Exploiting Localization in Matrix Computations
IV. Locality in Electronic Structure Computations

Michele Benzi

Department of Mathematics and Computer Science
Emory University

Atlanta, Georgia, USA

CIME-EMS Summer School in Applied Mathematics

Exploiting Hidden Structure in Matrix Computations:
Theory and Applications

Cetraro, 22-26 June, 2015

Outline

Outline

1. Motivation
2. The electronic structure problem
3. Density matrices
4. $\mathcal{O}(N)$ methods
5. Numerical examples

Motivation

The main purpose of this lecture is to present a sound theoretical foundation for a class of $\mathcal{O}(N)$ methods that are being developed by computational physicists and chemists for the solution of the N-body electronic structure problem. This problem is fundamental to quantum chemistry, solid state physics, material science, biology, etc.

The treatment is based on our general theory of decay in the entries of functions of sparse matrices. In particular, we study the asymptotic behavior of the off-diagonal matrix elements for $N \to \infty$.

While this work is primarily theoretical, our theory is being used to develop better algorithms for electronic structure computations.

Acknowledgements

- Paola Boito (U. of Limoges)
- Nader Razouk (former PhD student)
- Thanks to Matt Challacombe (T-Division, Los Alamos National Lab)

Main reference:

M. Benzi, P. Boito and N. Razouk, *Decay properties of spectral projectors with applications to electronic structure*, SIAM Rev., 55 (2013), pp. 3–64.

Outline

1. Motivation
2. The electronic structure problem
3. Density matrices
4. $\mathcal{O}(N)$ methods
5. Numerical examples

The Electronic Structure Problem

A fundamental problem in quantum chemistry and solid state physics is to determine the electronic structure of (possibly large) atomic and molecular systems. Knowledge of the electronic structure allows scientists to predict many of the properties of various substances and materials under different conditions.

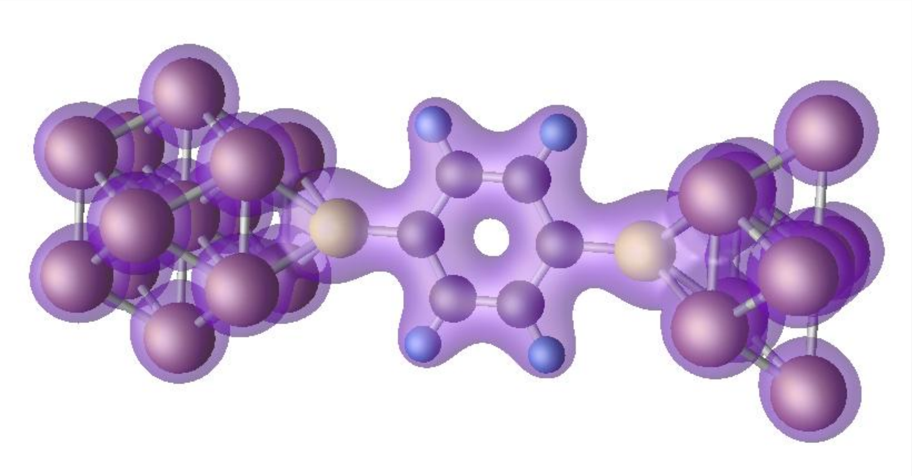
The problem amounts to computing the ground state (smallest eigenvalue and corresponding eigenfunction) of the many-body quantum-mechanical Hamiltonian (Schrödinger operator), $\mathcal{H}$.

Variationally, we want to minimize the Rayleigh quotient:

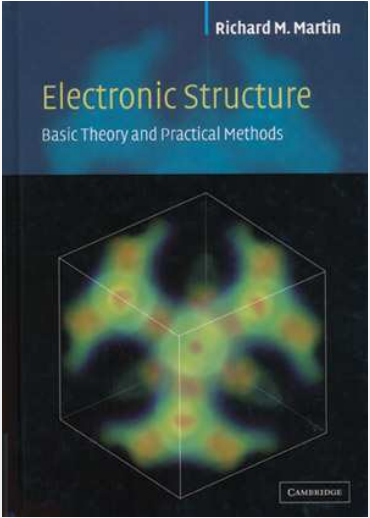
$$E_0 = \min_{\Psi \neq 0} \frac{\langle \mathcal{H}\Psi, \Psi \rangle}{\langle \Psi, \Psi \rangle} \quad \text{and} \quad \Psi_0 = \operatorname{argmin}_{\Psi \neq 0} \frac{\langle \mathcal{H}\Psi, \Psi \rangle}{\langle \Psi, \Psi \rangle}$$

where $\langle \cdot, \cdot \rangle$ denotes the L^2 inner product. Note that $\mathcal{H}$ is a self-adjoint, unbounded operator.

The Electronic Structure Problem (cont.)



The Electronic Structure Problem (cont.)



The Electronic Structure Problem (cont.)

In the Born-Oppenheimer approximation, the many-body Hamiltonian (in atomic units) is given by

$$\mathcal{H} = \sum_{i=1}^{N} \left(-\frac{1}{2}\Delta_i - \sum_{j=1}^{M} \frac{Z_j}{|\mathbf{x}_i - \mathbf{r}_j|} + \sum_{j \neq i}^{N} \frac{1}{|\mathbf{x}_i - \mathbf{x}_j|} \right)$$

where $\mathbf{r}_j =$ position of the jth nucleus, $\mathbf{x}_i =$ position of the ith electron, $N =$ number of electrons and $M =$ number of nuclei in the system.

The operator $\mathcal{H}$ acts on a suitable subspace $D(\mathcal{H}) \subset L^2(\mathbb{R}^{3N})$, consisting of antisymmetric wavefunctions:

$$\Psi(\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_N) = -\Psi(\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N).$$

This is because electrons are Fermions and therefore subject to Pauli's Exclusion Principle.

Note: To simplify notation, spin is ignored here. It can be easily incorporated into the formulas.

Two pioneers of quantum physics

Erwin Schrödinger (1887-1961) and Wolfgang Pauli (1900-1958)

The Electronic Structure Problem (cont.)

Unless N is very small, the curse of dimensionality makes this problem intractable.
In order to make the problem tractable various approximations have been introduced, most notably:

- Wavefunction methods (e.g., Hartree-Fock)
- Density Functional Theory (e.g., Kohn-Sham; Nobel Prize, 1998)
- Hybrid methods

In these approximations the original, linear eigenproblem $\mathcal{H}\Psi = E\Psi$ for the many-electrons Hamiltonian is replaced by a nonlinear one-particle eigenproblem of the form

$$\hat{H}(\psi_i) = \lambda_i \psi_i, \quad \langle \psi_i, \psi_j \rangle = \delta_{ij}, \quad 1 \leq i, j \leq N$$

where $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$. This problem is nonlinear because the operator $\hat{H}$ depends nonlinearly on the ψ_i.

The Electronic Structure Problem (cont.)

Informally speaking, in DFT the idea is to consider a “pseudo-particle" (single electron) moving in the electric field generated by the nuclei and by some average distribution of the other electrons. Starting with an initial guess of the charge density, a potential is formed and the corresponding one-particle eigenproblem is solved; the resulting charge density is used to define the new potential, and so on until the charge density no longer changes appreciably.

More formally, DFT reformulates the problem so that the unknown function is the electronic density

$$\rho(\mathbf{x}) = N \int_{\mathbb{R}^{3(N-1)}} |\Psi(\mathbf{x}, \mathbf{x}_2, \ldots, \mathbf{x}_N)|^2 \mathrm{d}\mathbf{x}_2 \cdots \mathrm{d}\mathbf{x}_N \,,$$

a scalar field on $\mathbb{R}^3$.

The function ρ minimizes a certain functional on $H^1(\mathbb{R}^3)$, the exact form of which is not known explicitly—this leads to various approximations.

The Electronic Structure Problem (cont.)

Various forms of the density functional have been proposed, the most successful being the Kohn-Sham model:

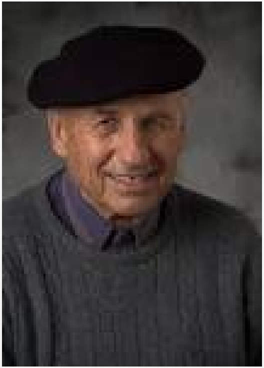
$$I_{KS}(\rho) = \inf \left\{ T_{KS} + \int_{\mathbb{R}^3} \rho V \, \mathrm{d}\mathbf{x} + \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(\mathbf{x})\rho(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} + E_{xc}(\rho) \right\},$$

where $\rho(\mathbf{x}) = \sum_{i=1}^{N} |\psi_i(\mathbf{x})|^2$, $T_{KS} = \frac{1}{2}\sum_{i=1}^{N} \int_{\mathbb{R}^3} |\nabla \psi_i|^2 \, \mathrm{d}\mathbf{x}$ is the kinetic energy term, V denotes the Coulomb potential, and E_{xc} denotes the exchange term that takes into account the interaction between electrons.

The infimum above is taken over all functions $\psi_i \in H^1(\mathbb{R}^3)$ such that $\langle \psi_i, \psi_j \rangle = \delta_{ij}$, where $1 \leq i, j \leq N$ and $\sum_{i=1}^{N} |\psi_i(\mathbf{x})|^2 = \rho$.

This I_{KS} is minimized with respect to ρ. Note that ρ, being the electron density, must satisfy $\rho > 0$ and $\int_{\mathbb{R}^3} \rho \, \mathrm{d}\mathbf{x} = N$.

Founders of Density Functional Theory



Walter Kohn (b. 1923) and Lu Jeu Sham (b. 1938)

The Electronic Structure Problem (cont.)

The Euler–Lagrange equations for the corresponding variational problem are the Kohn–Sham equations:

$$\hat{H}(\rho)\psi_i = \lambda_i \psi_i, \quad \langle \psi_i, \psi_j \rangle = \delta_{ij} \quad (1 \leq i, j \leq N)$$

with

$$\hat{H}(\rho) = -\frac{1}{2}\Delta + \hat{V}(\mathbf{x}, \rho)$$

where $\hat{V}$ denotes a (complicated) potential, and $\rho = \sum_{i=1}^{N} |\psi_i(\mathbf{x})|^2$.

Hence, the original intractable, linear eigenproblem for the many-body Hamiltonian $\mathcal{H}$ is reduced to tractable, nonlinear eigenproblem for the single-particle Hamiltonian $\hat{H}$.

But how do we solve the latter?

The Electronic Structure Problem (cont.)

The nonlinear problem can be solved by a ‘self-consistent field’ (SCF) iteration, leading to a sequence of linear eigenproblems

$$\hat{H}^{(k)}\psi_i^{(k)} = \lambda_i^{(k)}\psi_i^{(k)}, \quad \langle \psi_i^{(k)}, \psi_j^{(k)} \rangle = \delta_{ij}, \quad k = 1, 2, \ldots$$

($1 \le i, j \le N$), where each $\hat{H}^{(k)} = -\frac{1}{2}\Delta + \hat{V}^{(k)}$ is a one-electron linearized Hamiltonian:

$$\hat{V}^{(k)} = \hat{V}^{(k)}(\mathbf{x}, \rho^{(k-1)}), \quad \rho^{(k-1)} = \sum_{i=1}^{N} |\psi_i^{(k-1)}(\mathbf{x})|^2.$$

For insulators (non-metallic systems), convergence is usually fast.

Solution of each of the (discretized) linear eigenproblems above leads to a typical $\mathcal{O}(N^3)$ cost per SCF iteration. This is a major bottleneck for large systems.

The Electronic Structure Problem (cont.)

Fortunately, however, the actual eigenpairs $(\psi_i^{(k)}, \lambda_i^{(k)})$ are not needed, and diagonalization of the one-particle Hamiltonians can be avoided!

Indeed, at each SCF iteration one can compute instead the orthogonal projector P onto the occupied subspace

$$\mathcal{V}_{occ} = \mathsf{span}\{\psi_1, \ldots, \psi_N\}$$

corresponding to the N lowest eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$ of the current linearized, one-particle Hamiltonian.

All quantities of interest in electronic structure theory can be computed from P. For example, the total energy of the system can be computed once P is known, as well as the forces acting on the atoms.

Note: From here on, we focus on the computation of P within a single SCF iteration, and we use H to denote the corresponding linear, one-particle Hamiltonian operator or its discretization.

Outline

1. Motivation
2. The electronic structure problem
3. Density matrices
4. $\mathcal{O}(N)$ methods
5. Numerical examples

Density matrices

In quantum physics the spectral projector P is called the density operator corresponding to the occupied states of the system (von Neumann, 1927).

In actual computations, the single-particle Hamiltonians are replaced by matrices by Rayleigh-Ritz projection onto a finite-dimensional subspace spanned by a set of basis functions $\{\phi_i\}_{i=1}^{n}$, where n is a multiple of N. Typically, $n = C \cdot N$ where $C \geq 2$ is a moderate integer when linear combinations of GTOs (Gaussian-type orbitals) are used:

$$\phi(\mathbf{x}) = \phi\,(x, y, z) = C\,x^{n_x} y^{n_y} z^{n_z} \mathsf{e}^{-\alpha r^2}\,, \quad r = \sqrt{x^2 + y^2 + z^2}$$

(here C is a normalization constant). These basis functions are localized and this makes the discrete Hamiltonians essentially sparse.

Finite difference and finite element ("real space") methods, while also used, are less popular in chemistry. In some codes, plane waves are used. These often lead to full matrices.

Father of density matrix theory

John von Neumann (1903-1957)

Orthogonal vs. non-orthogonal representations

The Gramian matrix $S = (S_{ij})$, where

$$S_{ij} = \langle \phi_i, \phi_j \rangle = \int_{\mathbb{R}^3} \phi_i(\mathbf{x}) \phi_j(\mathbf{x}) \, \mathrm{d}\mathbf{x} \,,$$

is called the overlap matrix in electronic structure.

It is a dense matrix, but its entries fall off very rapidly for increasing separation. In the case of GTOs for a 1D system:

$$|S_{ij}| \approx \mathrm{e}^{-|i-j|^2}, \quad 1 \le i, j \le n \,.$$

Neglecting tiny entries leads to banded overlap matrices.

The density matrix P is the S-orthogonal projector onto the occupied subspace $\mathcal{V}_{occ}$. It is often computationally convenient to make a change of basis, from non-orthogonal to orthogonal.

Orthogonal vs. non-orthogonal representations (cont.)

Trasforming the Hamiltonian H to an orthogonal basis means performing a congruence trasformation:

$$H \to ZHZ^T, \quad \text{where } Z \text{ is such that } ZZ^T = S^{-1}.$$

It is important to observe that in actual computations it is not necessary to compute the product ZHZ^T explicitly, it can be kept in factored form. Also note that since S is banded and SPD, the entries of S^{-1}, and therefore of Z, exhibit exponential decay.

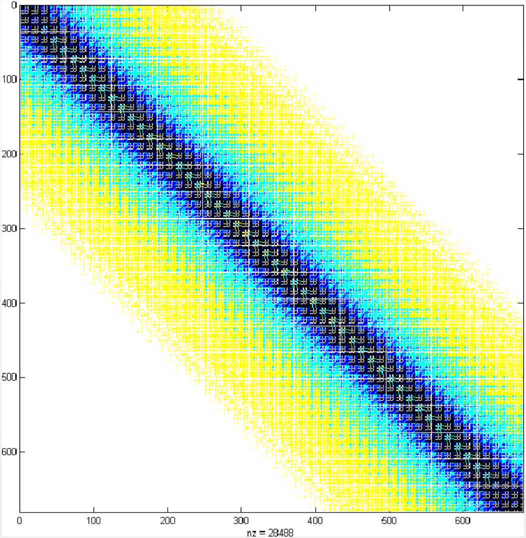
Common choices for Z are the inverse Cholesky factor of S or the inverse square root $S^{-1/2}$ (this is known as Löwdin orthogonalization).

Sparse approximations to the inverse Cholesky factor Z of S can be computed efficiently, e.g., using the AINV algorithm.

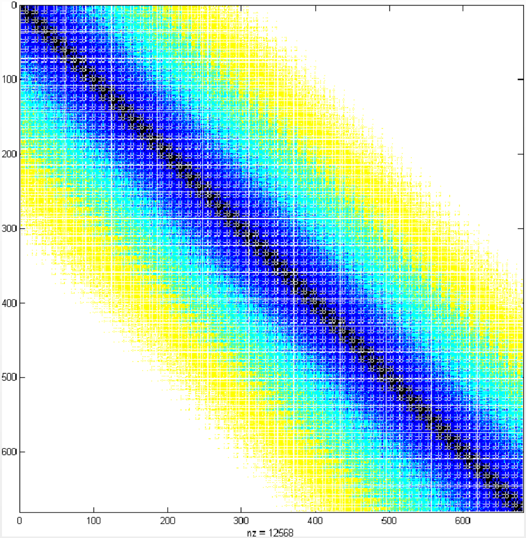
M. Benzi, C. Meyer and M. Tuma, SIAM J. Sci. Comput., 17:1135–1149, 1996.

M. Benzi and M. Tuma, SIAM J. Sci. Comput., 21:1851–1868, 2000.

Example: Hamiltonian for $C_{52}H_{106}$, GTO basis.



Example: Hamiltonian for $C_{52}H_{106}$, orthogonal basis



Density matrices (cont.)

Summarizing, the core computational task in electronic structure theory is the calculation of P, the spectral projector onto the subspace spanned by the N lowest eigenfunctions of H (occupied states):

$$P = \psi_1 \otimes \psi_1 + \cdots + \psi_N \otimes \psi_N = |\psi_1\rangle\langle\psi_1| + \cdots + |\psi_N\rangle\langle\psi_N|$$

where N is the number of electrons and $H\psi_i = \lambda_i \, \psi_i, \ i = 1, \ldots, N$.

Note that we can write $P = h(H)$ where h is the Heaviside (step) function

$$h(x) = \begin{cases} 1 & \text{if} \quad x < \mu \\ \frac{1}{2} & \text{if} \quad x = \mu \\ 0 & \text{if} \quad x > \mu \end{cases}$$

with $\lambda_N < \mu < \lambda_{N+1}$ (μ is the "Fermi level").

Outline

1. Motivation
2. The electronic structure problem
3. Density matrices
4. $\mathcal{O}(N)$ methods
5. Numerical examples

$\mathcal{O}(N)$ methods

Physicists have observed (long ago!) that the entries of the density matrix P decay away from the main diagonal. The decay rate is algebraic for metallic systems, and exponential (or faster) for insulators and also for semiconductors.

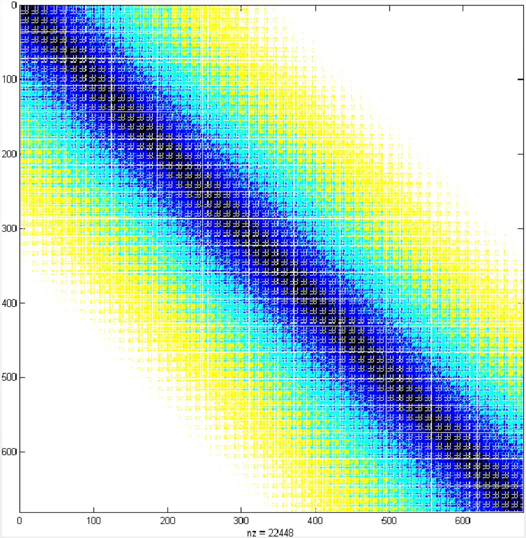
Hence, the density matrix is localized. This property has been called nearsightedness of electronic matter by W. Kohn, since each part of the system can only “see" nearby parts.

In other words, long-range correlations are absent, and local disturbances do not propagate too far.

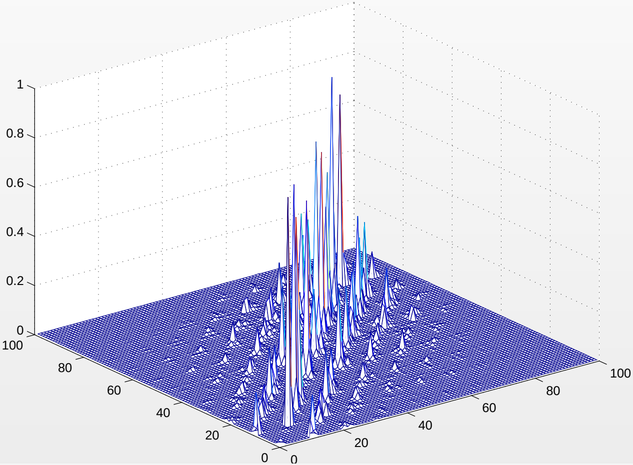
Nearsightedness implies that in the bulk limit ($N \to \infty$ while keeping the particle density constant) the number of entries P_{ij} with $|P_{ij}| > \varepsilon$ grows only linearly with N (and therefore with n), for any prescribed $\varepsilon > 0$.

For insulators and semiconductors, this makes $\mathcal{O}(N)$ methods possible.

Example: Density matrix for $C_{52}H_{106}$, orthogonal basis



Example: Plot of $|P_{ij}|$ where P is the density matrix for $H = -\frac{1}{2}\Delta + V$, random V, finite differences (2D lattice), $N = 10$ occupied states ("Anderson model").



$\mathcal{O}(N)$ methods (cont.)

The fact that for many systems of physical interest P is very nearly sparse allows the development of $\mathcal{O}(N)$ approximation methods. Some popular methods are:

1. Chebyshev expansion: $P = h(H) \approx \frac{c_0}{2} I + \sum_{k=1}^{n} c_k T_k(H)$
2. Rational expansions based on contour integration:

$$P = \frac{1}{2\pi i} \int_{\Gamma} (zI - H)^{-1} \mathrm{d}z \approx \sum_{k=1}^{q} w_k (z_k I - H)^{-1}$$

3. Density matrix minimization:

$$\mathrm{Tr}(PH) = \min, \quad \text{subject to} \quad P = P^* = P^2 \quad \text{and} \quad \mathrm{rank}(P) = N$$

4. Etc. (see, e.g.: C. Le Bris, *Acta Numerica*, 2005; Bowler & Miyazaki, *Rep. Progr. Phys.*, 2012)

All these methods can achieve $\mathcal{O}(N)$ scaling by exploiting "sparsity."

$\mathcal{O}(N)$ methods (cont.)

As mentioned, the fact that the density matrix P is exponentially localized means that for any $\varepsilon > 0$, P contains only $\mathcal{O}(N)$ elements with $|P_{ij}| > \varepsilon$, as $N \to \infty$.

It is possible to show that dropping small entries from P results in controllable errors in the quantities of interests (e.g., the total energy).

Furthermore, having *a priori* decay bounds means being able to identify the positions of the nonnegligible off-diagonal elements of P ("cut-off rules"). Hence, only $\mathcal{O}(N)$ elements of P need to be computed.

These two observations are the key to the development of linear scaling methods.

A mathematical foundation for $\mathcal{O}(N)$ methods

Until recently, one could not find in the literature any rigorously proved, general results on the decay behavior in P. Decay estimates are either non-rigorous (at least for a mathematician!), or valid only for very special cases. Some physicists even expressed doubts that purely mathematical arguments could be used to explain nearsightedness!

Using our decay results for matrix functions, one can obtain rigorous estimates for the rate of decay in the off-diagonal entries of P in the form of upper bounds on the entries of the density matrix, uniform in N.

Ideally, these bounds can be used to provide cut-off rules, that is, to decide which entries in P need not be computed, given a prescribed error tolerance.

As a result, the possibility of $\mathcal{O}(N)$ methods for gapped systems is now rigorously justified.

Intermezzo: two quotes from a physicist...

> *To obtain a linear scaling, the extended orbitals [i.e., the eigenfunctions of the one-particle Hamiltonian corresponding to occupied states] have to be replaced by the density matrix, whose physical behavior can be exploited to obtain a fast algorithm. This last point is essential. Mathematical and numerical analyses alone are not sufficient to construct a linear algorithm. They have to be combined with physical intuition.*

S. Goedecker, *Low complexity algorithms for electronic structure calculations*,
J. Comp. Phys., 118 (1995), pp. 261–268.

Intermezzo: two quotes from a physicist...

> *Even though* $\mathcal{O}(N)$ *algorithms contain many aspects of mathematics and computer science they have, nevertheless, deep roots in physics. Linear scaling is not obtainable by purely mathematical tricks, but it is based on an understanding of the concept of locality in quantum mechanics.*

S. Goedecker, *Linear scaling electronic structure methods*, Rev. Mod. Phys., 71 (1999), pp. 1085–1123.

… and one from a mathematician

> *The latter assumption [the exponential localization of the density matrix] is in some sense an* a posteriori *assumption, and not easy to analyse […] It is to be emphasized that the numerical analysis of the linear scaling methods overviewed above that would account for cut-off rules and locality assumptions, is not yet available.*

C. Le Bris, *Computational chemistry from the perspective of numerical analysis*, Acta Numer., 14 (2005), pp. 363–444.

A mathematical foundation for $\mathcal{O}(N)$ methods (cont.)

We begin by formalizing the problem at hand.

Let C be a fixed positive integer and let $n := C \cdot N$, where $N \to \infty$. We think of C as the number of basis functions per electron while N is the number of electrons.

Definition: A sequence of discrete Hamiltonians is a sequence of $n \times n$ Hermitian matrices $\{H_N\}$ such that

1. The matrices H_N have spectra uniformly bounded w.r.t. N: up to shifting and scaling, we can assume $\sigma(H_N) \subset [-1, 1]$ for all N.
2. The H_N are banded, with uniformly bounded bandwidth as $N \to \infty$. More generally, the H_N are sparse with maximum number of nonzeros per row bounded w.r.t. N.

These assumptions model the fact that the Hamiltonians have finite interaction range.

A mathematical foundation for $\mathcal{O}(N)$ methods (cont.)

Next, let H be a generic discrete Hamiltonian of order n. Denote the eigenvalues of H as

$$-1 \leq \lambda_1 \leq \ldots \leq \lambda_N < \lambda_{N+1} \leq \ldots \leq \lambda_n \leq 1 \, .$$

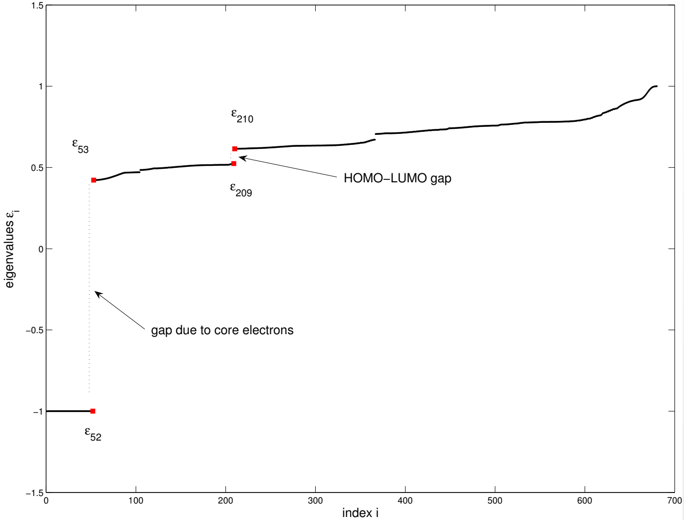
The spectral gap is then $\gamma_N = \lambda_{N+1} - \lambda_N$. In quantum chemistry this is known as the HOMO-LUMO gap, in solid state physics as the band gap.

Two cases are possible:

1. There exists $\gamma > 0$ such that $\gamma_N \geq \gamma$ for all N.
2. $\inf_N \gamma_N = 0$.

The first case corresponds to insulators and semiconductors, the second one to metallic systems.

Example: Spectrum of the Hamiltonian for $C_{52}H_{106}$.



Exponential decay in the density matrix for gapped systems

Our main result is the following

Theorem

Let $\{H_N\}$ be a sequence of discrete Hamiltonians of size $n = C \cdot N$, with C constant and $N \to \infty$. Let P_N denote the spectral projector onto the N occupied states associated with H_N. If there exists $\gamma > 0$ such that the gaps $\gamma_N \geq \gamma$ for all N, then there exists constants K and α such that

$$|[P_N]_{ij}| \leq K\,\mathrm{e}^{-\alpha d_N(i,j)} \quad (1 \leq i, j \leq N),$$

where $d_N(i,j)$ denotes the distance between node i and node j in the graph G_N associated with H_N. The constants K and α depend only on the gap γ (not on N) and are easily computable, with $\alpha = \mathcal{O}(\gamma)$ as $\gamma \to 0$.

Note: Explicit expressions for K and α can be found in the cited SIAM Review article.

Exponential decay in the density matrix for gapped systems (cont.)

Proof (sketch):

Recall that the density matrix P can be expressed as $P = h(H)$ where $h(x)$ is the step function that is equal to 1 on $[-1, \mu)$ and equal to 0 on $(\mu, 1]$, with μ in the gap between λ_N and λ_{N+1}. If the gap does not vanish as $N \to \infty$, then it is is possible to approximate h with an arbitrarily small error on $[-1, \lambda_N]$ and $[\lambda_{N+1}, 1]$ (uniformly in N!) by an analytic function $f(x)$. Since H has no eigenvalues in the gap, it follows that $\|P - f(H)\|_2$ can be made smaller than any prescribed quantity by a suitable choice of f (smooth). The decay results for $f(H)$ then imply that the entries in P decay at an exponential (or faster) rate.

Hence, the decay in the density matrix follows from our general theory of decay in the entries of analytic functions of sparse matrices.

Analytic approximations of the step function

If μ (the "Fermi level") is in the gap, $\lambda_N < \mu < \lambda_{N+1}$, the step function can be approximated by the Fermi-Dirac function:

$$h(x) = \lim_{\beta\to\infty} f_{FD}(x), \quad \text{where} \quad f_{FD}(x) = \frac{1}{1 + \mathrm{e}^{\beta(x-\mu)}}.$$

Here β can be interpreted as an inverse temperature, $\beta = (\kappa_B T)^{-1}$.

Other approximations of the step function are also in use, such as

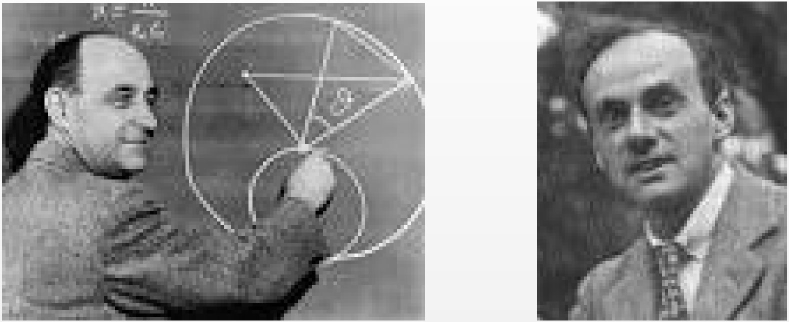
$$h(x) = \lim_{\beta\to\infty} \left[\frac{1}{2} + \frac{1}{\pi}\tan^{-1}(\beta\pi(x-\mu))\right] ,$$

$$h(x) = \lim_{\beta\to\infty} \mathrm{erfc}\left(-\beta(x-\mu)\right) ,$$

or

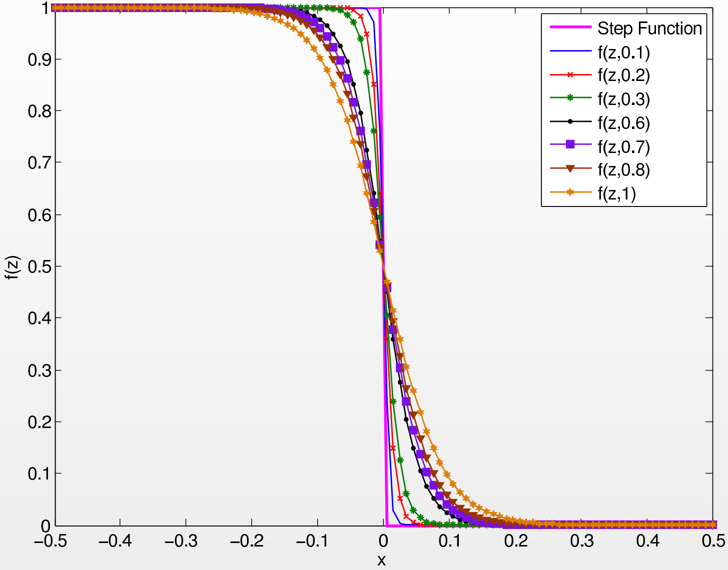
$$h(x) = \lim_{\beta\to\infty} \left[1 + \tanh(\beta((x-\mu))\right] .$$

Co-discoverers of the Fermi–Dirac distribution



Enrico Fermi (1901-1954) and Paul Adrien Maurice Dirac (1902-1984)

Fermi-Dirac approximation of step function ($\mu = 0$)



Fermi-Dirac approximation of step function

The actual choice of β in the Fermi-Dirac function is dictated by the size of the gap γ and by the approximation error.

What is important is that for any prescribed error, there is a maximum value of β that achieves the error for all N, provided the system has non-vanishing gap.

On the other hand, for metallic systems $\gamma \to 0$, hence $\beta \to \infty$ and the bounds blow up. Indeed, $f_{FD}(z)$ has poles at $z = \mu \pm \pi \mathbf{i}/\beta$, hence the distance from the poles of f to the spectrum tends to zero as $\beta \to \infty$.

There is still decay in the density matrix, but algebraic rather than exponential. Simple examples show it can be as slow as $\mathcal{O}(|i-j|^{-1})$.

In practice, γ is either known experimentally or can be estimated by computing the eigenvalues of a moderate-size Hamiltonian.

Dependence of decay rate on the spectral gap and on the temperature

In the physics literature, there has been some controversy on the precise dependence of the inverse correlation length α in the decay estimate

$$|[P_N]_{ij}| \leq c \cdot \mathrm{e}^{-\alpha\, d_N(i,j)}$$

on the spectral gap γ (for insulators) and on the electronic temperature T (for metals at positive temperature).

Our theory gives the following results:

1. $\alpha = c\gamma + \mathcal{O}(\gamma^3)$, for $\gamma \to 0+$ and $T = 0$;
2. $\alpha = \pi\kappa_B T + \mathcal{O}(T^3)$, for $T \to 0+$ (indep. of γ).

These asymptotics are in agreement with experimental and numerical results, as well as with physical intuition.

Outline

1. Motivation
2. The electronic structure problem
3. Density matrices
4. $\mathcal{O}(N)$ methods
5. Numerical examples

Decay bounds for the Fermi-Dirac approximation

Assume that H is m-banded and has spectrum in $[-1, 1]$, then

$$\left| \left[\left(I + \mathrm{e}^{\beta(H-\mu I)} \right)^{-1} \right]_{ij} \right| \leq K \mathrm{e}^{-\alpha|i-j|} \equiv K\, \lambda^{\frac{|i-j|}{m}}.$$

Note that K, λ depend only on β. In turn, β depends on γ and on the desired accuracy.
We have

$$\gamma \to 0^+ \quad \Rightarrow \quad \lambda \to 1^-$$

and

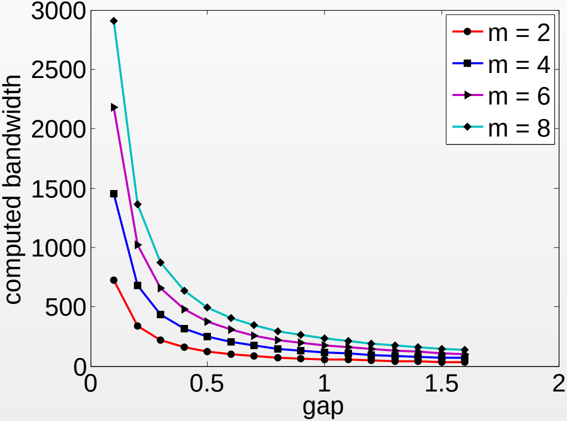
$$\gamma \to 1 \quad \Rightarrow \quad \lambda \to 0.872.$$

We choose β and $\hat{m}$ so as to guarantee an error $\|P - f_{FD}(H)\|_2 < 10^{-6}$.

We can regard γ^{-1} as a measure of the difficulty of the problem. It is also a condition number for the sensitivity of P to perturbations in H.

Computed bandwidth for approximations of P

$$f_{FD}(x) = \frac{1}{1 + \mathrm{e}^{\beta(x-\mu)}}$$



Approximation of $f_{FD}(H)$ by Chebyshev polynomials

Algorithm (Goedecker & Colombo, 1994) ▸ More

- We compute approximations of $f_{FD}(H)$ using Chebyshev polynomials
  - The degree of the polynomial can be estimated a priori
  - The coefficients of the polynomial can be pre-computed (indep. of N)
  - Estimates for the extreme eigenvalues of H are required
- The polynomial expansion is combined with a procedure that a priori determines a bandwidth or sparsity pattern for $f_{FD}(H)$ outside which the elements are so small that they can be neglected

Cost

This method is multiplication-rich; the matrices are kept sparse throughout the computation, hence $\mathcal{O}(N)$ arithmetic and storage requirements. The matrix polynomials can be efficiently evaluated by the Paterson-Stockmeyer algorithm.

Chebyshev expansion of Fermi-Dirac function

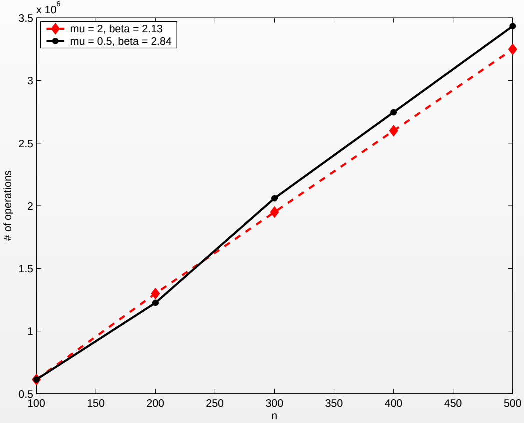
The bandwidth was computed prior to the calculation to be ≈ 20; here H is tridiagonal (toy example).

Table: Results for $f_{FD}(x) = \frac{1}{1+\mathrm{e}^{(\beta(x-\mu))}}$

	$\mu = 2,\ \beta = 2.13$			$\mu = 0.5,\ \beta = 1.84$		
N	error	k	$\hat{m}$	error	k	$\hat{m}$
100	$9e{-}06$	18	20	$6e{-}06$	18	22
200	$4e{-}06$	19	20	$9e{-}06$	18	22
300	$4e{-}06$	19	20	$5e{-}06$	20	22
400	$6e{-}06$	19	20	$8e{-}06$	20	22
500	$8e{-}06$	19	20	$8e{-}06$	20	22

Note: In the table, ‘error’ means relative error in the Frobenius norm.

Computation of Fermi-Dirac function



The $\mathcal{O}(N)$ behavior of Chebyshev's approximation to the Fermi–Dirac function $f_{FD}(H) = (\exp(\beta(H - \mu I)) + I)^{-1}$.

Chebyshev expansion of entropy-like function

Linear scaling is observed also for other matrix functions.

Examples: results for $H = H_N$ tridiagonal, SPD, $f(x) = x \ln x$

	$H \ln H$	Tr $(H \ln H)$		
N	rel. error	error	$\hat{m}$	k
100	$5e{-}07$	$3e{-}04$	20	9
200	$6e{-}07$	$8e{-}04$	20	9
300	$1e{-}07$	$3e{-}04$	20	10
500	$2e{-}07$	$5e{-}04$	20	10

In the Table, $\hat{m}$ is the estimated bandwidth and k is the number of terms in the Chebyshev expansion. Note the $\mathcal{O}(N)$ behavior in terms of cost.

Summary

- ‘Gapped’ systems, like insulators, exhibit strong localization
- Localization in $P = h(H)$, when present, can lead to fast approximation algorithms
- Our decay bounds for density matrices depend only on the gap γ and on the sparsity of H; they are, in a sense, universal
- These bounds can be useful in determining appropriate sparsity patterns (or bandwidths) that capture the nonnegligible entries in $P = h(H)$
- Constants in $\mathcal{O}(N)$ algorithms can be large. In practice, to beat sub-optimal, $\mathcal{O}(N^p)$ algorithms ($p = 2, 3$) algorithms N must be quite large.

Chebyshev approximation

For H with $\sigma(H) \subset [-1, 1]$ the Chebyshev polynomials are given by

$$T_{k+1}(H) = 2HT_k(H) - T_{k-1}(H),\ T_1(H) = H,\ T_0(H) = I.$$

Then $f(H)$ can be represented in a series of the form

$$f(H) = \sum_{k=0}^{\infty} c_k T_k(H).$$

The coefficients of the expansion are given by

$$c_k \approx \frac{2}{M} \sum_{j=1}^{M} f(\cos(\theta_j)) \cos((k-1)\theta_j),$$

where $\theta_j = \pi(j - \frac{1}{2})/M$. ▸ Back

The N-independence of the error

The mth truncation error without dropping can be written as

$$\|e_m(H)\| = \|f(H) - \sum_{k=0}^{m} c_k T_k(H)\|.$$

For x in $[-1, 1]$ we have that $|T_k(x)| \leq 1$ for $k = 1, 2, \ldots$. Then

$$\|e_m(H)\| = \| \sum_{k=m+1}^{\infty} c_k T_k(H)\| \leq \sum_{k=m+1}^{\infty} |c_k|.$$

▸ Back