<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *SIAM Journal on Scientific Computing*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

Rubensson, E H. (2012)

Controlling errors in recursive Fermi–Dirac operator expansions with applications in electronic structure theory

SIAM Journal on Scientific Computing, 34: B1-B23

<https://doi.org/10.1137/11083352X>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-167975>

CONTROLLING ERRORS IN RECURSIVE FERMI-DIRAC OPERATOR EXPANSIONS WITH APPLICATIONS IN ELECTRONIC STRUCTURE THEORY

EMANUEL H. RUBENSSON*

Abstract. Recursive Fermi–Dirac operator expansion is an efficient way to compute one-electron density matrices in electronic structure theory. The convergence is rapid and depends only weakly on the conditioning of the problem and, for many systems, the computational cost increases only linearly with system size. In this article, errors introduced when evaluating the recursive expansion are analyzed and schemes to control the forward error are proposed. The error has previously been analyzed for explicit schemes working at zero electronic temperature [J. Chem. Phys. 128 (2008), 074106]. Here, implicit schemes [Phys. Rev. B, 68 (2003), 233104] working at zero or finite temperature are treated. The proposed schemes for error control are demonstrated by tight-binding as well as density functional theory electronic structure calculations on several test systems. Condition numbers for the problem of computing the density matrix are derived, giving quantitative insight into under what circumstances a temperature dependent formulation results in better conditioning. It is shown that for the considered recursive expansions, the number of matrix-matrix multiplications needed to compute the density matrix increases only with the squared logarithm of the condition number of the problem.

1. Introduction. Calculation of the one-electron density matrix D for a given effective Hamiltonian matrix F is a key operation in electronic structure calculations using for example Hartree–Fock [65] or Kohn–Sham density functional theory [32]. The elements of F are given by

$$F_{ij} = \int \phi_i(r) \mathcal{H} \phi_j(r) dr \quad (1.1)$$

where $\{\phi_i\}_{i=1}^N$ is a set of basis functions and $\mathcal{H}$ is the effective Hamiltonian operator. The density matrix is given by the Fermi–Dirac operator: $D = f_{\text{FD}}(F)$,

$$f_{\text{FD}}(x) = \frac{1}{e^{\beta(x-\mu)} + 1}, \quad (1.2)$$

where μ is the chemical potential also known as the Fermi level, and $\beta = 1/(k_B T)$ with k_B being the Boltzmann constant and T the electronic temperature. Let $\{\lambda_i\}$ be the eigenvalues of F . Application of the scalar Fermi–Dirac function to the matrix F is defined by

$$D = f_{\text{FD}}(F) = V f_{\text{FD}}(\Lambda) V^T \quad (1.3)$$

where

$$F = V \Lambda V^T \quad (1.4)$$

is an eigendecomposition of F and $f_{\text{FD}}(\Lambda)$ is a diagonal matrix with nonzero entries $f_{\text{FD}}(\lambda_1), f_{\text{FD}}(\lambda_2), \dots, f_{\text{FD}}(\lambda_N)$. To actually carry out an eigendecomposition to compute the density matrix is often inefficient. The computational cost scales cubically with system size and optimization for modern heterogeneous computer architectures is difficult. Also, standard diagonalization routines typically spend a lot of effort

*Division of Scientific Computing, Department of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden

computing accurate individual eigenvectors; as noted by Bekas, Kokiopoulou, and Saad, such efforts do not necessarily result in a more accurate density matrix [4]. Much research has therefore focused on the development of alternative methods for the construction of the density matrix [11, 20, 61].

Alternative methods include various kinds of minimization schemes [9, 24, 25, 33, 37, 48, 67] and schemes where an expansion of the Fermi–Dirac operator is built up either serially by a Chebyshev series [1, 21, 22] or recursively by density matrix purification [38, 42, 43, 46, 49] or sign matrix methods [8, 41]. The computational kernel of these methods is matrix-matrix multiplication. Using sparse matrix libraries [10, 14, 27, 55, 62] together with ways to bring about matrix sparsity [54] these methods can for many systems deliver the density matrix at a computational cost that increases only linearly with system size.

Linear scaling is possible provided that the average number of non-negligible elements per row in the density matrix does not increase with system size. It is well known that for non-metallic systems (insulators), and for metals at finite temperatures, the magnitude of the density matrix elements decays exponentially with distance, see for example [5, 15, 19]. It should thus be possible to find a sparse approximation of the density matrix.

Other alternatives to diagonalization include domain decomposition methods such as the divide-and-conquer and fragment molecular orbital methods [2, 31, 66, 68]. In these methods, the problem is divided into smaller problems whose solutions are merged to give an approximate solution to the entire problem.

In this article, we shall focus on recursive Fermi–Dirac operator expansion methods such as the aforementioned density matrix purification and sign matrix methods. These methods construct approximations of the density matrix by recursive application of low-order polynomial or rational functions. More specifically, we shall in this article analyze errors introduced when evaluating the recursive expansion and propose schemes to control the forward error as measured by

$$\|D - \tilde{D}\| \quad (1.5)$$

where $\tilde{D}$ is the approximate density matrix given by the recursive expansion. Throughout this article, $\|\cdot\|$ denotes the Frobenius matrix norm, unless stated otherwise.

The error has previously been analyzed for the case of zero electronic temperature [56]. In the zero temperature limit the Fermi–Dirac function tends to a step function and the density matrix is given by

$$D = \theta(\mu I - F) \quad (1.6)$$

where

$$\theta(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{otherwise} \end{cases} \quad (1.7)$$

is the Heaviside function and application to matrix arguments is defined as for the Fermi–Dirac function. This means that the density matrix is the matrix for orthogonal projection onto the so-called occupied subspace spanned by the eigenvectors that correspond to eigenvalues of F smaller than the chemical potential.

In case of zero temperature the error can be analyzed in terms of the canonical angles between the exact and approximate occupied subspaces [56, 57]. This is possible because the density matrix at zero temperature is a projection matrix. The primary

purpose of the present work is to analyze the error introduced in the implicit recursive Fermi–Dirac operator expansions, proposed by Niklasson [43], working at any (zero or finite) electronic temperature. The previous analysis for explicit schemes working at zero temperature [56] cannot be straightforwardly applied here since finite electronic temperatures may result in fractional occupation numbers. That is, eigenvalues of the density matrix that lie in $]0, 1[$.

The error analysis is used to select stopping criteria for the conjugate gradient iterations employed in each iteration of the implicit recursive expansion. It should be possible to combine the error analysis of this work with ways to bring about matrix sparsity as needed to achieve linear scaling. However, this has not been done in the present work.

Expressing the density matrix in terms of the Fermi–Dirac or Heaviside functions as above may give the impression that the number of electrons is determined by the chemical potential. However, in practical calculations it is often the other way around; the chemical potential is automatically adjusted so that the desired number of electrons given as input to the program is obtained. In other words, the chemical potential should be such that *the trace condition*

$$\text{Tr}[D] = N_{\text{occ}}, \quad (1.8)$$

where N_{occ} is the number of occupied orbitals, is fulfilled.

Satisfying the trace condition is not always straightforward and several ways to achieve correct occupation count have been suggested. In minimization methods the trace condition can be handled by a modification of the objective function [40]. In recursive polynomial expansion methods the purifying polynomials can be chosen based on the trace condition as in for example canonical [49] and trace correcting [42] purification. If a good starting guess is available, the chemical potential can be found by a Newton–Raphson search [45]. Another approach is based on the observation that the occupation count depends only on the diagonal of the density matrix which may be cheaper to compute than the whole density matrix [36].

It is assumed throughout this article that the basis set is orthonormal. In practice, non-orthogonal basis sets for example built up from Slater or Gaussian-type functions are commonly used. In this case, an overlap matrix S with elements

$$S_{ij} = \int \phi_i(r) \phi_j(r) dr \quad (1.9)$$

enters the calculation. The density matrix is then given by

$$\begin{aligned} F_{\perp} &= Z^T F_S Z \\ D_{\perp} &= f_{\text{FD}}(F_{\perp}) \\ D_S &= Z D_{\perp} Z^T \end{aligned} \quad (1.10)$$

where Z is an inverse factor of the overlap matrix: $Z^T S Z = I$ and the subscripts S and $\perp$ have been used to indicate when matrices are in non-orthogonal and orthogonal bases respectively. An inverse factor Z can be computed using for example inverse Cholesky decomposition [6, 7, 12, 13, 40, 63] or methods based on iterative refinement [28, 44, 53].

This article is organized as follows. In the following section we will derive condition numbers for the problem of computing the density matrix. In Section 3 the recursive Fermi–Dirac operator expansion methods are described. The key results

of this article are presented in Section 4. In that section, the error is analyzed and schemes for error control are proposed. In Section 5 example tight-binding and density functional theory calculations are presented. The article ends with a discussion and concluding remarks in Section 6.

2. Conditioning. In this section we are looking for a measure of how difficult it is to compute the density matrix. We therefore want to generalize the notion of condition numbers of scalar valued functions to functions of matrices, and the particular case of the Fermi–Dirac function applied to a matrix. Condition numbers for the zero temperature case and for general functions of matrices have previously been suggested by Rubensson, Rudberg, and Salek [56] and Higham [26], respectively. Here, we define the condition number as

$$\kappa = \lim_{h \rightarrow 0} \sup_{A: \|A\| = \Delta\epsilon} \frac{\|f_{\text{FD}}(F + hA) - f_{\text{FD}}(F)\|}{h} \quad (2.1)$$

where $\Delta\epsilon = \lambda_{\max} - \lambda_{\min}$ is the spectral width of F . Compared to previously suggested condition numbers, we have chosen to let $\|A\| = \Delta\epsilon$ instead of $\|A\| = \|F\|$. This has the consequence that the condition number in (2.1) is invariant to shifts of the eigenspectrum of F . Since we want a measure of how difficult the problem is to solve, we do not want something that can be solved by a simple shift to influence the condition number.

It is shown in Appendix A that

$$\kappa = \Delta\epsilon \max_{j,k} |f_{\text{FD}}'[\lambda_j, \lambda_k]| \quad (2.2)$$

where we recall that $\{\lambda_i\}$ are the eigenvalues of F . Here, we used the notation

$$f[a, b] = \begin{cases} \frac{f(b) - f(a)}{b - a} & \text{if } a \neq b, \\ f'(a) & \text{otherwise.} \end{cases} \quad (2.3)$$

Assume that there is a gap ξ between the two eigenvalues closest above and below the chemical potential μ . Then we can let the temperature go to zero and get that

$$\lim_{T \rightarrow 0} \kappa = \frac{\Delta\epsilon}{\xi}. \quad (2.4)$$

The gap ξ is often referred to as the band gap or the Highest Occupied Molecular Orbital - Lowest Unoccupied Molecular Orbital (HOMO-LUMO) gap. For larger temperatures, the following bound is useful:

$$\kappa \leq \Delta\epsilon \max_s |f_{\text{FD}}'(s)| = \frac{\beta\Delta\epsilon}{4}. \quad (2.5)$$

In Figure 1 it is shown how the condition number depends on temperature and eigenvalue distribution. Inequality (2.5) is sharp in the sense that it approaches an equality as the eigenvalue gap at μ vanishes. This can be seen in Figure 1 where for example at a temperature of 30000 K, $\kappa/\Delta\epsilon \leq \beta/4 \approx 10^{-1} \text{ eV}^{-1}$.

We can conclude that for systems with small gaps ξ , use of a finite temperature Fermi–Dirac distribution gives an improved conditioning of the density matrix construction problem compared to the zero temperature step function.

FIG. 1. Absolute condition numbers $\kappa/\Delta\epsilon$ for different temperatures and varying eigenvalue gap ξ at μ .

3. Recursive Fermi–Dirac operator expansions. A general description of recursive expansion procedures for approximation of $D = f_{\text{FD}}(F)$ including numerical errors made in each iteration is given by

$$\begin{aligned} X_0 &= f_0(F) \\ \tilde{X}_0 &= X_0 + E_0 \\ X_i &= f_i(\tilde{X}_{i-1}) \\ \tilde{X}_i &= X_i + E_i \end{aligned} \tag{3.1}$$

where $i = 1, 2, \dots, n$. Here, f_0 is a linear transformation of F such that the eigenvalues of X_0 end up in the $[0, 1]$ interval in reverse order and f_i , $i = 1, 2, \dots, n$ is a sequence of low-order polynomial or rational functions such that

$$f^n(x) = f_n(f_{n-1}(\dots f_2(f_1(x)) \dots)) \tag{3.2}$$

maps all eigenvalues sufficiently close to their desired values. The error in each step i is written as an explicit perturbation E_i . Sources of errors include approximate evaluation of the functions f_i and removal of matrix elements deemed negligible.

Different choices of the f_i are possible. At zero temperature, low order polynomials with fixed points at zero and one can be used. Several such polynomial-based schemes, in the following referred to as *explicit schemes*, have been proposed. Examples include canonical purification by Palser and Manolopoulos [49] and trace correcting purification by Niklasson [42].

Another choice is to use rational expressions of the form

$$g_k(x) = \frac{x^k}{x^k + (1-x)^k}. \tag{3.3}$$

These functions are equivalent to the so-called principal Padé iteration functions

$$\frac{(1+x)^k - (1-x)^k}{(1+x)^k + (1-x)^k} \tag{3.4}$$

FIG. 2. Approximation of Fermi-Dirac distributions using equation (3.5). Panel (a): Fermi-Dirac distributions $\frac{1}{e^{4k(x-0.5)} + 1}$ (solid red lines) and rational approximations $g_k(1-x)$ (dashed blue lines) for $k \in \{2, 4, 8, 16\}$. Panel (b): Error in approximation measured by $\max_{x \in [0,1]} |g_k(1-x) - 1/(e^{4k(x-0.5)} + 1)|$ for increasing k .

for matrix sign function evaluation [26, 29]. We note that the matrix sign function up to scaling and shifting is equivalent to the Heaviside step function.

Niklasson importantly showed that $g_k(1-x)$ on the $[0, 1]$ interval is an approximation to the Fermi-Dirac distribution function with $\beta = 4k$ and $\mu = 0.5$ [43]:

$$g_k(1-x) \approx \frac{1}{e^{4k(x-0.5)} + 1}. \quad (3.5)$$

The quality of the approximation increases with k , see Figure 2. Furthermore, $g_k(x)$ satisfies a semi-group property $g_{ab}(x) = g_a(g_b(x))$ [29, 30]. This means that a recursive expansion

$$g_{m^n}(x) = \underbrace{g_m(g_m(\dots g_m(g_m(x)) \dots))}_{g_m \text{ applied } n \text{ times}} \quad (3.6)$$

can be used to efficiently calculate $g_k(x)$ for large $k = m^n$. In the following we shall use the notation $g_m^n(x) = g_{m^n}(x)$.

When the rational functions (3.3) are used in the recursive expansion procedure (3.1) with

$$f_i(x) = g_m(x), \quad i = 1, 2, \dots, n, \quad (3.7)$$

the set of linear systems

$$\left[\tilde{X}_{i-1}^m + (I - \tilde{X}_{i-1})^m \right] X_i = \tilde{X}_{i-1}^m \quad (3.8)$$

has to be solved in each iteration. Therefore, we will refer to such schemes as *implicit schemes*. The initial linear transformation can be chosen as

$$f_0(F) = \alpha_0(\mu I - F) + 0.5I. \quad (3.9)$$

In case of zero temperature, the recursive expansion approaches a step function for any initial scaling factor α_0 [26, 29]. In other words, the iterations are globally

convergent provided that the band gap is nonzero. However, the choice of initial scaling does affect the number iterations needed to reach the desired accuracy. A reasonable initial scaling is given by

$$\alpha_0 = \frac{1}{2} \min \left(\frac{1}{\mu - \lambda_{\min}}, \frac{1}{\lambda_{\max} - \mu} \right) \quad (3.10)$$

which makes all eigenvalues of X_0 end up in the $[0, 1]$ interval.

In case of a finite temperature T , the initial scaling factor α_0 and the number of iterations n should be chosen such that

$$\alpha_0 = \frac{\beta}{4k} = \frac{1}{4kT k_B} \quad (3.11)$$

where $k = m^n$. The relation (3.11) can be derived by considering (1.2), (3.5), and (3.9). Furthermore, k should be sufficiently large so that all eigenvalues of $X_0 = f_0(F)$ end up in the $[0, 1]$ interval. This is satisfied for

$$k \geq \frac{\beta}{2} \max(\mu - \lambda_{\min}, \lambda_{\max} - \mu). \quad (3.12)$$

4. Analysis and error control. In the following, the forward error $\|D - \tilde{X}_n\|$ will be analyzed and schemes for error control proposed. An analysis for explicit schemes working at zero temperature has been presented elsewhere [56]. Here, implicit schemes working at zero or finite temperature are treated.

First, we distinguish between errors due to a finite number of iterations n and other errors:

$$\|D - \tilde{X}_n\| \leq \|D - g_m^n(X_0)\| + \|g_m^n(X_0) - \tilde{X}_n\|. \quad (4.1)$$

4.1. Errors due to a finite number of iterations. The first term on the right hand side of (4.1), measuring errors in eigenvalues due to a finite number of iterations n , can be controlled by choosing n sufficiently large.

At zero temperature,

$$\|D - g_m^n(X_0)\| \leq \sqrt{N} \max_j |\theta(\mu - \lambda_j) - g_m^n(f_0(\lambda_j))|, \quad (4.2)$$

where N is the matrix dimension. Assume that the initial scaling given by (3.10) is used and that μ is located in the middle of the band gap. Then,

$$\max_j |\theta(\mu - \lambda_j) - g_m^n(f_0(\lambda_j))| = |\theta(-\xi/2) - g_m^n(f_0(\mu + \xi/2))| \quad (4.3)$$

$$= g_m^n(0.5 - \alpha_0 \xi/2) \quad (4.4)$$

$$= \frac{1}{1 + \left(\frac{1 + \alpha_0 \xi}{1 - \alpha_0 \xi} \right)^k} \quad (4.5)$$

$$\leq \frac{1}{1 + e^{2k\alpha_0 \xi}} \quad (4.6)$$

where again $k = m^n$. Note that the maximum in (4.3) is degenerate and attained also for $\lambda_j = \mu - \xi/2$ which gives

$$|\theta(\mu - \lambda_j) - g_m^n(f_0(\lambda_j))| = |\theta(\xi/2) - g_m^n(f_0(\mu - \xi/2))| \quad (4.7)$$

$$= 1 - g_m^n(0.5 + \alpha_0 \xi/2) \quad (4.8)$$

$$= g_m^n(0.5 - \alpha_0 \xi/2). \quad (4.9)$$

Thus, choosing the number of iterations n so that

$$k \geq \frac{1}{2\alpha_0\xi} \log \left(\frac{1-\varepsilon}{\varepsilon} \right) \quad (4.10)$$

gives an error

$$\max_j |\theta(\mu - \lambda_j) - g_m^n(f_0(\lambda_j))| \leq \varepsilon. \quad (4.11)$$

At finite temperature, we can take advantage of the smoothness of f_{FD} which makes the following bound useful,

$$\|D - g_m^n(X_0)\| \leq \sqrt{N} \max_{s \in C(F)} |f_{\text{FD}}(s) - g_m^n(f_0(s))| \quad (4.12)$$

where $C(F)$ is the closed convex hull of the eigenvalues of F . Assume that k fulfills (3.12) so that all eigenvalues of X_0 are in the $[0, 1]$ interval. A polynomial fit to the curve in Figure 2(b) gives the following demand on k :

$$k \geq e^{(\log(\varepsilon)-b)/a} \quad (4.13)$$

where $\max_{s \in C(F)} |f_{\text{FD}}(s) - g_m^n(f_0(s))| \leq \varepsilon$ is the requested accuracy, $a \approx -2.0077$, and $b \approx -2.2387$. The number of iterations n should be chosen sufficiently large so that both (3.12) and (4.13) are fulfilled. Once the number of iterations is determined, the initial scaling factor is given by (3.11), which gives the desired temperature.

4.2. Errors due to matrix perturbations. The second term on the right hand side of (4.1), measuring errors due to the perturbations E_i in each step, is more tricky. Let us first split up the error between the iterations. Since

$$g_m^{n-i}(\tilde{X}_i) = g_m^{n-i-1}(X_{i+1}), \quad i = 0, 1, 2, \dots, n-2, n-1, \quad (4.14)$$

we have that

$$\|g_m^n(X_0) - \tilde{X}_n\| \leq \|g_m^n(X_0) - g_m^n(\tilde{X}_0)\| + \|g_m^{n-1}(X_1) - \tilde{X}_n\| \quad (4.15)$$

$$\leq \sum_{i=0}^1 \|g_m^{n-i}(X_i) - g_m^{n-i}(\tilde{X}_i)\| + \|g_m^{n-2}(X_2) - \tilde{X}_n\| \quad (4.16)$$

$$\leq \dots \quad (4.17)$$

$$\leq \sum_{i=0}^n \|g_m^{n-i}(X_i) - g_m^{n-i}(\tilde{X}_i)\|. \quad (4.18)$$

We may now consider the error made in a single step i . Using recent matrix function perturbation theory by Gil' [18] we have that

$$\|g_m^{n-i}(X_i) - g_m^{n-i}(\tilde{X}_i)\| \leq \|E_i\| \max_{j,k} |g_m^{n-i}[\eta_j, \tilde{\eta}_k]| \quad (4.19)$$

where $\{\eta_j\}$ and $\{\tilde{\eta}_k\}$ are the eigenvalues of X_i and $\tilde{X}_i$ respectively, see Corollary 2.3 in [18].

At zero temperature,

$$\max_{j,k} |g_m^{n-i}[\eta_j, \tilde{\eta}_k]| = \max_{j,k} \left| \frac{g_m^{n-i}(\tilde{\eta}_k) - g_m^{n-i}(\eta_j)}{\tilde{\eta}_k - \eta_j} \right| = \frac{1}{\hat{\xi}_i} \quad (4.20)$$

FIG. 3. Eigenspectra of X_i and $\tilde{X}_i$. The perturbed gap in (4.20) is $\hat{\xi}_i = \min(p, q)$.

where $\hat{\xi}_i$ is the perturbed band gap in iteration i , see Figure 3. This is in agreement with previous results for explicit schemes [56] based on Davis and Kahan's bounds on perturbations of subspaces [16, 17]. We note that the perturbation of the gap is bounded by the spectral norm of the error matrix:

$$|\hat{\xi}_i - \xi_i| \leq \|E_i\|_2, \quad (4.21)$$

where ξ_i is the band gap of X_i . This is a direct consequence of Weyl's theorem on eigenvalue movement [64].

Regardless of temperature, we may use that

$$\max_{j,k} |g_m^{n-i}[\eta_j, \tilde{\eta}_k]| \leq \max_s \left| \frac{d}{ds} g_m^{n-i}(s) \right| = m^{n-i}. \quad (4.22)$$

This bound is somewhat easier to use than (4.20) since information about the band gap is not needed. The bound (4.22) is particularly sharp at temperatures that give occupation numbers close to 0.5.

Here it could be noted that (4.20) and (4.22) are directly related to the condition number for the zero temperature case (2.4) and the bound of the condition number for finite temperatures (2.5), respectively. For example, for the zero temperature case we have that $1/\hat{\xi}_0 \approx 1/(\xi\alpha_0) \approx \Delta\epsilon/\xi = \kappa$. Thus, by (4.19), the condition number can for recursive expansion methods be seen as an amplification factor that determines the worst possible growth of any errors that occur during the calculation.

Using (4.22) we get

$$\|g_m^n(X_0) - \tilde{X}_n\| \leq \sum_{i=0}^n \|E_i\| m^{n-i} \quad (4.23)$$

for the second term on the right hand side of (4.1). Assume that an accuracy of $\|D - \tilde{X}_n\| \leq \gamma$ is desired. Let us choose n large enough so that $\|D - g_m^n(X_0)\| \leq \gamma/2$. Then, the desired accuracy is guaranteed by (4.1) and (4.23) provided that

$$\|E_i\| \leq \frac{\gamma}{2(n+1)m^{n-i}}, \quad i = 0, 1, \dots, n. \quad (4.24)$$

The system of linear equations (3.8) can be solved using an iterative linear system solver [3] to obtain an approximate solution $\tilde{X}_i$ with residual matrix

$$R_i = [\tilde{X}_{i-1}^m + (I - \tilde{X}_{i-1})^m] \tilde{X}_i - \tilde{X}_{i-1}^m. \quad (4.25)$$

Note that the coefficient matrix is symmetric positive definite, and therefore we can apply the conjugate gradient method. In order to stop the iterations of the linear system solver in due time, we need a relationship between the residual R_i and the error matrix E_i .

For the exact solution X_i to the set of linear systems we have that

$$[\tilde{X}_{i-1}^m + (I - \tilde{X}_{i-1})^m]X_i = \tilde{X}_{i-1}^m \quad (4.26)$$

and therefore,

$$R_i = [\tilde{X}_{i-1}^m + (I - \tilde{X}_{i-1})^m](\tilde{X}_i - X_i) \quad (4.27)$$

and

$$E_i = \tilde{X}_i - X_i = [\tilde{X}_{i-1}^m + (I - \tilde{X}_{i-1})^m]^{-1}R_i. \quad (4.28)$$

Since

$$\|[\tilde{X}_{i-1}^m + (I - \tilde{X}_{i-1})^m]^{-1}\|_2 = \max_j \frac{1}{\tilde{\eta}_j^m + (1 - \tilde{\eta}_j)^m} \leq 2^{m-1} \quad (4.29)$$

where $\{\tilde{\eta}_j \in [0, 1]\}$ are the eigenvalues of $\tilde{X}_{i-1}$, we have that

$$\|E_i\| \leq 2^{m-1}\|R_i\|. \quad (4.30)$$

Here we have used that $\|AB\| \leq \|A\|_2\|B\|$ for every unitarily invariant norm $\|\cdot\|$, see for example Theorem 3.9 in Chapter II of [64]. Thus, the desired accuracy is guaranteed if

$$\|E_0\| \leq \frac{\gamma}{2(n+1)m^n} \quad (4.31)$$

and

$$\|R_i\| \leq \frac{\gamma}{2^m(n+1)m^{n-i}}, \quad i = 1, 2, \dots, n. \quad (4.32)$$

This is the key result of this article, making it possible to control the forward error of implicit recursive Fermi–Dirac operator expansions at any temperature.

5. Examples. In this section the error control of the previous section is demonstrated. Calculations using the implicit recursive expansion scheme with $m = 2$ are performed using the linear conjugate gradient method to solve the set of linear systems in each iteration. The number of iterations and convergence criteria for the conjugate gradient iterations are chosen as described in the previous section. In the conjugate gradient method, the previous matrix $\tilde{X}_{i-1}$ is used as starting guess when solving for $\tilde{X}_i$. Computational cost is measured by the number of matrix-matrix multiplications needed for the entire expansion. In the conjugate gradient iterations, N matrix-vector multiplications are counted as one matrix-matrix multiplication.

5.1. Tight binding calculations. We consider here two tight-binding models that have previously been used in example calculations by Lin, Lu, Car, and E [35]. Thus, the results presented here can be directly compared to the results of that work. The first model tight-binding Hamiltonian

$$\mathcal{H} = -t \sum_{\langle i,j \rangle} c_i^\dagger c_j \quad (5.1)$$

μ	κ	$\ D - D^2\ $	γ	$\ D - \tilde{X}_n\ $	ΔE	n	mmul
0	25.4	9.9×10^{-14}	10^{-2}	2.3×10^{-5}	2.0×10^{-9}	10	61
			10^{-4}	1.2×10^{-6}	2.0×10^{-9}	10	72
			10^{-6}	3.9×10^{-9}	1.6×10^{-15}	12	101
5.44	50.3	2.7×10^{-2}	10^{-2}	1.8×10^{-5}	5.0×10^{-8}	11	77
			10^{-4}	6.2×10^{-7}	3.1×10^{-9}	11	92
			10^{-6}	1.2×10^{-8}	1.1×10^{-10}	12	116
10.88	50.8	1.1×10^{-6}	10^{-2}	2.3×10^{-4}	2.7×10^{-6}	11	78
			10^{-4}	1.4×10^{-7}	9.5×10^{-11}	11	92
			10^{-6}	9.4×10^{-10}	1.7×10^{-12}	12	114

TABLE 1

Performance of the implicit scheme applied to the tight binding Hamiltonian matrix given by (5.1). We count the number of matrix-matrix multiplications (mmul) for different values of chemical potential μ and desired accuracy $\|D - \tilde{X}_n\| \leq \gamma$. The condition number κ , a measure of idempotency $\|D - D^2\|$, the relative error in energy $\Delta E = |\text{Tr}[DF] - \text{Tr}[\tilde{X}_n F]|/|\text{Tr}[DF]|$, and the number of recursive expansion iterations n are also included in the table.

where $c_i^\dagger$ and c_j are creation and annihilation operators in second quantization notation and i and j label neighboring sites on a simple cubic lattice with $N = 10 \times 10 \times 10$ sites and periodic boundary conditions. This gives a $N \times N$ Hamiltonian matrix with 6 nonzero entries per row corresponding to hopping between nearest neighbors. We choose the hopping parameter and temperature as in the work by Lin et al. [34, 35]: $t = 2.2676$ eV and $T = 100$ K. We investigate the performance for a set of chemical potentials chosen as in [35]. Results for the implicit recursive expansion method using the schemes for error control presented in the previous section are shown in Table 1.

Noteworthy is that the density matrix for this example is close to being idempotent ($D = D^2$) when $\mu = 0$ eV and when $\mu = 10.88$ eV. This is an effect of low temperature and large eigenvalue gap around μ . When the density matrix is idempotent, explicit recursive expansion schemes designed for the zero temperature case can be used. To get a feeling for the computational savings that are possible when the result is idempotent, we apply four different explicit schemes to this tight-binding Hamiltonian. The used methods are grand-canonical [39, 49] and trace-correcting [42] purification. We also employ the accelerated nonmonotonic variants of these two methods described in [52]. The results are shown in Table 2.

The explicit schemes use up to almost an order of magnitude fewer matrix-matrix multiplications than the implicit scheme. The number of multiplications is for the explicit schemes insensitive to the requested accuracy. Lower or higher accuracy demands result in a similar number of multiplications. However, with a lower accuracy demand, each multiplication can be carried out with lower accuracy as described for example in [56], resulting in a reduced computational cost.

The second tight-binding Hamiltonian comes from a disordered Anderson model. This Hamiltonian is obtained by adding on-site random energies ε_i to the tight-binding Hamiltonian in (5.1):

$$\mathcal{H} = -t \sum_{\langle i,j \rangle} c_i^\dagger c_j + \sum_i \varepsilon_i c_i^\dagger c_i. \quad (5.2)$$

The on-site energies ε_i are taken to be uniformly distributed in $[-1.13, 1.13]$. Results are shown in Table 3.

As already mentioned, the examples in this subsection were taken from [35] where an algorithm based on a multipole representation of the Fermi–Dirac operator was

μ	mmul			
	McW	TC2	McW-S	TC2-S
0	26	24	16	14
10.88	36	24	20	13

TABLE 2

Performance of explicit recursive expansion schemes applied to the tight binding Hamiltonian matrix given by (5.1). (This is the same Hamiltonian matrix as was used for the results presented in Table 1.) The applied schemes are grand-canonical (McW) and trace-correcting (TC2) purification and their accelerated variants (McW-S and TC2-S). The table shows number of matrix-matrix multiplications (mmul) needed to reach an accuracy of $\|D - \tilde{X}_n\| \leq 10^{-6}$.

μ	κ	$\ D - D^2\ $	γ	$\ D - \tilde{X}_n\ $	ΔE	n	mmul
0	41.8	1.5×10^{-13}	10^{-2}	1.3×10^{-4}	1.3×10^{-8}	10	67
			10^{-4}	3.6×10^{-7}	3.3×10^{-13}	10	85
			10^{-6}	1.4×10^{-8}	1.6×10^{-13}	12	118
5.44	754	3.3×10^{-1}	10^{-2}	4.3×10^{-4}	2.1×10^{-7}	11	88
			10^{-4}	4.0×10^{-6}	1.9×10^{-9}	11	112
			10^{-6}	2.0×10^{-8}	3.0×10^{-13}	12	144
10.88	71.3	3.4×10^{-4}	10^{-2}	1.6×10^{-4}	3.4×10^{-6}	11	85
			10^{-4}	1.3×10^{-6}	1.8×10^{-9}	11	109
			10^{-6}	2.2×10^{-8}	9.3×10^{-11}	12	140

TABLE 3

Performance of the implicit scheme applied to the Anderson tight binding Hamiltonian matrix given by (5.2). We count the number of matrix-matrix multiplications (mmul) for different values of chemical potential μ and desired accuracy $\|D - \tilde{X}_n\| \leq \gamma$. The condition number κ , a measure of idempotency $\|D - D^2\|$, the relative error in energy $\Delta E = |\text{Tr}[DF] - \text{Tr}[\tilde{X}_n F]|/|\text{Tr}[DF]|$, and the number of recursive expansion iterations n are also included in the table.

presented. A comparison between that work and the results presented in tables 1 and 3 indicates that the implicit recursive expansion scheme is substantially more efficient than the multipole based method. For the tight-binding examples considered here, the multipole based method used 300–380 matrix-matrix multiplications with variations depending mainly on the requested accuracy [35]. It should be noted that the overall computational cost for both methods depends on the cost to carry out the matrix-matrix multiplications. If sparse matrix algebra and dropping of small matrix elements were used, the performance would depend on the matrix sparsity that can be brought about in intermediate steps of the two methods.

We recall from Section 2 that the condition number $\kappa \leq \beta\Delta\epsilon/4$. For both Hamiltonian matrices considered in this subsection we have that $\beta\Delta\epsilon/4 \approx 789$. We note that the condition numbers in tables 1 and 3 are much smaller than $\beta\Delta\epsilon/4$ when the density matrix is close to being idempotent.

5.2. Kohn–Sham density functional theory calculations. In this section we consider examples from self-consistent field Kohn–Sham density functional theory calculations. As discussed in Section 2, it is important to consider how the conditioning of the problem affects the performance of the method. For systems with fractional occupation numbers, the condition number depends mainly on the temperature and the spectral width of the effective Hamiltonian matrix, see (2.5). Thus, different condition numbers can be obtained by varying either the temperature or the spectral width. In the following, we try both approaches. Kohn–Sham matrices and overlap matrices used in this subsection were generated with the Ergo program [59, 60] run-

ning self-consistent field density functional theory calculations using the local density approximation (LDA) and standard Gaussian basis sets. Inverse Cholesky factors of the overlap matrices were used for congruence transformations to orthogonal basis, see (1.10).

We start by varying the temperature for a small graphene sheet depicted in Figure 4(f) whose atomic coordinates were obtained by replicating a perfect honeycomb pattern with C-C distance 1.42 Å and H terminations with C-H distance 1.01 Å. Results are shown in Figure 4.

Figure 5 shows results for a set of hydrides HX, where X is one of H, F, Cl, Br, and I. For these systems, Kohn–Sham matrices were generated at fixed electronic temperature. Since the spectral width of the Kohn–Sham matrix generally increases with the largest atom number in the system, we get a varying condition number also for this set of calculations.

For the graphene sheet, the density matrix is essentially idempotent ($D \approx D^2$) for temperatures up to 316.2 K, see the rightmost eigenspectrum in Figure 4(a). This means that temperatures up to 316.2 K give the same result as for zero temperature. In Figure 4(a) one can also see that for increasing temperature, the density matrix approaches the identity matrix scaled with the occupation factor: $D = (N_{\text{occ}}/N)I$ which depends on the number of electrons and basis functions in the system. Temperatures like 316.2 K were used in order to get approximately equidistant spacing in the logarithmic plots ($316.2 \approx 10^{2.5}$).

The calculations were performed with three different requested accuracies: $\|D - \tilde{X}_n\| \leq \gamma \in \{10^{-2}, 10^{-4}, 10^{-6}\}$. As can be seen in figures 4(b) and 5(b), the errors end up below the requested accuracy as expected.

The number of recursive expansion iterations is determined by inequalities (3.12) and (4.13). At low temperatures (large condition numbers), inequality (3.12) is the most restrictive of the two and therefore the number of iterations

$$n \approx \frac{\log(\beta \max(\mu - \lambda_{\min}, \lambda_{\max} - \mu)/2)}{\log m} = \frac{\log(\beta \Delta\epsilon/C_1(\mu))}{\log m} \quad (5.3)$$

where $C_1(\mu) \in [2, 4]$ is a constant depending on the chemical potential μ . At high temperatures (small condition numbers) inequality (4.13) is the most restrictive and the number of iterations

$$n \approx \frac{C_2(\gamma)}{\log m} \quad (5.4)$$

where $C_2(\gamma)$ is a constant depending on the requested accuracy γ . This explains the results presented in figures 4(c) and 5(c).

It is shown in Appendix B that the number of conjugate gradient iterations needed to reach the accuracy required by (4.24) is bounded by

$$\frac{1}{2} \sqrt{\kappa_2} \log \left(2 \sqrt{\kappa_2} \delta \frac{2(n+1)m^{n-i}}{\gamma} \right) \quad (5.5)$$

where κ_2 is the spectral condition number of the coefficient matrix and $\delta = \|\tilde{X}_{i-1} - X_i\|$ is the initial error in the conjugate gradient iterations. (We recall that $\tilde{X}_{i-1}$ is used as starting guess when solving for X_i .) The condition number κ_2 is bounded:

$$\kappa_2 = \frac{\max_j \tilde{\eta}_j^m + (1 - \tilde{\eta}_j)^m}{\min_j \tilde{\eta}_j^m + (1 - \tilde{\eta}_j)^m} \leq \frac{1}{0.5^{m-1}} = 2^{m-1} \quad (5.6)$$

where we recall that $\{\tilde{\eta}_j \in [0, 1]\}$ are the eigenvalues of $\tilde{X}_{i-1}$. Therefore, the number of conjugate gradient iterations needed in a single recursive expansion iteration is bounded by

$$\begin{aligned} \frac{1}{2}\sqrt{2^{m-1}} \log \left(4\sqrt{2^{m-1}}\delta \frac{(n+1)m^{n-i}}{\gamma} \right) = \\ \frac{1}{2}\sqrt{2^{m-1}} \left(\log \left(\frac{4\sqrt{2^{m-1}}\delta}{\gamma} \right) + \log(n+1) + (n-i) \log(m) \right). \end{aligned} \quad (5.7)$$

Note that in each conjugate gradient iteration we perform N matrix-vector multiplications since we need to solve for all N columns of X_i .

We also need to count matrix-matrix multiplications needed to evaluate the coefficient matrix and the right hand side matrix. This amounts to evaluating matrix polynomials which can be done with the Paterson–Stockmeyer method using $\mathcal{O}(\sqrt{m})$ matrix-matrix multiplications [50]. Counting N matrix-vector multiplications as one matrix-matrix multiplication we have that the total number of matrix-matrix multiplications in the recursive expansion is bounded by

$$\mathcal{O}(n^2). \quad (5.8)$$

The analysis above may give the impression that the number of matrix-matrix multiplications for a fixed requested accuracy should be proportional to the squared number of iterations. The results presented in figures 4(d) and 5(d) are not fully in agreement with such a conclusion. The reason is mainly that, although bounded, the condition numbers of the linear system coefficient matrices vary with temperature and eigenvalue distribution. For example, at extremely high temperatures, the coefficient matrices are close to being scaled identity matrices with condition numbers equal to 1. This leads to very few conjugate gradient iterations being needed.

In figures 4(e) and 5(e) we can see that $\kappa \approx \beta\Delta\epsilon/4$ in cases with fractional occupation, just as anticipated in Section 2. Assume that the chemical potential μ lies in the middle of the eigenspectrum so that $C_1(\mu) = 4$. Then, (5.3) and (5.8) imply that the computational cost measured by the number of matrix-matrix multiplications for low temperatures is bounded by $\mathcal{O}((\log(\kappa))^2)$.

6. Discussion and concluding remarks. In this work, it was suggested that the number of iterations n should be chosen sufficiently large such as to guarantee a certain accuracy in eigenvalues. However, one motivation for using a finite temperature Fermi–Dirac distribution is to improve the self-consistent field convergence [51]. Then, the temperature is lowered in each self-consistent field iteration and set to zero before the end of the self-consistent field process. In such cases, it is not likely that high accuracy in fractional occupation numbers is important and the condition on the number of iterations given by (4.13) may then be relaxed.

The forward error was in this paper bounded in terms of the Frobenius norm. As has been discussed previously, use of the Frobenius norm whose value grows intrinsically with system size is unsatisfactory when very large systems are to be treated [58]. We recall from (4.1) that the total forward error was split in two parts. In case of the spectral norm, bounds for the first part ($\|D - g_m^n(X_0)\|$) can be obtained by replacing the factor $\sqrt{N}$ with 1 in (4.2) and (4.12). The results for the second part ($\|g_m^n(X_0) - \tilde{X}_n\|$) are expected to hold also for the spectral norm. However, to the author’s best knowledge, the bound in (4.19) has for the spectral norm only been proven for the zero temperature case [17, 56].

FIG. 4. Graphene sheet: Results for the implicit scheme with $m = 2$ when the number of iterations and convergence criteria for conjugate gradient iterations are chosen as described in Section 4. The effective Hamiltonian matrices come from converged density functional theory self-consistent field calculations on a small graphene sheet—depicted in Panel (f)—using the local density approximation (LDA) and the Gaussian basis set 6-311G**. The effective Hamiltonian matrices which were essentially dense with dimension 588×588 were computed for temperatures between 316.2 and $3.162 \cdot 10^8$ K to obtain varying condition number. The spectral width $\Delta\epsilon$ of the Kohn–Sham matrices varied between 921 and 963 eV depending on temperature. The implicit scheme was run for three different requested accuracies: $\|D - \tilde{X}_n\| \leq \gamma \in \{10^{-2}, 10^{-4}, 10^{-6}\}$. Panel (a) shows the eigenspectrum of the density matrix, Panel (b) the error in the density matrix $\|D - \tilde{X}_n\|$, Panel (c) the number of recursive expansion iterations, Panel (d) the number of matrix-matrix multiplications, and Panel (e) the condition number κ .

FIG. 5. Hydrides: Results for the implicit recursive expansion with $m = 2$ when the number of iterations and convergence criteria for conjugate gradient iterations are chosen as described in Section 4. The effective Hamiltonian matrices come from converged density functional theory self-consistent field calculations on a set of hydrides HX , $X \in \{H, F, Cl, Br, I\}$ using the local density approximation (LDA) and the Gaussian basis set 3-21G. The matrix dimensions were 4, 11, 15, 24, and 33 for HH, HF, HCl, HBr, and HI, respectively. The temperature was fixed to 100000 K in all calculations. The temperature was chosen large enough to get fractional occupation numbers for the set of test systems. The implicit scheme was run for three different requested accuracies: $\|D - \tilde{X}_n\| \leq \gamma \in \{10^{-2}, 10^{-4}, 10^{-6}\}$. Panel (a) shows the eigenspectrum of the density matrix, Panel (b) the error in the density matrix $\|D - \tilde{X}_n\|$, Panel (c) the number of recursive expansion iterations, Panel (d) the number of matrix-matrix multiplications, and Panel (e) the condition number κ . The spectral widths ($\Delta\epsilon$) of the Kohn-Sham matrices were about 45, 740, 2740, 13000, and 31400 eV for HH, HF, HCl, HBr, and HI, respectively. The largest condition number is therefore obtained for the iodine hydride molecule (rightmost points) whereas the smallest condition number is obtained for the hydrogen molecule (leftmost points).

The requirement for E_i and for the residual R_i given by (4.24) and (4.32) may seem restrictive for large n . The requirement reflects the fact that the problem becomes increasingly difficult to solve when the temperature decreases, and n must be increased. For systems with a large bandgap, the bound in (4.22) is not sharp and in such cases it could be better to use an algorithm that computes the left hand side of (4.22) and uses that information to decide how large $\|E_i\|$ is allowed to be. Systems with small gaps give large condition numbers for low temperatures and it is then natural that the requirement for E_i is restrictive.

The set of systems in (3.8) consists of N linear systems that share a coefficient matrix, often referred to as a multiple right hand sides linear system. The linear systems are extremely well conditioned with spectral condition numbers $\kappa_2 \leq 2^{m-1}$. Many possibilities to speed up the solution of these equations exist. The N systems can be solved in parallel. It is also possible to shift the work from matrix-vector multiplications to more efficient matrix-matrix multiplications. Alternatively, some method specialized for the solution of multiple right hand sides linear systems may be used, see for example [47]. Preconditioning techniques such as incomplete Cholesky factorization could also be used; perhaps combined with iterative refinement [44] between the recursive expansion iterations. Neither of the approaches above were tested here. However, we took advantage of the fact that $\tilde{X}_{i-1}$ is a good starting guess when solving for X_i [43]. For the tight-binding calculations in Section 5.1, this resulted in a reduction of the number of matrix-matrix multiplications of around 20% compared to calculations employing random starting guesses.

As can be seen in Table 2, the convergence of explicit recursive polynomial expansion schemes can be accelerated by use of a scale and fold technique [52]. This type of acceleration is possible also for implicit schemes working at zero electronic temperature. In this case, the accelerated iteration is given by

$$f_i(x) = g_m(\alpha_i(x - 0.5) + 0.5) \quad (6.1)$$

where $\alpha_i \geq 1$ determines the amount of scaling.

In conclusion, ways to control the forward error of recursive Fermi–Dirac operator expansions were proposed. The proposed schemes guarantee that the requested accuracy in terms of the difference to the exact result is achieved. The new schemes were here used to select stopping criteria for the conjugate gradient iterations in each recursive expansion iteration. However, the analysis may also be used to control errors due to neglect of small matrix elements to achieve matrix sparsity as needed in linear scaling methods. Condition numbers for the problem of computing the density matrix were derived and it was shown that the computational cost of the recursive schemes increases only with the squared logarithm of the condition number.

Appendix A. Derivation of condition numbers.

THEOREM A.1. *Let X be a Hermitian matrix with eigenvalues $\{\lambda_i\}$ and spectral width $\Delta\epsilon$ and let f be a function defined on $\mathbb{R}$ and differentiable on the eigenspectrum of X . Then,*

$$\kappa = \lim_{h \rightarrow 0} \sup_{A: \|A\| = \Delta\epsilon} \frac{\|f(X + hA) - f(X)\|}{h} = \Delta\epsilon \max_{j,k} |f[\lambda_k, \lambda_j]| \quad (\text{A.1})$$

where $\|\cdot\|$ is the Frobenius matrix norm.

LEMMA A.2. *Let X be a Hermitian matrix with eigenvalues $\{\lambda_i\}$ and spectral width $\Delta\epsilon$ and let f be a function defined on $\mathbb{R}$ and differentiable on the eigenspectrum*

of X . Then, there exists a matrix A such that $\|A\| = \Delta\epsilon$ and

$$\lim_{h \rightarrow 0} \frac{\|f(X + hA) - f(X)\|}{h} = \Delta\epsilon \max_{j,k} |f[\lambda_k, \lambda_j]| \quad (\text{A.2})$$

where $\|\cdot\|$ is either the spectral or Frobenius matrix norm.

Proof of Lemma A.2. Let

$$X = \sum_i \lambda_i x_i x_i^* \quad (\text{A.3})$$

be the eigendecomposition of X and let

$$\tilde{X} = \tilde{\lambda}_j \tilde{x}_j \tilde{x}_j^* + \lambda_k \tilde{x}_k \tilde{x}_k^* + \sum_{i \neq j,k} \lambda_i x_i x_i^* \quad (\text{A.4})$$

where j and k are chosen such as to maximize $|f[\lambda_j, \lambda_k]|$ and where $\tilde{\lambda}_j = \lambda_j + h\delta$ and

$$\tilde{x}_k = x_k \cos \theta + x_j \sin \theta, \quad (\text{A.5})$$

$$\tilde{x}_j = -x_k \sin \theta + x_j \cos \theta. \quad (\text{A.6})$$

Case 1. Assume that $\lambda_k = \lambda_j$. Choose $\theta = 0$ and $\delta = \Delta\epsilon$ and let

$$A = \frac{\tilde{X} - X}{h} = \Delta\epsilon x_j x_j^*. \quad (\text{A.7})$$

Then, $\|A\| = \Delta\epsilon$ and

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{\|f(X + hA) - f(X)\|}{h} &= \lim_{h \rightarrow 0} \frac{\|f(\tilde{X}) - f(X)\|}{h} \\ &= \lim_{h \rightarrow 0} \frac{\|(f(\lambda_j + \Delta\epsilon h) - f(\lambda_j))x_j x_j^*\|}{h} \\ &= \lim_{h \rightarrow 0} \frac{|f(\lambda_j + \Delta\epsilon h) - f(\lambda_j)|}{h} \\ &= \Delta\epsilon \lim_{h \rightarrow 0} \frac{|f(\lambda_j + \Delta\epsilon h) - f(\lambda_j)|}{\Delta\epsilon h} \\ &= \Delta\epsilon |f'(\lambda_j)|. \end{aligned} \quad (\text{A.8})$$

Since j and k are chosen such as to maximize $f[\lambda_j, \lambda_k]$ and since $\lambda_k = \lambda_j$,

$$\Delta\epsilon |f'(\lambda_j)| = \Delta\epsilon \max_{j,k} |f[\lambda_k, \lambda_j]| \quad (\text{A.9})$$

and

$$\lim_{h \rightarrow 0} \frac{\|f(X + hA) - f(X)\|}{h} = \Delta\epsilon \max_{j,k} |f[\lambda_k, \lambda_j]|. \quad (\text{A.10})$$

Case 2. Assume that $\lambda_k \neq \lambda_j$. Choose $\delta = 0$ and

$$\theta = \arcsin \left(\frac{\Delta\epsilon}{\tau} \frac{h}{\lambda_k - \lambda_j} \right) \quad (\text{A.11})$$

where $\tau = 1$ in case of spectral norm and $\tau = \sqrt{2}$ in case of the Frobenius norm. First, we note that

$$\begin{aligned}\tilde{x}_j \tilde{x}_j^* - x_j x_j^* &= x_k x_k^* \sin^2 \theta + x_j x_j^* \cos^2 \theta - x_k x_j^* \sin \theta \cos \theta - x_j x_k^* \cos \theta \sin \theta - x_j x_j^* \\ &= x_k x_k^* (1 - \cos^2 \theta) - x_j x_j^* \sin^2 \theta - x_k x_j^* \sin \theta \cos \theta - x_j x_k^* \cos \theta \sin \theta \\ &= -(\tilde{x}_k \tilde{x}_k^* - x_k x_k^*)\end{aligned}\quad (\text{A.12})$$

which gives

$$\begin{aligned}\tilde{X} - X &= \lambda_j(\tilde{x}_j \tilde{x}_j^* - x_j x_j^*) + \lambda_k(\tilde{x}_k \tilde{x}_k^* - x_k x_k^*) \\ &= (\lambda_k - \lambda_j)(\tilde{x}_k \tilde{x}_k^* - x_k x_k^*).\end{aligned}\quad (\text{A.13})$$

Furthermore,

$$f(X) = \sum_i f(\lambda_i) x_i x_i^* \quad (\text{A.14})$$

$$f(\tilde{X}) = f(\lambda_j) \tilde{x}_j \tilde{x}_j^* + f(\lambda_k) \tilde{x}_k \tilde{x}_k^* + \sum_{i \neq j, k} f(\lambda_i) x_i x_i^*. \quad (\text{A.15})$$

Let $A = (\tilde{X} - X)/h$. Then, by (A.11) and (A.13) we have that

$$\begin{aligned}A &= \frac{\tilde{X} - X}{h} = \frac{\lambda_k - \lambda_j}{h} (\tilde{x}_k \tilde{x}_k^* - x_k x_k^*) \\ &= \frac{\Delta\epsilon}{\tau \sin \theta} (\tilde{x}_k \tilde{x}_k^* - x_k x_k^*).\end{aligned}\quad (\text{A.16})$$

The nonzero singular values of $\tilde{x}_k \tilde{x}_k^* - x_k x_k^*$ are $\{\sin \theta, \sin \theta\}$, see for example Theorem 5.5 in Chapter I of [64]. Therefore, we have that $\|\tilde{x}_k \tilde{x}_k^* - x_k x_k^*\| = \tau \sin \theta$ and by (A.16) that $\|A\| = \Delta\epsilon$.

Finally, by (A.12) and (A.13) we have that

$$\begin{aligned}\lim_{h \rightarrow 0} \frac{\|f(X + hA) - f(X)\|}{h} &= \lim_{h \rightarrow 0} \frac{\|f(\lambda_j)(\tilde{x}_j \tilde{x}_j^* - x_j x_j^*) + f(\lambda_k)(\tilde{x}_k \tilde{x}_k^* - x_k x_k^*)\|}{h} \\ &= \lim_{h \rightarrow 0} \frac{\|(f(\lambda_k) - f(\lambda_j))(\tilde{x}_k \tilde{x}_k^* - x_k x_k^*)\|}{h} \\ &= \lim_{h \rightarrow 0} \frac{\|\tilde{X} - X\|}{h} \left| \frac{f(\lambda_k) - f(\lambda_j)}{\lambda_k - \lambda_j} \right| \\ &= \Delta\epsilon \max_{j, k} |f[\lambda_k, \lambda_j]|. \quad \square\end{aligned}\quad (\text{A.17})$$

Proof of Theorem A.1. By Lemma A.2,

$$\kappa \geq \Delta\epsilon \max_{j, k} |f[\lambda_j, \lambda_k]|. \quad (\text{A.18})$$

Using recent matrix perturbation theory by Gil' [18] we have that

$$\kappa = \lim_{h \rightarrow 0} \sup_{A: \|A\| = \Delta\epsilon} \frac{\|f(X + hA) - f(X)\|}{h} \quad (\text{A.19})$$

$$= [\text{by Corollary 2.3 in [18]}] \quad (\text{A.20})$$

$$\leq \lim_{h \rightarrow 0} \sup_{A: \|A\| = \Delta\epsilon} \frac{h \|A\| \max_{j, k} |f[\lambda_j, \lambda_k]|}{h} \quad (\text{A.21})$$

$$= \Delta\epsilon \max_{j, k} |f[\lambda_j, \lambda_k]|. \quad (\text{A.22})$$

Thus, $\kappa = \Delta\epsilon \max_{j,k} |f[\lambda_j, \lambda_k]|$. $\square$

Appendix B. Convergence of the conjugate gradient method. Given a linear system $Ax = b$ and a sequence of conjugate gradient iterates x_j , $j = 0, 1, \dots$, let

$$a_j = x - x_j \quad (\text{B.1})$$

be the error vector after j iterations. The error is bounded, see for example [23]:

$$\|a_j\|_2 \leq 2\sqrt{\kappa_2} \left(\frac{\sqrt{\kappa_2} - 1}{\sqrt{\kappa_2} + 1} \right)^j \|a_0\|_2. \quad (\text{B.2})$$

Here, κ_2 is the spectral condition number of A .

We would like to have a bound on the number of iterations needed to reach an accuracy

$$\|a_j\|_2 \leq \varepsilon \|a_0\|_2. \quad (\text{B.3})$$

Let

$$\varepsilon = 2\sqrt{\kappa_2} \left(\frac{\sqrt{\kappa_2} - 1}{\sqrt{\kappa_2} + 1} \right)^j. \quad (\text{B.4})$$

Then we get that

$$j = \frac{1}{2}\sqrt{\kappa_2} \frac{\log\left(\frac{2\sqrt{\kappa_2}}{\varepsilon}\right)}{\log\left(\left(\frac{\sqrt{\kappa_2}+1}{\sqrt{\kappa_2}-1}\right)^{\sqrt{\kappa_2}/2}\right)}. \quad (\text{B.5})$$

Since

$$\log\left(\left(\frac{\sqrt{\kappa_2}+1}{\sqrt{\kappa_2}-1}\right)^{\sqrt{\kappa_2}/2}\right) \geq 1 \quad (\text{B.6})$$

we have that

$$j \leq \frac{1}{2}\sqrt{\kappa_2} \log\left(\frac{2\sqrt{\kappa_2}}{\varepsilon}\right). \quad (\text{B.7})$$

In the recursive Fermi operator expansion, we need an accuracy such that the norm of the error matrix

$$\|E_i\| \leq \frac{\gamma}{2(n+1)m^{n-i}} \quad (\text{B.8})$$

in each recursive expansion iteration i , see (4.24).

For the sake of analysis, the set of linear systems to be solved in each recursive expansion iteration can be seen as a single system with a $N^2 \times N^2$ coefficient matrix consisting of N disconnected systems with different right hand sides. Then, we have that to fulfill (B.8) we need to reach an accuracy

$$\|E_i\| = \|a_j\|_2 \leq \varepsilon \|a_0\|_2 = \frac{\gamma}{2(n+1)m^{n-i}}. \quad (\text{B.9})$$

By (B.7) and (B.9) we have that the number of conjugate gradient iterations is bounded by

$$\frac{1}{2}\sqrt{\kappa_2}\log\left(2\sqrt{\kappa_2}\|a_0\|_2\frac{2(n+1)m^{n-i}}{\gamma}\right). \quad (\text{B.10})$$

Acknowledgements. I am grateful to Gunilla Kreiss, Anders M. N. Niklasson, Elias Rudberg, and Sara Zahedi who all in different ways have helped me during the preparation of the manuscript for this article. Support from the Swedish Research Council under Grant No. 623-2009-803 is gratefully acknowledged. The presented calculations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

REFERENCES

- [1] R. BAER AND M. HEAD-GORDON, *Chebyshev expansion methods for electronic structure calculations on large molecular systems*, J. Chem. Phys., 107 (1997), pp. 10003–10013.
- [2] M. BARRAULT, E. CANCÈS, W. W. HAGER, AND C. LE BRIS, *Multilevel domain decomposition for electronic structure calculations*, J. Comput. Phys., 222 (2007), pp. 86–109.
- [3] R. BARRETT, M. BERRY, T. F. CHAN, J. DEMMEL, J. DONATO, J. DONGARRA, V. EIJKHOUT, R. POZO, C. ROMINE, AND H. V. DER VORST, *Templates for the solution of linear systems: building blocks for iterative methods*, SIAM, Philadelphia, PA, 1994.
- [4] C. BEKAS, E. KOKIOPOULOU, AND Y. SAAD, *Computation of large invariant subspaces using polynomial filtered lanczos iterations with applications in density functional theory*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 397–418.
- [5] M. BENZI, P. BOITO, AND N. RAZOUK, *Decay properties of spectral projectors with applications to electronic structure*, Tech. Rep. 2010-025, Emory University Math/CS, November 2010.
- [6] M. BENZI, R. KOUHIA, AND M. TUMA, *Stabilized and block approximate inverse preconditioners for problems in solid and structural mechanics*, Comput. Meth. Appl. Mech. Eng., 190 (2001), pp. 6533–6554.
- [7] M. BENZI, C. D. MEYER, AND M. TUMA, *A sparse approximate inverse preconditioner for the conjugate gradient method*, SIAM J. Sci. Comput., 17 (1996), pp. 1135–1149.
- [8] G. BEYLKIN, N. COULT, AND M. J. MOHLENKAMP, *Fast spectral projection algorithms for density-matrix computations*, J. Comput. Phys., 152 (1999), pp. 32–54.
- [9] D. R. BOWLER AND M. J. GILLAN, *Density matrices in $O(N)$ electronic structure calculations: theory and applications*, Comp. Phys. Commun., 120 (1999), pp. 95–108.
- [10] D. R. BOWLER, T. MIYAZAKI, AND M. J. GILLAN, *Parallel sparse matrix multiplication for linear scaling electronic structure calculations*, Comp. Phys. Commun., 137 (2001), pp. 255–273.
- [11] ———, *Recent progress in linear scaling ab initio electronic structure techniques*, J. Phys.: Condens. Matter, 14 (2002), pp. 2781–2781.
- [12] K. BRANDHORST AND M. HEAD-GORDON, *Fast sparse Cholesky decomposition and inversion using nested dissection matrix reordering*, J. Chem. Theory Comput., 7 (2011), pp. 351–368.
- [13] M. CHALLACOMBE, *A simplified density matrix minimization for linear scaling self-consistent field theory*, J. Chem. Phys., 110 (1999), pp. 2332–2342.
- [14] ———, *A general parallel sparse-blocked matrix multiply for linear scaling SCF theory*, Comp. Phys. Commun., 128 (2000), p. 93.
- [15] J. D. CLOIZEAUX, *Energy bands and projection operators in a crystal: Analytic and asymptotic properties*, Phys. Rev., 135 (1964), pp. A685–A697.
- [16] C. DAVIS AND W. M. KAHAN, *Some new bounds on perturbation of subspaces*, Bull. Amer. Math. Soc., 75 (1969), pp. 863–868.
- [17] ———, *The rotation of eigenvectors by a perturbation. III*, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.
- [18] M. I. GIL’, *Perturbations of functions of diagonalizable matrices*, Electron. J. Linear Al., 20 (2010), p. 303.
- [19] S. GOEDECKER, *Decay properties of the finite-temperature density matrix in metals*, Phys. Rev. B, 58 (1998), pp. 3501–3502.

- [20] S. GOEDECKER, *Linear scaling electronic structure methods*, Rev. Mod. Phys., 71 (1999), pp. 1085–1123.
- [21] S. GOEDECKER AND L. COLOMBO, *Efficient linear scaling algorithm for tight-binding molecular dynamics*, Phys. Rev. Lett., 73 (1994), pp. 122–125.
- [22] S. GOEDECKER AND M. TETER, *Tight-binding electronic-structure calculations and tight-binding molecular dynamics with localized orbitals*, Phys. Rev. B, 51 (1995), pp. 9455–9464.
- [23] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations, third edition*, The Johns Hopkins University Press, Baltimore and London, 1996.
- [24] P. D. HAYNES AND M. C. PAYNE, *Corrected penalty-functional method for linear-scaling calculations within density-functional theory*, Phys. Rev. B, 59 (1999), pp. 12173–12176.
- [25] T. HELGAKER, H. LARSEN, J. OLSEN, AND P. JØRGENSEN, *Direct optimization of the AO density matrix in Hartree-Fock and Kohn-Sham theories*, Chem. Phys. Lett., 327 (2000), pp. 397–403.
- [26] N. J. HIGHAM, *Functions of matrices : theory and computation*, SIAM, Philadelphia, 2008.
- [27] N. D. M. HINE, P. D. HAYNES, A. A. MOSTOFI, AND M. C. PAYNE, *Linear-scaling density-functional simulations of charged point defects in Al_2O_3 using hierarchical sparse matrix algebra*, J. Chem. Phys., 133 (2010), p. 114111.
- [28] B. JANSÍK, S. HØST, P. JØRGENSEN, AND J. OLSEN, *Linear-scaling symmetric square-root decomposition of the overlap matrix*, J. Chem. Phys., 126 (2007), p. 124104.
- [29] C. KENNEY AND A. J. LAUB, *Rational iterative methods for the matrix sign function*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 273–291.
- [30] C. S. KENNEY AND A. J. LAUB, *A hyperbolic tangent identity and the geometry of Padé sign function iterations*, Numer. Algorithms, 7 (1994), pp. 111–128.
- [31] K. KITaura, E. IKEO, T. ASADA, T. NAKANO, AND M. UEBAYASI, *Fragment molecular orbital method: an approximate computational method for large molecules*, Chem. Phys. Lett., 313 (1999), pp. 701 – 706.
- [32] W. KOHN AND L. J. SHAM, *Self-consistent equations including exchange and correlation effects*, Phys. Rev., 140 (1965), pp. A1133–A1138.
- [33] X.-P. LI, R. W. NUNES, AND D. VANDERBILT, *Density-matrix electronic-structure method with linear system-size scaling*, Phys. Rev. B, 47 (1993), pp. 10891–10894.
- [34] L. LIN. private communication.
- [35] L. LIN, J. LU, R. CAR, AND W. E, *Multipole representation of the Fermi operator with application to the electronic structure analysis of metallic systems*, Phys. Rev. B, 79 (2009), p. 115133.
- [36] L. LIN, J. LU, L. YING, R. CAR, AND W. E, *Fast algorithm for extracting the diagonal of the inverse matrix with application to the electronic structure analysis of metallic systems*, Commun. Math. Sci., 7 (2009), pp. 755–777.
- [37] F. MAURI, G. GALLI, AND R. CAR, *Orbital formulation for electronic-structure calculations with linear system-size scaling*, Phys. Rev. B, 47 (1993), pp. 9973–9976.
- [38] D. A. MAZZIOTTI, *Towards idempotent reduced density matrices via particle-hole duality: Mcweeny’s purification and beyond*, Phys. Rev. E, 68 (2003), p. 066701.
- [39] R. McWEENY, *The density matrix in self-consistent field theory. I. Iterative construction of the density matrix*, Proc. R. Soc. London Ser. A, 235 (1956), pp. 496–509.
- [40] J. M. MILLAM AND G. E. SCUSERIA, *Linear scaling conjugate gradient density matrix search as an alternative to diagonalization for first principles electronic structure calculations*, J. Chem. Phys., 106 (1997), pp. 5569–5577.
- [41] K. NÉMETH AND G. E. SCUSERIA, *Linear scaling density matrix search based on sign matrices*, J. Chem. Phys., 113 (2000), pp. 6035–6041.
- [42] A. M. N. NIKLASSON, *Expansion algorithm for the density matrix*, Phys. Rev. B, 66 (2002), p. 155115.
- [43] ———, *Implicit purification for temperature-dependent density matrices*, Phys. Rev. B, 68 (2003), p. 233104.
- [44] ———, *Iterative refinement method for the approximate factorization of a matrix inverse*, Phys. Rev. B, 70 (2004), p. 193102.
- [45] ———, *A note on the Pulay force at finite electronic temperatures*, J. Chem. Phys., 129 (2008), p. 244107.
- [46] ———, *Density matrix methods in linear scaling electronic structure theory*, in Linear-scaling techniques in computational chemistry and physics, R. Zalesny, M. G. G. Papadopoulos, P. G. G. Mezey, and J. Leszczynski, eds., vol. 13 of Challenges and advances in computational chemistry and physics, Springer Netherlands, 2011, pp. 439–473.
- [47] D. P. O’LEARY, *The block conjugate gradient algorithm and related methods*, Linear Algebra Appl., 29 (1980), pp. 293 – 322.

- [48] P. ORDEJÓN, D. A. DRABOLD, M. P. GRUMBACH, AND R. M. MARTIN, *Unconstrained minimization approach for electronic computations that scales linearly with system size*, Phys. Rev. B, 48 (1993), pp. 14646–14649.
- [49] A. H. R. PALSER AND D. E. MANOLOPOULOS, *Canonical purification of the density matrix in electronic-structure theory*, Phys. Rev. B, 58 (1998), pp. 12704–12711.
- [50] M. S. PATERSON AND L. STOCKMEYER, *On the number of nonscalar multiplications necessary to evaluate matrix polynomials*, SIAM J. Comput., 2 (1973), pp. 60–66.
- [51] A. D. RABUCK AND G. E. SCUSERIA, *Improving self-consistent field convergence by varying occupation numbers*, J. Chem. Phys., 110 (1999), pp. 695–700.
- [52] E. H. RUBENSSON, *Nonmonotonic recursive polynomial expansions for linear scaling calculation of the density matrix*, J. Chem. Theory Comput., 7 (2011), pp. 1233–1236.
- [53] E. H. RUBENSSON, N. BOCK, E. HOLMSTRÖM, AND A. M. N. NIKLASSON, *Recursive inverse factorization*, J. Chem. Phys., 128 (2008), p. 104105.
- [54] E. H. RUBENSSON AND E. RUDBERG, *Bringing about matrix sparsity in linear scaling electronic structure calculations*, J. Comput. Chem., 32 (2011), pp. 1411–1423.
- [55] E. H. RUBENSSON, E. RUDBERG, AND P. SALEK, *A hierarchic sparse matrix data structure for large-scale Hartree-Fock/Kohn-Sham calculations*, J. Comput. Chem., 28 (2007), pp. 2531–2537.
- [56] ———, *Density matrix purification with rigorous error control*, J. Chem. Phys., 128 (2008), p. 074106.
- [57] ———, *Rotations of occupied invariant subspaces in self-consistent field calculations*, J. Math. Phys., 49 (2008), p. 032103.
- [58] ———, *Truncation of small matrix elements based on the Euclidean norm for blocked data structures*, J. Comput. Chem., 30 (2009), pp. 974–977.
- [59] E. RUDBERG, E. H. RUBENSSON, AND P. SALEK, *Ergo (version 3.50); a program for large-scale electronic structure calculations*, 2011.
- [60] ———, *Kohn-Sham density functional theory electronic structure calculations with linearly scaling computational time and memory usage*, J. Chem. Theory Comput., 7 (2011), p. 340.
- [61] Y. SAAD, J. R. CHELIKOWSKY, AND S. M. SHONTZ, *Numerical methods for electronic structure calculations of materials*, SIAM Review, 52 (2010), pp. 3–54.
- [62] C. SARAVANAN, Y. SHAO, R. BAER, P. N. ROSS, AND M. HEAD-GORDON, *Sparse matrix multiplications for linear-scaling electronic structure calculations in an atom-centered basis set using multiatom blocks*, J. Comput. Chem., 24 (2003), pp. 618–622.
- [63] S. SCHWEIZER, J. KUSSMANN, B. DOSER, AND C. OCHSENFELD, *Linear-scaling cholesky decomposition*, J. Comput. Chem., 29 (2008), pp. 1004–1010.
- [64] G. W. STEWART AND J. SUN, *Matrix perturbation theory*, Academic Press, San Diego, 1990.
- [65] A. SZABO AND N. S. OSTLUND, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, Dover Publications, Inc., Mineola, New York, 1996.
- [66] L.-W. WANG, *Novel computational methods for nanostructure electronic structure calculations*, Annu. Rev. Phys. Chem., 61 (2010), pp. 19–39.
- [67] V. WEBER AND J. HUTTER, *A smooth l_1 -norm sparseness function for orbital based linear scaling total energy minimization*, J. Chem. Phys., 128 (2008), p. 064107.
- [68] W. YANG, *Direct calculation of electron density in density-functional theory*, Phys. Rev. Lett., 66 (1991), pp. 1438–1441.