Linear Scaling Algorithm for Density-Functional Theory with Optimally Localized Wave Functions

Carlos J. García-Cervera
cgarcia@math.ucsb.edu

Mathematics Department
University of California, Santa Barbara

Quantum-Classical Modeling of Chemical Phenomena
CSCAMM, Maryland, March 8-11, 2010

Joint work with:

- Jianfeng Lu, currently at Courant Institute (NYU).
- Yulin Xuan, (previously at UCSB).
- Weinan E, Mathematics department and PACM, Princeton.

Supported by NSF CAREER award.

Outline

1. Orthogonal formulation of Kohn-Sham.
2. Non-Orthogonal formulation of Kohn-Sham.
3. Localization in Quantum Mechanics.
4. Localized Subspace Iteration (LSI)
   - Localization.
   - Filtering.
   - Computation of the electronic density.
   - Estimation of the Fermi energy.
5. Examples.
6. Performance and convergence properties.

Kohn-Sham Density Functional Theory

$$E_{KS}[\{\psi_j\}] = 2\sum_{j=1}^{N}(-\frac{1}{2})\int \psi_j\,(\Delta\psi_j)\;dx$$
$$+\frac{1}{2}\int\int \frac{\rho(x)\rho(y)}{|x-y|}\,dx\,dy + \int V_{ext}(x)\rho(x)\,dx + \int \rho\varepsilon(\rho),$$

where

$$\rho(x) = 2\sum_{j=1}^{N}|\psi_j(x)|^2$$

Kohn-Sham DFT:

$$\min_{(\psi_i,\psi_j)=\delta_{ij}} E_{KS}[\{\psi_j\}]$$

- Code implemented using Troullier-Martins nonlocal pseudopotentials in Kleinman-Bylander form.
- LDA not a limitation. GGA could be used.

Kohn-Sham Density Functional Theory

The Euler-Lagrange equations for the Kohn-Sham energy functional are

$$\left(-\frac{1}{2}\Delta + V_{eff}[\rho]\right)\psi_i = \lambda_i \psi_i, \quad i = 1, \ldots, N,$$

where

$$\rho = 2\sum_{i=1}^{N} |\psi_i|^2, \qquad \int \psi_i \psi_j = \delta_{ij},$$

and

$$V_{eff}[\rho](x) = \int \frac{\rho(y)}{|x-y|}\, dy + V_{ext}(x) + \varepsilon(\rho(x)) + \rho(x)\varepsilon'(\rho(x)).$$

Typical Kohn-Sham approach:

1. Given a potential V_{eff}, diagonalize and orthogonalize to obtain the wave functions.
2. Given the new wave functions, update the density, and compute new V_{eff}.
3. **Diagonalization is** $O(N^3)$.

Non-Orthonormal Formulation of Kohn-Sham DFT[1]

Given N linearly independent wave functions, $\{\psi_j\}$, define the overlap matrix:

$$\mathbf{S}_{jk} = \int \psi_j \psi_k.$$

Then,

$$E_{KS}[\{\psi_j\}] = 2 \sum_{j,k} \left(-\frac{1}{2}\right) (\mathbf{S}^{-1})_{jk} \int \psi_j \left(\Delta \psi_k\right) \, dx + \frac{1}{2} \int \int \frac{\rho(x)\rho(y)}{|x-y|} \, dx \, dy + \int V_{ext}(x)\rho(x) \, dx + \int \rho \varepsilon(\rho),$$

where

$$\rho(x) = 2 \sum_{jk} \psi_j(x) (\mathbf{S}^{-1})_{jk} \psi_k(x)$$

[1] W. Yang '97

Advantages of the Non-Orthogonal formulation

- $\{\psi_j\}$ not orthogonal.
- Invariant under nonsingular linear transformations: Let $\widetilde{\Psi} = \Psi R$, with $R \in \mathbb{R}^{N\times N}$, invertible. Then,

$$E_{KS}[\widetilde{\Psi}] = E_{KS}[\Psi].$$

- The emphasis is therefore on the subspace spanned by $\{\psi_j\}$.
- Nonorthogonal wave functions have better localization properties.

Linear scaling methods for Kohn-Sham DFT

A number of linear scaling methods have appeared in the literature (Goedecker '99):

1. Orbital Minimization (Mauri, Galli, Car '93, W. Yang '97, C. Yang, J.C. Meza and L.W. Wang '06, Burger and Yang '07, W. Gao and W. E '08).
2. Density Matrix Minimization (Li, Nunes, Vanderbilt, '93).
3. Fermi Operator Expansion (Goedecker '94; Lin, Lu, Car, E '09).
4. Divide and Conquer (Yang '91, L.W. Wang, Z. Zhao, J. Meza '06, Barrault, Cancès, Hager, Le Bris '07).

An interesting $O(N^3)$ algorithm:

- **Subspace Iteration** (Zhou, Saad, Tiago, Chelikowsky '06).
- New algorithm: Similar to the subspace iteration method of Zhou, Saad, Tiago, and Chelikowsky '06, **but**, we avoid diagonalization and orthogonalization (which is $O(N^3)$) (CJGC, Lu, E, '07, CJGC, Lu, Xuan, E '08).

Linear scaling methods for Kohn-Sham DFT

Guiding principles:

- In the non-orthogonal formulation, the emphasis is on the **subspace generated by the wave functions**.
- We want to generate the optimal eigenspace of the self-consistent Hamiltonian: **Filtering out the high end of the spectrum**.
- Localization is key for linear scaling: **We choose a localized basis for this subspace**.
- We use finite differences, and **real-space formulation**.

We **avoid**:

- Diagonalization and orthogonalization.
- Using a basis set.
- Using plane waves.
- Using a *supercell* for non-periodic problems.

Localization in Quantum-Mechanics

- Related to *Nearsightedness:* A small disturbance in a molecule only has a local effect in the electron density (W. Kohn, '96).
- Consider a Hamiltonian with a periodic potential in a crystalline solid:

$$\mathbf{H} = -\frac{1}{2}\Delta + V(x). \tag{1}$$

Floquet-Bloch theorem: The eigenfunctions have the form

$$\psi_{n,k}(x) = e^{ik\cdot x} u_{n,k}(x), \tag{2}$$

where $u_{n,k}$ has the periodicity of V, and k belongs to the reciprocal lattice.

- $\psi_{n,k}$ is the Bloch function associated to wave vector k and band index n.

Wannier Functions

From the Bloch functions, we construct the Wannier function for the n-th band as:

$$W_n(x, R) = \frac{V}{(2\pi)^3} \int_{BZ} e^{-ikR} \psi_{k,n}(x)\, dk. \tag{3}$$

- Wannier functions are not unique: The Bloch functions can be multiplied by an arbitrary phase.
- Wannier functions are translation invariant: $W_n(x, R) = W_n(x - R)$.
- With this definition, they form an orthonormal basis.

Wannier Functions (II)

To illustrate the previous definitions, consider the following one-dimensional model:

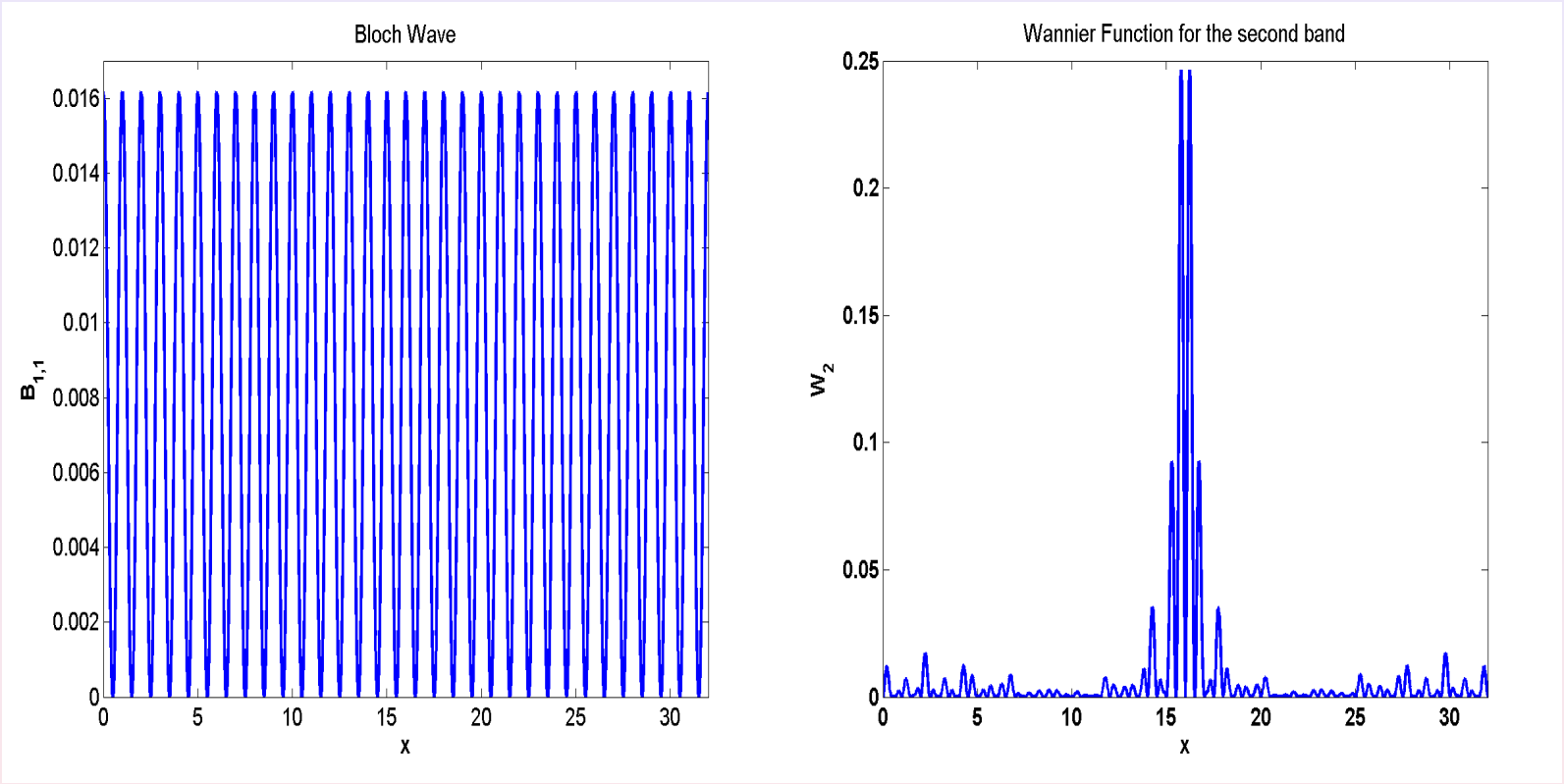
$$\mathbf{H} = -\frac{1}{2}\frac{d^2}{dx^2} + V(x), \tag{4}$$

where

$$V(x) = -a \sum_{i=-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-i)^2/(2\sigma^2)}. \tag{5}$$

- The parameter a represents the strength of the potential.
- The parameter σ represents the width of the potential.

Bloch waves and Wannier functions



Wannier Functions and Localization

- Wannier functions have good localization properties (W. Kohn '59, des Cloizeaux '63-'64, E. Prodan & W. Kohn '05, G. Panati '06-'07, Jianfeng Lu '09).
- Wannier functions have been used for numerical computations, e.g., *Maximally* Localized Wannier Functions (Marzari and Vanderbilt, '97).
- In general, localized wave functions have been used to design $O(N)$ methods, typically
  - As basis sets,
  - Via truncation,
  - Or both.
- OM and DMM include some form of truncation.
- For the study of solids, the definition must be extended to non-orthogonal wave functions, and non-periodic systems.
- Localized Wannier functions can be constructed for elastically deformed solids (W. E, J. Lu '09).

Optimally localized wave functions[1]

Given $\{\psi_j\}$, define

$$V = span\{\psi_j\}.$$

The optimally localized wave function, or generalized non-orthogonal Wannier function, ϕ^*, is the minimizer of

$$\inf_{\phi \in V, \|\phi\|_2 = 1} \int w(x)|\phi(x)|^2 \, dx.$$

- Generalizes the Maximally Localized Wannier Functions of Marzari and Vanderbilt (1997).
- An alternative procedure is the *Frobenius Localization* (Weiguo Gao and Weinan E '08).
- One can show that the non-orthogonal wave functions have exponential decay[2].
- The *best* weight function: $w(x) = |x - c|^{2p}$.

[2] Jianfeng Lu '08

[1] Weinan E, Tiejun Li, and Jianfeng Lu, '07

Algorithm for Localization[1]

1. Given a set of wave functions, $\{\psi_j\}_{j=1}^N$, centered at the locations $\{b_j\}_{j=1}^N$, respectively.
2. We obtain an optimally localized basis by minimizing

$$F[\phi] = \frac{\int_{\mathbb{R}^3} |x - b_j|^{2p} |\phi(x)|^2 \, dx}{\int_{\mathbb{R}^3} |\phi(x)|^2 \, dx},$$

   among functions ϕ of the form

$$\phi(y) = \sum_{k=1}^{r} \alpha_k \psi_k(x).$$

3. Minimization leads to

$$\mathbf{W}a = \lambda \mathbf{S}a.$$

- **Only a fixed number r of functions involved, so this is $O(N)$**
- The localized functions span the same space.

[1] Weinan E, Tiejun Li, Jianfeng Lu, '07; CJGC, Jianfeng Lu, Weinan E, '07; CJGC, Jianfeng Lu, Yulin Xuan, Weinan E, '08

Filtering Step

Goal: To improve the subspace by removing components in the high end of the spectrum of the Hamiltonian

$$H = -\frac{1}{2}\Delta + V_{eff}[\rho].$$

Power Method

The simplest filter is probably the Power Method (Parlett, '98):

1. Given an initial vector $\mathbf{v}^0$.
2. For $k \geq 0$, define
   2.1
   $$\mathbf{v}^{k+1} = \frac{\mathbf{H}\mathbf{v}^k}{\|\mathbf{H}\mathbf{v}^k\|}, \tag{6}$$
   2.2 $\mu^{k+1} = (\mathbf{v}^{k+1})^T \cdot \mathbf{H}\mathbf{v}^{k+1}$.
3. Repeat until $|\mu^{k+1} - \mu^k| \leq \text{Tolerance}$.

Convergence: If $\mathbf{H}\psi_i = \lambda_i \psi_i$, and $|\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_N|$, then

$$\frac{1}{\|\mathbf{H}\mathbf{v}\|}\mathbf{H}\mathbf{v} = \psi_N + O\left(\left|\frac{\lambda_{N-1}}{\lambda_N}\right|\right). \tag{7}$$

- Note that when applied to a subspace, the space collapses to a one-dimensional space.

Subspace Iteration

The Subspace Iteration generalizes the Power Method to a subspace (Parlett, '98):

1. Given an initial space V_0 of dimension $M < N$, for each $k \geq 1$:
   - 1.1 Calculate $W_k = \mathbf{H}V_k$.
   - 1.2 Orthogonalize the basis (QR decomposition, for example): $W_k = Q_k R_k$.
   - 1.3 Let $V_k = Q_k$.
2. Repeat until convergence.

- The orthogonalization step is necessary in order to ensure the linear independence of the vectors in the new space.
- If $\mathbf{H}\psi_i = \lambda_i \psi_i$, and $|\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_N|$, then the subspace iteration converges with rate of convergence

$$\tau = \frac{\lambda_M}{\lambda_{M+1}} < 1. \tag{8}$$

Polynomial Filtering

- Filtering improves the rate of convergence of the subspace iteration.
- If the polynomial P splits the spectrum of $\mathbf{H}$, in the sense that

$$P(\lambda_i) \leq P(\lambda_M), \quad i = 1, \ldots, M, \tag{9}$$
$$P(\lambda_j) \geq P(\lambda_{M+1}), \quad j = M+2, \ldots, M, \tag{10}$$

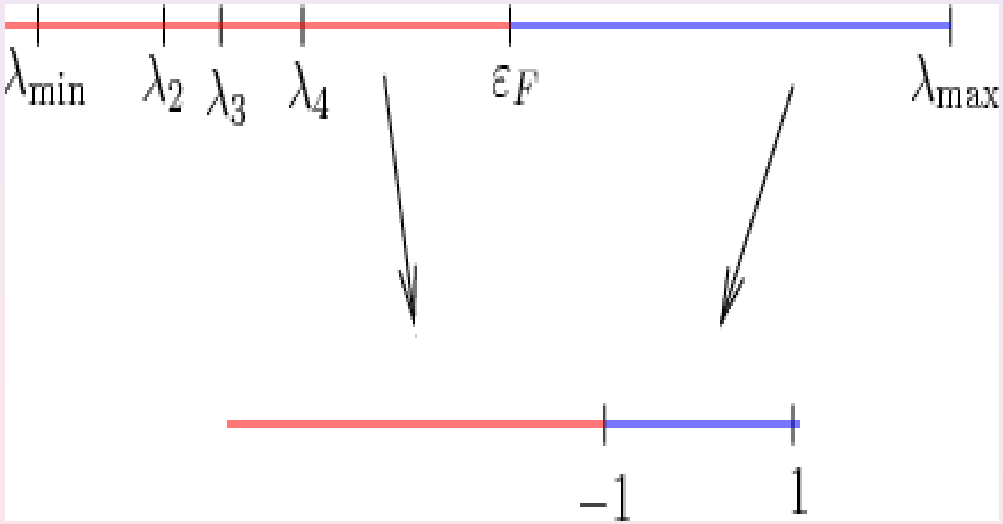
  the rate of convergence of the polynomial filtered subspace iteration is

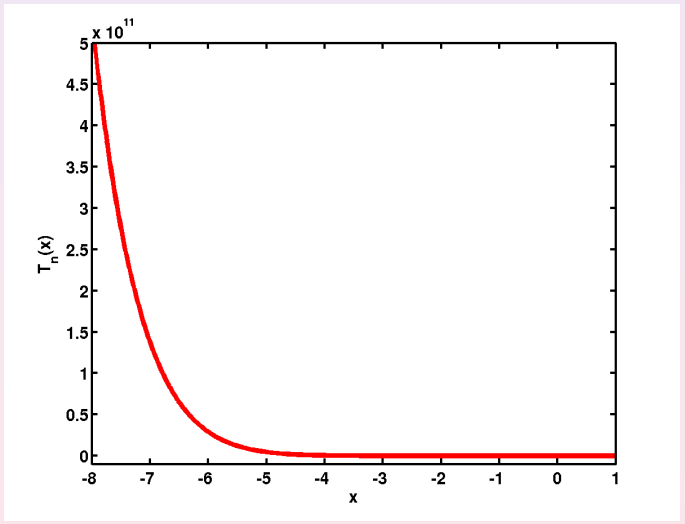
$$\kappa = \left| \frac{P(\lambda_M)}{P(\lambda_{M+1})} \right|. \tag{11}$$

- No diagonalization necessary.

Chebyshev Filter

- Optimal choice: Chebyshev polynomial. $T_n(H)$.
  - Recursive: $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$.
  - $$T_n(x) = \begin{cases} \cos(n \cos^{-1} x) & if\,|x| \leq 1, \\ (-1)^n \cosh(n \cosh^{-1} |x|) & if\,|x| \geq 1, \end{cases}$$





Chebyshev Filter

- In the context of electronic structure analysis, subspace iteration has been used by Zhou, Saad, Tiago, and Chelikowsky, '06.
- The orthogonalization step leads to an $O(N^3)$ method.
- We replace the orthogonalization step with a localization step, achieving $O(N)$.
- The Fermi energy must be estimated (no diagonalization is used).

Estimation of the Fermi energy

- Given the wave functions Ψ, we know that $\Phi = \Psi \mathbf{S}^{-1/2}$ are orthogonal (*Löwdin transformation*).
- The Ritz matrix is

$$\mathbf{R} = \Phi^T \mathbf{H} \Phi = \mathbf{S}^{-1/2} \Psi^T \mathbf{H} \Psi \mathbf{S}^{-1/2}$$

- We estimate the Fermi energy by the maximum eigenvalue of the Ritz matrix.
- The eigenvalues of $\mathbf{R}$ are the same as the eigenvalues of $\mathbf{S}^{1/2} \mathbf{R} \mathbf{S}^{-1/2} = \Psi^T \mathbf{H} \Psi \mathbf{S}^{-1}$.
- We can use the Power method.
- Note that we do not need $\mathbf{S}^{-1}$, only $\mathbf{w} = \mathbf{S}^{-1} \mathbf{v}$, which can be obtained by solving

$$\mathbf{S} \mathbf{w} = \mathbf{v}.$$

- $\mathbf{S}$ is sparse and localized, and $\Psi^T \mathbf{H} \Psi$ is sparse: The Fermi energy can be estimated in $O(N)$.

Computation of the Electronic Density

$$\rho(x) = 2 \sum_{jk} \psi_j(x) (\mathbf{S}^{-1})_{jk} \psi_k(x)$$

- Computing $\mathbf{S}^{-1}$ directly is $O(N^3)$.
- Instead, we use the Newton-Schultz iteration to solve

$$\mathbf{DSD} - \mathbf{D} = 0.$$

- $\mathbf{S}$ and $\mathbf{S}^{-1}$ are localized near the diagonal.
- Exploiting sparsity, computation is $O(N)$ (Jansik, Host, Jorgensen, Olsen, and Helgaker '07, Rubensson and Salek '05).
- Alternatively, a pseudoinverse can be used (W. Yang '97).

Linear Scaling Algorithm for Kohn-Sham[1]

1: Given (localized) wave functions Ψ_0.
2: **repeat** {(*Self-Consistency Loop (SCF)*)}
3: Compute electronic density: ρ
4: Compute effective potential: $V_{\text{eff}}[\rho]$.
5: **repeat** {(*Localized Subspace Iteration*)}
6: Estimate Fermi energy.
7: Filtering Step: $\Phi = T_n(H)\Psi$.
8: Localization Step: Localize ψ_r for $r = 1, \cdots, k$.
9: Truncation beyond cut-off radius.
10: **until** Convergence of LINEAR iteration
11: Update electronic density (mixing).
12: **until** $\|\rho_{k+1} - \rho_k\|_1 \leq Tol$.

- Similar to the subspace iteration method of Zhou, Saad, Tiago, and Chelikowsky '06, **but**, we avoid diagonalization and orthogonalization (which is $O(N^3)$).
- In practice, only one filtering step is performed.

[1] CJGC, Jianfeng Lu, Weinan E '07; CJGC, Jianfeng Lu, Yulin Xuan, Weinan E, '08

Numerical Examples

$$E_{\mathrm{KS}}[\{\psi_j\}] = 2\sum_{j,k}(\mathbf{S}^{-1})_{jk}\int_{\mathbb{R}^3}\psi_j\left(-\frac{1}{2}\Delta\psi_k\right)\,d\boldsymbol{x} + E_{\mathrm{XC}}[\rho]$$
$$+\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{(\rho-m)(\boldsymbol{x})(\rho-m)(\boldsymbol{y})}{|\boldsymbol{x}-\boldsymbol{y}|}\,d\boldsymbol{x}\,d\boldsymbol{y} + E_{\mathrm{PS}}[\{\psi_i\}], \quad (12)$$

- Electron density:

$$\rho(\boldsymbol{x}) = 2\sum_{jk}\psi_j(\boldsymbol{x})(\mathbf{S}^{-1})_{jk}\psi_k(\boldsymbol{x}), \quad (13)$$

- Ionic function:

$$m(\boldsymbol{x}) = \sum_{j=1}^{N_a} m^a(\boldsymbol{x}-\boldsymbol{R}_j), \quad (14)$$

- Exchange and Correlation: Ceperley and Alder '80, as parameterized by Perdew and Zunger, '81.

Numerical Examples (II)

- Pseudopotential energy:

$$E_{\mathrm{PS}}[\{\psi_i\}] = 2\sum_{j,k}(\mathbf{S}^{-1})_{jk}\int_{\mathbb{R}^3}\psi_j(\boldsymbol{x})\widehat{V}_{PS}\psi_k(\boldsymbol{x})\,d\boldsymbol{x}. \qquad (15)$$

- Norm conserving Troullier-Martins pseudopotential in the Kleinman-Bylander form:

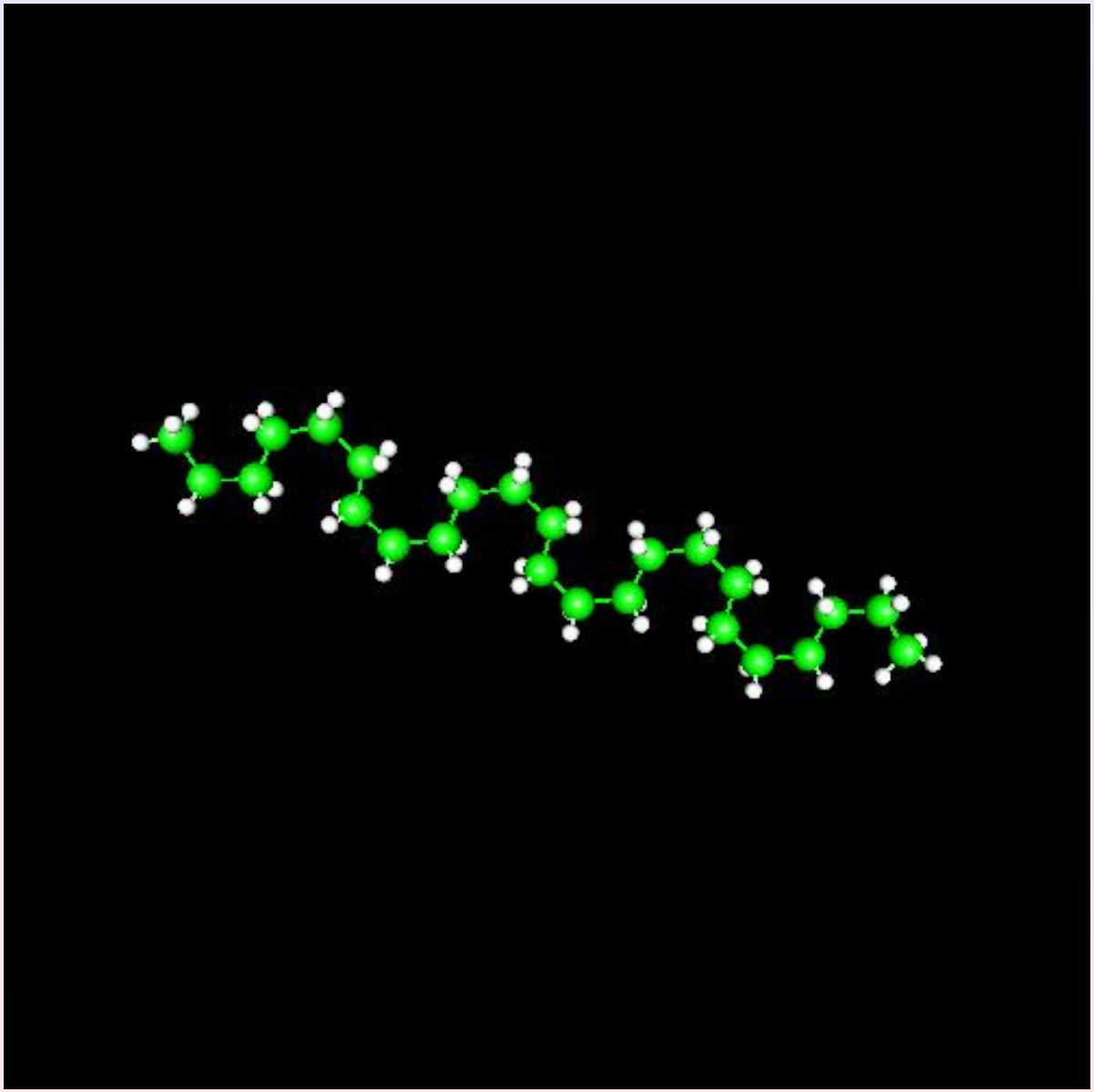
$$\widehat{V}_{\mathrm{PS}}\psi(\boldsymbol{x}) = \sum_{j=1}^{N_a}\Bigg(V^j_{\mathrm{Local}}(\boldsymbol{x}-\boldsymbol{R}_j)\psi(\boldsymbol{x}) + \sum_{l=0}^{l_{max}}\sum_{m=-l}^{l}\int_{\mathbb{R}^3}\beta^j_{lm}(\boldsymbol{y}-\boldsymbol{R}_j)\psi(\boldsymbol{y})\,d\boldsymbol{y}\;\beta^j_{lm}(\boldsymbol{x}-\boldsymbol{R}_j)\Bigg), \qquad (16)$$

- m^a decays exponentially.
- V_{Local}, β_{lm}, are compactly supported.

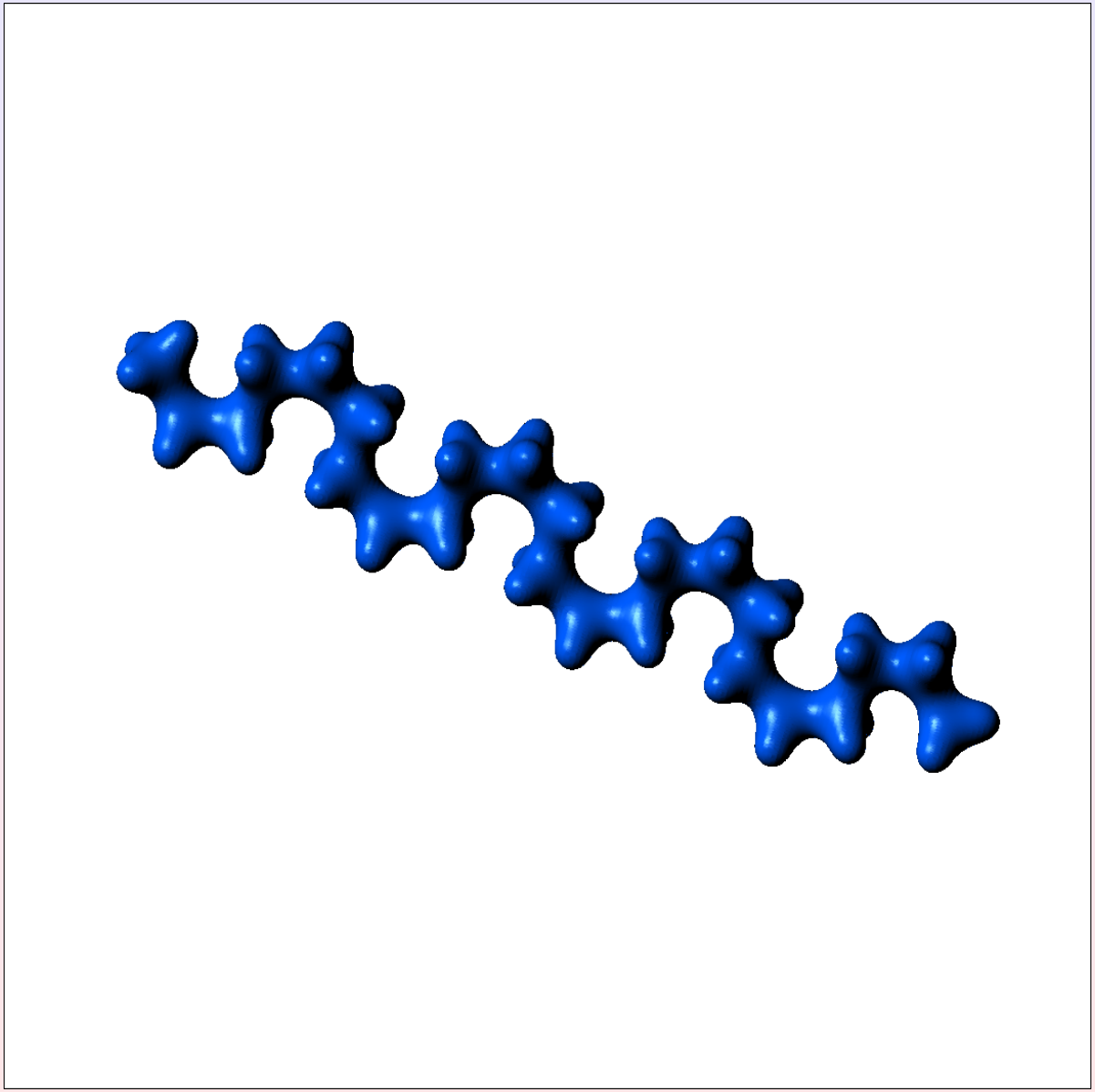
Details of the implementation

- Finite differences.
- Sparse representation for the wave functions and pseudopotential components.
- Coulomb term is approximated as a discrete convolution, evaluated using the Fast Fourier Transform (FFT).
- Self-consistent iteration: Linear Mixing (not a limitation).
- Pseudopotential:
  - Hydrogen: local component for the $1s$ orbital and no nonlocal components.
  - Carbon: We choose the $2p$ pseudopotential to be the local component. The nonlocal pseudopotential is therefore the $2s$ component.
- Pseudopotentials generated using code by Paolo Gianozzi and his collaborators.

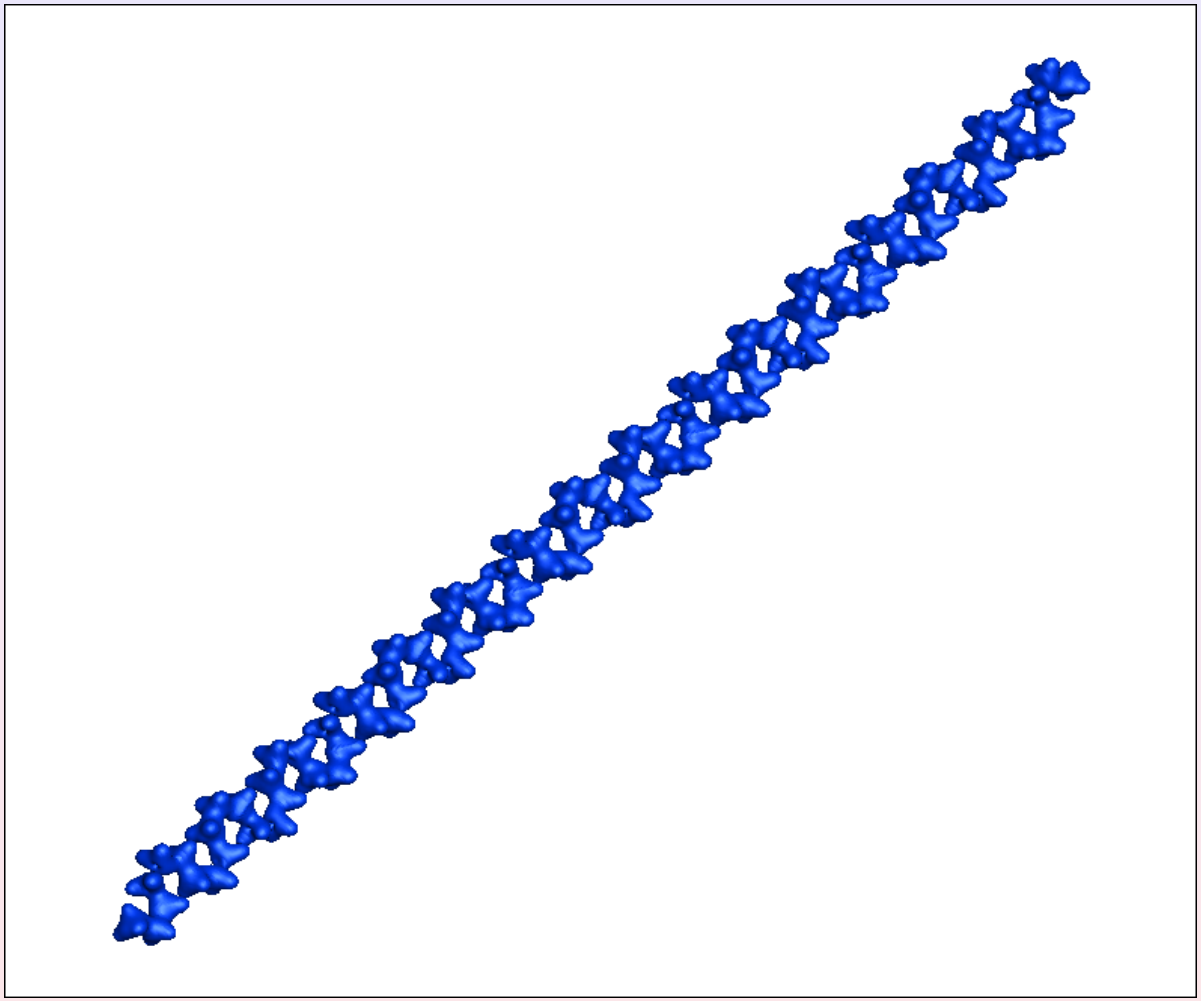
Example: Alkane - $CH_3(CH_2)_{10}CH_3$ (74 atoms)



Example: Alkane - $CH_3(CH_2)_{10}CH_3$ - Density



Example: Alkane - Molecular chain (290 atoms)



LSI Timings

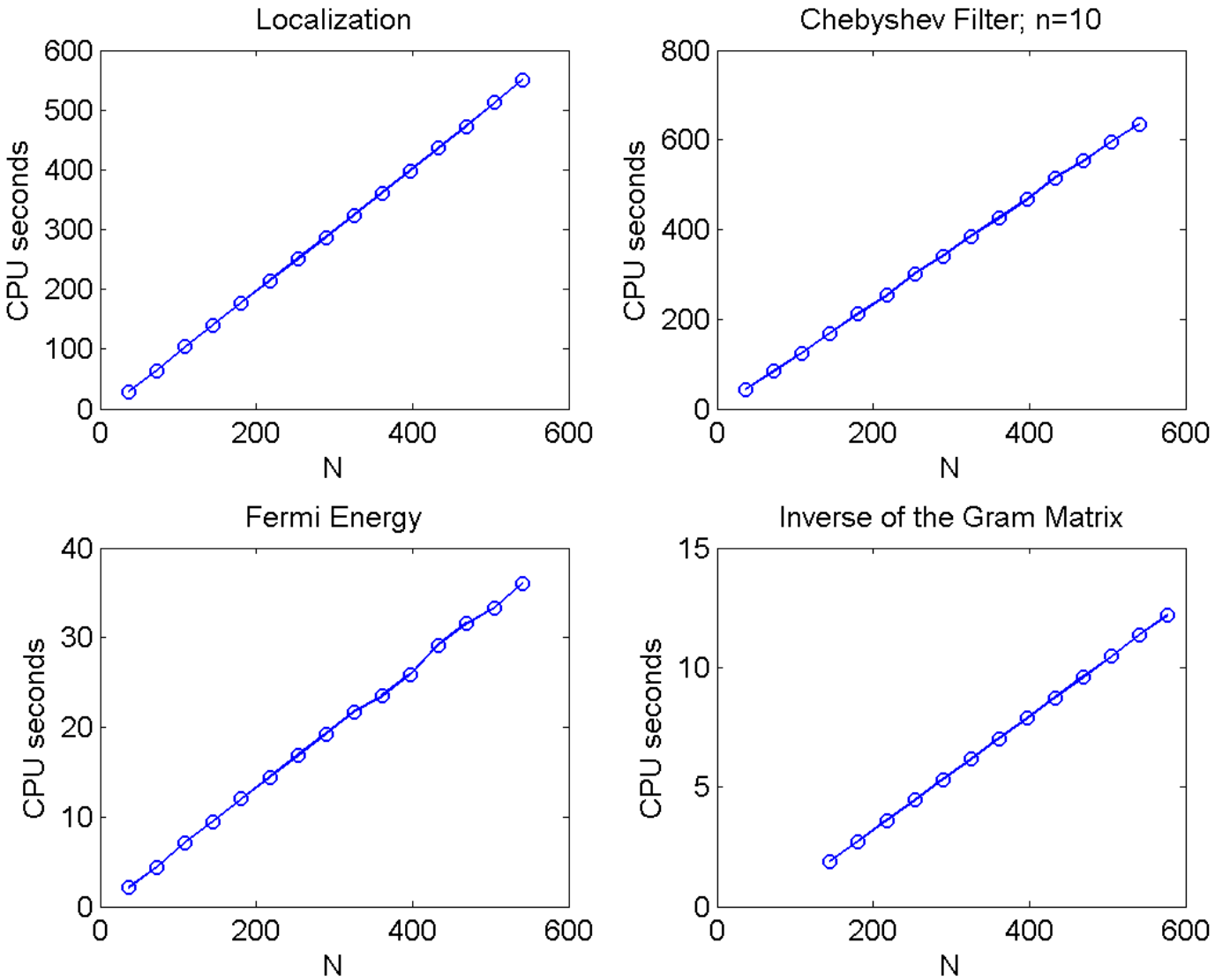


Figure: Timings obtained with the LSI code. Linear scaling is observed.

Convergence Properties: 1d model problem

- Consider an infinite array of atoms on a line with unit spacing: $X_i = i$, for $i \in \mathbb{Z}$.
- Each atom has one valence electron and we ignore spin degeneracy.
- The electrons are non-interacting: the electronic structure of the system is determined by solving linear eigenvalue problems

$$H\psi_i = \epsilon_i \psi_i \tag{17}$$

- Hamiltonian given by

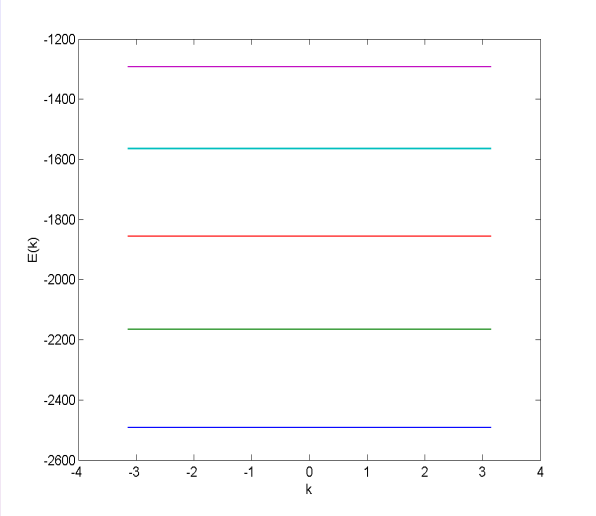
$$H = -\frac{1}{2}\frac{d^2}{dx^2} + V(x). \tag{18}$$

- Effective potential V is a sum of Gaussian wells located at the atom sites:

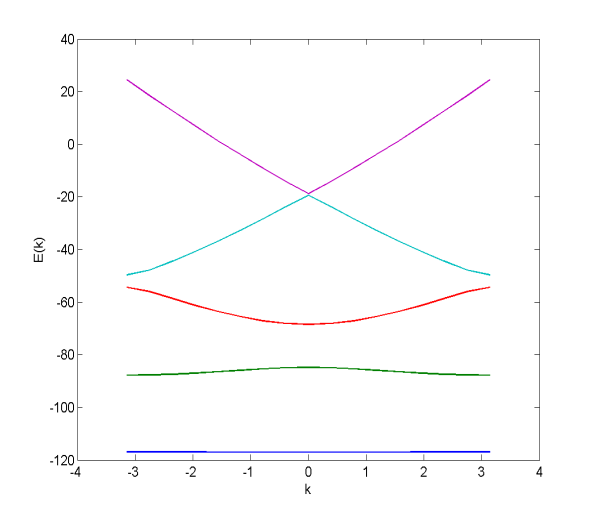
$$V(x) = -\sum_{i\in\mathbb{Z}} \frac{a}{\sqrt{2\pi\sigma^2}} \exp(-(x - X_i)^2/2\sigma^2). \tag{19}$$

- Two parameters:
  - a characterizes the depth of the wells.
  - σ characterizes its width.

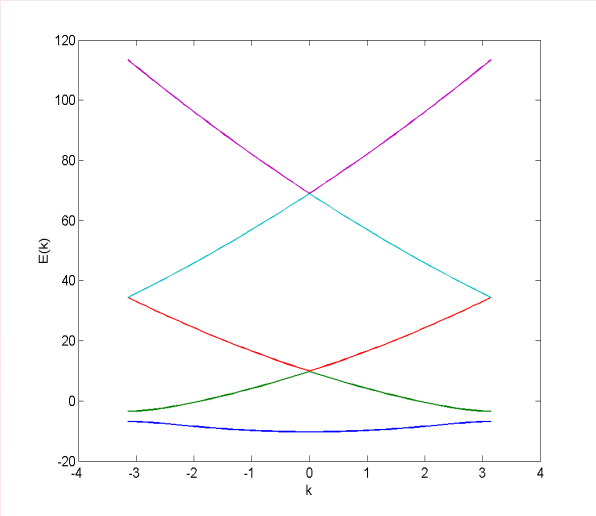
1d model problem: Band structure



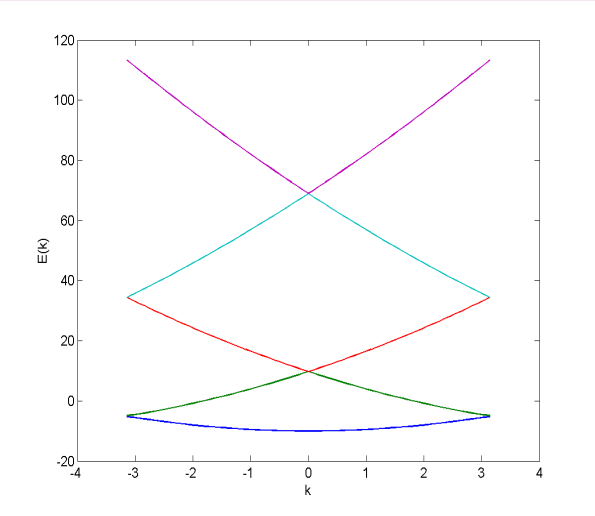
(a) $a = 1000;\ \sigma = 0.15.$



(b) $a = 100;\ \sigma = 0.3.$

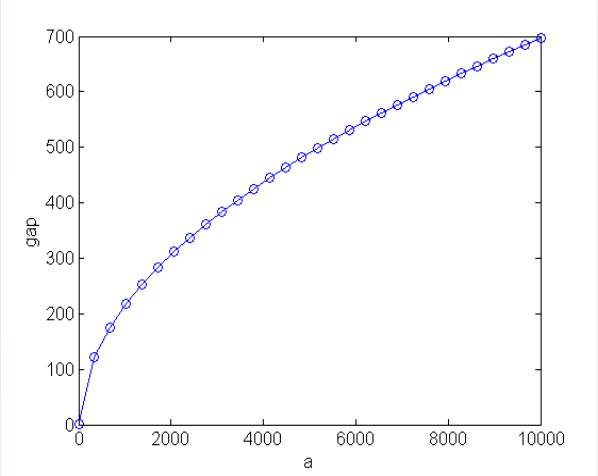


(c) $a = 10;\ \sigma = 0.3.$

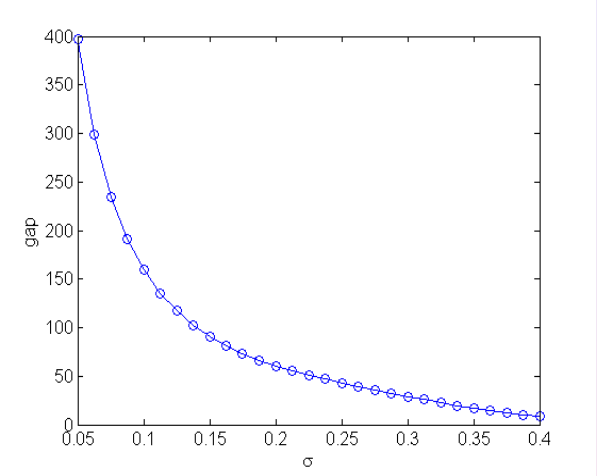


(d) $a = 10;\ \sigma = 0.45.$

1d model problem: Band gap



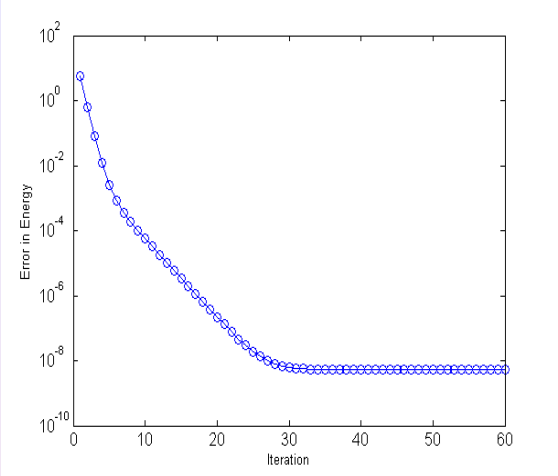
(e) Band gap as function of a.



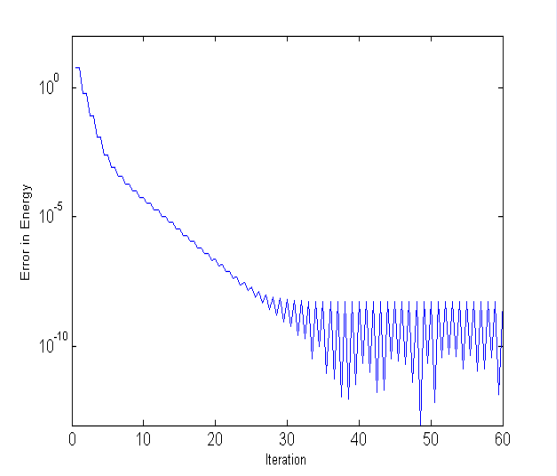
(f) Band gap as function of σ.

- The gap is proportional to $\sqrt{a}/\sigma$.
- By changing parameters, we may change the model from a well gapped insulator to a metal-like system.

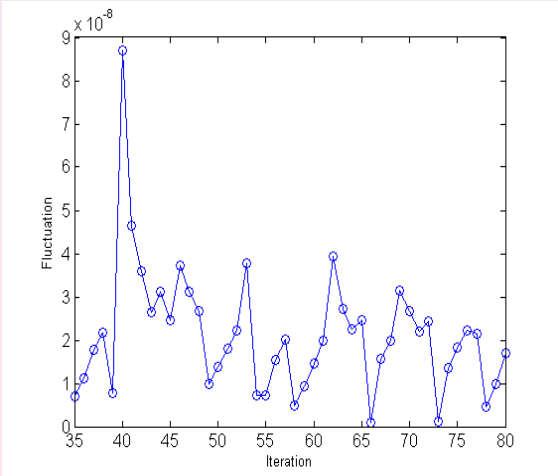
The Issue of Convergence



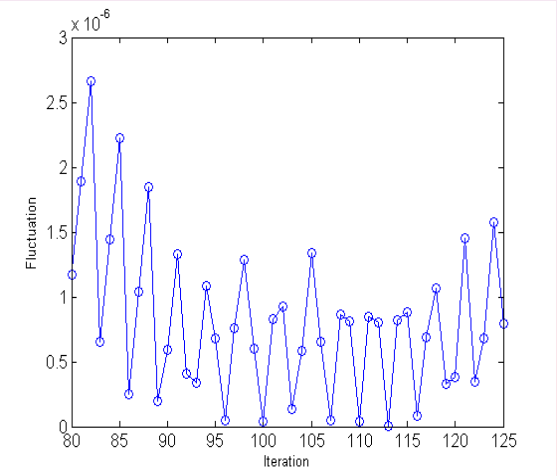
(g) Error in energy.



(h) Error in energy before and after truncation.



(i) Remanent fluctuation (semiconductor case).



(j) Remanent fluctuation (metal-like case).

The Issue of Convergence (II)

- Each LSI iteration contains three steps:
  1. Starting from a given subspace, the filtering step makes the subspace closer to the occupied subspace by filtering out the higher spectrum.
  2. The localization step finds a better representation while keeping the subspace unchanged.
  3. Truncation of the localized basis.
- Without truncation, the LSI iteration will converge.
- After truncation, the subspace deviates from the correct occupied subspace. As a result, the iteration process might not converge.
- This is generic to linear scaling algorithms involving truncation, including variational methods such as orbital minimization (Weiguo Gao, CJGC, Jianfeng Lu, Weinan E, '08).
- For LSI, initially the error decays exponentially and then starts to fluctuate around a value that is small but different from the round-off error.
- We call these fluctuations remanent fluctuation, and the error between the numerical solution and the true minimizer (without truncation) remanent error.

Local Error Estimate

- Denote by V_0 the true occupied subspace, and by V_n the subspace in the n-th step.
- The error is $d(V_n, V_0)$.
- We use F, L and T to represent the filtering, localization and truncation steps.
- δ quantifies the error caused by truncating the localized representation of the true occupied subspace:

$$\delta = d(TL(V_0), V_0). \tag{20}$$

Local Error Estimate (II)

- If V_n is sufficiently close to V_0,

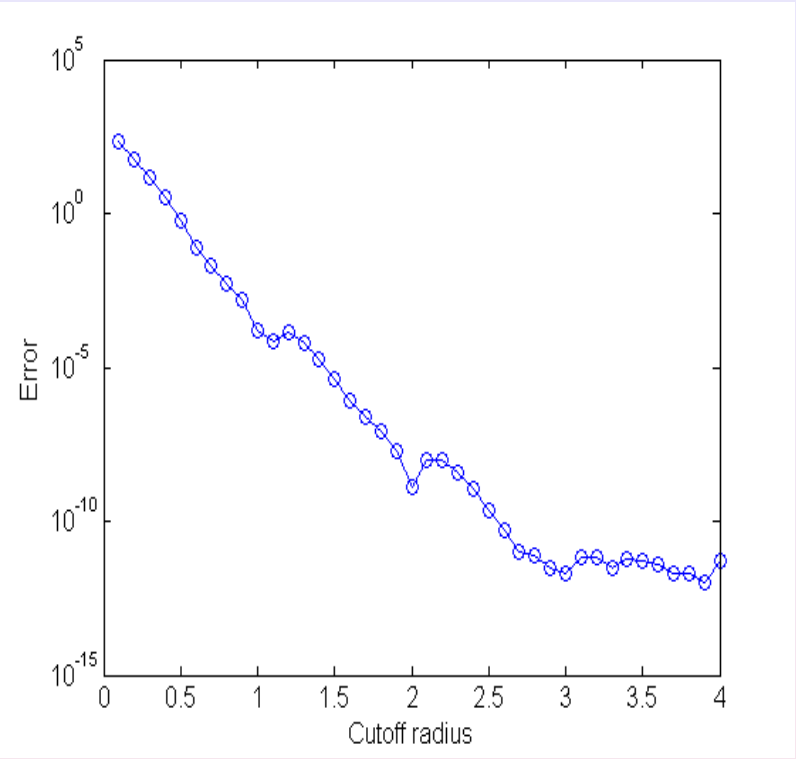
$$e_{n+1} = d(TLF(V_n), V_0) \leq d(TLF(V_n), F(V_n)) + d(F(V_n), V_0) \leq C\delta + \lambda e_n. \quad (21)$$

- Therefore,

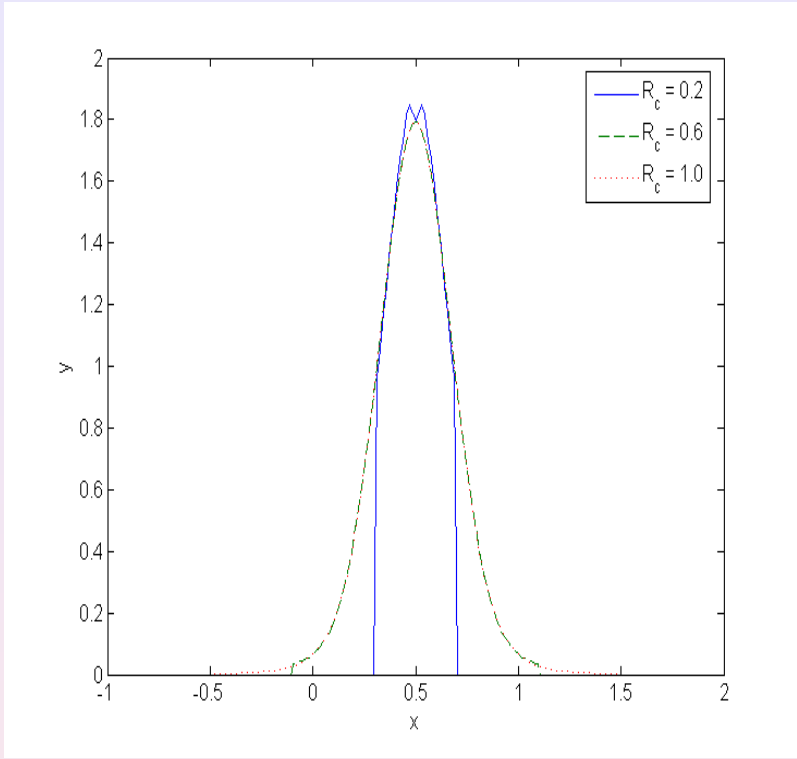
$$e_\infty \leq C\delta/(1-\lambda). \quad (22)$$

- This guarantees that V_n stays close to V_0.

Performance of the LSI



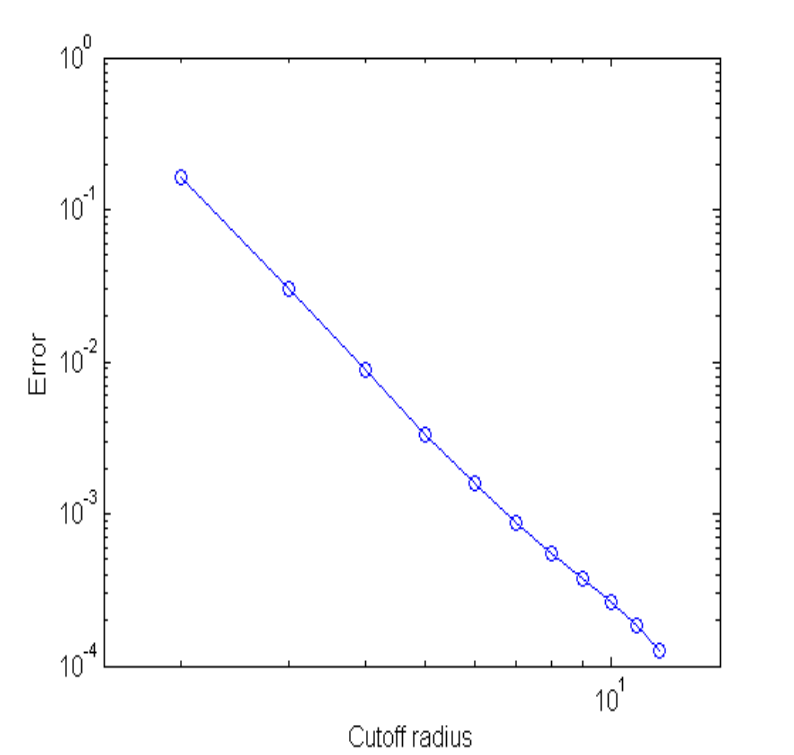
(k) Error in the energy as a function of the cut-off radius.



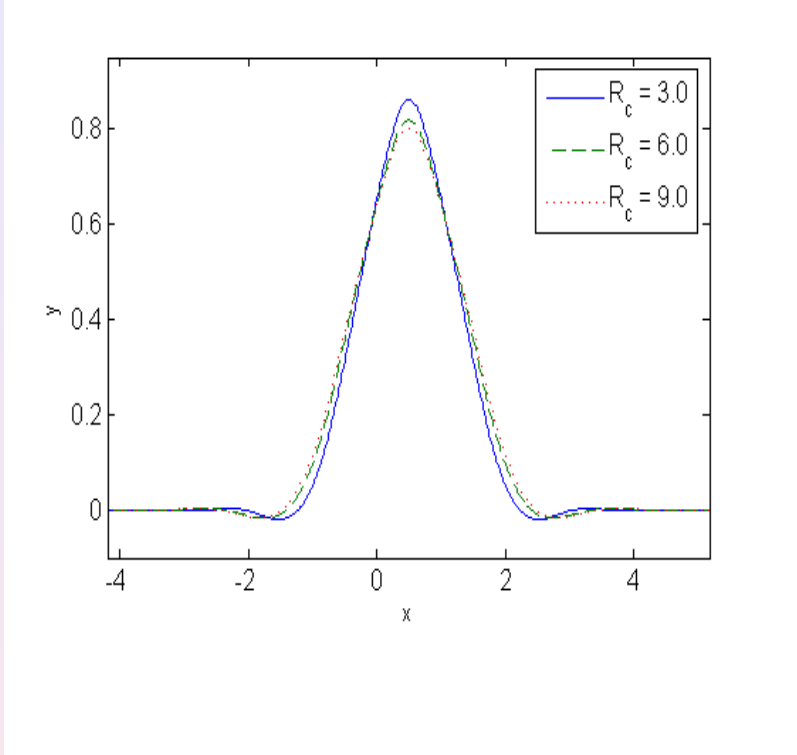
(l) Wave functions for three values of the cut-off radius.

Figure: The effect of the cut-off radius on the accuracy for the insulator case. (k) Logarithmic plot of the error in the energy for different cut-off radii; (l) The resulting wave functions for cut-off radii 0.2, 0.6 and 1.0.

Performance of the LSI (II)



(a) Error in the energy as a function of the cut-off radius.



(b) Wave functions for three cut-off radii.

Figure: The effect of the cut-off radius on the accuracy for the metallic case. (a) Logarithmic plot of the error in the energy for different cut-off radii; (b) The resulting wave functions for cut-off radii 3.0, 6.0 and 9.0.

Comments and conclusions

- We have presented an efficient linear scaling methodology for Kohn-Sham DFT.
- We have introduced a 1d model as a benchmark for linear scaling methods.
- Convergence of a numerical algorithm is typically understood in terms of the discretization size, and the number of iterations. In that sense, linear scaling methods with truncation (typically) do not converge.
- The size of the remanent error decreases when either the cut-off radius, the band gap, or the order of the Chebyshev filter is increased.
- Localization must be done before truncation.

Thank you!