
A Direct Constrained Minimization Algorithm for Solving the Kohn-Sham Problem

Chao Yang

Joint work with J. Meza & L. W. Wang

Computational Research Division

Lawrence Berkeley National Laboratory

Berkeley, California, USA

Outline

- Finite-dimensional Kohn-Sham (KS) equations
- The self-consistent field (SCF) iteration
- Direct constrained minimization (DCM)
- Examples

Finite Dimensional KS Problem

$$\min_{X^* X = I_{n_e}} E_{KS}(X) \equiv E_{kinetic}(X) + E_{ion}(X) + E_{Hartree}(X) + E_{xc}(X),$$

where

$$X$$

$$E_{kinetic} = \frac{1}{2} \text{trace}(X^* L X)$$

$$E_{ion} = \text{trace}(X^* V_{ion} X)$$

$$E_{Hartree} = \frac{1}{2} \rho(X)^T L^\dagger \rho(X)$$

$$E_{xc} = \rho(X)^T (\mu_{xc}[\rho(X)])$$

$$\rho(X) = \text{diag}(X X^*)$$

The KKT Condition

- KKT condition

$$\begin{aligned}\nabla_X \mathcal{L}(X, \Lambda) &= 0; \\ X^* X &= I_{n_e}.\end{aligned}$$

- Kohn-Sham equation

$$\begin{aligned}H(X)X &= X\Lambda, \\ X^* X &= I_{n_e}.\end{aligned}$$

- Kohn-Sham Hamiltonian

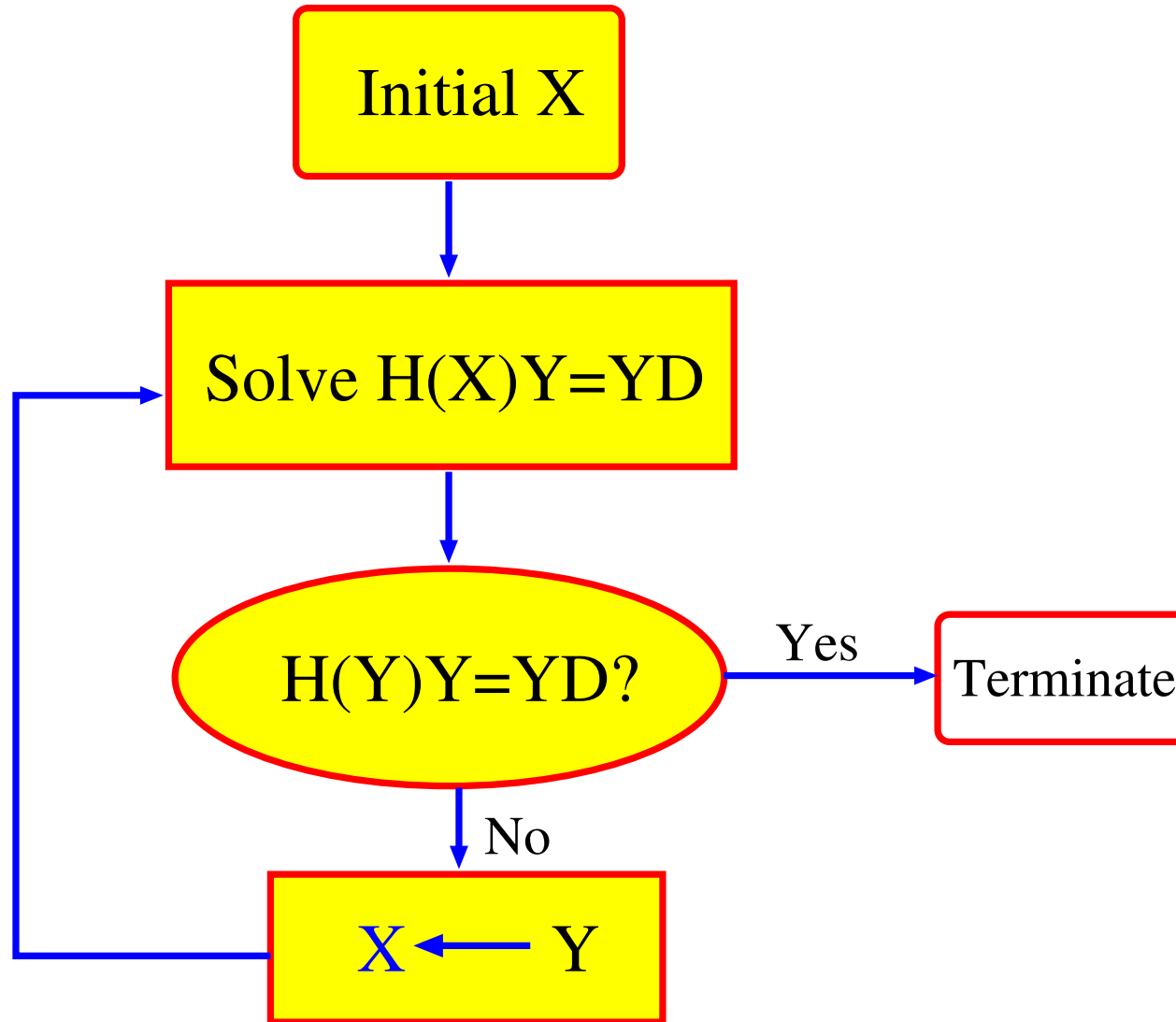
$$H(X) = \frac{1}{2}L + V_{ion} + \text{Diag}(L^\dagger \rho(X)) + \text{Diag}(g_{xc}(\rho(X)))$$

Solving the KS Problem

- Two approaches:
 - Work with the KS equation
 - Self-Consistent Field Iteration
 - Minimize the total energy directly
 - Direct Constrained Minimization
- Unique in KS Problem:
 - The invariance property of the KS problem

$$\begin{aligned} E(XQ) &= E(X) \\ H(XQ) &= H(X) \end{aligned} \quad \text{for any } Q^*Q = I_{n_e}$$

The SCF Iteration



The convergence of SCF

- SCF does not always converge;

$$\lim_{i \rightarrow \infty} \|\rho(X^{(i+1)}) - \rho(X^{(i)})\| \neq 0$$

- $E(x)$ may not decrease in SCF;
- For HF problems, one can show subsequence convergence;

$$\lim_{i \rightarrow \infty} \|\rho(X^{(i+1)}) - \rho(X^{(i-1)})\| = 0$$

(Cancès & Le Bris 2000)

- Conditions under which SCF becomes a contraction map
(Prodan 2003, 2005)

Toy Example

$$E(x) = \frac{1}{2}x^T Lx + \frac{\alpha}{4}\rho(x)^T L^{-1}\rho(x)$$

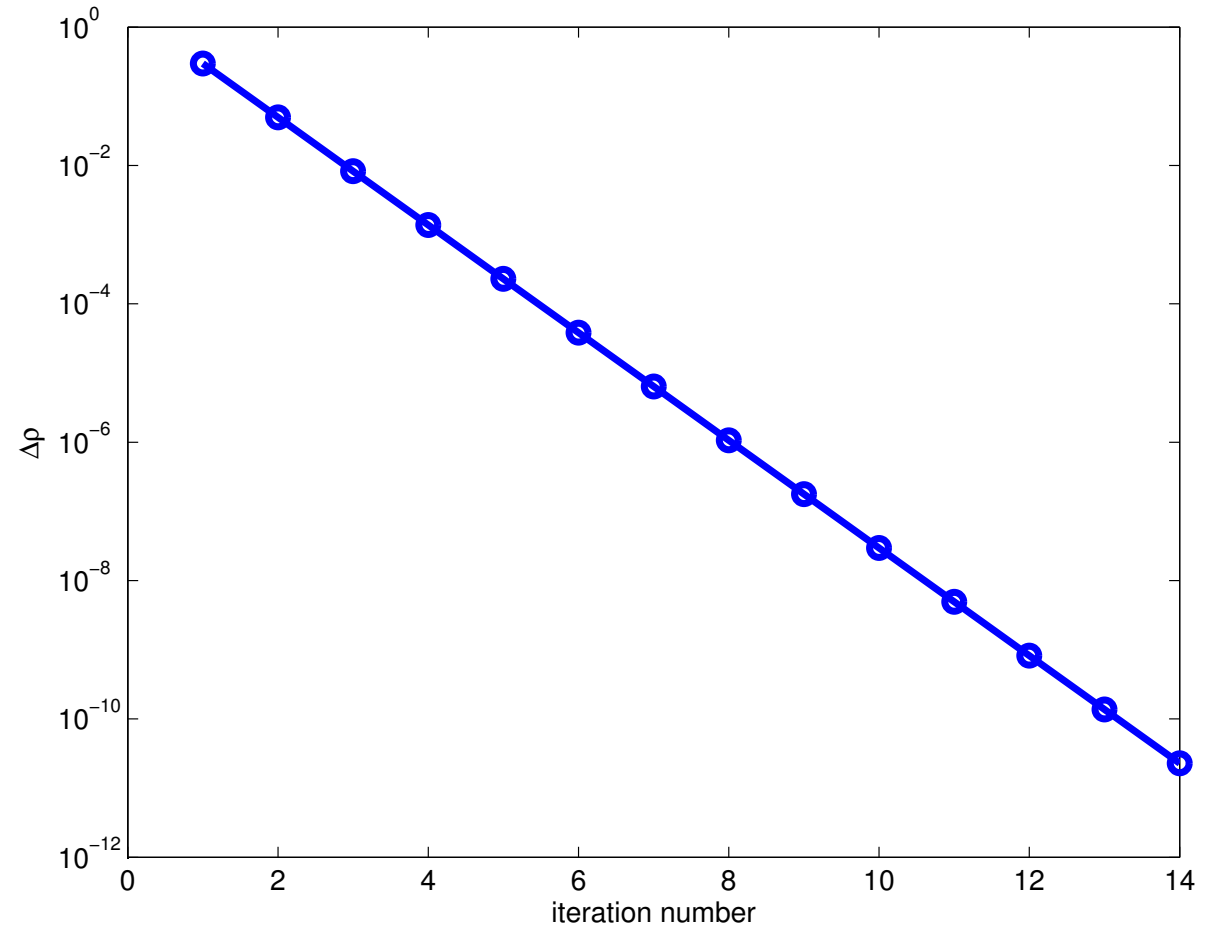
$$L = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \rho(x) = \begin{pmatrix} x_1^2 \\ x_2^2 \end{pmatrix}$$

$$\begin{array}{l} \min E(x) \\ \text{s.t. } x_1^2 + x_2^2 = 1 \end{array}$$

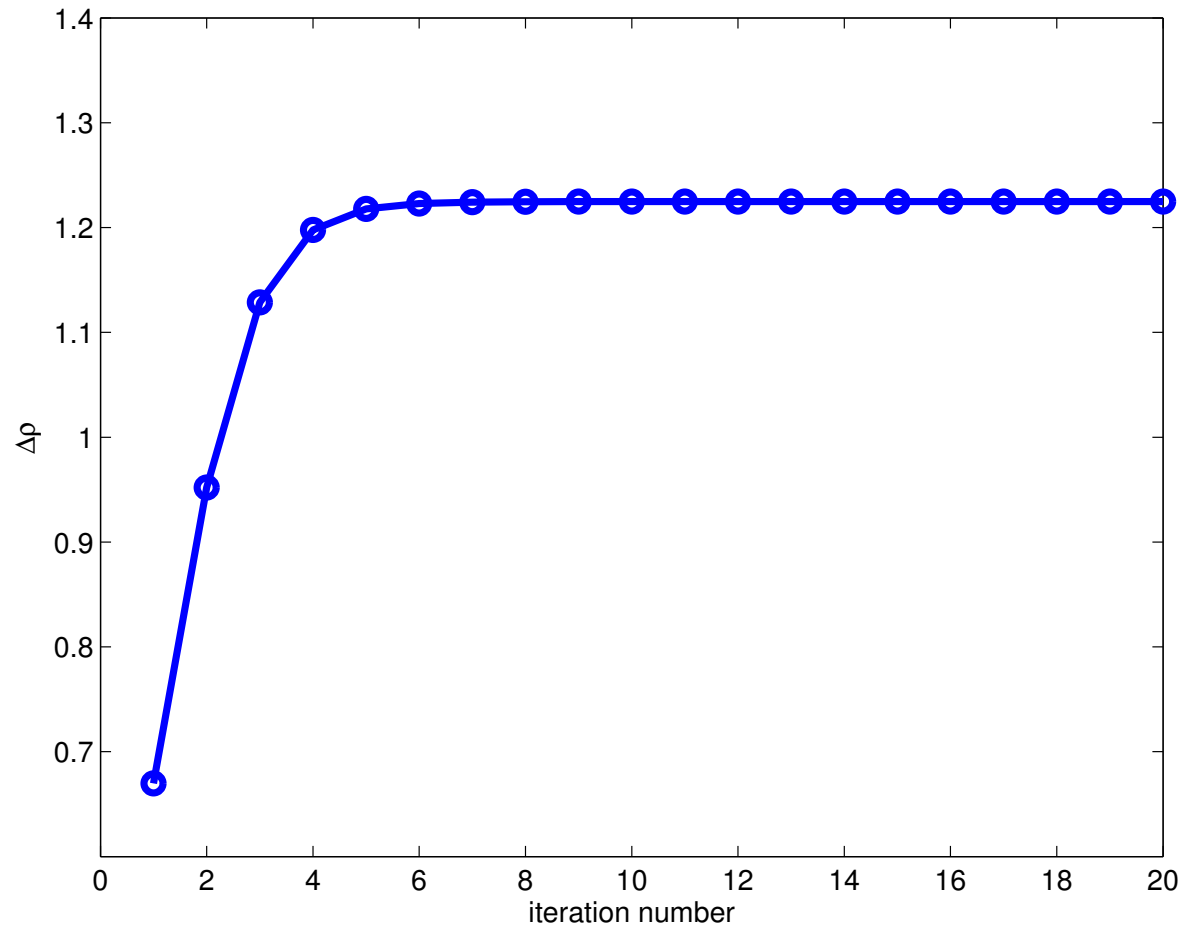
$$\left[L + \alpha \mathbf{Diag}(L^{-1}\rho(x)) \right] x = \lambda_1 x$$

SCF Converges when $\alpha = 2.0$

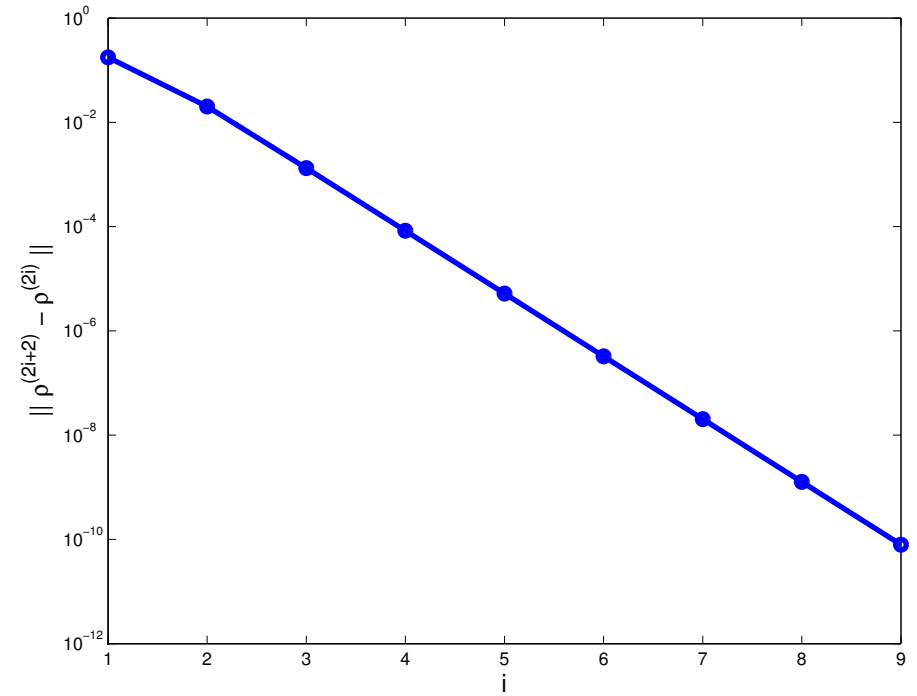
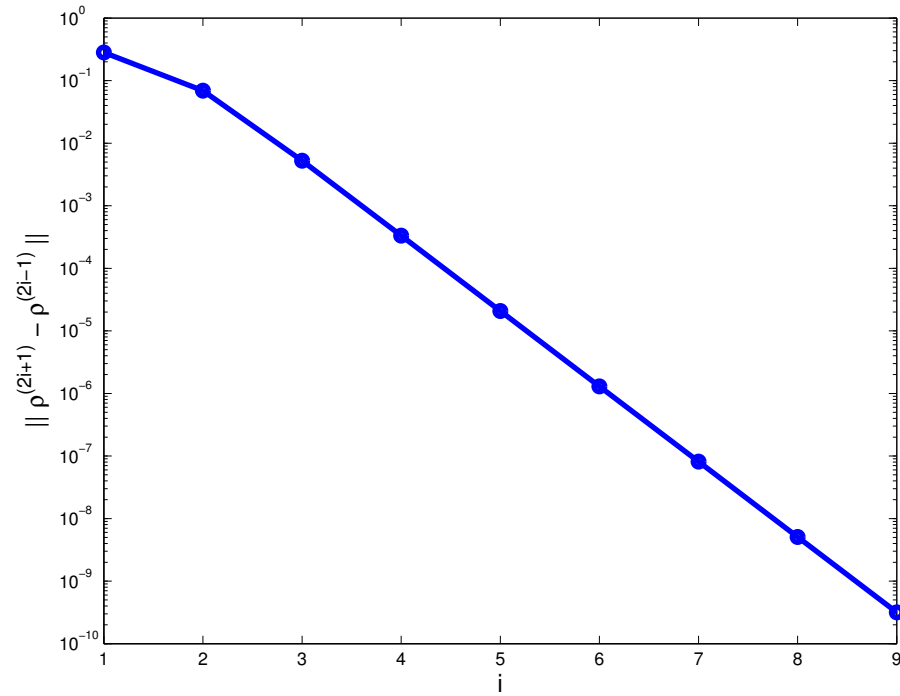
$$\Delta\rho^{(i)} = \|\rho^{(i)} - \rho^{(i-1)}\|$$



SCF fails when $\alpha = 12.0$



Subsequence Convergence



odd
subsequence
(Cancès & Le Bris 2000)

even
subsequence

Why does SCF fail?

- SCF minimizes a sequence of surrogate models

- Objective:

- $E(x) = \frac{1}{2}x^T Lx + \frac{\alpha}{4}\rho(x)^T L^{-1}\rho(x)$

- $E_{sur}(x) = \frac{1}{2}x^T H(x^{(i)})x,$

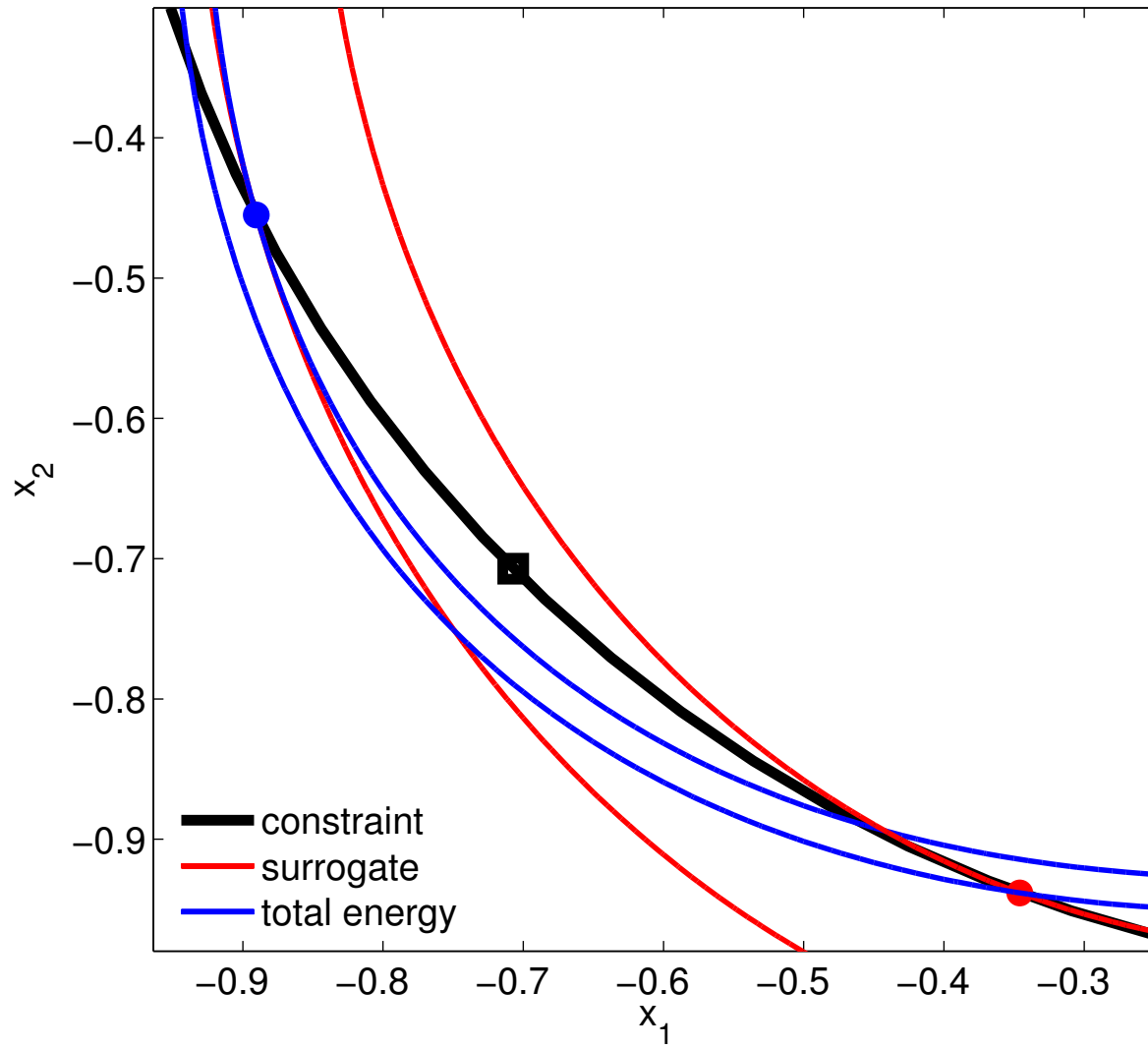
- Gradient:

- $\nabla E(x) = H(x)x$

- $\nabla E_{sur}(x) = H(x^{(i)})x$

$$\nabla E(x^{(i)}) = \nabla E_{sur}(x^{(i)})$$

SCF Moves Too Far



Improving SCF

- Use **Trust Region** to restrict the update of the x in a small neighborhood of the gradient matching point
 - Level-shifting (Saunders & Hillier 1973)
 - (Cancès & Le Bris 2000)
 - TRSCF (Thogersen et al 2004, Francisco, Martinez, Martinez 2006, Yang, Meza & Wang 2007)
- Construct a better surrogate
 - Standard TR model requires Hessian (too expensive!)
 - **Charge mixing** to minimize lack of self consistency (related to Broyden methods)

Trust Region Subproblem

- Solve

$$\begin{aligned} & \min E_{sur}(x) \\ \text{s.t. } & x^T x = 1, \quad \|xx^T - x^{(i)}(x^{(i)})^T\|_F^2 \leq \Delta \end{aligned}$$

(constraint is invariant under unitary transformation)

- Equivalent to solving

$$\begin{aligned} \left[H(x^{(i)}) - \sigma x^{(i)}(x^{(i)})^T \right] x &= \lambda_1 x \\ x^T x &= 1 \end{aligned}$$

σ is a penalty parameter (Lagrange multiplier for the trust region constraint)

How to Choose σ (heuristic)?

- The convergence of SCF depends on the gap between λ_{n_e} and λ_{n_e+1} . (Yang, Gao & Meza 2007)
- If X is the solution to $H(X)X = X\Lambda$, where $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_{n_e}$ are the smallest n_e eigenvalues of $H(X)$ that appear in Λ , then the eigenvalues of $H(X) - \sigma XX^*$ are

$$\lambda_1 - \sigma, \lambda_2 - \sigma, \dots, \lambda_{n_e} - \sigma, \lambda_{n_e+1}, \dots, \lambda_n.$$

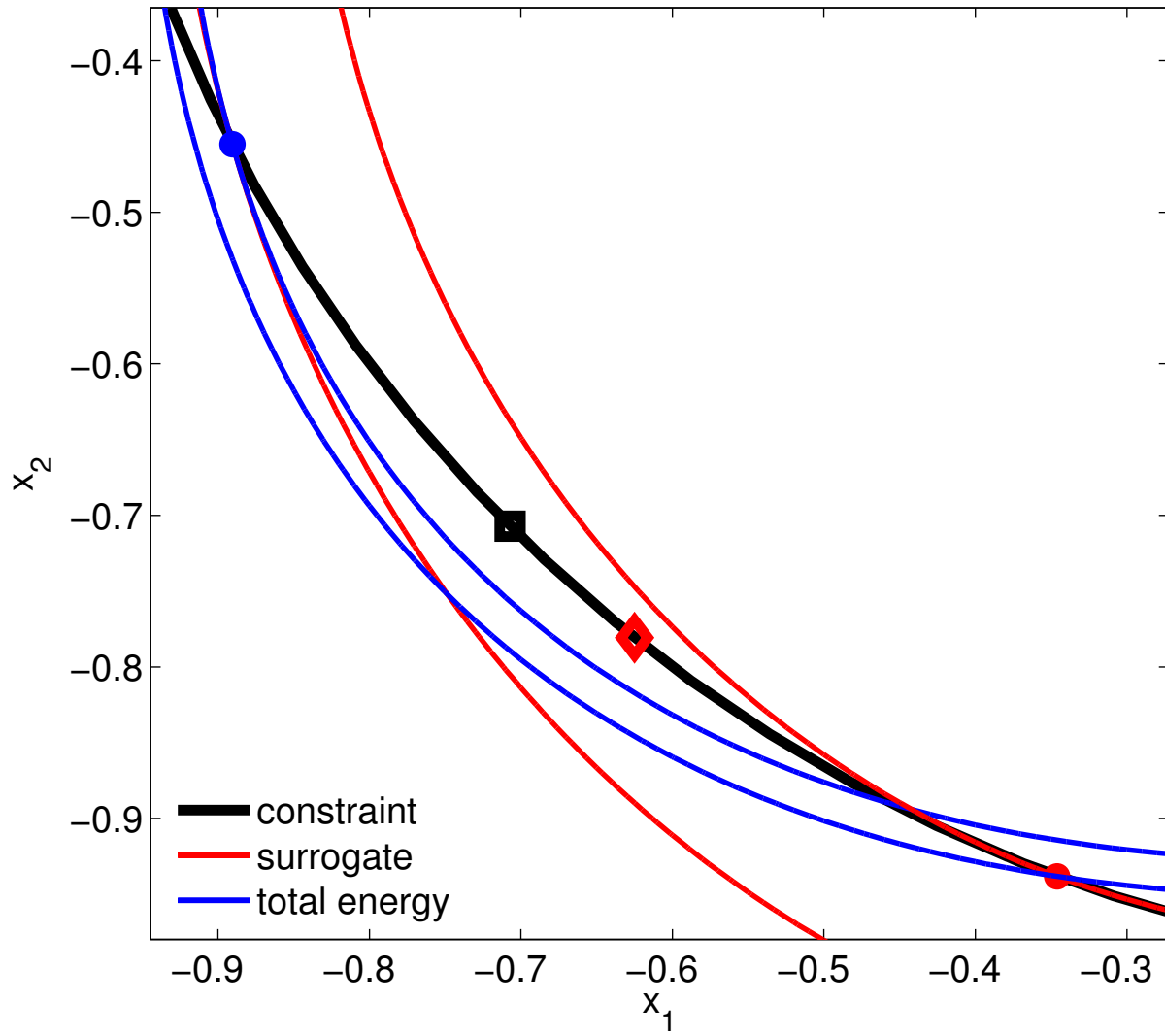
- Strategy: choose σ to open up the gap between $\lambda_{n_e}(\hat{H}(X^{(i)}))$ and $\lambda_{n_e+1}(\hat{H}(X^{(i)}))$, where

$$\hat{H}(X^{(i)}) = H(X^{(i)}) - \sigma X^{(i)}(X^{(i)})^*.$$

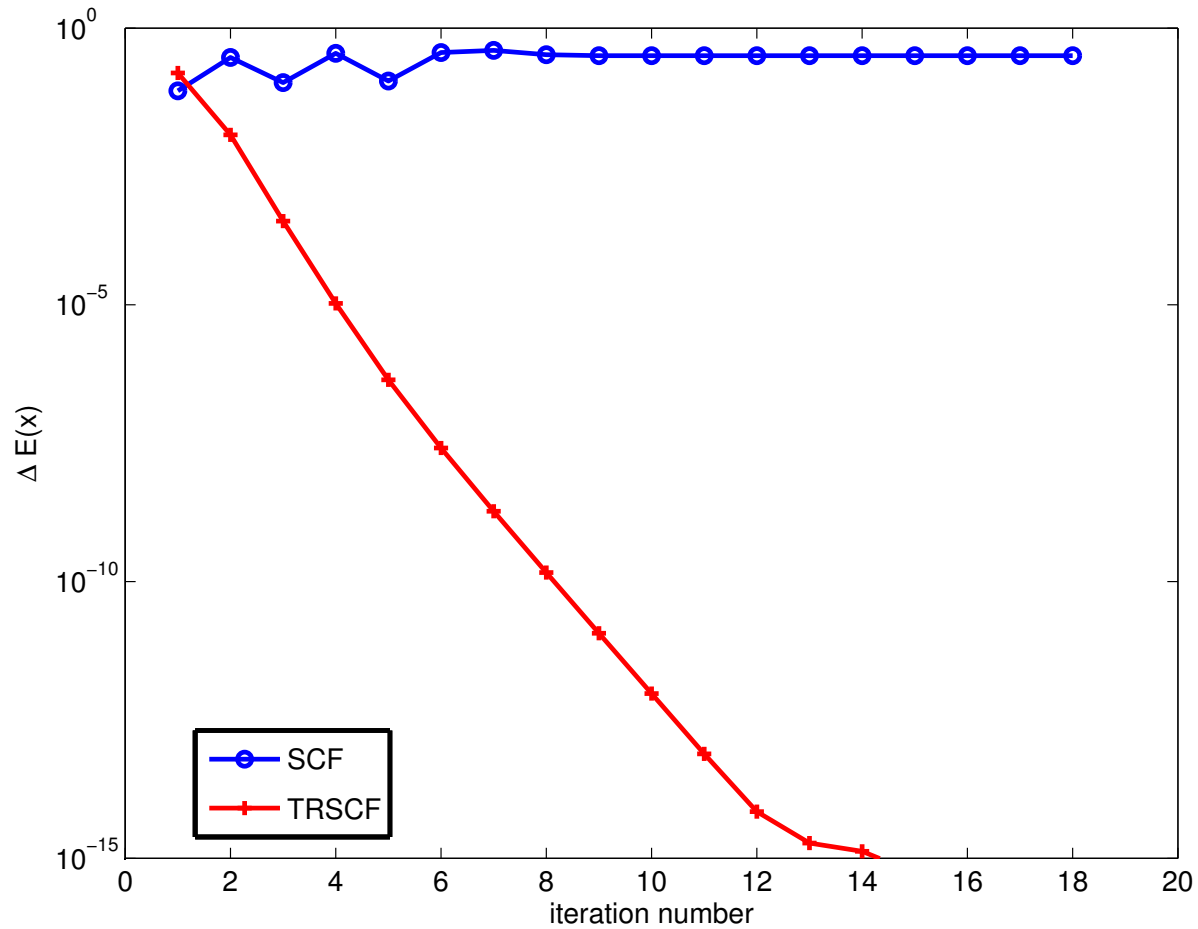
A Recipe for Choosing σ

```
% H = Hamiltonian(X); C = X*X'; sigma = 0;
if (Etot > Etot0)
    repeat: increase sigma
            eig(H-sigma*C)
    until gap(n_e) = max(gaps)
end;
while (Etot > Etot0 & ...)
    increase sigma again;
    eig(H-sigma*C);
end;
Etot0 = Etot;
```

Trust Region SCF (TRSCF)



SCF vs TRSCF



Charge mixing

- Simple mixing

$$\rho^{(i+1)} \leftarrow \tau \rho_{in}^{(i)} + (1 - \tau) \rho_{out}^{(i)}, \quad 0 < \tau < 1.$$

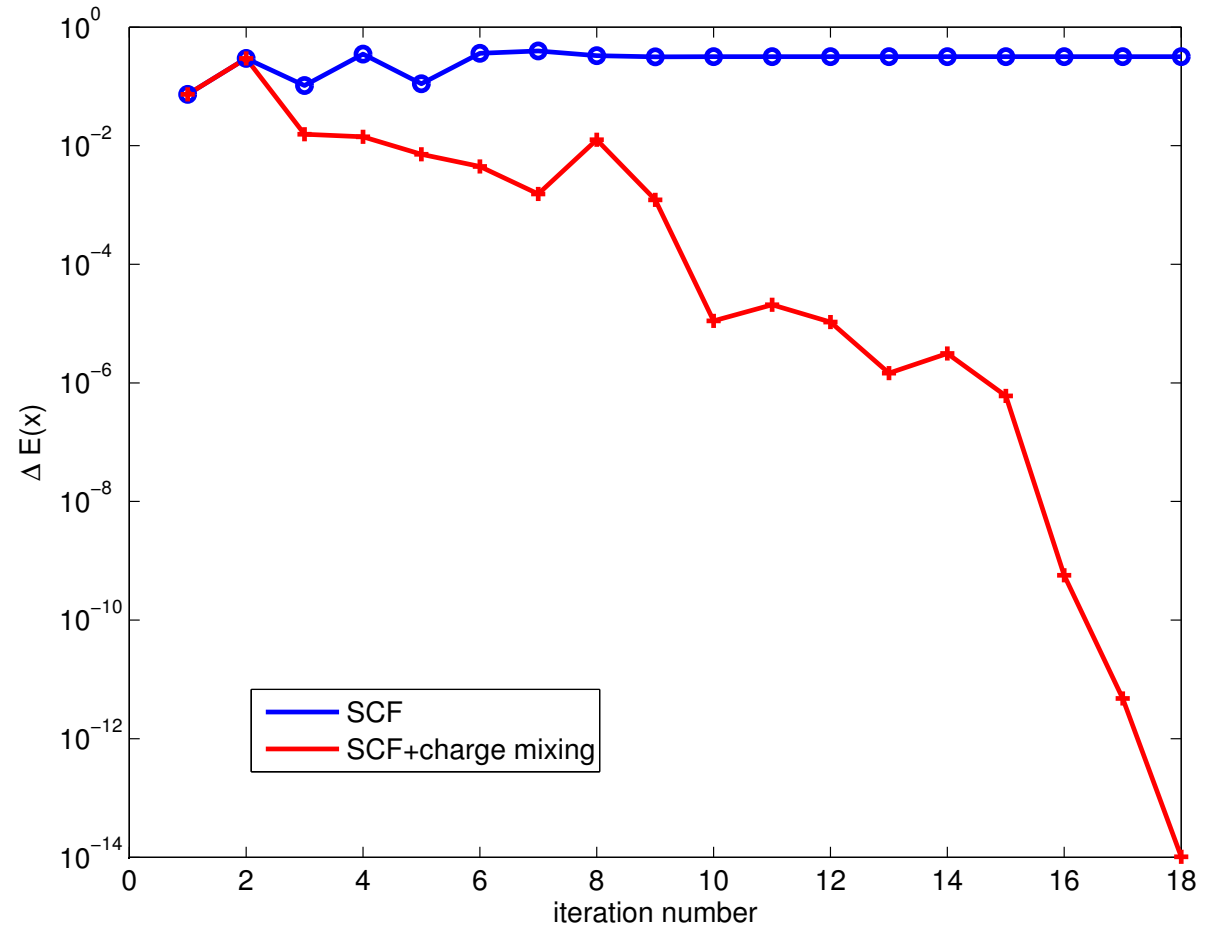
- Pulay mixing (Direct Inversion of Iterative Subspace)

$$\rho^{(i+1)} = \sum_{j=1}^i \alpha_j \rho^{(j)}, \quad \sum_{j=1}^i \alpha_j = 1$$

- Broyden mixing
- Anderson mixing

SCF vs SCF+charge mixing

$$\Delta E(x^{(i)}) = \|E(x^{(i)}) - E_{min}\|$$



Another View of the KS problem

- If $\lambda_{n_e} < \lambda_{n_e+1}$, there exists a filter function $\phi(t)$ such that

$$\phi(\lambda) = \begin{cases} 1, & \text{for } \lambda = \lambda_1, \lambda_2, \dots, \lambda_{n_e}, \\ 0, & \text{for } \lambda = \lambda_{n_e+1}, \lambda_{n_e+2}, \dots, \lambda_n. \end{cases}$$

- In this case

$$\rho \equiv \text{diag}(X_{n_e} X_{n_e}^*) = \text{diag}(\phi(H(\rho))) = X \phi(\Lambda) X^*,$$

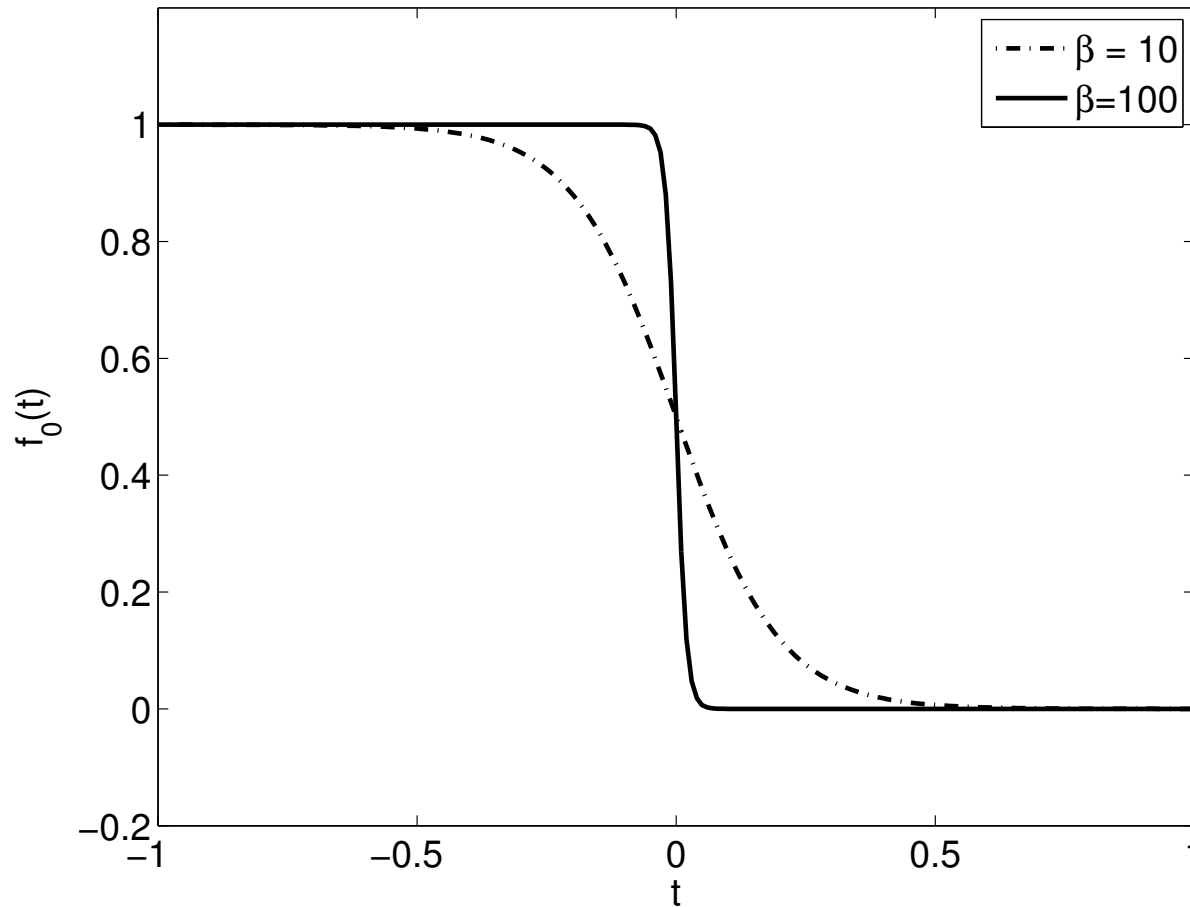
where $X = (X_{n_e}, \dots)$.

- Solving the KS problem is equivalent to solving

$$r(\rho) \equiv \rho - \text{diag}(\phi(H(\rho))) = 0, \|\rho\|_1 = n_e.$$

Choice of ϕ

Fermi-Dirac distribution $\phi(t) \equiv f_{\mu,\beta}(t) = \frac{1}{1+e^{\beta(t-\mu)}}$, $\beta > 0$.



KS Nonlinear Equations

- $n + 1$ equations, $n + 1$ unknowns

$$F(\rho, \mu) \equiv \begin{pmatrix} r(\rho, \mu) \\ \nu(\rho, \mu) \end{pmatrix} \equiv \begin{pmatrix} \rho - \text{diag}[f_{\beta, \mu}(H(\rho))] \\ \text{trace}[f_{\mu, \beta}(H(\rho))] - n_e \end{pmatrix} = 0.$$

- Newton's method

$$\begin{pmatrix} \rho^{(k+1)} \\ \mu^{(k+1)} \end{pmatrix} = \begin{pmatrix} \rho^{(k)} \\ \mu^{(k)} \end{pmatrix} - J_k^{-1} \begin{pmatrix} r(\rho^{(k)}, \mu^{(k)}) \\ \nu(\rho^{(k)}, \mu^{(k)}) \end{pmatrix}.$$

- We can focus on

$$\rho^{(k+1)} = \rho^{(k)} - \frac{\partial r}{\partial \rho} r^{(k)},$$

because ...

Broyden's Method

- $J = \partial r / \partial \rho$ cannot be evaluated analytically
- Setting $J = I$ gives the simple SCF iteration

$$\rho^{(k+1)} = \rho^{(k)} - \left(\rho^{(k)} - \text{diag}[f_{\beta,\mu}(H(\rho^{(k)}))] \right)$$

- Setting $J = \tau I$ gives the simple mixing.
- Broyden's method: successive approximation to J_k or J_k^{-1}

Let $s_k = \rho^{(k)} - \rho^{(k-1)}$, $y_k = r^{(k)} - r^{(k-1)}$, find B or C such

that

$\begin{aligned} \min_B \quad & \frac{1}{2} \ B - B_k\ _F^2, \\ \text{s.t} \quad & B s_k = y_k, \end{aligned}$	or	$\begin{aligned} \min_C \quad & \frac{1}{2} \ C - C_k\ _F^2, \\ \text{s.t} \quad & s_k = C y_k, \end{aligned}$
--	----	--

where $B_k \sim J_k$ (good) and $C_k \sim J_k^{-1}$ (bad).

Updating Formula

- Let $S_k = (s_1, s_2, \dots, s_{k-1})$, $Y_k = (y_1, y_2, \dots, y_{k-1})$ where $s_{j-1} = \rho^{(j)} - \rho^{(j-1)}$ and $y_{j-1} = r^{(j)} - r^{(j-1)}$

- Good Broyden:

$$B = B_k + (Y_k - B_k S_k) S_k^\dagger.$$

Sherman-Morrison-Woodbury:

$$C = C_k + (S_k - C_k Y_k) (S_k^T C_k Y_k)^{-1} S_k^T C_k.$$

- Bad Broyden:

$$C = C_k + (S_k - C_k Y_k) Y_k^\dagger.$$

Practical Issues

- Good Broyden is more expensive when $\text{rank}(S_k) > 1$
- Choice of step length τ and scaling of C_1

$$\rho^{(k+1)} = \rho^{(k)} + \tau C_{k+1} r_k$$

$0 < \tau < 1$ appears to work better. (Marks & Luke 2008)

- Limited-memory and restart: Resetting C_k to τI every m iterations. $m = 1$ gives *Anderson* mixing

$$\rho^{(k+1)} = \rho^{(k)} + \tau r_k + (S_k - \tau Y_k) Y_k^\dagger r_k$$

(Eyert 1996, Fang & Saad 2007)

- Accuracy of function evaluation r_k , optimal rank of S_k ?

Pulay mixing

- Objective:

$$\min_{\rho \in \mathcal{S}} \|f_{\beta, \mu}(\rho) - \rho\|_2^2,$$

where $\mathcal{S} = \{\rho \mid \rho = \sum_{j=1}^k \alpha_j \rho_j, \sum_{j=1}^k \alpha_j = 1.\}$.

- Taylor expansion at the (unknown) solution ρ_* :

$$\begin{aligned} f_{\beta}(\rho) - \rho &= \sum_{j=1}^k \alpha_j \left[(\rho_* - \rho_j) - J_*(\rho_* - \rho_j) \right] + \mathcal{O}(\|\rho - \rho_*\|^2) \\ &= \sum_{j=1}^k \alpha_j r_j + \sum_{j=1}^k \mathcal{O}(\|\rho_j - \rho_*\|^2) + \mathcal{O}(\|\rho - \rho_*\|^2) \end{aligned}$$

where $r_j = f_{\beta, \mu}(\rho_j) - \rho_j$.

Pulay = Anderson with $\tau = 0$

- If we ignore 2nd order terms,

$$\min_{\rho \in \mathcal{S}} \|f_{\beta, \mu}(\rho) - \rho\|_2^2, \iff \min_{a^T e = 1} \|R_k a\|_2^2,$$

where $a = (\alpha_1, \alpha_2, \dots, \alpha_k)^T$, $e = (1, 1, \dots, 1)^T$ and $R_k = (r_1, r_2, \dots, r_k)$.

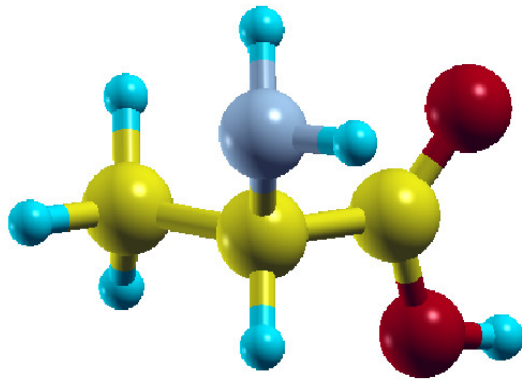
- Let $a = (-\gamma_1, \gamma_1 - \gamma_2, \dots, \gamma_{k-1} - \gamma_{k-2}, 1 - \gamma_{k-1})^T$.

$$\min_{a^T e = 1} \|R_k a\|_2^2 \iff \min_g \|Y_k g - r_k\|_2^2,$$

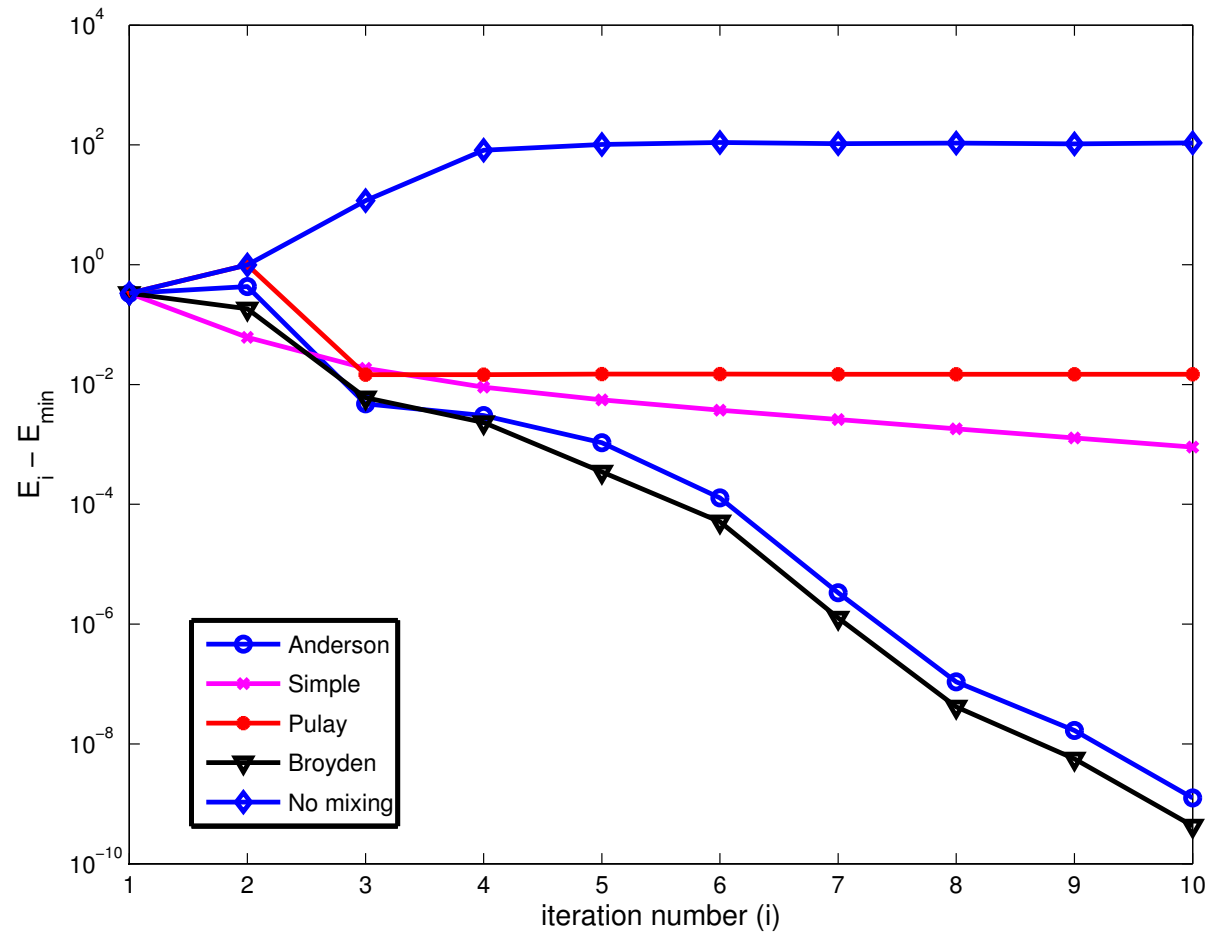
where $Y_k = R_k(:, 2 : k) - R_k(:, 1 : k - 1)$ and $g = (\gamma_1, \gamma_2, \dots, \gamma_{k-1})^T$.

$$\rho^{(k+1)} = \rho^{(k)} + \underbrace{\tau r_k}_{\text{Pulay}} + \underbrace{(S_k - \tau Y_k)}_{\text{Anderson}} Y_k^\dagger r_k$$

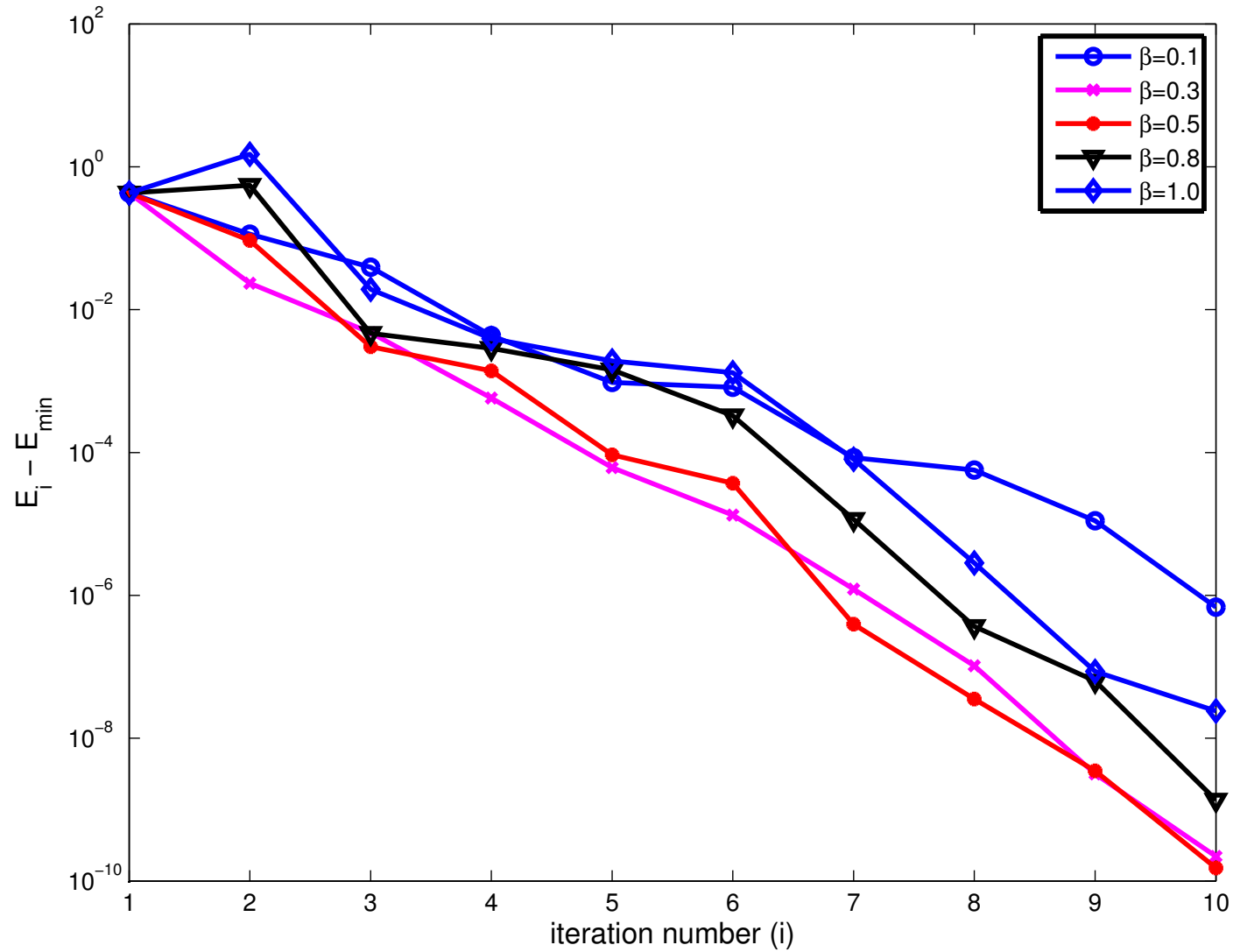
Comparison of Charge Mixing



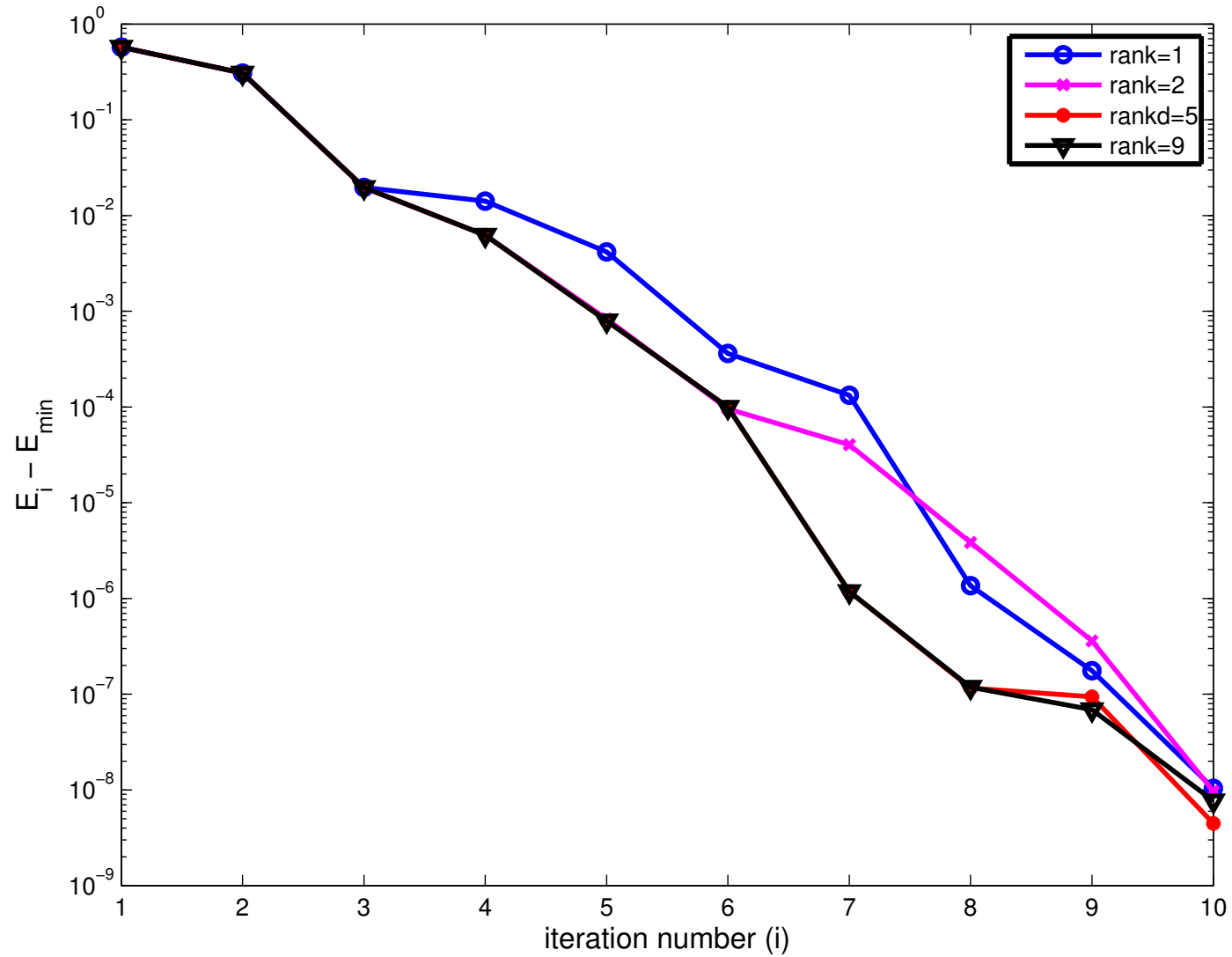
alanine



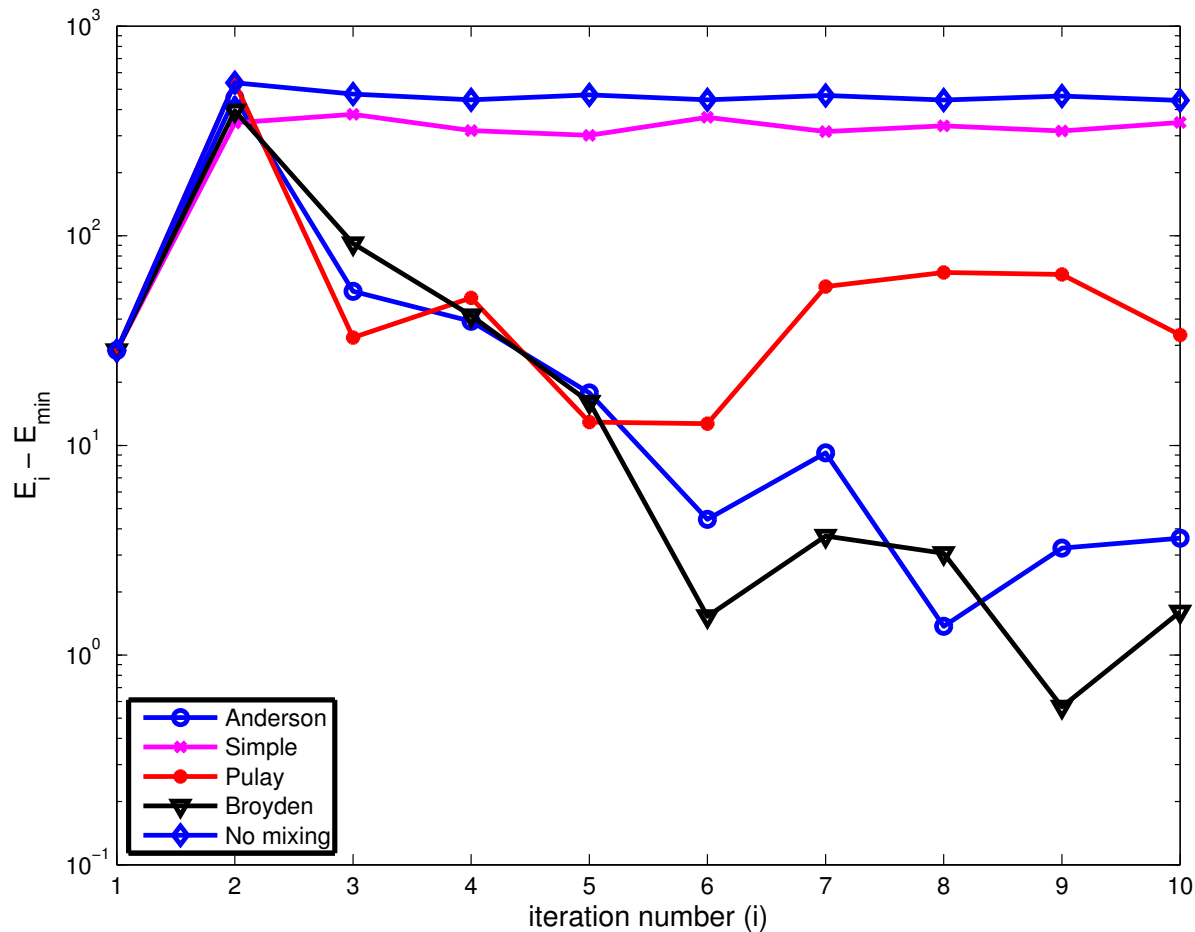
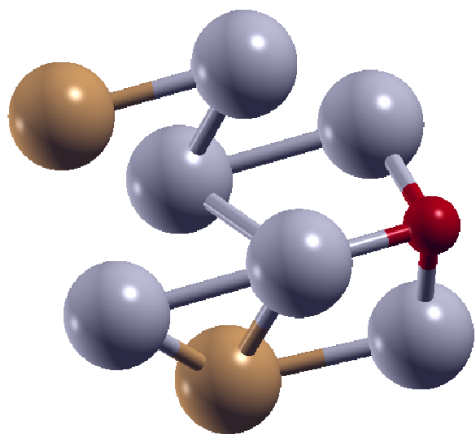
Scaling Matters



Rank Matters



Charge Mixing can Fail



Direct Constrained Minimization

$$\min_{X^* X = I_{n_e}} E(X) \equiv E_{kinetic}(X) + E_{ion}(X) + E_{Hartree}(X) + E_{xc}(X),$$

- Simple schemes
 - Determine a search direction $P^{(i)}$;
 - Perform line search $X^{(i+1)} \leftarrow X^{(i)} + P^{(i)} G_i$ so that $E(X^{(i+1)}) < E(X^{(i)})$;
 - Normalize so that $(X^{(i+1)})^* X^{(i+1)} = I_{n_e}$;
- Use standard constrained minimization solver (e.g., sequential quadratic programming).
- Minimize over Grassmann manifold (Edelman et al. 1998, Voorhis & Head-Gordon 2002)

A Nonlinear CG-like Algorithm

- Assume $X^{(i)}$ is the current approximation;

- Update

$$X^{(i+1)} = X^{(i)}G_1 + P^{(i-1)}G_2 + R^{(i)}G_3;$$

- $P^{(i-1)}$ previous search direction;

- $R^{(i)} = K^{-1}(H^{(i)}X^{(i)} - X^{(i)}\Theta^{(i)});$

- choose G_1 , G_2 and G_3 so that

- $(X^{(i+1)})^* X^{(i+1)} = I_{n_e};$

- $E(X^{(i+1)}) < E(X^{(i)});$

Extension of LOBPCG (Knyazev) to nonlinear

EV

Subspace Minimization

- Let $V = (X^{(i)}, P^{(i-1)}, R^{(i)})$; $X^{(i+1)} = VG$, for some G ;
- Solve

$$\min_{G^T V^T V G = I_e} E(VG)$$

- Equivalent to solving

$$\begin{aligned}\hat{H}(G)G &= BG\Omega \\ G^T BG &= I_{n_e}\end{aligned}$$

where $B = V^T V$ and $\hat{H}(G) = V^T (H(\rho(VG)))V$;

- Trust region constraint:

$$\|GBG^T - G^{(i)}B(G^{(i)})^T\|_F \leq \Delta$$

DCM Algorithm

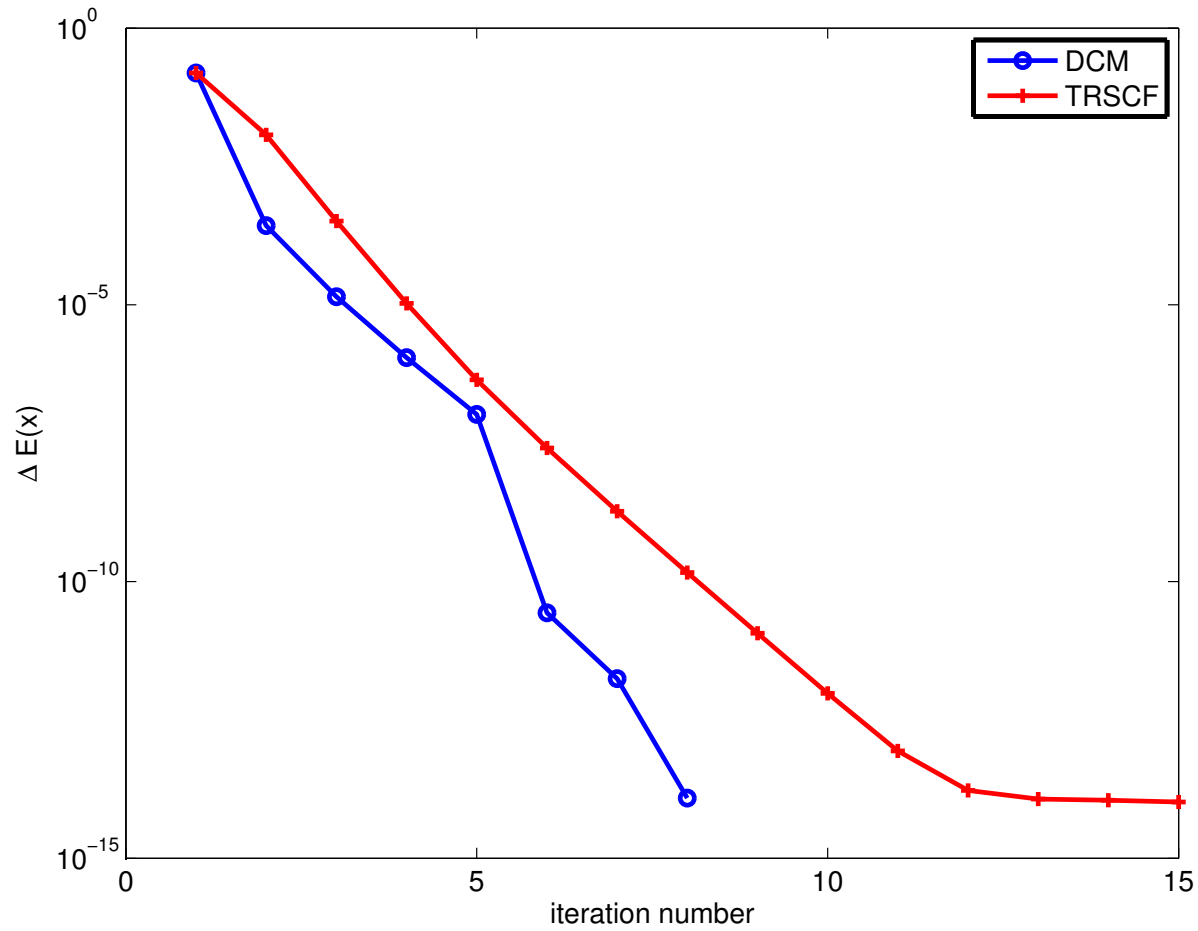
Input: Initial guess $X^{(0)} \in \mathbb{C}^{n \times n_e}$, pseudopotential etc;

Output: X such that $E_{KS}(X)$ is minimized

1. $P^{(0)} = \square, i = 0;$
2. while (not converged)
 - (a) $\Theta^{(i)} = X^{(i)*} H^{(i)} X^{(i)};$
 - (b) $R^{(i)} = H^{(i)} X^{(i)} - X^{(i)} \Theta^{(i)};$
 - (c) Set $V = (X^{(i)}, P^{(i-1)}, K^{-1} R^{(i)});$
 - (d) Solve $\min_{G^* V^* V G = I_k} E_{tot}(VG)$ iteratively;
 - (e) $X^{(i+1)} = VG; P^{(i+1)} = VG(n_e + 1 : 3n_e, :);$
 - (f) $i \leftarrow i + 1;$

(Yang, Meza, Wang, JCP 2006, SISC 2007)

DCM vs. TRSCF



MATVEC Count

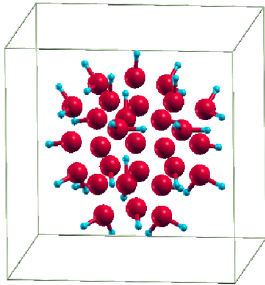
DCM

- Outer iteration
 - n_e MATVEs/iter
- Inner iteration (q times)
 - Update Hamiltonian
 $Y^*[\text{Diag}(\dots)]Y$
- Overall: $n_e + q$
MATVECs/outer iter

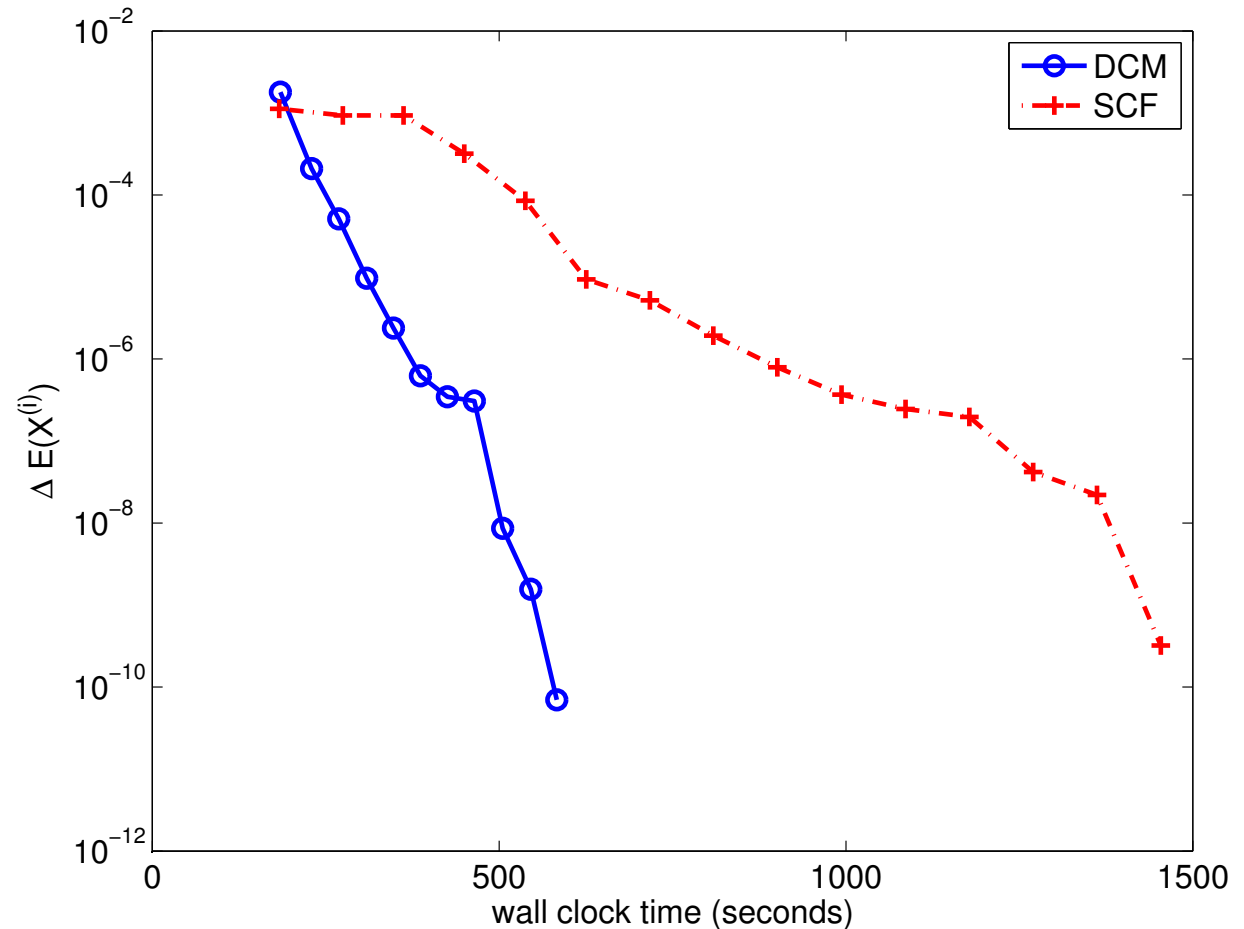
SCF

- Outer iteration:
 - Update Hamiltonian
 $\text{Diag}(\dots\rho\dots)$
- Inner iteration: (p times)
 - n_e MATVECs/iter
- Overall: $n_e \times p$
MATVECs/outer iter

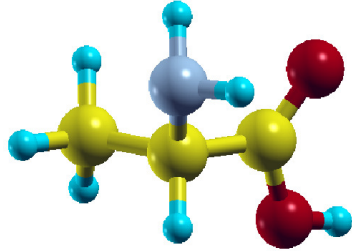
Example 1: $Si_{29}H_{36}$



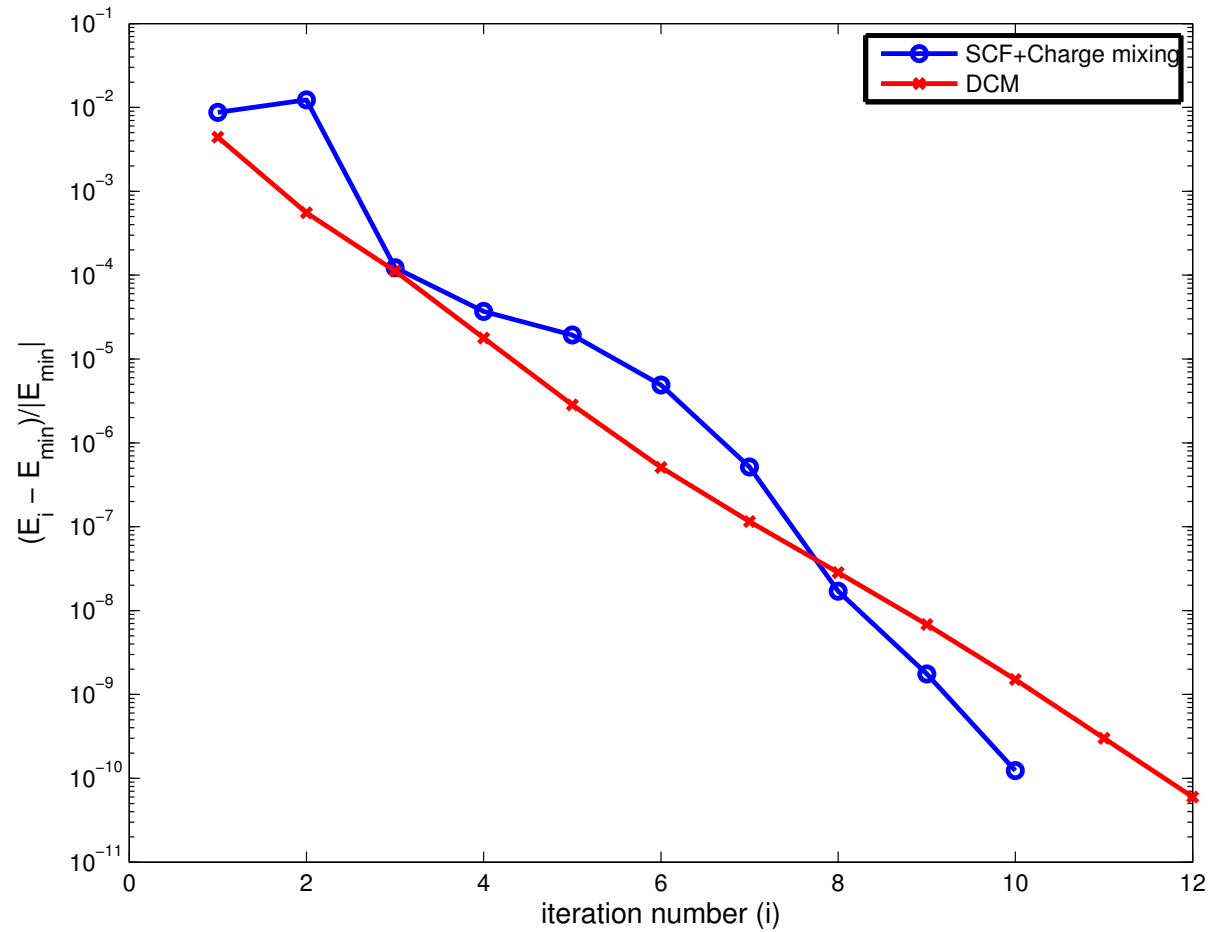
- supercell:
 $25.65 \times 25.65 \times 25.65$
- sampling grid:
 $96 \times 96 \times 96$ (ecut=25 Ryd)
- 10 PCG iterations in each SCF outer iteration. (PETOT)
- 3 inner SCF iteration in each DCM outer iteration



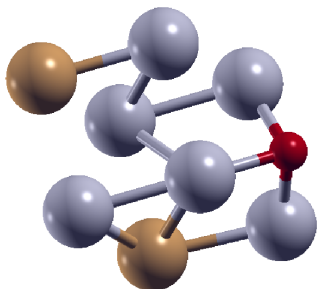
Example 2: Alanine



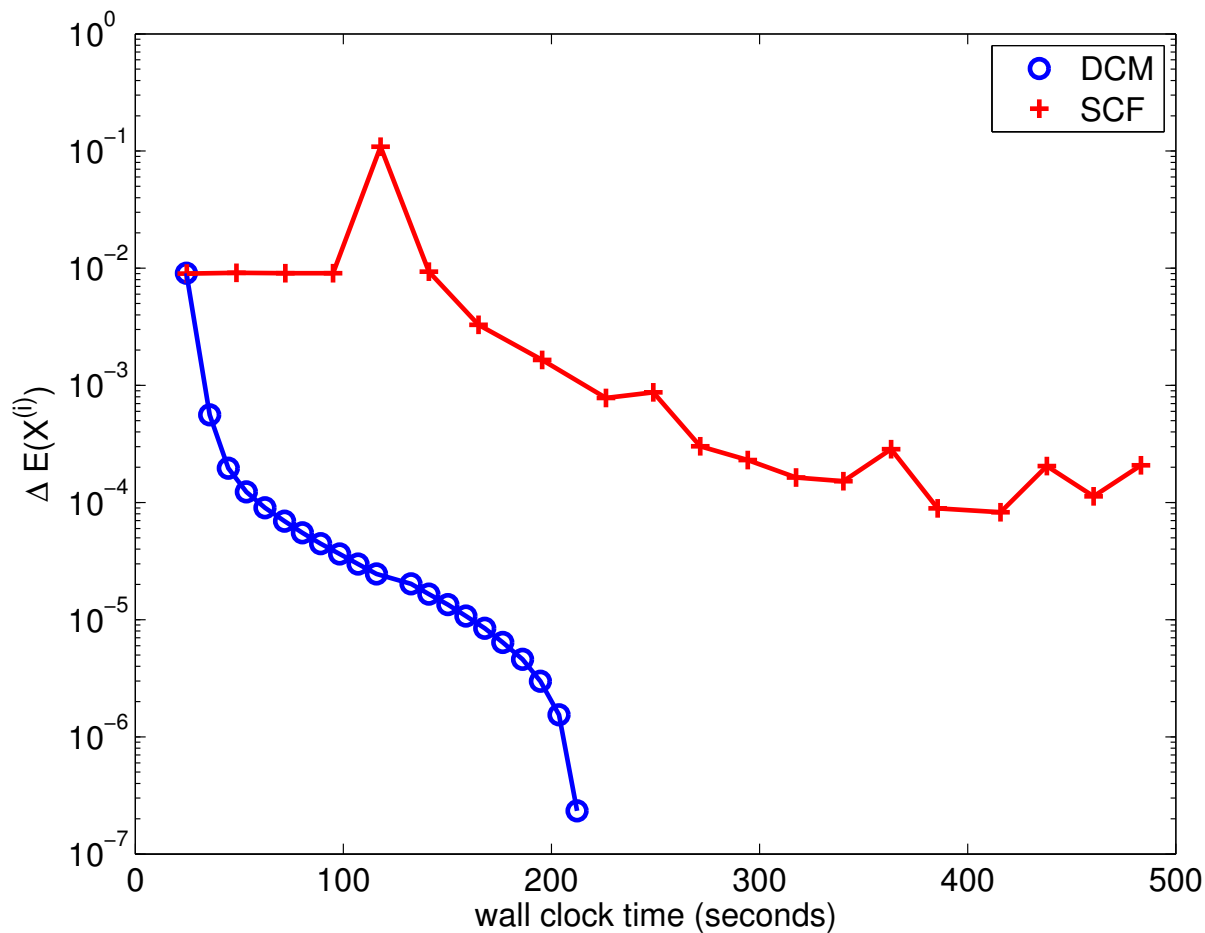
- supercell: $20 \times 15 \times 20$
- sampling grid:
 $96 \times 48 \times 96$ (ecut=25 Ryd)
- 10 PCG iterations in each SCF outer iteration.
- 3 inner SCF iteration in each DCM outer iteration



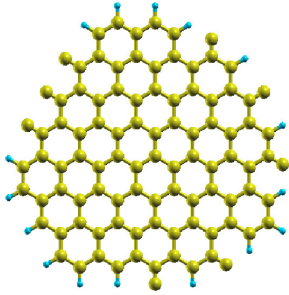
Example 3: Pt_6Ni_2O



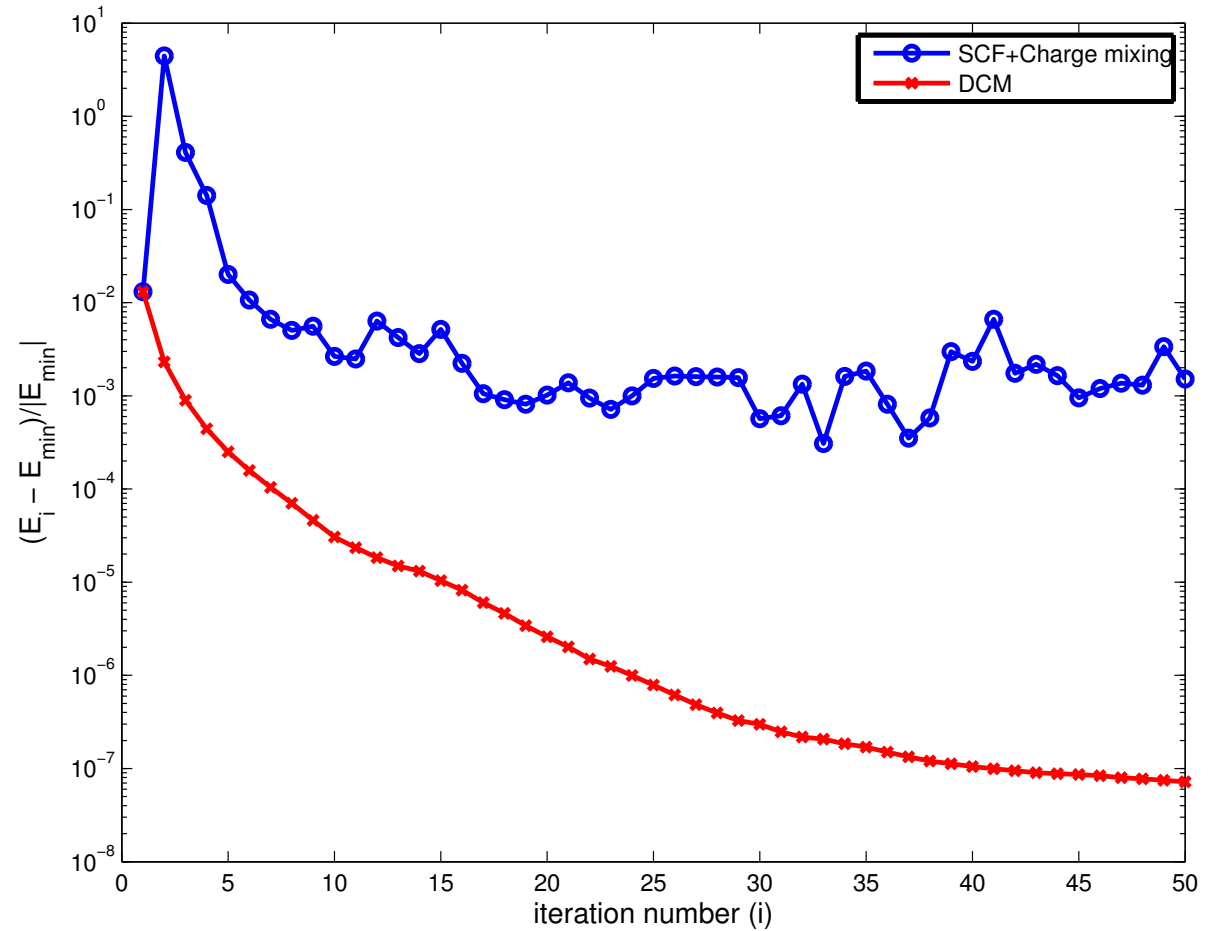
- supercell:
 $19.6 \times 10.7 \times 9.2$
- sampling grid:
 $96 \times 48 \times 48$
- 10 PCG iterations in each SCF outer iteration.(PETOT)
- 5 inner SCF iteration in each DCM outer iteration



Example 4: Graphene



- supercell: $40 \times 40 \times 5$
- sampling grid: $114 \times 114 \times 15$
- 10 PCG iterations in each SCF outer iteration.
- 5 inner SCF iteration in each DCM outer iteration



Summary

- Solving the Kohn-Sham problem from the optimization & nonlinear Equation points of views;
- Techniques for improving the convergence of SCF
 - Trust Region
 - Charge mixing
- Direct Constrained Minimization