, L} \right| + \left\| \rho^{\sigma} - \rho^{\sigma, L} \right\|_{H_{\text{per}}^{s+2}} \leq C \sigma^{-\frac{5d}{3}} L^{-\frac{d}{3}} e^{-\eta \sigma^{2/3} L^{4/3}}. \quad (5.19)$$

**Proof** The Fermi level is estimated as outlined above. For the energy (the proof is similar for the density), we have

$$\left| E^{\sigma} - E^{\sigma, L} \right| \leq \left| E^{\sigma}(\varepsilon_F^{\sigma}) - E^{\sigma}(\varepsilon_F^{\sigma, L}) \right| + \left| E^{\sigma}(\varepsilon_F^{\sigma, L}) - E^{\sigma, L}(\varepsilon_F^{\sigma, L}) \right|. \quad (5.20)$$

The second term of (5.20) is exactly in the scope of Lemma 5.8 when we consider the function

$$F_{\sigma}^E(\mathbf{k}) := \sum_{n \in \mathbb{N}^*} \varepsilon_{n\mathbf{k}} f^{\sigma}(\varepsilon_{n\mathbf{k}} - \varepsilon^{\sigma, L}) = \text{Tr}_{L_{\text{per}}^2} \left[ H_{\mathbf{k}} f^{\sigma} \left( H_{\mathbf{k}} - \varepsilon^{\sigma, L} \right) \right].$$

Following the lines of Lemma 5.9 together with Lemmas A.1–A.2, we see that the first term is exponentially (resp. superexponentially) small. The first term of (5.20) can be evaluated by noticing that

$$\left| E^\sigma(\varepsilon_F^\sigma) - E^\sigma(\varepsilon_F^{\sigma,L}) \right| \leq \left( \max_{\varepsilon \in [\varepsilon_F - \delta_0/2, \varepsilon_F + \delta_0/2]} |\partial_\varepsilon E^\sigma| \right) \left| \varepsilon_F^\sigma - \varepsilon_F^{\sigma,L} \right|.$$

The proof follows.  $\square$

## 5.4 Total error

We finally combine Lemmas 5.5, 5.6, 5.7 and 5.10 to obtain our final result.

**Theorem 5.11** *Consider a smearing method of order  $p \geq 1$ . Under Assumption 1 and 2, there exist  $C \in \mathbb{R}_+$  and  $\eta > 0$  such that, for all  $0 < \sigma \leq \sigma_0$  and all  $L \in \mathbb{N}^*$ , it holds that*

$$|\varepsilon_F^{\sigma,L} - \varepsilon_F| + |E^{\sigma,L} - E| + \|\rho^{\sigma,L} - \rho\|_{H_{\text{per}}^{s+2}} \leq C(\sigma^{p+1} + \sigma^{-(d+1)} e^{-\eta\sigma L}). \quad (5.21)$$

**Remark 5.12** (Choosing  $\sigma$  adaptively) In practice, only the parameter  $L$  is relevant when considering CPU time. The numerical parameter  $\sigma$  can be chosen freely with no extra numerical cost, and can be optimized with respect to  $L$ . For instance, the choice  $\sigma \propto L^{-1}$  has been recommended for practical calculations in [1], based on a heuristic argument. According to our previous theorem, this is not enough: the right-hand side of (5.21) does not tend to zero when  $L \rightarrow \infty$  and  $\sigma = C/L$ . Still, choosing the slightly different scaling  $\sigma \propto \log(L)L^{-1}$  leads to a decay of the error proportional to  $L^{-(p+1)}$ , up to log factors. Our results therefore broadly support those of [1].

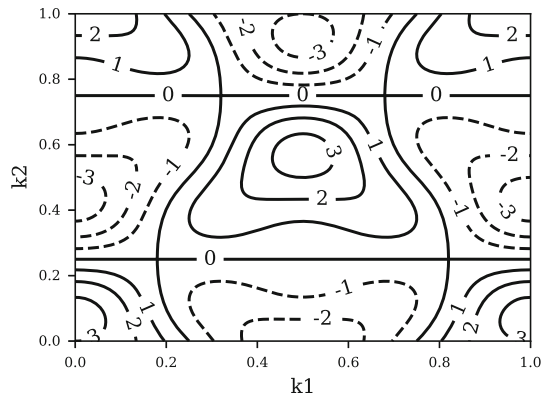
**Remark 5.13** (Superexponential convergence) The choice  $\sigma \propto \log(L)/L$  gives  $\sigma^2 L \propto \log(L)^2/L$ , which goes to 0 as  $L \rightarrow \infty$ . In particular, the condition  $\sigma^2 L \geq 4$  is not satisfied for large  $L$ . The super-exponential scaling result is therefore irrelevant for numerical purposes when we want to compute zero-temperature quantities.

## 6 Numerical results

The aim of this section is to present some numerical results on toy test cases to illustrate the convergence properties of the methods analyzed in the previous sections. We also present some examples where Assumption 1 or Assumption 2 are not valid. In these cases, we numerically observe a degraded convergence rate.

In Sect. 4, some results about interpolation methods are presented for different interpolation orders. In Sect. 6.2, numerical tests are performed using some of the smearing methods described in Sect. 5.

**Fig. 5** Isolines of the band structure (6.1) used in cases 1 and 2 (Fermi level  $\approx 1.7275$  and 0 respectively)



## 6.1 Interpolation methods

We consider two-dimensional toy test cases, with  $\mathcal{B} = (-1/2, 1/2)^2$ . For a given  $L \in \mathbb{N}^*$ , we consider a uniform  $L \times L$  discretization grid of  $\mathcal{B}$  as described in Sect. 4, and use B-splines of order 1 and 2 as interpolation operators. In all computations in this section the numerical quadratures are performed up to an error of  $10^{-6}$ , and so we display error curves only above that threshold.

**Case 1** Let us first consider a two-dimensional example where Assumptions 1 and 2 are satisfied (case 1 in the following). We consider one analytical band:

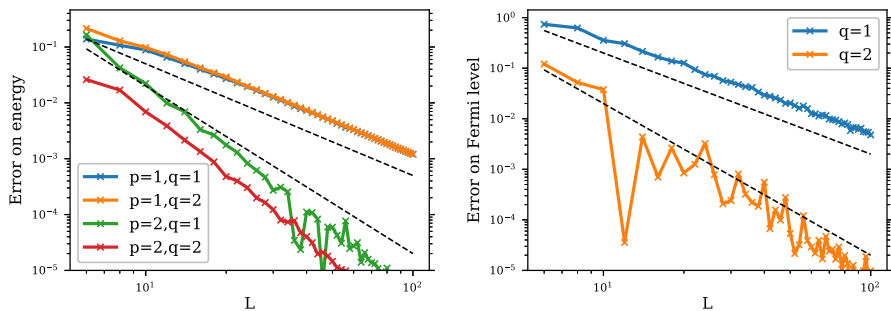
$$\forall \mathbf{k} := (k_1, k_2) \in \mathcal{B}, \quad \varepsilon_{1\mathbf{k}} = 3 \cos(2\pi k_1) \cos(2\pi k_2) + \sin(4\pi k_1) \cos(4\pi k_2) \quad (6.1)$$

represented on Fig. 5. The number of electrons per unit cell is chosen to be equal to  $N = 0.85$  so that the exact Fermi level is approximately  $\varepsilon_F \approx 1.7275$ . We plot the error on the energy  $|E - E^{L,p,q}|$  and the Fermi level  $|\varepsilon_F - \varepsilon_F^{L,q}|$  as a function of  $L$  on Fig. 6, for different interpolation schemes (we recall that  $p$  is the order used to interpolate the integrand, and  $q$  the order used to interpolate  $\varepsilon_{n\mathbf{k}}$  for the purposes of determining the integration region and Fermi level).

We see that the energy for the methods with  $p = 1$  converge converges as  $1/L^2$ , irrespective of  $q$ , and that the Fermi level converges as  $1/L^2$  for  $q = 1$ , as expected from our estimates. For the higher-order methods, the convergence is more erratic, and it is difficult to determine the order precisely. However, it can be seen that the  $p = 2, q = 2$  method only marginally improves the error on the energy compared to the  $p = 2, q = 1$  method, as expected from our estimates.

**Cases 2 and 3** We now consider two other two-dimensional examples where either Assumption 1 or Assumption 2 is violated.

In the test violating Assumption 2 (case 2 in the following), only one band is considered, with the same analytic expression (6.1) as in case 1. However, the number of electrons is now chosen to be equal to  $N = 0.5$  so that the exact Fermi level is



**Fig. 6** Case 1: errors made on the energy (left) and the Fermi level (right). The two dashed lines represent convergence rates in  $L^{-2}$  and  $L^{-3}$  respectively. The  $p = 1, q = 1$  and  $p = 1, q = 2$  curves are almost identical in the left plot

equal to  $\varepsilon_F = 0$ . There are saddle points of the function  $\varepsilon_{1\mathbf{k}}$  at level  $\varepsilon_F = 0$ , as can be seen in Fig. 5. This produces a van Hove singularity in the density of states at the Fermi level, violating Assumption 1. Note however that the singularity for a saddle point is relatively mild in 2D:  $\mathcal{D}(\varepsilon)$  diverges logarithmically near  $\varepsilon_F$ .

The test case violating Assumption 1 (case 3 in the following) is the standard tight-binding model of graphene [4], where band crossings occur on the Fermi surface.

We denote for all  $\mathbf{k} = (k_1, k_2) \in \mathcal{B}$

$$c_1(\mathbf{k}) = \frac{k_1 + k_2}{3}, \quad c_2(\mathbf{k}) = \frac{k_1 - 2k_2}{3}, \quad c_3(\mathbf{k}) = \frac{-2k_1 + k_2}{3}.$$

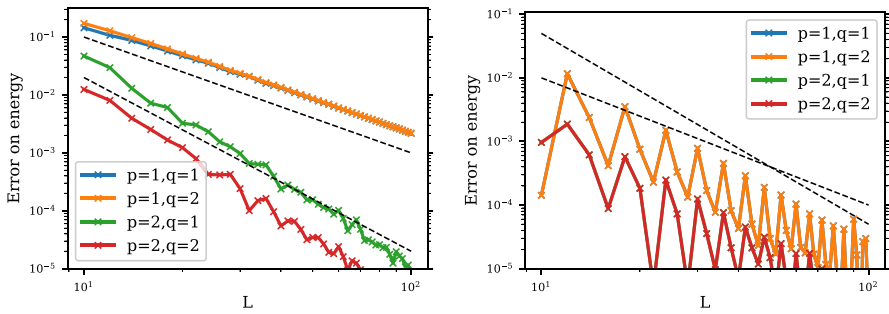
For all  $\mathbf{k} \in \mathcal{B}$ , we then define the  $2 \times 2$  Hermitian matrix  $H(\mathbf{k})$  as follows

$$H(\mathbf{k}) := \begin{pmatrix} 0 & \sum_{j=1}^3 e^{-2\pi i c_j(\mathbf{k})} \\ \sum_{j=1}^3 e^{2\pi i c_j(\mathbf{k})} & 0 \end{pmatrix}.$$

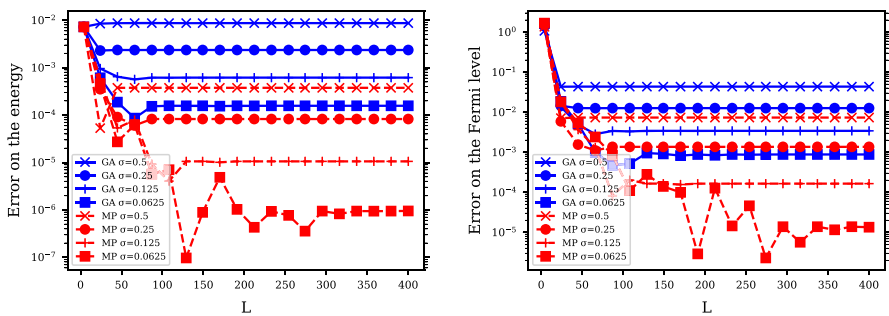
It can be checked that, while  $H$  is not periodic, for all  $\mathbf{K} \in \mathcal{R}^*$ ,  $H(\mathbf{k} + \mathbf{K})$  is unitarily equivalent to  $H(\mathbf{k})$ , so that the eigenvalues of  $H(\mathbf{k})$  are periodic on  $\mathcal{B}$ . In this test case, the number of electrons per unit cell is chosen to be equal to  $N = 1$  so that the exact Fermi level is  $\varepsilon_F = 0$ . The two bands  $\varepsilon_{1\mathbf{k}}$  and  $\varepsilon_{2\mathbf{k}}$  touch at the Fermi level at two non-equivalent points of the Brillouin zone (the Dirac points  $\mathbf{K}$  and  $\mathbf{K}'$ ), violating Assumption 2.

In Fig. 7 we plot the errors between the exact and approximate energies  $|E - E^{L,p,q}|$  as a function of  $L$  using the same interpolation schemes as described above.

The different errors for case 2 seem to decay at a rate similar to that in case 1. We attribute this to the relatively mild singularity of this case. Case 3 however displays a very different behavior, the results depending on the position of the Dirac points  $\mathbf{K}$  and  $\mathbf{K}'$  on the grid and therefore displaying an oscillating pattern.



**Fig. 7** Errors made on the energy for interpolation methods on cases 2 (left) and 3 (right). The two dashed lines represent convergence rates in  $L^{-2}$  and  $L^{-3}$  respectively. On the right plot, the results are independent of  $q$



**Fig. 8** Error on the energy  $|E - E^{\sigma,L}|$  (left) and Fermi level  $|\varepsilon_F - \varepsilon_F^{\sigma,L}|$  (right) for case 1 as a function of  $L$  for different smearing schemes and different values of  $\sigma$

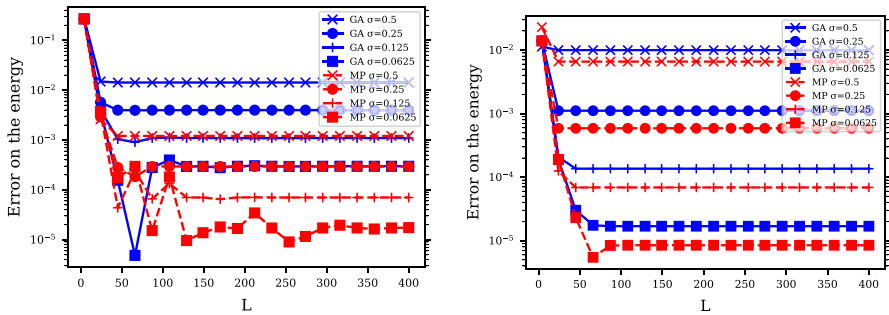
## 6.2 Smearing methods

The aim of this section is to present numerical results for some smearing methods presented and analyzed in Sect. 5, for the three two-dimensional test cases presented in Sect. 4. We consider here the Gaussian (denoted by GA, of order 1) and the first Methfessel–Paxton method (denoted by MP, of order 3) smearing schemes.

Let us begin with the first test case, namely the one-band model (6.1) for which Assumptions 1 and 2 are satisfied. The errors on the energy and on the Fermi level are plotted on Fig. 8 as a function of  $L$  and for different values of  $\sigma$ .

From this we obtain the following conclusions:

- The error as  $L \rightarrow \infty$  decreases as  $\sigma$  decreases. As  $\sigma$  is reduced by a factor of 2, the error  $|E^{\sigma,\infty} - E|$  is reduced by a factor of about 4 (for GA) and by a factor of about 8 (for MP1), suggesting that the asymptotic regime is not yet reached for the MP1 method (we would expect a factor 16 since the MP1 method is of order 3).
- As  $\sigma$  is reduced, so is the speed of convergence of  $E^{\sigma,L}$  to  $E^{\sigma,\infty}$ . The number of points  $L$  required to achieve convergence of  $E^{\sigma,L}$  to  $E^{\sigma,\infty}$  scales approximately linearly as  $1/\sigma$ , as expected from the  $e^{-\eta\sigma L}$  term in Theorem 5.11.
- On this toy example and in the parameter regimes considered here, the method that gives the lowest error for a given  $L$  seems to be that with lowest smearing and



**Fig. 9** Error on the energy  $|E - E^{\sigma,L}|$  for cases 2 (left) and 3 (right) as a function of  $L$  for different smearing schemes and different values of  $\sigma$

highest order, giving the impression that smearing is not advantageous. This is of course not the case in more realistic examples with lower values of  $L$ , where there is a non-zero optimal smearing.

We also present numerical results obtained on the two degenerate cases presented in Sect. 4 where Assumption 1 or Assumption 2 are violated. Errors  $|E^{\sigma,L} - E|$  are plotted as a function of  $L$  for cases 1 and 2 in Fig. 9. In contrast to before, we see that the Methfessel–Paxton scheme is not able to achieve a higher order than the Gaussian smearing. This is because of the lack of regularity of the density of states at the Fermi level in these two cases.

## 7 Conclusion

We have presented an *a priori* error analysis of quadrature rules and smearing methods for  $\mathbf{k}$ -point integration in the Brillouin zone. Our conclusions justify several non-obvious schemes, and give rigorous bounds allowing one to choose the smearing parameter optimally.

Our analysis is concerned with linear periodic Schrödinger operators; as such, a number of extensions are necessary to covers the framework of density functional theory, which is the main motivation underlying this work. The nonlinearity of the Kohn–Sham equations is expected to give rise to difficulties, which can probably be addressed using the tools developed in [5]. Another source of error not considered here is the space (or plane-wave) discretization used in numerical simulations of periodic Schrödinger operators. This has a strong impact on our estimates, which rely on the fact that the map  $\mathbb{R}^d \ni \mathbf{k} \mapsto (i + H_{\mathbf{k}})^{-1} \in \mathcal{B}(L^2_{\text{per}})$  is smooth and pseudo-periodic. This cannot be ensured at the discrete level for a standard Galerkin discretization, see Remark 2.2. We plan to explore all these issues in a forthcoming paper.

In this work, we only considered simple ground state properties. Several quantities of interest do not fit into this framework, such as response functions, which involve integrals over Fermi surfaces, derivatives with respect to  $\mathbf{k}$  of occupied Bloch states, or unoccupied Bloch states. The error estimates in this case are expected to be sub-

stantially different, in accordance to the common observation that these quantities converge slowly in practice as functions of  $L$ .

To establish our results, we have relied on Assumptions 1 and 2, which mathematically define what is a “simple” metal (at least for the properties we have considered). For such metals, the asymptotic behavior is universal, and the scaling laws only depend on the interpolation or smearing used. It would be interesting to explore what happens in the case of a semimetal, or when a van Hove singularity is present at (or close to) the Fermi surface.

**Acknowledgements** The authors are grateful to Gus Hart, Volker Blum and Nicola Marzari for interesting discussions. This work was supported in part by ARO MURI Award W911NF-14-1-0247. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant Agreement No. 810367).

## Appendix: Analytic properties of the integrand

The goal of this appendix is to study the analytic properties of the functions  $F_{\varepsilon,\sigma}^{\mathcal{N}}$ ,  $F_{\varepsilon,\sigma}^E$  and  $F_{\varepsilon,\sigma}^{\rho,W}$  defined respectively by

$$\begin{aligned} F_{\varepsilon,\sigma}^{\mathcal{N}}(\mathbf{k}) &:= \mathrm{Tr}_{L_{\mathrm{per}}^2} [f^\sigma (H_{\mathbf{k}} - \varepsilon)], \\ F_{\varepsilon,\sigma}^E(\mathbf{k}) &:= \mathrm{Tr}_{L_{\mathrm{per}}^2} [H_{\mathbf{k}} f^\sigma (H_{\mathbf{k}} - \varepsilon)], \\ F_{\varepsilon,\sigma}^{\rho,W}(\mathbf{k}) &:= \mathrm{Tr}_{L_{\mathrm{per}}^2} [W f^\sigma (H_{\mathbf{k}} - \varepsilon)]. \end{aligned}$$

We prove the following two results. The first one studies the analytic properties of these functions near the real line.

**Lemma A.1** (Analyticity near the real line) *Let  $f^1$  be any of the smearing functions (5.6)–(5.6'''). Then, there exist  $C \in \mathbb{R}_+$  and  $Y > 0$  such that, for all  $\varepsilon \in [\varepsilon_F - \delta_0, \varepsilon_F + \delta_0]$  and all  $0 < \sigma \leq \sigma_0$ , the maps  $F_{\varepsilon,\sigma}^X$  admits an analytic continuation on  $S_{\sigma Y}$ , and it holds that*

$$\sup_{\mathbf{z} \in S_{\sigma Y}} |F_{\varepsilon,\sigma}^{\mathcal{N}}(\mathbf{z})| + \sup_{\mathbf{z} \in S_{\sigma Y}} |F_{\varepsilon,\sigma}^E(\mathbf{z})| + \sup_{\mathbf{z} \in S_{\sigma Y}, \|W\|_{H_{\mathrm{per}}^{-(s+2)}=1}} |F_{\varepsilon,\sigma}^{\rho,W}(\mathbf{z})| \leq C \sigma^{-1}.$$

Our second result studies the analytic properties on the entire complex plane.

**Lemma A.2** (Analyticity on  $\mathbb{C}$ ) *Let  $f^1$  be one of the Gaussian-type smearing functions (5.6')–(5.6'''). Then, the maps  $F_{\varepsilon,\sigma}^X$  are entire, and there exists  $C \in \mathbb{R}_+$  and  $\eta > 0$  such that, for all  $Y \geq 1$ , all  $\varepsilon \in [\varepsilon_F - \delta_0, \varepsilon_F + \delta_0]$  and all  $0 < \sigma \leq \sigma_0$ ,*

$$\sup_{\mathbf{z} \in S_{\sigma Y}} |F_{\varepsilon,\sigma}^{\mathcal{N}}(\mathbf{z})| + \sup_{\mathbf{z} \in S_{\sigma Y}} |F_{\varepsilon,\sigma}^E(\mathbf{z})| + \sup_{\mathbf{z} \in S_{\sigma Y}, \|W\|_{H_{\mathrm{per}}^{-(s+2)}=1}} |F_{\varepsilon,\sigma}^{\rho,W}(\mathbf{z})| \leq C e^{\eta \frac{Y^4}{\sigma^2}}.$$

Let us first highlight the idea of the proofs of these lemmas. First, we will obtain bounds that only depend on  $\|V_{\mathrm{per}}\|_{L^\infty}$ , so that we can absorb  $\varepsilon \in (\varepsilon_F - \delta_0, \varepsilon_F + \delta_0)$  in

$V_{\text{per}}$ . Without loss of generality, we take  $\varepsilon = 0$  and drop the subscript  $\varepsilon$ . In addition, to shorten the presentation, we only do the proof for the energy. In the following,  $F_\sigma$  denotes  $F_{\varepsilon=0,\sigma}^E$ .

We wish to study the analytic properties of the function

$$F_\sigma(\mathbf{k}) = \text{Tr}_{L^2_{\text{per}}} \left[ H_{\mathbf{k}} f^1(H_{\mathbf{k}}/\sigma) \right],$$

where  $f^1$  is one of the smearing functions (5.6)–(5.6'''). Formally,  $H_{\mathbf{k}}$  admits an analytic continuation to the whole complex plane: when  $\mathbf{z} = \mathbf{k} + i\mathbf{y} \in \mathbb{C}^d$ , we set

$$H_{\mathbf{z}} = -\frac{1}{2}\Delta_{\mathbf{z}} + V_{\text{per}},$$

where

$$-\Delta_{\mathbf{z}} = (-i\nabla + \mathbf{z})^2 = (-i\nabla + \mathbf{k})^2 + 2i\mathbf{y} \cdot (-i\nabla + \mathbf{k}) - |\mathbf{y}|^2. \quad (\text{A.1})$$

The operator  $H_{\mathbf{z}}$  is not self-adjoint (it is not even a normal operator), so it is difficult to compute the analytical extension  $F_\sigma(\mathbf{z})$  of  $F_\sigma(\mathbf{k})$ . We will use a representation in terms of contour integrals. Formally, it holds that

$$\begin{aligned} \text{Tr}_{L^2_{\text{per}}} \left[ H_{\mathbf{z}} f^1(H_{\mathbf{z}}/\sigma) \right] &= \text{Tr}_{L^2_{\text{per}}} \left[ \frac{1}{2\pi i} \oint_{\mathcal{C}} \lambda f^1(\lambda/\sigma) \left( \frac{1}{\lambda - H_{\mathbf{z}}} \right) d\lambda \right] \\ &= \frac{1}{2\pi i} \oint_{\mathcal{C}} \lambda f^1(\lambda/\sigma) \text{Tr}_{L^2_{\text{per}}} \left[ \frac{1}{\lambda - H_{\mathbf{z}}} \right] d\lambda \end{aligned}$$

for some (infinite) contour  $\mathcal{C}$  enclosing the spectrum of  $H_{\mathbf{z}}$ . Unfortunately, we cannot commute the trace and the integral in the last line whenever the dimension  $d$  is greater than 1. The reason is that the operator  $(\lambda - H_{\mathbf{z}})^{-1}$  is not trace-class when  $d \geq 2$  (see (2.2)). Instead, we consider the contour integral<sup>1</sup>

$$G_\sigma(\mathbf{z}) = \frac{1}{2\pi i} \int_{\mathcal{C}} \left[ \lambda f^1(\lambda/\sigma) (\lambda + \Sigma)^2 \right] \text{Tr}_{L^2_{\text{per}}} \left( \frac{1}{(H_{\mathbf{z}} + \Sigma)^2} \frac{1}{\lambda - H_{\mathbf{z}}} \right) d\lambda, \quad (\text{A.2})$$

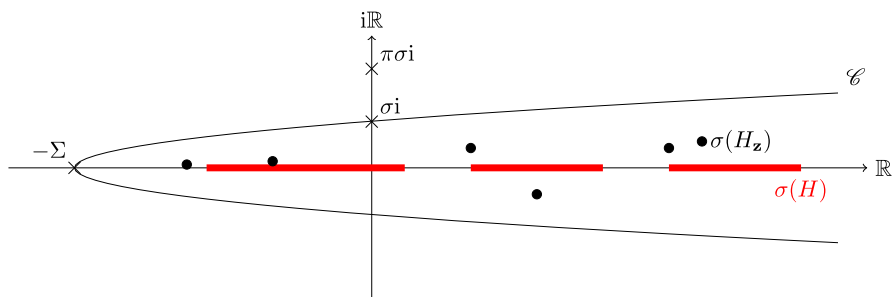
where  $\Sigma \in \mathbb{R}$  is a well-chosen shift, and prove that  $G_\sigma$  is an analytic continuation of  $F_\sigma$ .

We prove Lemmas A.1 and A.2 in the following two sections. In both cases we prove that there exists appropriate contours such that  $G_\sigma = F_\sigma$  using a perturbation argument. When  $\mathbf{z} = \mathbf{k} + i\mathbf{y} \in \mathbb{C}^3$  with  $|\mathbf{y}|$  small, we see  $H_{\mathbf{z}}$  as a perturbation of  $H_{\mathbf{k}}$ , while when  $\mathbf{y}$  is large, we see it as a perturbation of the free operator  $-\frac{1}{2}\Delta_{\mathbf{z}}$ .

<sup>1</sup> In the case of the density, we can take for instance

$$G_\sigma^W(\mathbf{z}) := \frac{1}{2\pi i} \int_{\mathcal{C}} \left[ f^1(\lambda/\sigma) (\lambda + \Sigma)^2 \right] \text{Tr}_{L^2_{\text{per}}} \left( \frac{1}{(H_{\mathbf{z}} + \Sigma)^{(s+4)/2}} W \frac{1}{(H_{\mathbf{z}} + \Sigma)^{(s+4)/2}} \frac{1}{\lambda - H_{\mathbf{z}}} \right) d\lambda,$$

with the integrand trace-class from (5.15).



**Fig. 10** Spectrum of the operator  $H_{\mathbf{z}}$  for  $\mathbf{z} = \mathbf{k} + i\mathbf{y}$ , and the contour  $\mathcal{C}$  that encloses it, while avoiding the poles of the Fermi–Dirac function at  $(2\mathbb{Z} + 1)\sigma i$

### Proof of Lemma A.1

We introduce  $\mathcal{C}$  the parabolic contour in the complex plane defined by

$$\mathcal{C} := \left\{ \lambda \in \mathbb{C}, |\operatorname{Im} \lambda|^2 = \sigma^2 \left( 1 + \frac{\operatorname{Re} \lambda}{\Sigma} \right) \right\}, \quad \text{where we set } \Sigma := \|V\|_{L^\infty} + 1. \quad (\text{A.3})$$

It is the (unique) parabola that passes through the points  $-\Sigma$  and  $\pm i\sigma$ . In particular, it does not encounter the poles of the Fermi–Dirac function at  $(2\mathbb{Z} + 1)\pi\sigma i$ . Let us prove that  $\mathcal{C}$  encloses the spectrum of  $H_{\mathbf{k}+i\mathbf{y}}$  for  $\mathbf{y}$  small enough (see Fig. 10).

**Lemma A.3** *There exist  $C \in \mathbb{R}^+$  and  $Y > 0$  such that, for all  $0 < \sigma \leq \sigma_0$ , all  $\mathbf{z} = \mathbf{k} + i\mathbf{y} \in S_{\sigma Y}$ , and all  $\lambda \in \mathcal{C}$ , we have the following estimates:*

$$\|(H_{\mathbf{z}} - \lambda)^{-1}\|_{\mathcal{B}} \leq C\sigma^{-1}, \quad (\text{A.4})$$

$$\|(H_{\mathbf{z}} + \Sigma)^{-1}\|_{\mathfrak{S}_2} \leq C, \quad (\text{A.5})$$

$$\operatorname{Tr}_{L^2_{\text{per}}} \left( \frac{1}{(H_{\mathbf{z}} + \Sigma)^2} \frac{1}{\lambda - H_{\mathbf{z}}} \right) \leq C\sigma^{-1}, \quad (\text{A.6})$$

$$\partial_{\mathbf{z}} \operatorname{Tr}_{L^2_{\text{per}}} \left( \frac{1}{(H_{\mathbf{z}} + \Sigma)^2} \frac{1}{\lambda - H_{\mathbf{z}}} \right) \leq C\sigma^{-1}. \quad (\text{A.7})$$

**Proof** From (A.1), it holds that

$$H_{\mathbf{z}} = \frac{1}{2}(-i\nabla + \mathbf{k})^2 + i\mathbf{y} \cdot (-i\nabla + \mathbf{k}) - \frac{1}{2}|\mathbf{y}|^2 + V_{\text{per}} = \left( H_{\mathbf{k}} - \frac{1}{2}|\mathbf{y}|^2 \right) + i\mathbf{y} \cdot (-i\nabla + \mathbf{k}).$$

For  $Y \leq 1$ , the spectrum of the self-adjoint operator  $H_{\mathbf{k}} - \frac{1}{2}|\mathbf{y}|^2$  is contained in  $[-\Sigma + \frac{1}{2}, +\infty)$ , which is disjoint from  $\mathcal{C}$ . In particular, for  $\lambda \in \mathcal{C}$ , we have

$$(\lambda - H_{\mathbf{z}}) = \left( \lambda - \left( H_{\mathbf{k}} - \frac{1}{2}|\mathbf{y}|^2 \right) \right) \left( 1 + i\mathbf{y} \cdot \left( \lambda - \left( H_{\mathbf{k}} - \frac{1}{2}|\mathbf{y}|^2 \right) \right)^{-1} (i\nabla + \mathbf{k}) \right). \quad (\text{A.8})$$

Let us evaluate the norm of the normal operator  $(\lambda - (H_{\mathbf{k}} - \frac{1}{2}|\mathbf{y}|^2))^{-1}$ . From (A.3), we have

$$\begin{aligned} \left\| \left( \lambda - (H_{\mathbf{k}} - \tfrac{1}{2}|\mathbf{y}|^2) \right)^{-1} \right\|_B &= \text{dist} \left( \lambda, \text{Spec}(H_{\mathbf{k}} - \tfrac{1}{2}|\mathbf{y}|^2) \right)^{-1} \\ &\leq \left[ (\text{Im}\lambda)^2 + (-\Sigma + \tfrac{1}{2} - \text{Re}\lambda)_+^2 \right]^{-1/2} \\ &= \left[ \sigma^2 \left( 1 + \frac{\text{Re}\lambda}{\Sigma} \right) + (-\Sigma + \tfrac{1}{2} - \text{Re}\lambda)_+^2 \right]^{-1/2} \leq C\sigma^{-1}, \end{aligned}$$

where  $x_+ := \max(x, 0)$  and where  $C \in \mathbb{R}_+$  is a constant that depends only on  $\Sigma$  and  $\sigma_0$ . The last inequality is obtained by optimizing over all  $\text{Re}\lambda \geq -\Sigma$ .

From the fact that  $H_{\mathbf{k}}$  is a bounded perturbation of  $\frac{1}{2}(-i\nabla + \mathbf{k})^2$ , and using similar calculations, we easily get that there exists  $C \in \mathbb{R}_+$  that depends only on  $\Sigma$  and  $\sigma_0$  such that

$$\|i\mathbf{y} \cdot (\lambda - (H_{\mathbf{k}} - \tfrac{1}{2}|\mathbf{y}|^2))^{-1} \cdot (-i\nabla + \mathbf{k})\|_B \leq C\sigma^{-1}|\mathbf{y}|. \quad (\text{A.9})$$

As a consequence, for  $Y \leq 1/(2C)$  and  $|\mathbf{y}| \leq \sigma Y$ , the operator on the right parenthesis of (A.8) is invertible, and its inverse is bounded in norm by 2. Inverting (A.8) leads to (A.4).

Inequality (A.5) is proved in a similar way (notice that the operator  $(\Sigma - (H_{\mathbf{k}} - \frac{1}{2}|\mathbf{y}|^2))^{-1}$  is Hilbert–Schmidt by (2.2)). Inequality (A.6) is a consequence of (A.4) and (A.5), together with the operator inequality  $|\text{Tr}(B^2 A)| \leq \|B\|_{\mathfrak{S}_2}^2 \|A\|_{\mathcal{B}}$ .

We finally prove (A.7). For all  $\mu$  in the resolvent set of  $H_{\mathbf{z}}$ , we have

$$\partial_{\mathbf{z}} \left( \frac{1}{H_{\mathbf{z}} - \mu} \right) = - \left( \frac{1}{H_{\mathbf{z}} - \mu} \right) \partial_{\mathbf{z}} H_{\mathbf{z}} \left( \frac{1}{H_{\mathbf{z}} - \mu} \right) = - \frac{1}{H_{\mathbf{z}} - \mu} (-i\nabla + \mathbf{z}) \frac{1}{H_{\mathbf{z}} - \mu}.$$

We then use similar arguments and the fact that the operator  $(-i\nabla + \mathbf{z})(H_{\mathbf{z}} - \lambda)^{-1}$  is bounded.  $\square$

We can now prove the analyticity of  $G_{\sigma}$  defined in (A.2).

**Lemma A.4** *There exist  $C \in \mathbb{R}_+$  and  $Y > 0$  such that, for all  $0 < \sigma \leq \sigma_0$ , the function  $G_{\sigma}$  is analytic on  $S_{\sigma Y}$ , and*

$$\sup_{\mathbf{z} \in S_{\sigma Y}} |G_{\sigma}(\mathbf{z})| \leq C\sigma^{-1}.$$

**Proof** From the previous Lemma A.3, we already see that  $G_{\sigma}$  is analytic. Let us prove the bound. It holds that

$$|G_{\sigma}(\mathbf{z})| \leq C\sigma^{-1} \int_{\mathcal{C}} |f^1(\sigma^{-1}\lambda)\lambda(\lambda + \Sigma)^2| |\mathrm{d}\lambda|.$$

To evaluate the last integral, we parametrize the contour  $\mathcal{C}$  with  $\lambda(t) := \Sigma(t^2 - 1) + i\sigma t$  for  $t \in \mathbb{R}$ , so that

$$|\lambda| = \left( \Sigma^2(t^2 - 1)^2 + \sigma^2 t^2 \right)^{1/2} \leq C(t^2 + 1) \quad \text{and} \quad |d\lambda| = \sqrt{4t^2 \Sigma^2 + \sigma^2} dt \leq C(|t| + 1) dt, \quad (\text{A.10})$$

for some  $C \in \mathbb{R}_+$  that depends only on  $\Sigma$  and  $\sigma_0$ . We obtain

$$|G_\sigma(\mathbf{z})| \leq C\sigma^{-1} \left( \int_{\mathbb{R}} \left| f^1 \left( \sigma^{-1} \Sigma(t^2 - 1) + it \right) \right| (1 + |t|^7) dt \right).$$

Let us prove that the last integral is uniformly bounded for  $0 < \sigma \leq \sigma_0$ . We prove this result in full details for the Fermi–Dirac smearing  $f^1(x) = (1 + e^x)^{-1}$ , the other cases being similar. We split the integral in the regions  $|t| \leq \pi/2$  and  $|t| > \pi/2$ . For  $|t| \leq \pi/2$ , it holds that  $\cos t \geq 0$ , so that

$$\begin{aligned} \left| f^1 \left( \sigma^{-1} \Sigma(t^2 - 1) + it \right) \right| &= \left| 1 + e^{\sigma^{-1} \Sigma(t^2 - 1) + it} \right|^{-1} \leq \left| \operatorname{Re} \left( 1 + e^{\sigma^{-1} \Sigma(t^2 - 1) + it} \right) \right|^{-1} \\ &= \left| 1 + e^{\sigma^{-1} \Sigma(t^2 - 1)} \cos t \right|^{-1} \leq 1. \end{aligned}$$

We deduce that the integral over  $|t| \leq \pi/2$  is uniformly bounded for  $0 < \sigma \leq \sigma_0$ . For  $t \geq \pi/2 > 1$ , it holds that  $(t^2 - 1) > 0$ , so that

$$\left| 1 + e^{\sigma^{-1} \Sigma(t^2 - 1) + it} \right|^{-1} \leq \left| e^{\sigma^{-1} \Sigma(t^2 - 1)} - 1 \right|^{-1} \leq \left| e^{\sigma_0^{-1} \Sigma(t^2 - 1)} - 1 \right|^{-1},$$

where the right-hand side no longer depends on  $0 < \sigma \leq \sigma_0$ . Finally, we check that the integral

$$\int_{|t| \geq \frac{\pi}{2}} \left| e^{\sigma_0^{-1} \Sigma(t^2 - 1)} - 1 \right|^{-1} (1 + |t|^7) dt$$

is absolutely convergent, and independent of  $0 < \sigma \leq \sigma_0$ . This ends the proof of Lemma A.4.  $\square$

We now prove, as claimed, that  $G_\sigma$  is an analytic extension of  $F_\sigma$ .

**Lemma A.5** *For all  $0 < \sigma \leq \sigma_0$  and all  $\mathbf{k} \in \mathbb{R}^d$ , it holds that  $G_\sigma(\mathbf{k}) = F_\sigma(\mathbf{k})$ .*

**Proof** We first approximate  $H_{\mathbf{k}}$  by a finite-rank operator, then apply the Cauchy residual formula, and pass to the limit. Recall that  $H_{\mathbf{k}}$  is a self-adjoint operator with spectral decomposition (2.1). For  $Q \in \mathbb{N}^*$ , we introduce the truncated operator

$$H_{\mathbf{k}}^Q := \sum_{n=1}^Q \varepsilon_{n\mathbf{k}} |u_{n\mathbf{k}}\rangle \langle u_{n\mathbf{k}}|.$$

We have

$$|G_\sigma(\mathbf{k}) - F_\sigma(\mathbf{k})| \leq \left| G_\sigma(\mathbf{k}) - \operatorname{Tr}_{L^2_{\text{per}}} \left( H_{\mathbf{k}}^Q f^\sigma(H_{\mathbf{k}}^Q) \right) \right| \quad (\text{A.11})$$

$$+ \left| \text{Tr}_{L^2_{\text{per}}} \left( H_{\mathbf{k}}^Q f^\sigma(H_{\mathbf{k}}^Q) \right) - \text{Tr}_{L^2_{\text{per}}} \left( H_{\mathbf{k}} f^\sigma(H_{\mathbf{k}}) \right) \right|. \quad (\text{A.12})$$

We first focus on (A.12). Using the asymptotic (2.2) and the decay properties of  $f^\sigma$ , it holds that

$$\text{Tr}_{L^2_{\text{per}}} \left( H_{\mathbf{k}} f^\sigma(H_{\mathbf{k}}) \right) - \text{Tr}_{L^2_{\text{per}}} \left( H_{\mathbf{k}}^Q f^\sigma(H_{\mathbf{k}}^Q) \right) = \sum_{n \geq Q} \varepsilon_{n\mathbf{k}} f^\sigma(\varepsilon_{n\mathbf{k}}) \xrightarrow{Q \rightarrow \infty} 0. \quad (\text{A.13})$$

We now focus on the right-hand side of (A.11). For  $M \geq \varepsilon_{Q\mathbf{k}} + 1$  we denote by  $\mathcal{C}_M$  the positively oriented *closed* contour defined by

$$\mathcal{C}_M := \{\lambda \in \mathcal{C}, \text{Re} \lambda \leq M\} \cup \left[ M + i\sigma \left( 1 + \frac{M}{\Sigma} \right), M - i\sigma \left( 1 + \frac{M}{\Sigma} \right) \right].$$

The contour  $\mathcal{C}_M$  is obtained by truncating the parabola  $\mathcal{C}$  to the region  $\text{Re} \lambda \leq M$  and closing the contour by a segment. For all  $M \geq \varepsilon_{Q\mathbf{k}} + 1$ , this contour encloses the spectrum of  $H_{\mathbf{k}}^Q$ , so that, from the Cauchy residual formula,

$$\text{Tr}_{L^2_{\text{per}}} \left( H_{\mathbf{k}}^Q f^\sigma(H_{\mathbf{k}}^Q) \right) = \frac{1}{2\pi i} \oint_{\mathcal{C}_M} \left[ f^\sigma(\lambda) \lambda (\lambda + \Sigma)^2 \right] \text{Tr}_{L^2_{\text{per}}} \left( \frac{1}{(H_{\mathbf{k}}^Q + \Sigma)^2} \frac{1}{\lambda - H_{\mathbf{k}}^Q} \right) d\lambda.$$

As  $M \rightarrow \infty$ , and using the same arguments as in the proofs of Lemmas A.3 and A.4, we see that the right-hand side converges to the integral over the full contour  $\mathcal{C}$ . Moreover, we have the point-wise convergence

$$\begin{aligned} \forall \lambda \in \mathcal{C}, \quad \lim_{Q \rightarrow \infty} \text{Tr}_{L^2_{\text{per}}} \left( \frac{1}{(H_{\mathbf{k}}^Q + \Sigma)^2} \frac{1}{\lambda - H_{\mathbf{k}}^Q} \right) &= \lim_{Q \rightarrow \infty} \sum_{n=1}^Q \frac{1}{(\varepsilon_{n\mathbf{k}} + \Sigma)^2 (\lambda - \varepsilon_{n\mathbf{k}})} \\ &= \sum_{n=1}^{\infty} \frac{1}{(\varepsilon_{n\mathbf{k}} + \Sigma)^2 (\lambda - \varepsilon_{n\mathbf{k}})} = \text{Tr}_{L^2_{\text{per}}} \left( \frac{1}{(H_{\mathbf{k}} + \Sigma)^2} \frac{1}{\lambda - H_{\mathbf{k}}} \right). \end{aligned}$$

We conclude from the dominated convergence theorem that  $\text{Tr}_{L^2_{\text{per}}} \left( H_{\mathbf{k}}^Q f^\sigma(H_{\mathbf{k}}^Q) \right)$  converges to  $G_\sigma(\mathbf{k})$  as  $Q \rightarrow \infty$ . The proof of Lemma A.5 follows.  $\square$

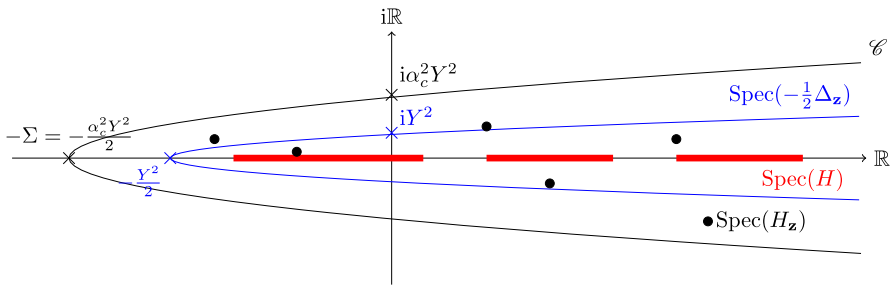
Combining Lemma A.5 together with Lemma A.4 ends the proof of Lemma A.1.

## Proof of Lemma A.2

We now focus on Gaussian-type smearing functions. The idea of the proof is similar to previously. We need however to re-define  $G_\sigma$  for large  $|\mathbf{y}|$  (i.e. choose an appropriate contour). In the sequel, we fix  $Y \geq 1$  and provide estimates uniform in  $\mathbf{z} \in S_Y$ .

Looking at the operator  $-\frac{1}{2}\Delta_{\mathbf{z}}$  in Fourier basis, we see that its spectrum is

$$\sigma \left( -\frac{1}{2}\Delta_{\mathbf{z}} \right) := \left\{ \frac{1}{2}(\mathbf{K} + \mathbf{z})^2 \right\}_{\mathbf{K} \in \mathcal{R}^*} \left\{ \frac{1}{2}(|\mathbf{K} + \mathbf{k}|^2 - |\mathbf{y}|^2) + i\mathbf{y} \cdot (\mathbf{K} + \mathbf{k}) \right\}_{\mathbf{K} \in \mathcal{R}^*},$$



**Fig. 11** The spectra of the operator  $-\frac{1}{2}\Delta_{\mathbf{z}}$  and  $H_{\mathbf{z}}$ , and the contour  $\mathcal{C}$  that encloses them

hence is contained in the parabolic set  $(\text{Im}\lambda)^2 \leq Y^2(2\text{Re}\lambda + Y^2)$ . In the sequel, we take an parabolic integration contour that encloses this region. More specifically, for  $\alpha > 1$  and  $Y \geq 1$ , we introduce the dilated contour

$$\mathcal{C}_{\alpha,Y} := \left\{ \lambda \in \mathbb{C}, \text{Im}\lambda^2 = \alpha^2 Y^2 (2\text{Re}\lambda + \alpha^2 Y^2) \right\}.$$

As  $\alpha$  increases, the distance between  $\mathcal{C}_{\alpha,Y}$  and  $\sigma(-\frac{1}{2}\Delta_{\mathbf{z}})$  goes to  $+\infty$ . Since the operator  $-\frac{1}{2}\Delta_{\mathbf{z}}$  is normal, this implies that there exists  $\alpha_c \geq 1$  such that

$$\forall Y \geq 1, \quad \forall \mathbf{z} \in S_Y, \quad \forall \lambda \in \mathcal{C}_{\alpha_c,Y}, \quad \left\| (\lambda + \frac{1}{2}\Delta_{\mathbf{z}})^{-1} \right\|_{\mathcal{B}} \leq (2\|V\|_{L^\infty})^{-1}.$$

The contour  $\mathcal{C} := \mathcal{C}_{\alpha_c,Y}$  is our integration contour for  $\mathbf{z} \in S_Y$  (see Fig. 11).

For  $\lambda \in \mathcal{C}$ , it holds that (compare with (A.8))

$$(\lambda - H_{\mathbf{z}}) = (\lambda + \frac{1}{2}\Delta_{\mathbf{z}} - V) = (\lambda + \frac{1}{2}\Delta_{\mathbf{z}}) \left( 1 - (\lambda + \frac{1}{2}\Delta_{\mathbf{z}})^{-1} V \right).$$

As in Lemma A.3 we deduce the following inequalities. We do not repeat the proof, as it is similar. We denote by  $\Sigma := \frac{\alpha_c^2 Y^2}{2}$  for the sake of clarity.

**Lemma A.6** *There exists  $C \in \mathbb{R}_+$  such that, for all  $Y \geq 1$ , all  $\mathbf{z} \in S_Y$ , and all  $\lambda \in \mathcal{C}$ , it holds that*

$$\begin{aligned} \| (H_{\mathbf{z}} - \lambda)^{-1} \|_{\mathcal{B}} &\leq C, \quad \| (H_{\mathbf{z}} + \Sigma)^{-1} \|_{\mathfrak{S}_2} \leq C, \\ \text{Tr}_{L^2_{\text{per}}} \left( \frac{1}{(H_{\mathbf{z}} + \Sigma)^2} \frac{1}{\lambda - H_{\mathbf{z}}} \right) &\leq C, \quad \text{and} \quad \partial_{\mathbf{z}} \text{Tr}_{L^2_{\text{per}}} \left( \frac{1}{(H_{\mathbf{z}} + \Sigma)^2} \frac{1}{\lambda - H_{\mathbf{z}}} \right) \leq C. \end{aligned}$$

We now define, for  $0 < \sigma \leq \sigma_0$  and  $Y \geq 1$ , the function defined on  $S_Y$  by

$$G_{\sigma,Y}(\mathbf{z}) := \frac{1}{2\pi i} \int_{\mathcal{C}} \left[ \lambda f^1(\lambda/\sigma)(\lambda + \Sigma)^2 \right] \text{Tr}_{L^2_{\text{per}}} \left( \frac{1}{(H_{\mathbf{z}} + \Sigma)^2} \frac{1}{\lambda - H_{\mathbf{z}}} \right) d\lambda. \quad (\text{A.14})$$

Before stating a bound on  $G_{\sigma,Y}$ , we need the following technical lemma on the growth of  $f^1$ :

**Lemma A.7** (Growth of  $f^1$  in the complex plane) *When  $f^1$  is one of the Gaussian-type smearing functions (5.6''–5.6'''), then  $f^1$  is entire, and there exists  $C \in \mathbb{R}_+$  and  $q \geq 0$  such that*

$$\forall x, y \in \mathbb{R}, \quad \left| f^1(x + iy) \right| \leq \begin{cases} C(1 + (x^2 + y^2)^q) \left( e^{y^2 - x^2} \right) & \text{if } x \geq 0, \\ C(1 + (x^2 + y^2)^q) \left( 1 + e^{y^2 - x^2} \right) & \text{if } x < 0. \end{cases} \quad (\text{A.15})$$

**Proof** We first handle the case when  $f^1$  is the Gaussian smearing function (in which case we can choose  $q = 0$ ). We have

$$f^1(x) = \frac{1}{2} (1 - \operatorname{erf}(x)) = \frac{1}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt.$$

First recall that for  $x \in \mathbb{R}$ , we have  $0 < f^1(x) < 1$ . Moreover, for  $x \geq 1$ , it holds that

$$0 \leq f^1(x) \leq \frac{1}{\sqrt{\pi}} \int_x^\infty t e^{-t^2} dt = \frac{1}{2\sqrt{\pi}} e^{-x^2}.$$

This already proves (A.15) for  $y = 0$ . Then, we notice that an analytic continuation of  $f^1$  is given by

$$f^1(x + iy) = \frac{1}{\sqrt{\pi}} \left( \int_x^\infty e^{-t^2} dt + i \int_0^y e^{-(x+it)^2} dt \right) = f^1(x) + i \frac{e^{-x^2}}{\sqrt{\pi}} \int_0^y e^{-2ixt} e^{t^2} dt.$$

Lemma A.7 then follows from the inequalities

$$\left| f^1(x + iy) \right| \leq \left| f^1(x) \right| + \frac{e^{-x^2}}{\sqrt{\pi}} \int_{[0,y]} e^{t^2} dt,$$

together with the fact that

$$\int_{[0,y]} e^{t^2} dt = \int_{[0,1]} e^{t^2} dt + \int_{[1,y]} e^{t^2} dt \leq \int_{[0,1]} e^{t^2} dt + \int_{[1,y]} t e^{t^2} dt \leq C e^{y^2}.$$

The case of the Methfessel–Paxton and cold smearing schemes follows immediately by noting that they differ from the Gaussian smearing function by terms of the form  $x^n e^{-x^2}$ , and the fact that

$$\left| z^n e^{-z^2} \right| = (x^2 + y^2)^{n/2} e^{y^2 - x^2}.$$

□

We are now in position to prove estimates on  $G_{\sigma,Y}$  defined on (A.14).

**Lemma A.8** *There exists  $C \in \mathbb{R}_+$  and  $\eta > 0$  such that, for all  $0 < \sigma \leq \sigma_0$  and all  $Y \geq 1$ , the function  $G_{\sigma,Y}$  is analytic on  $S_Y$ , and*

$$\sup_{\mathbf{z} \in S_Y} |G_{\sigma,Y}(\mathbf{z})| \leq C e^{\eta \frac{Y^4}{\sigma^2}}.$$

**Proof** For the sake of clarity, we do the proof when  $f^1$  is the Gaussian smearing function (i.e.  $q = 0$ ). From Lemma A.6, it holds that

$$|G_{\sigma,Y}(\mathbf{z})| \leq C \int_{\mathcal{C}} |\lambda f^1(\lambda/\sigma)(\lambda + \Sigma)^2| |\mathrm{d}\lambda|.$$

We parametrize the contour  $\mathcal{C}$  with  $\lambda(t) := \frac{\alpha_c^2 Y^2}{2} ((t^2 - 1) + i2t)$ , so that (compare with (A.10))

$$|\lambda| \leq CY^2(t^2 + 1) \quad \text{and} \quad |\mathrm{d}\lambda| \leq CY^2(|t| + 1)dt.$$

Then,

$$|G_{\sigma,Y}(\mathbf{z})| \leq CY^8 \int_{\mathbb{R}} \left| f^1 \left( \frac{\alpha_c^2 Y^2}{2\sigma} (t^2 - 1 + i2t) \right) \right| (1 + |t|^7) dt.$$

We split this integral in regions where  $|t| \geq 1$ , and  $|t| \leq 1$ . When  $|t| \leq 1$ , it holds that  $\operatorname{Re} \lambda(t) \leq 0$ . Together with the second inequality of Lemma A.7 and the inequalities

$$\forall -1 \leq t \leq 1, \quad (1 + |t|^7) \leq 2, \quad \text{and} \quad 4t^2 - (t^2 - 1)^2 \leq 4,$$

we get

$$\begin{aligned} \int_{[-1,1]} \left| f^1 \left( \frac{\alpha_c^2 Y^2}{2\sigma} (t^2 - 1 + i2t) \right) \right| (1 + |t|^7) dt &\leq 2C \int_{[-1,1]} \left[ 1 + \exp \left( \frac{\alpha_c^4 Y^4}{4\sigma^2} (4t^2 - (t^2 - 1)^2) \right) \right] dt \\ &\leq 4C \left[ 1 + e^{\frac{\alpha_c^4 Y^4}{\sigma^2}} \right] \leq 8C e^{\frac{\alpha_c^4 Y^4}{\sigma^2}}. \end{aligned}$$

For  $|t| \geq 1$ , we use the first inequality of Lemma A.7, and obtain

$$\begin{aligned} \int_{[-1,1]^c} \left| f^1 \left( \frac{\alpha_c^2 Y^2}{2\sigma} (t^2 - 1 + i2t) \right) \right| (1 + |t|^7) dt &\leq 2C \int_1^\infty \exp \left( \frac{\alpha_c^4 Y^4}{4\sigma^2} (4t^2 - (t^2 - 1)^2) \right) \\ &\quad (1 + |t|^7) dt. \end{aligned}$$

When  $t \geq 1$ , we have the inequalities

$$(1 + |t|^7) \leq 2|t|^7 \quad \text{while} \quad 6t^2 \leq \frac{1}{2}(t^4 + 6^2) \quad \text{so that} \quad 4t^2 - (t^2 - 1)^2 \leq -\frac{t^4}{2} + 17.$$

As a result,

$$\int_1^\infty \exp \left( \frac{\alpha_c^4 Y^4}{4\sigma^2} (4t^2 - (t^2 - 1)^2) \right) (1 + |t|^7) dt \leq 2 \exp \left( \frac{17\alpha_c^4 Y^4}{4\sigma^2} \right) \int_0^\infty \exp \left( \frac{-\alpha_c^4 Y^4}{8\sigma^2} t^4 \right) t^7 dt.$$

We finally perform the change of variable  $u = \frac{\alpha_c Y}{8^{1/4} \sqrt{\sigma}} t$  and get

$$\int_0^\infty \exp\left(\frac{-\alpha_c^4 Y^4}{8\sigma^2} t^4\right) t^7 dt = \left(\frac{8^{1/4} \sqrt{\sigma}}{\alpha_c Y}\right)^8 \int_0^\infty e^{-u^4} u^7 du,$$

where the right-hand side is uniformly bounded for  $0 < \sigma \leq \sigma_0$  and  $Y \geq 1$ . Combining all the inequalities concludes the proof of Lemma A.8  $\square$

Lemma A.2 is a consequence of the Lemma A.8 and the following one, whose proof is similar to the one of Lemma A.5.

**Lemma A.9** *For all  $0 < \sigma \leq \sigma_0$ , all  $Y \geq 1$ , and all  $\mathbf{k} \in \mathbb{R}^d$ , it holds that  $G_{\sigma,Y}(\mathbf{k}) = F_\sigma(\mathbf{k})$ .*

## References

1. Björkman, T., Granöas, O.: Adaptive smearing for Brillouin zone integration. *Int. J. Quantum Chem.* **111**(5), 1025–1030 (2011)
2. Blöchl, P.E., Jepsen, O., Andersen, O.K.: Improved tetrahedron method for Brillouinzone integrations. *Phys. Rev. B* **49**(23), 16223–16233 (1994)
3. Boon, M.H., Methfessel, M.S., Mueller, F.M.: Singular integrals over the Brillouin zone: the analytic-quadratic method for the density of states. *J. Phys. C* **19**(27), 5337 (1986)
4. Castro Neto, A.H., et al.: The electronic properties of graphene. *Rev. Mod. Phys.* **81**(1), 109 (2009)
5. Cancés, É., Chakir, R., Maday, Y.: Numerical analysis of the planewave discretization of some orbital-free and Kohn-Sham models. *ESAIM M2AN* **46**(2), 341–388 (2012)
6. De Vita, A., Gillan, M.J.: The ab initio calculation of defect energetics in aluminium. *J. Phys. Condens. Matter* **3**(33), 6225 (1991)
7. Dirac, P.A.M.: On the theory of quantum mechanics. *Proc. R. Soc. Lond. A* **112**, 661–677 (1926)
8. Federer, H.: Curvature measures. *Trans. Am. Math. Soc.* **93**(3), 418–491 (1959)
9. Fermi, E.: Sulla quantizzazione del gas perfetto monoatomico. *Rend. Lincei* **3**, 145–149 (1926)
10. Fefferman, Ch., Weinstein, M.: Honeycomb lattice potentials and Dirac points. *J. Am. Math. Soc.* **25**(4), 1169–1220 (2012)
11. Gontier, D., Lahbabi, S.: Convergence rates of supercell calculations in the reduced Hartree–Fock model. *ESAIM M2AN* **50**(5), 1403–1424 (2016)
12. Henk, J.: Integration over two-dimensional Brillouin zones by adaptive mesh refinement. *Phys. Rev. B* **64**(3), 035412 (2001)
13. Kawamura, M., Gohda, Y., Tsuneyuki, Sh: Improved tetrahedron method for the Brillouin-zone integration applicable to response functions. *Phys. Rev. B* **89**(9), 094515 (2014)
14. Kuchment, P.: An overview of periodic elliptic operators. *Bull. Am. Math. Soc.* **53**(3), 343–414 (2016)
15. Marzari, N., et al.: Thermal contraction and disordering of the Al (110) surface. *Phys. Rev. Lett.* **82**(16), 3296 (1999)
16. Marzari, N.: Ab initio molecular dynamics for metallic systems. PhD thesis (1996)
17. Methfessel, M.S., Boon, M.H., Mueller, F.M.: Analytic-quadratic method of calculating the density of states. *J. Phys. C* **16**(27), L949 (1983)
18. Mermin, N.D.: Thermal properties of the inhomogeneous electron gas. *Phys. Rev.* **137**(5A), A1441 (1965)
19. Morgan, W.S., et al.: Efficiency of generalized regular k-point grids. [arXiv:1804.04741](https://arxiv.org/abs/1804.04741) (2018)
20. Monkhorst, H.J., Pack, J.D.: Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**(12), 5188 (1976)
21. Methfessel, M., Paxton, A.T.: High-precision sampling for Brillouin-zone integration in metals. *Phys. Rev. B* **40**(6), 3616 (1989)
22. Osher, S., Fedkiw, R.P.: Level set methods: an overview and some recent results. *J. Comput. Phys.* **169**(2), 463–502 (2001)

23. Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: algorithms based on Hamilton–Jacobi formulations. *J. Comput. Phys.* **79**(1), 12–49 (1988)
24. Pickard, C.J., Payne, M.C.: Extrapolative approaches to Brillouin-zone integration. *Phys. Rev. B* **59**(7), 4685 (1999)
25. Reed, M., Simon, B.: *Methods of Modern Mathematical Physics. In: Analysis of Operators*, vol. IV. Academic Press (1978)
26. Sethian, J.A.: *Level Set Methods and Fast Marching Methods*. Cambridge Monographs on Applied and Computational Mathematics, 2nd edn. Cambridge University Press, Cambridge (1999)
27. Suryanarayana, Ph: On spectral quadrature for linear-scaling density functional theory. *Chem. Phys. Lett.* **584**, 182–187 (2013)
28. Trefethen, L.N., Weideman, J.A.C.: The exponentially convergent trapezoidal rule. *SIAM Rev.* **56**(3), 385–458 (2014)
29. von Neumann, J., Wigner, E.P.: Über das Verhalten von Eigenwerten bei adiabatischen Prozessen. In: *The Collected Works of Eugene Paul Wigner: Part A: The Scientific Papers*, pp. 294–297. Springer, Berlin (1993)
30. Zaharioudakis, D.: Quadratic and cubic tetrahedron methods for Brillouin zone integration. *Comput. Phys. Commun.* **167**(2), 85–89 (2005)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.