A mathematical analysis of stochastic density functional theory

Telluride Science Workshop:
Stochastic Electronic Structure Methods
June 4, 2025

Michael Lindsey

UC Berkeley

Background on DFT

Density functional theory (DFT):
effective single-particle Hamiltonian in terms of unknown functional

$$v_{\mathrm{hxc}}[\rho](x) := \frac{\delta E_{\mathrm{hxc}}[\rho]}{\delta \rho(x)}$$

Non-interacting Hamiltonian:

$$H := -\frac{1}{2}\Delta + v_{\mathrm{ext}}$$

Seek **self-consistent** orthonormal $\{\psi_j\}_{j=1}^{N}$:

$$(H + v_{\mathrm{hxc}}[\rho])\,\psi_j = \varepsilon_j\,\psi_j, \quad j = 1,\ldots,N$$

$$P(x,y) = \sum_{j=1}^{N} \psi_j(x)\,\psi_j(y)$$

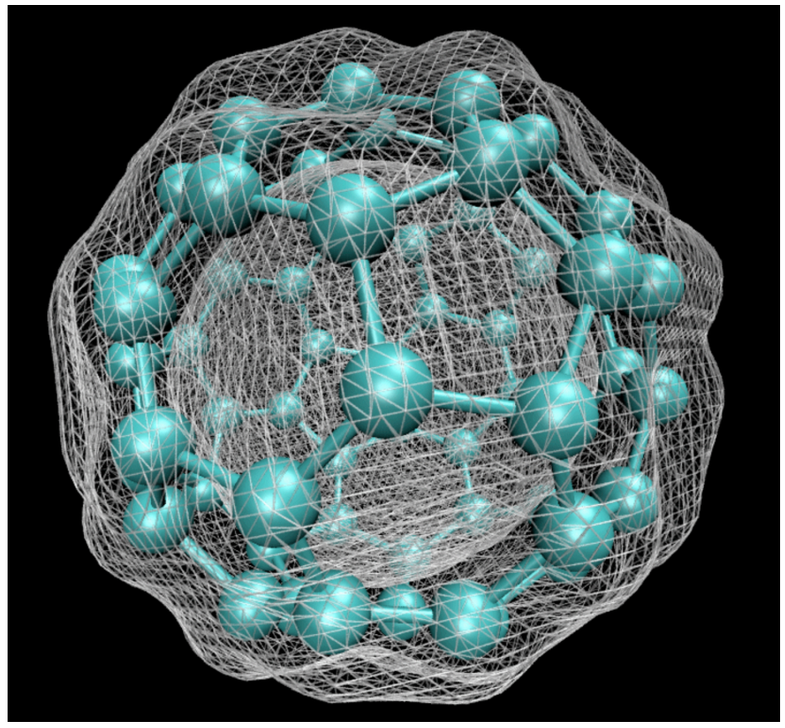
$$\rho(x) = P(x,x)$$

Self-consistent equations including exchange and correlation effects
W Kohn, LJ Sham
Physical review, 1965 • APS

Abstract
From a theory of Hohenberg and Kohn, approximation methods for treating an inhomogeneous system of interacting electrons are developed. These methods are exact for systems of slowly varying or high density. For the ground state, they lead to self-consistent equations analogous to the Hartree and Hartree-Fock equations, respectively. In these equations the exchange and correlation portions of the chemical potential of a uniform electron gas appear as additional effective potentials.(The exchange portion of

SHOW MORE

☆ Save Cite Cited by 72111 Related articles All 29 versions



Electron density of C_{60} molecule computed via DFT (source: Isaac Tamblyn)

Background on DFT

Consider for simplicity a grid discretization with M points:

$$(H + v_{\mathrm{hxc}}[\rho])\,\psi_j = \varepsilon_j\,\psi_j, \quad j = 1,\ldots,N$$

$$P = \sum_{j=1}^{N} \psi_j\,\psi_j^{\top}$$

$$\rho = \mathrm{diag}(P)$$

We can state results for a general (orthonormal) quantum chemistry basis $\{\phi_j\}_{j=1}^{M}$, in terms of which

$$\psi_j(x) = \sum_{i=1}^{M} U_{ij}\,\phi_i(x),$$

where $U \in \mathbb{R}^{M\times N}$ has orthonormal columns

Removing particle-number dependence

- Complexity of solving self-consistent equation is at least $O(MN^2)$
  - Orthogonalization of N orbitals $\psi_j,\ j = 1,\ldots,N$

- **Idea:** avoid orbitals and work with density operator

$$P = f_\beta(H + v_{\mathrm{hxc}}[\rho] - \mu\mathbf{I})$$
$$\rho = \mathrm{diag}(P)$$

$$f_\beta(x) = \frac{1}{1 + e^{\beta x}} \quad \text{(Fermi-Dirac)}$$

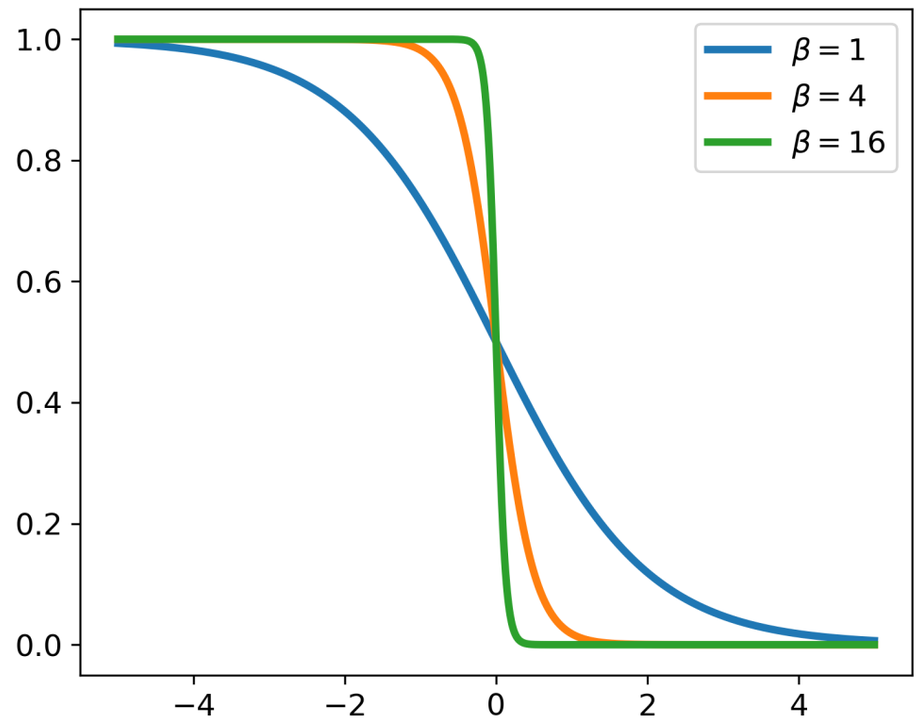
- $\beta \to \infty$ yields step, but works for arbitrary β

- **PEXSI** [Lin et al]: can approximate (PEX)

$$f_\beta(x) \approx \sum_i w_p\,(z_p - x)^{-1}$$

hence only need diagonal of inverses (SI)

  - Complexity dictated by Cholesky fill-in
  - $O(M)$, $O(M^{1.5})$, $O(M^2)$ in $d = 1, 2, 3$



Sorry, just Hartree!

$$\rho = \mathrm{diag}\left[f_\beta (H + v_\mathrm{h}[\rho] - \mu \mathbf{I}) \right]$$

Hartree theory defined by

$$E_\mathrm{hxc}[\rho] = E_\mathrm{h}[\rho] = \frac{1}{2}\rho^\top V \rho \qquad v_\mathrm{hxc}[\rho] = v_\mathrm{h}[\rho] = \mathrm{diag}^*[V\rho]$$

where V is the Coulomb kernel (or arbitrary PSD)

Why?	**Stylized conclusion:** solving the self-consistent Hartree theory is as easy as estimating the Hartree energy
(1) convex (2) quadratic	

(Algorithm is the same for general v_hxc)

Practical conclusion (?) : Converge Hartree as an inner loop

What is optimal scaling?

Wanted:

$\tilde{O}(M)$ scaling to achieve a fixed error ε, independent of the problem

(including correlation length, chemical potential $\mu \leftrightarrow$ particle number $N = \mathrm{Tr}[P]$)

Error dependence?

$\tilde{O}(M\varepsilon^{-2})$, where ε is "relative free energy error"

β dependence?

Thermodynamic limit (fixed basis resolution): additional factor of $O(\sqrt{\beta})$

Complete basis set limit: additional factor of $O(\beta)$

Determining μ (given N) ?

Additional factor of ε^{-1} (estimate dual objective on μ grid and search)

Diagonal estimator

Can we estimate the diagonal of $P = f_\beta\left(H + V_{\mathrm{hxc}}[\rho] - \mu\mathbf{I}\right)$ using only matrix-vector multiplications?

Generic approach: use identity

$$\mathrm{diag}[P] = \mathbb{E}_{z\sim\mathcal{N}}\left[z \odot Pz\right]$$

Can view this as a Hutchinson estimator via

$$P_{ii} = \mathrm{Tr}[e_i e_i^\top P] = \mathbb{E}_{z\sim\mathcal{N}}[z^\top (e_i e_i^\top P) z]$$

However, the variance is bad!

$$\mathrm{Var}(\hat{P}_{ii}) = \|P_{:,i}\|_2^2$$

(depends on correlation length)

Diagonal estimator

Alternative approach (always better)

Instead use identity

$$\mathrm{diag}[P] = \mathbb{E}_{z\sim\mathcal{N}}\left[(P^{1/2}z)\odot(P^{1/2}z)\right]$$

Can view this as a Hutchinson estimator via

$$P_{ii} = \mathrm{Tr}\left[P^{1/2}e_i e_i^{\top}P^{1/2}\right] = \mathbb{E}_{z\sim\mathcal{N}}\left[z^{\top}(P^{1/2}e_i e_i^{\top}P^{1/2})\,z\right]$$

Now the variance is good!

$$\mathrm{Var}(\hat{P}_{ii}) = 2\,(P_{ii})^2$$

(relative variance = 2; sub-exponential concentration [Meyer et al (2021)])

Stochastic DFT

[Drabold and Sankey, *PRL* (1993)] [Baer, Neuhauser, and Rabani *PRL* (2013)] [Fabian et al, *WIREs* (2019)]

$$\rho = \mathrm{diag}\left[f_\beta (H + v_\mathrm{h}[\rho] - \mu \mathbf{I}) \right]$$

Stochastic DFT:

To perform SCF, we compute "stochastic orbitals"

$$\psi^{(s)} := f_\beta^{1/2}(H_\mathrm{eff})\, z^{(s)} \Big/ \sqrt{S} \qquad (z^{(s)} \sim \mathcal{N},\ s = 1, \ldots, S)$$

in terms of which

$$\hat{P} = \sum_s \psi^{(s)}\, \psi^{(s)\,\mathsf{T}} \qquad \hat{\rho} = \sum_s \psi^{(s)} \odot \psi^{(s)}$$

(equivalent to our diagonal estimator)

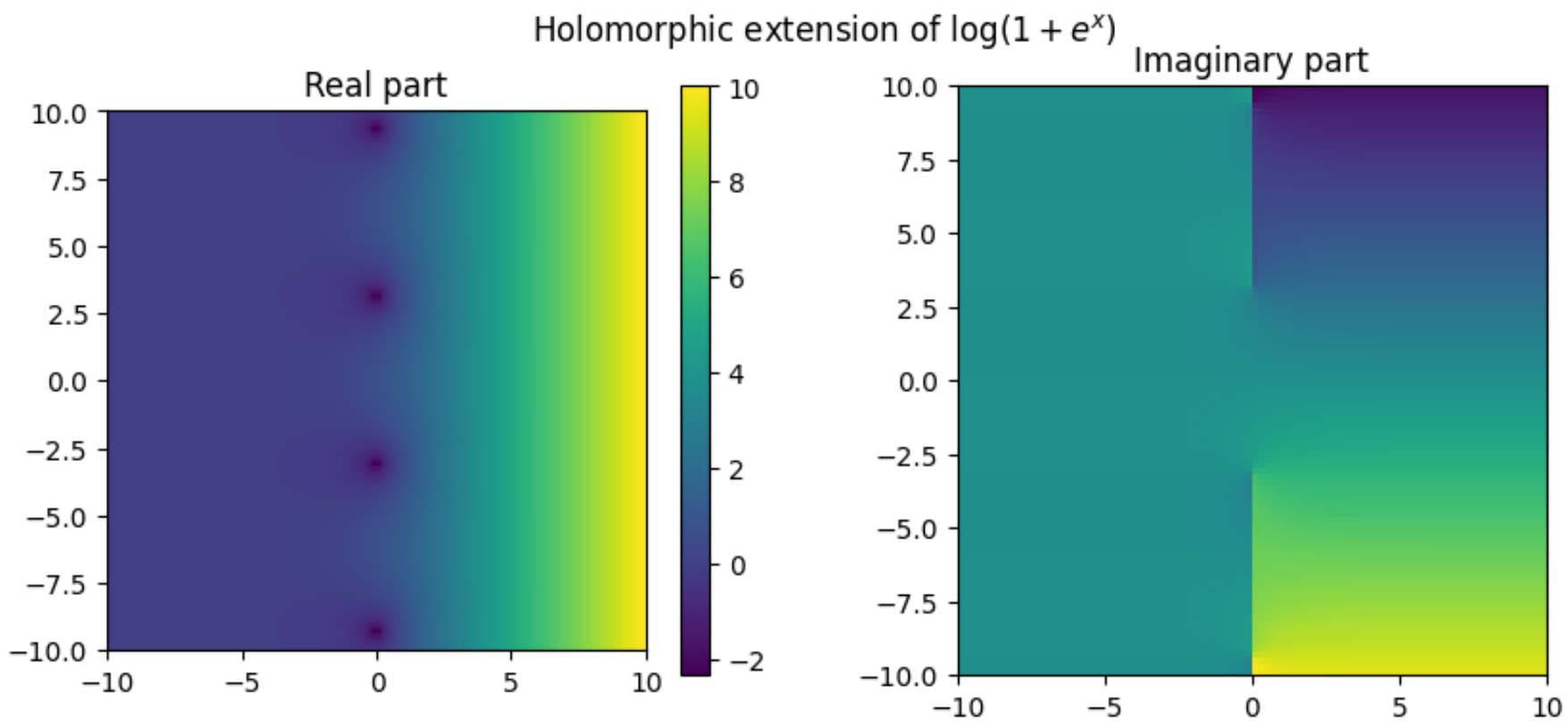
How to compute matvecs by $f_\beta^{1/2}(H_\mathrm{eff})$?

Chebyshev expansion with $O\left(\sqrt{\beta\, \|H_\mathrm{eff}\|}\right)$ terms

Works well for many basis sets but not robust to complete basis set limit

Pole expansion works for $f_\beta^{1/2}$

- **Recall:** $f_1(z) = (1 + e^z)^{-1}$ has poles on the imaginary axis, holomorphic on $i(-\pi, \pi)$
- For an appropriate choice of branch, so is $f_1^{1/2}$
- Can reduce to good choice of branch for $\log(1 + e^z)$ [*also necessary to estimate entropy*]
- Therefore pole expansion is possible with same contour as PEXSI

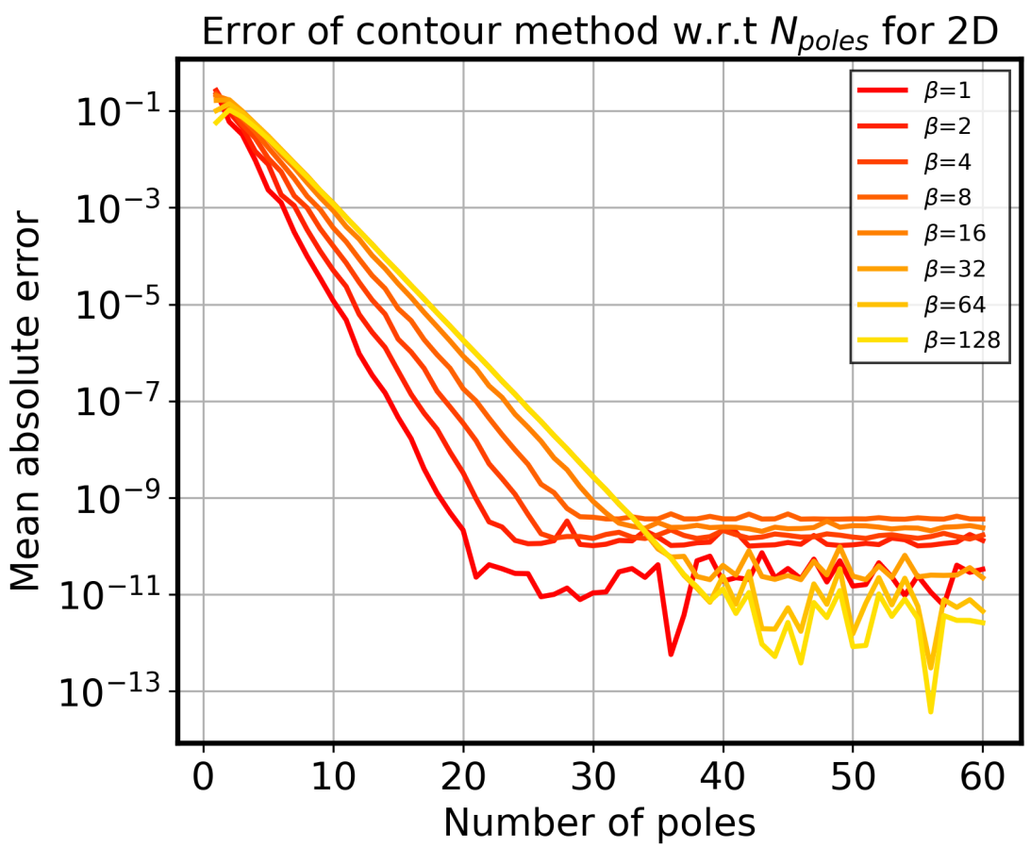


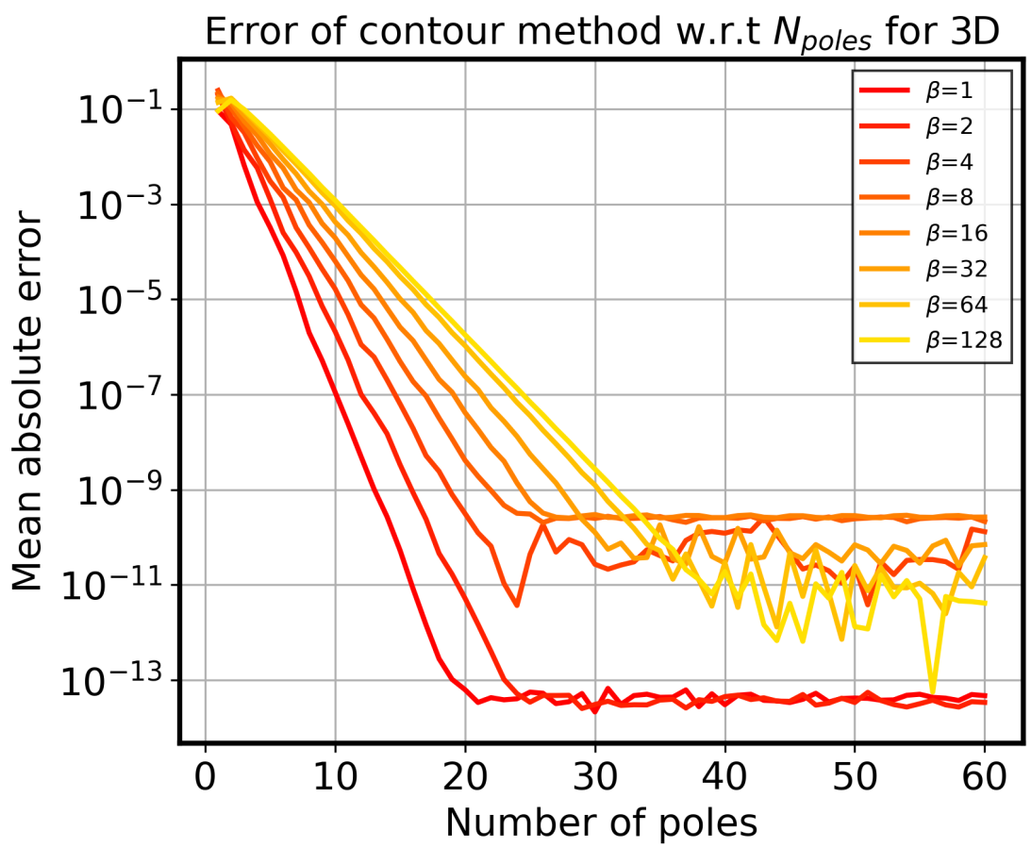
Pole expansion works for $f_\beta^{1/2}$

- Require $O\left(\log\left(\beta\,\|H_{\text{eff}}\|\right) + \log(1/\varepsilon)\right)$ poles for ε-accuracy

- Linear solves needed to implement the matrix function can be preconditioned with the Laplacian (so iterations remain bounded in complete basis set limit)

$$f_\beta^{1/2}(H_{\text{eff}})\,x = \sum_p w_{\beta,p}\left(z_{\beta,p}\,\mathbf{I} - H_{\text{eff}}\right)^{-1} x$$





Optimization formulation

$$\begin{aligned} \underset{X \in \mathbb{R}^{M \times M}}{\text{minimize}} \quad & E(X) + \beta^{-1} S_{\mathrm{FD}}(X) \\ \text{subject to} \quad & 0 \preceq X \preceq \mathbf{I} \end{aligned}$$

$$E(X) = \mathrm{Tr}\,[CX] + E_{\mathrm{h}}(X)$$

$$E_{\mathrm{h}}(X) = \frac{1}{2}\mathrm{diag}(X)^{\top} V\,\mathrm{diag}(X)$$

$$\begin{aligned} S_{\mathrm{FD}}(X) \;=\; & \mathrm{Tr}[X \log X] \\ & + \mathrm{Tr}[(\mathbf{I} - X)\log(\mathbf{I} - X)] \end{aligned}$$

- X is a variable for the density matrix, and the "cost matrix" is $C = H - \mu\mathbf{I}$
  - We are bringing the notation closer to semidefinite programming (SDP)
- Optimality condition is equivalent to self-consistency condition stated earlier
- Fermi-Dirac entropy acts as soft barrier for $\{0 \preceq X \preceq \mathbf{I}\}$
- We'll approach this problem via **mirror descent** viewing S_{FD} as Bregman potential
  - In fact, the algorithm will be the same as SCF with mixing

Mirror descent

- **Bregman potential:** $S(x)$, convex on $\mathcal{X}$
- **Bregman divergence:** $D(y\|x) = S(y) - S(x) - \langle \nabla S(x), y - x \rangle$
  - $D(y\|x) = 0$ if and only if $x = y$
- **Mirror map:** $\nabla S : \mathcal{X} \to \mathcal{X}^*$, where $\mathcal{X}^*$ is the domain of the convex conjugate S^*
  - **Inverse** $\nabla S^* : \mathcal{X}^* \to \mathcal{X}$
- **Update rule** for minimizing an objective $f : \mathcal{X} \to \mathbb{R}$:

$$\nabla S(x_{t+1}) = \nabla S(x_t) - \eta_t \nabla f(x_t)$$

- **Typical initialization** is "maximum entropy state" $x_0 = \operatorname{argmin}_{x \in \mathcal{X}} S(x)$
  - Then $D(x\|x_0) = S(x) - S(x_0)$, often bounded as $O(\log n)$ where n is dimension
- Can view as a **proximal algorithm**

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle}_{\text{linearization}} + \frac{1}{\eta_t} D(x\|x_t) \right\}$$

Typical usage, new usage

- Major use case:
  - **Domain:** $\mathcal{X} = \{x \in \mathbb{R}^n \ : \ x \geq 0, \mathbf{1} \cdot x = 1\}$ [*probability simplex*]
  - **Bregman potential:** $S(X) = x \cdot \log(x)$ [*Shannon entropy*]
  - **Bregman divergence:** $D(y\|x)$ is the *KL-divergence*
  - **Mirror map:** $\nabla S(x) = \log(x) \ + \ \text{const}\,.$
    - **Inverse:** $\nabla S^*(x^*) \propto \exp(x^*)$
    - **Mirror descent:** take usual gradients but do gradient descent in "log-space"
- Minor use case:
  - Quantum analogues (density operators, von Neumann entropy, etc.)
- Our use case:
  - $\mathcal{X} = \{0 \preceq X \preceq \mathbf{I}\}$ [*fermion density matrices*] , $\mathcal{X}^* = \{\text{symmetric matrices}\}$
  - $S(X) = S_{\text{FD}}(X)$ [*Fermi-Dirac entropy*]
  - $\nabla S^*(X^*) = f_1(-X^*)$ [*Fermi-Dirac occupation function*]

Typical convergence ingredients

- Many alternative assumptions give different convergence guarantees
  - e.g., strong convexity of objective, strong smoothness of objective, etc.
- **Relevant assumptions:**
  - Gradient bound $\|\nabla f(x)\| \leq c$ in some norm
  - Strong convexity of Bregman potential with respect to dual norm
    - i.e., $z^\top \nabla^2 S(x)\, z \geq \|z\|_*^2$ for all $x \in \mathcal{X}$ and $z \in \mathcal{X}^*$
- Extends to **proximal mirror descent**:

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + g(x) + \frac{1}{\eta_t} D(x \| x_t) \right\}$$

where g is possibly nonsmooth and not linearized

- Then typical mirror descent **regret bounds** guarantee objective error $\tilde{O}\left(c/\sqrt{T}\right)$, where T is the number of iterations

Connection to SCF, proximal interpretation

$$\underset{X \in \mathcal{X}}{\text{minimize}}\ F_\beta(X), \qquad F_\beta(X) := E(X) + \beta^{-1} S_{\mathrm{FD}}(X)$$

Abstract algorithm (proximal mirror descent)

$$X_{t+1} = \underset{X \in \mathcal{X}}{\operatorname{argmin}} \left\{ E(X_t) + \left\langle G_t - \mu \mathbf{I}_n, X - X_t \right\rangle + \frac{1}{\beta} S_{\mathrm{FD}}(X) + \frac{1}{\eta_t} D(X \| X_t) \right\}$$

where $G_t = C + v_h[\hat{\rho}_t]$ is our gradient estimator

Via mirror map, view as iteration of **effective Hamiltonians** H_t satisfying $X_t = f_\beta(H_t)$

$$H_{t+1} = (1 - \alpha_t) H_t + \alpha_t G_t$$

where $\alpha_t = (\eta_t + \beta)^{-1} \eta_t \in (0,1)$

If $H_0 = H$, precisely SCF with mixing parameter α_t and stochastic Hartree potential !

Preparing our convergence result

- **Strong convexity**
  - Can show: S_{FD} is $(2/M)$-strongly convex with respect to trace/nuclear norm
- **Gradient bound**
  - $c = \|C\| + c_{\mathrm{h}}$, where c_{h} is a uniform bound on the Hartree potential
- **Dealing with stochastic gradients**
  - Need a subexponential concentration bound for G_t conditioned on X_t
  - Yields a subexponential martingale difference sequence in proof of regret bound, whose sum admits Bernstein type concentration bound for sum
  - Self-averaging of stochastic estimation error
- **Initialization**
  - Take maximum entropy $X_0 = \mathbf{I}/2$, so $D(X\|X_0) = O(\log M)$

Convergence result

Theorem (informal): For time horizon T and step size $\eta_t = O(T^{-1/2})$, with high prob. the Cesàro mean satisfies

$$\frac{F_\beta(\overline{X}_T)}{M} \leq \frac{F_\beta(X_\star)}{M} + \tilde{O}\left(\frac{c_{\mathrm{h}}}{\sqrt{T}}\right) + \frac{\|C\|}{T}$$

Notes

- Slow convergence of regret bound is the same as the slow convergence of stochastic estimator
  - *Converging the Hartree approximation is as fast as estimating the Hartree energy !*
- Free energy error per basis function is satisfactory in thermodynamic limit, not in complete basis set limit
- Neither is $\|C\|$

Modifications for complete basis set

- **Rough idea:** want to show that eigenvalue growth of C confines you to a small subset of $\mathscr{X}$ on which $\mathrm{Tr}[X] = O(1)$ as $M \to \infty$
  - Assuming for simplicity that eigs. of C grow at least linearly: $\mathrm{Tr}[X] = O(\mu)$
  - On this set, strong convexity holds with parameter $1/O(\mu)$
- **Also,** take initial condition $H_0 = C = H - \mu\mathbf{I}$ (non-interacting Hamiltonian)
  - The iteration becomes $H_t = C + \mathrm{diag}^*[\, v_t \,]$, where

$$v_{t+1} = (1 - \alpha_t) v_t + \alpha_t V \hat{\rho}_t$$

    which is the ordinary SCF for the Hartree potential (with stochastic density)
  - Removes the pesky $\|C\|$ term in our error bound
- **The price we pay:** have to worry about $D(X_\star \| X_0)$ which is no longer $\tilde{O}(1)$
  - But can bound $D(X_\star \| X_0) = O(\beta\, c_{\mathrm{h}}\, \mu)$
  - Introduces some temperature-dependence

Complete basis convergence result

Theorem (informal): For time horizon T and step size $\eta_t = O(T^{-1/2})$, with high prob. the Cesàro mean satisfies

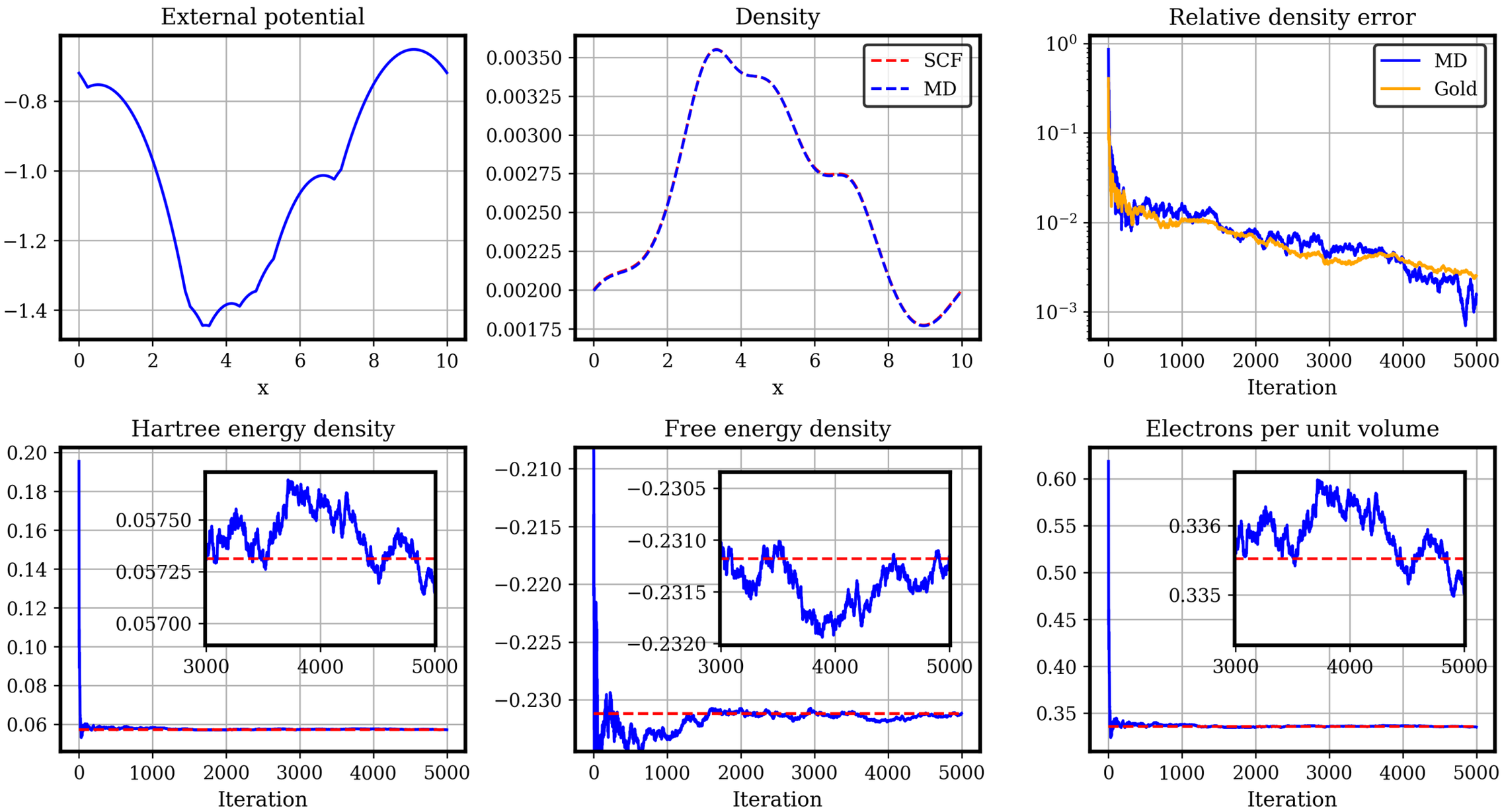
$$F_\beta(\overline{X}_T) \leq F_\beta(X^\star) + \tilde{O}\left(\frac{\mu\, c_{\mathrm{h}}\sqrt{\beta\, c_{\mathrm{h}}}}{\sqrt{T}}\right)$$

Notes

- $\|C\|$ is removed, but β is introduced
- μ is like an effective particle number
- Bound controls "free energy / Hartree energy"

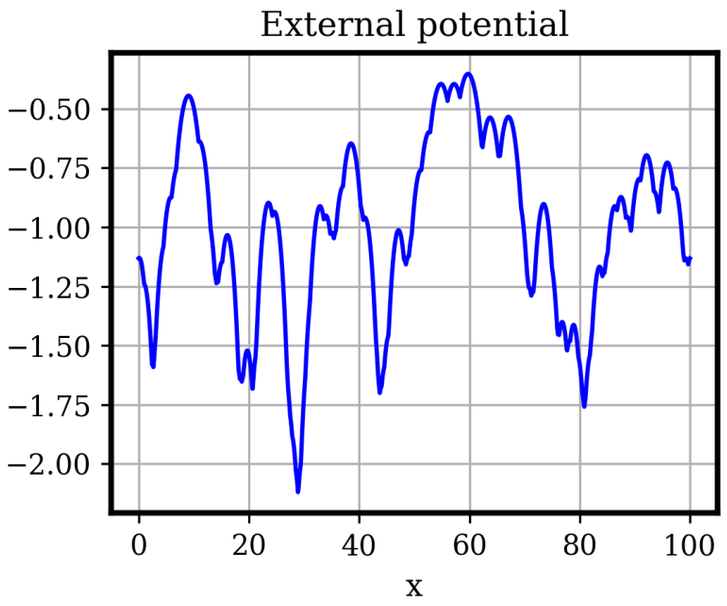
Sample numerical results

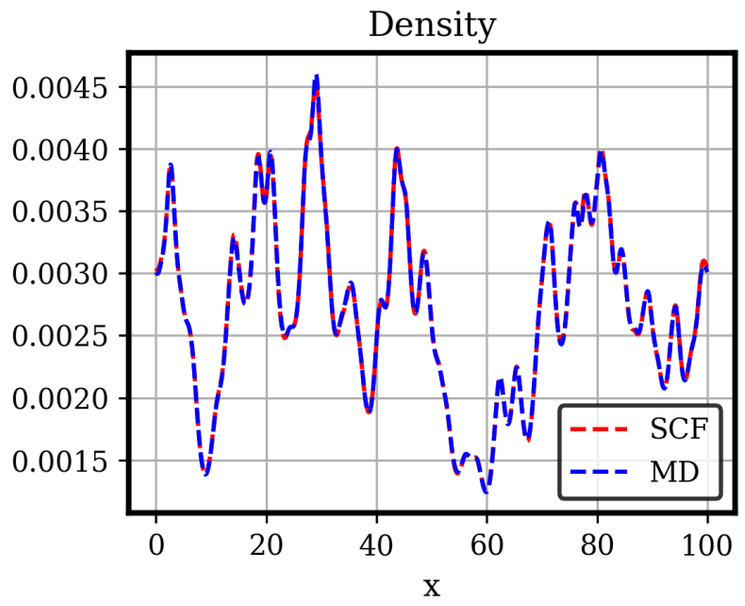
$d = 1,\ M = 1281,$ box length $L = 10$

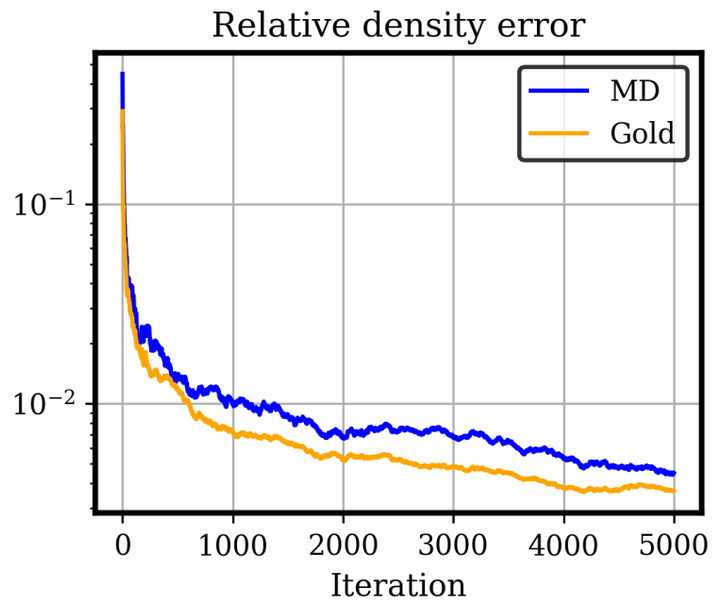


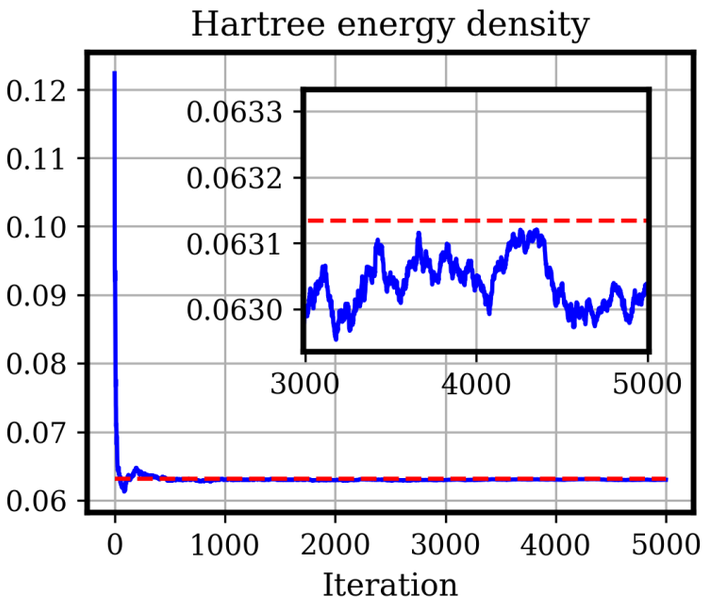
Sample numerical results

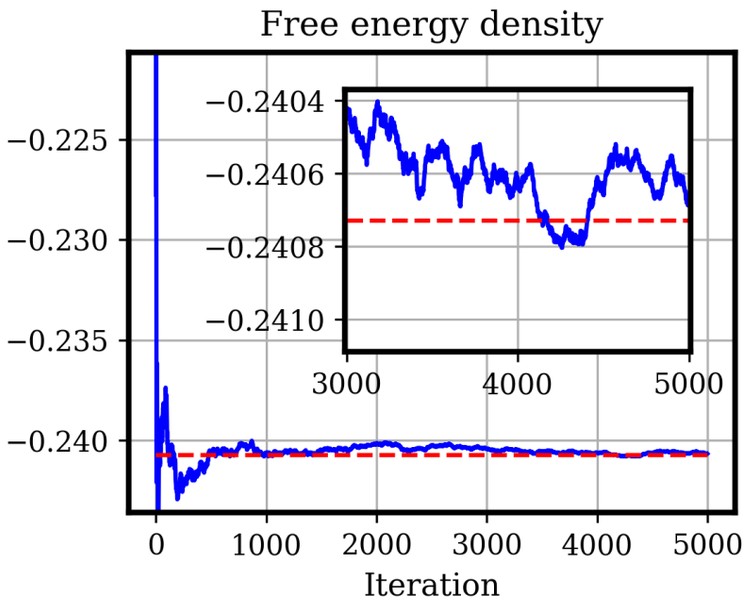
$d = 1,\ M = 12801,$ box length $L = 100$

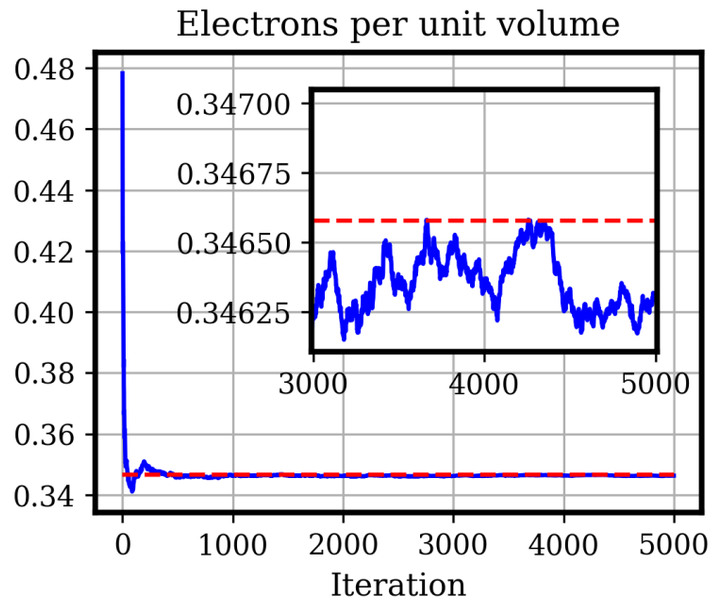






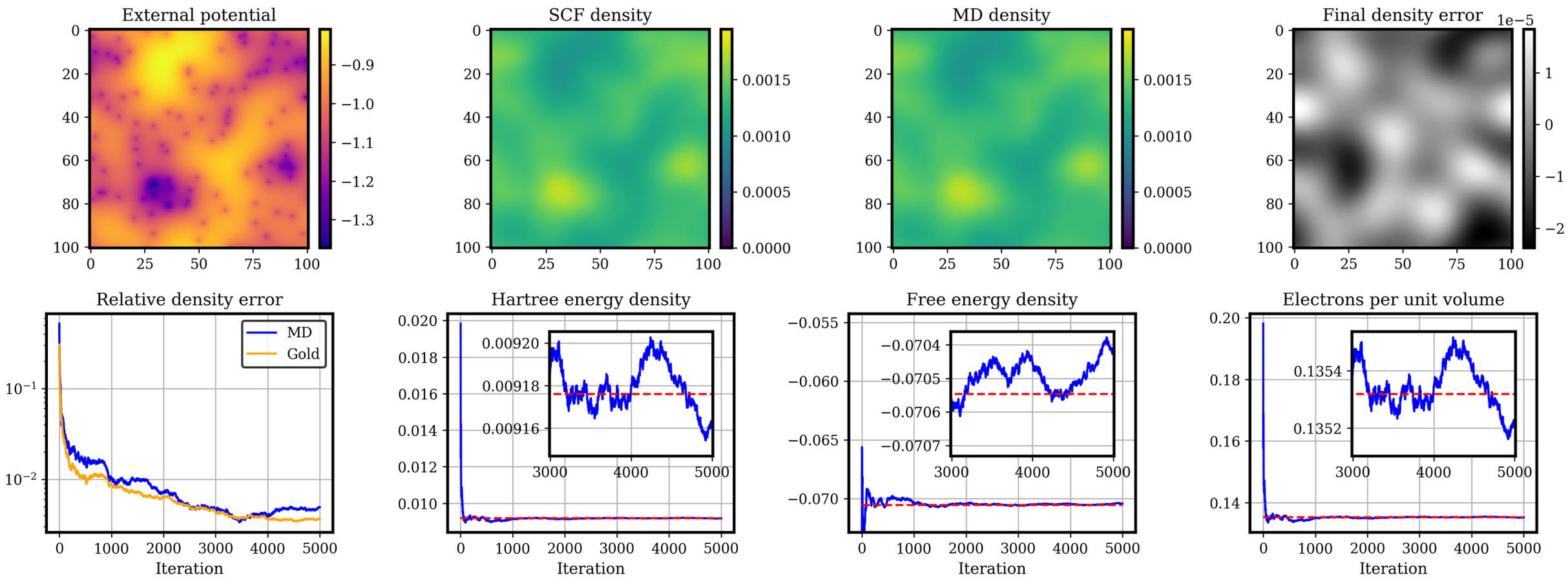






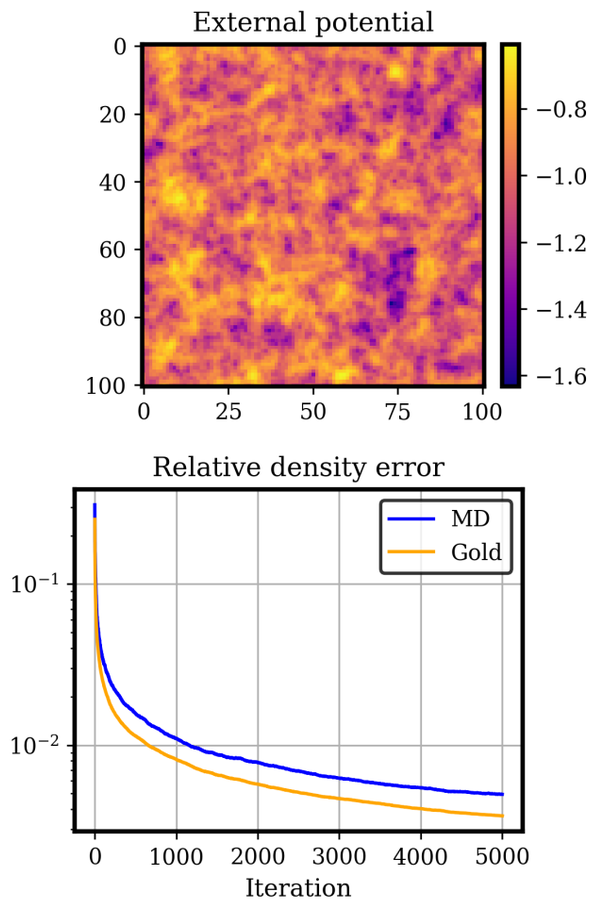
Sample numerical results

$d = 2,\ M = 101^2,$ box length $L = 10$



Sample numerical results

$d = 2,\ M = 101^2$, box length $L = 100$



Cut for time

- **General basis**
  - Get gradient estimator with concentration via THC factorization of 2-electron integrals
- **Chemical potential fitting**
  - Brute force approach: compute dual objective on a chemical potential grid and search
  - Perhaps more can be said

Conclusions

- I'd say we have achieved our goal of "**optimal $\tilde{O}(M)$ scaling**" for the Hartree theory (subject to some quibbles), by integrating several disparate tools
  - Trace estimators / concentration bounds
  - Pole expansion
  - Mirror descent regret bounds
- **Several practical questions**
  - Practical conclusions for more general functionals
  - Rank-structured gradient estimators to reduce variance
- Stay tuned for **preprint**
  - **Yuhang Cai** and Michael Lindsey, "Stochastic Hartree theory in the thermodynamic and complete basis set limits at arbitrary temperature."
- See also my related recent work on **SDP**
  - "Fast randomized entropically regularized semidefinite programming" [arXiv:2303.12133]
  - Analysis coming soon as well (joint with **Yuhang**)
- **P.S. I'm hiring (postdocs / interns)**