Re-anchoring quantum Monte Carlo with tensor-train sketching

Yuehaw Khoo
CAM Initiative and Department of Statistics
University of Chicago
Joint work with Ziang Yu (U.Chicago) and Shiwei Zhang (Flatiron Institute).

06/10/2025, IMSI

Intro to quantum ground state problem I

- For some particles $x_1, \dots x_d \in \Omega$. Want to find its stable configuration in a potential well $U : \Omega^d \to \mathbb{R}$.
- Classical minimization problem:

$$\min_{x \in \Omega^d} U(x), \quad x = [x_1^T \cdots x_d^T]$$

  - Not hard to get some local optimizer. Just gradient descent.

- Equivalent to a measure optimization problem

$$\min_{\mu} \int U(x)\mu(dx), \quad \text{s.t. } \mu \geq 0, \int \mu(dx) = 1$$

  - Minimizer: $\mu^\star = \delta(\cdot - x^\star)$.
  - Size: $|\Omega|^d$.

Intro to quantum ground state problem II

- When temperature $T \neq 0$, need to add some entropy
- Entropic minimization problem:

$$\min_{\mu} \int U(x)\mu(dx) + T \int \ln \mu(x)\mu(dx), \quad \text{s.t. } \mu \geq 0, \ \int \mu(dx) = 1$$

  - Free energy in stat. mech.
  - Temperature gives particles some kinetic energy $\rightarrow$ Entropy.
  - Particles' positions are spread out. $\mu^\star$ not a dirac-delta.

- Another parameterization of the problem

$$\min_{\Psi} \int U(x)|\Psi(x)|^2 dx + T \int |\Psi(x)|^2 \ln |\Psi(x)|^2 dx, \quad \text{s.t. , } \|\Psi\|_{L_2} = 1$$

  - Connection to quantum mechanics (probability is the square of some function). Born rule.
  - Hadamard parameterization in optimization over probability simplex (Li-McKinsey-Yin 21).

Intro to quantum ground state problem III

- Quantum mechanics: A different penalty to "spread out" the particles

$$\min_{\Psi} \int U(x)|\Psi(x)|^2 dx + \hbar \int |\nabla \Psi(x)|^2 dx, \quad \text{s.t. ,} \ \|\Psi\|_{L_2} = 1$$

  - First term: Potential energy
  - Second term: Kinetic energy ($\hbar$: Planck constant)
  - $\Psi : \Omega^d \to \mathbb{C}$: Wavefunction.

- Unlike stat. mech., even when temperature is zero, there is always kinetic energy spreading out particles (uncertainty principle).
- Solve an eigenvalue problem

$$E_0 := \min_{\Psi} \frac{\langle \Psi, H\Psi \rangle}{\langle \Psi, \Psi \rangle}$$

  - H: Hamiltonian. Hermitian. Size $\mathbb{C}^{|\Omega|^d \times |\Omega|^d}$
  - Denote lowest eigenstate as Ψ_0.
  - Curse-of-dim.

Function approximation approaches I:

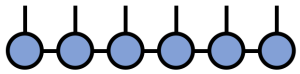
- Represent solution Ψ as Ψ_θ.
- Solve

$$E_0 \approx \min_\theta \frac{\langle \Psi_\theta, H\Psi_\theta \rangle}{\langle \Psi_\theta, \Psi_\theta \rangle}$$

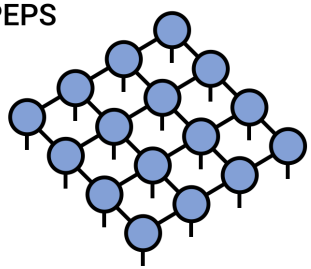
- Accuracy limited by approximation error.
- Methods: Meanfield, Perturbation theory, Exponential ansatz, Tensor-network (White 92), Deep neural-network (Carleo-Troyer 17)
  - Similar to variational inference methods (Blei 16) to solve entropic minimization problem.
- Meanfield has $O(d)$ storage complexity. Others are worse in both storage and computational complexities.
  - Lots of efforts in making them closer to linear scaling

Function approximation approaches II:

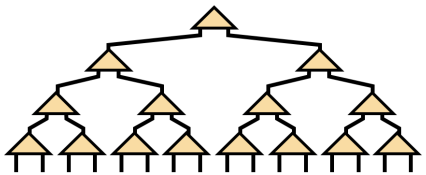
Matrix Product State / Tensor Train



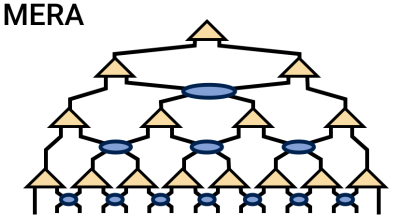
PEPS



Tree Tensor Network / Hierarchical Tucker

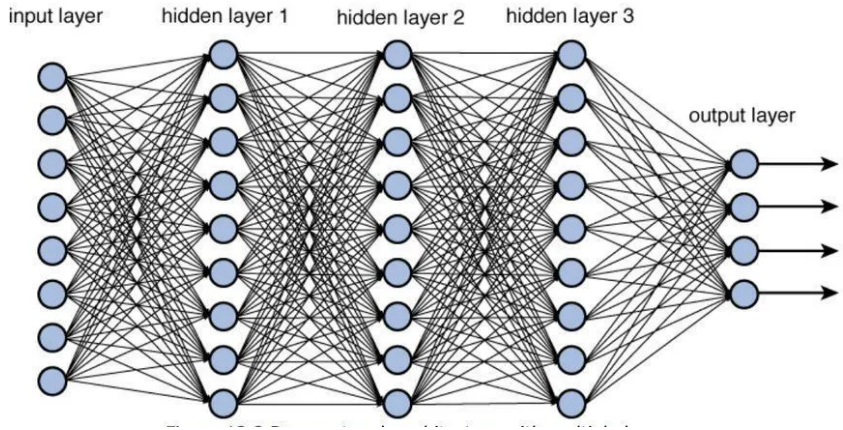


MERA



https://tensornetwork.org/

R. Parmar – Towards Data Science



Monte-Carlo approaches I:

- Ground state Ψ_0 satisfies $H\Psi_0 = E_0\Psi_0$.
  - Energy is calculated with some "trial wavefuction" Ψ_{tr}:

$$E_0 = \frac{\langle \Psi_{\text{tr}}, H\Psi_0 \rangle}{\langle \Psi_{\text{tr}}, \Psi_0 \rangle}$$

- Approximate $\Psi_0 \approx \sum_{k=1}^{N} \Phi_k$ (empirical distribution of a wavefunction). Each Φ_k very simple. Storage $O(Nd)$.

- Energy is calculated as

$$E_0 \approx \frac{\langle \Psi_{\text{tr}}, H\sum_{k=1}^{N} \Phi_k \rangle}{\langle \Psi_{\text{tr}}, \sum_{k=1}^{N} \Phi_k \rangle}$$

  - Large variance!

Monte-Carlo approaches II:

- Variance can blow-up without sophisticated importance sampling.
- Denominator can go to zero.

Questions:

- Function approximation approaches:
  - Advantages: Deterministic. No/low variance.
  - Disadvantages: Approximation error. High computational cost.
- Monte-Carlo approaches:
  - Advantages: Cheap.
  - Disadvantages: High variance.
- Can we have the best of both worlds?

Intro to auxilliary field quantum Monte-Carlo I:

- Imaginary time evolution to get ground state:

$$\Psi_0 = \lim_{\tau \to \infty} \exp(-\tau H)\Phi^{(0)}$$

- Power method compute top eigenvector of $\exp(-\Delta\tau H)$:

$$\exp(-\tau H)\Phi^{(0)} = \exp(-\Delta\tau H)\cdots\exp(-\Delta\tau H)\Phi^{(0)}$$

- Suppose there is a decomposition of propagator (mixture model for operator):

$$\exp(-\Delta\tau H) = \mathbb{E}_{y\sim P} B(y),\ B(y) \in \mathbb{C}^{|\Omega|^d \times |\Omega|^d}.$$

  - $y \in \mathbb{R}^d$, P is a distribution on $\mathbb{R}^d$.

- Approximate $\exp(-\Delta\tau H)$ by sampling operators $B(y)$:

$$\Phi^{(n)} = B(y^{(n)})\cdots B(y^{(1)})\Phi^{(0)}, \quad y^{(1)}, \cdots, y^{(n)} \overset{\text{i.i.d.}}{\sim} P$$

- Then $\Psi_0 = \mathbb{E}\Phi^{(n)}$

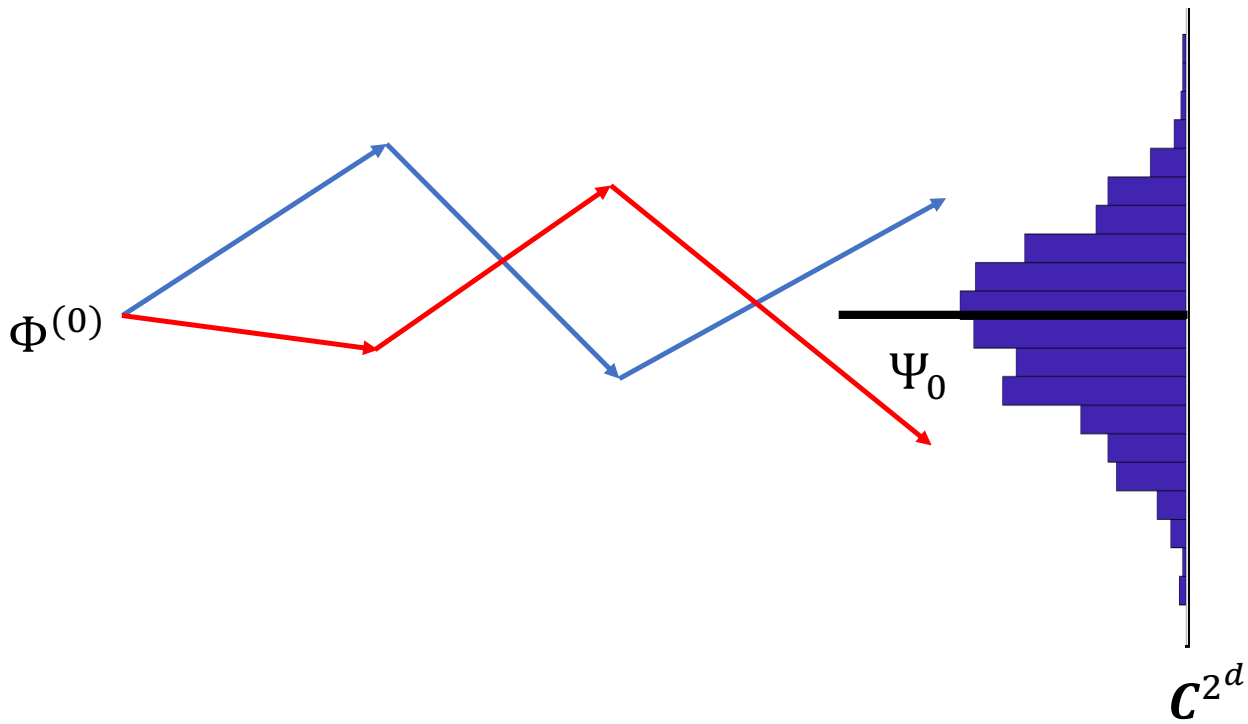
Intro to auxilliary field quantum Monte-Carlo II:

- Three necessary ingredients:
  - Walker $\Phi^{(n)}$ is a tensor product functions (e.g. $g_1 \otimes \cdots \otimes g_d$)
  - $B(y)$ is a tensor product operators (e.g. $O_1 \otimes \cdots \otimes O_d$)
  - $B(y)\Phi^{(n)} = O_1 g_1 \otimes \cdots \otimes O_d g_d$ can be done fast ($O(d)$ complexity) and stays separable.

- An analogy of classical Ising model for d spins $s \in \{\pm 1\}^d$
  - Let $\mu(s) \propto \exp\left(-\frac{1}{2}\sum_{i,j=1}^{d} J_{ij} s_i s_j\right)$.
  - If J positive semidefinite, Hubbard-Stratonovich transformation.
  - $\mu(s) \propto \int_{\mathbb{R}^d} P(y) \exp(i < y, s >) dy$, $P(y)$ is a Gaussian (Fourier transform of Gaussian is Gaussian).
  - Then $\mu(s) \approx \exp(i < y, s >)$ where $y \sim P$.
    $\exp(i < y, s >) = \exp(i y_1 s_1) \cdots \exp(i y_d s_d)$ is separable function.

Quantum Monte-Carlo II:

- **Pros**: Each mat-vec in $B(y^{(n)}) \cdots B(y^{(1)})\Phi^{(0)}$ has $O(d)$ complexity.
- **Cons**: Brownian motion in $\mathbb{C}^{2^d}$. Chance of finding Ψ_0 is $O\left(\frac{1}{2^d}\right)$. A very rare event.
  - $\langle \Psi_0, B(y^{(n)}) \cdots B(y^{(1)})\Phi^{(0)} \rangle \approx 0$



Importance sampling I

- Importance sampling: Guide walkers to land on trial wavefunction $\Psi_{\text{tr}} \approx \Psi_0$.

(phys.org)

Importance sampling II

- Change sampling from P to $P^{(n)}$

$$\begin{aligned}\mathbb{E}_{y\sim P} B(y)\Phi^{(n)} &= \mathbb{E}_{y\sim P^{(n)}} \frac{P(y)B(y)}{P^{(n)}(y)}\Phi^{(n)} \\ &= \mathbb{E}_{y\sim P^{(n)}} \tilde{B}^{(n)}(y)\Phi^{(n)}\end{aligned}$$

  - $P^{(n)}(y) = \frac{\langle \Psi_{\mathrm{tr}}, B(y)\Phi^{(n)}\rangle P(y)}{\mathcal{N}^{(n)}}$. $\mathcal{N}^{(n)}$: Normalizing constant.

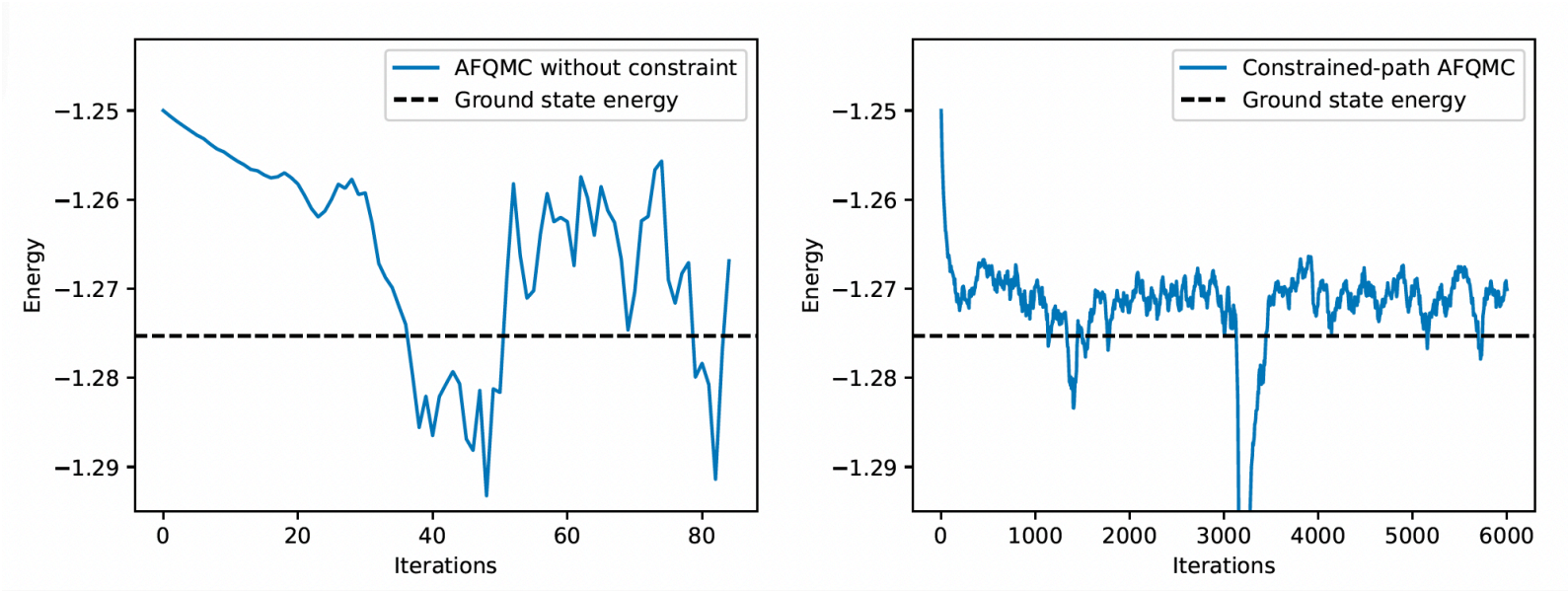
- $P^{(n)}(y)$ can be negative. If so, do **constrained path approximation** (Zhang-Carlson-Gubernatis 97):

$$P^{(n)}(y) = \frac{\max\{\langle \Psi_{\mathrm{tr}}, B(y)\Phi^{(n)}\rangle, 0\}P(y)}{\mathcal{N}^{(n)}}.$$

  - Space of opposite sign cannot be visited. Introducing some bias.

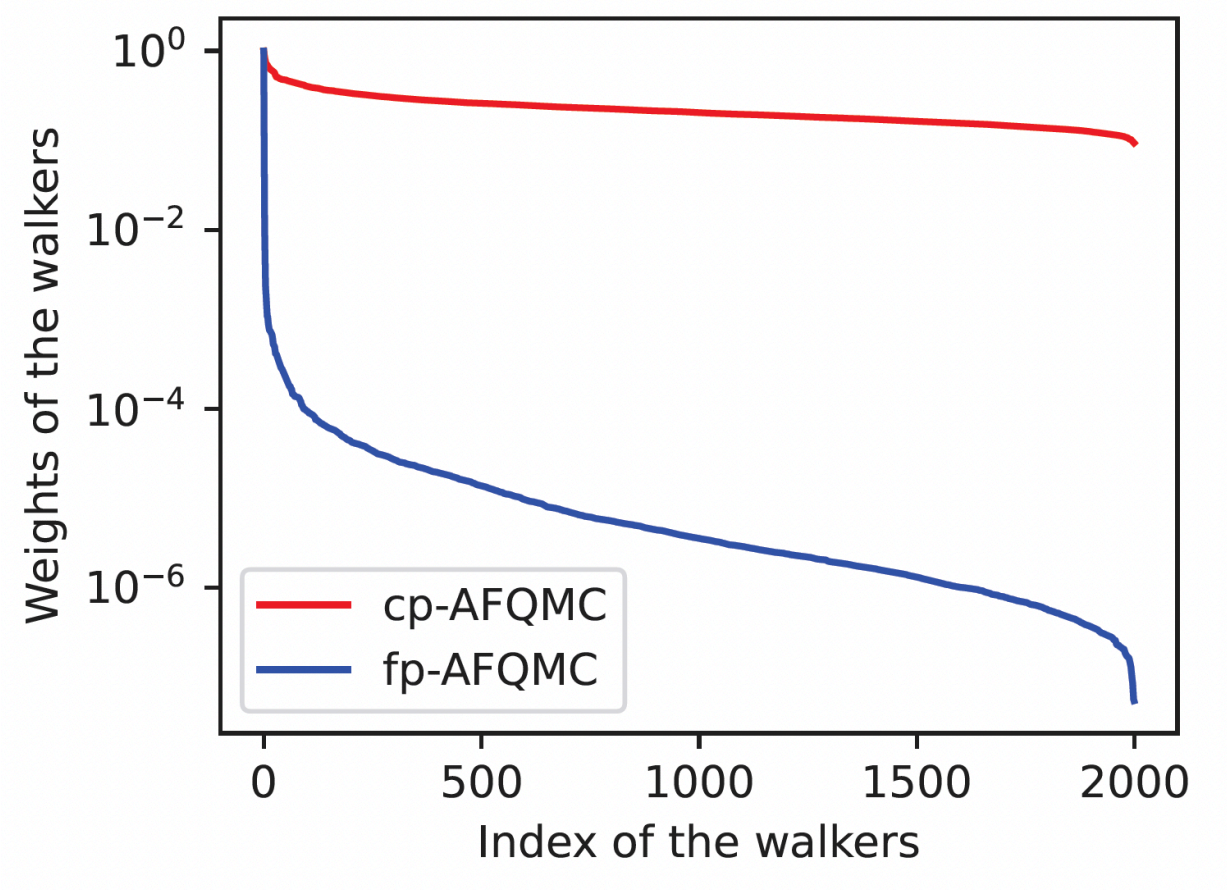
Importance sampling III

- Transverse field Ising model. 16 spins.



Importance sampling IV

- Overlap $\langle \Psi_{\text{tr}}, \Phi^{(n)} \rangle$



Importance sampling V

- Bias depends on the choice of Ψ_{tr}. Usually closer to ground state Ψ_0 the better.
- An overlooked fact: Variance also depends on how close Ψ_{tr} to ground state Ψ_0.
- A good trial wavefunction Ψ_{tr} is crucial!

Theoretical results:

- Let Ψ_0 be the ground state of H with eigenvalues $E_0 < E_1 \leq E_2 \cdots$.
- Standard results of one step of power method:

$$\left| \frac{\langle \Psi_{\text{tr}}, H \exp(-\Delta\tau H)\Phi^{(n)} \rangle}{\langle \Psi_{\text{tr}}, \exp(-\Delta\tau H)\Phi^{(n)} \rangle} - E_0 \right| \lesssim e^{-\Delta\tau(E_1 - E_0)} \left| \frac{\langle \Psi_{\text{tr}}, H\Phi^{(n)} \rangle}{\langle \Psi_{\text{tr}}, \Phi^{(n)} \rangle} - E_0 \right|$$

- Now replace $\exp(-\Delta\tau H)$ with a sample $\widetilde{B}^{(n)}(y)$.
- Theorem (Yu-Zhang-K. 2024): Assuming there is no bias in the random walk. For $\tan(\angle(\Psi_0, \Psi_{\text{tr}}))$ small enough:

$$\left| \frac{\langle \Psi_{\text{tr}}, H\widetilde{B}^{(n)}(y)\Phi^{(n)} \rangle}{\langle \Psi_{\text{tr}}, \widetilde{B}^{(n)}(y)\Phi^{(n)} \rangle} - E_0 \right| \lesssim e^{-\Delta\tau(E_1 - E_0)} \left| \frac{\langle \Psi_{\text{tr}}, H\Phi^{(n)} \rangle}{\langle \Psi_{\text{tr}}, \Phi^{(n)} \rangle} - E_0 \right|$$

$$+ \underbrace{\|H - E_0 I\|_2 \tan(\angle(\Psi_0, \Psi_{\text{tr}}))}_{\text{magnitude}} \underbrace{\frac{\|\widetilde{B}^{(n)}(y) - e^{-\Delta\tau H}\|_2}{\|e^{-\Delta\tau E_0}\|_2} \tan(\angle(\Psi_0, \Phi^{(n)}))}_{\text{random fluctuation}}$$

- Better Ψ_{tr} gives smaller variance.

Intuition:

- Due to importance sampling, overlap variance

$$\mathrm{Var}_{y\sim P^{(n)}}(\langle \Psi_{\mathrm{tr}}, \tilde{B}^{(n)}(y)\Phi^{(n)}\rangle) = 0.$$

- Therefore, when no bias, get faithful overlap:

$$\langle \Psi_{\mathrm{tr}}, \tilde{B}^{(n)}(y)\Phi^{(n)}\rangle = \langle \Psi_{\mathrm{tr}}, \exp(-\Delta\tau H)\Phi^{(n)}\rangle$$

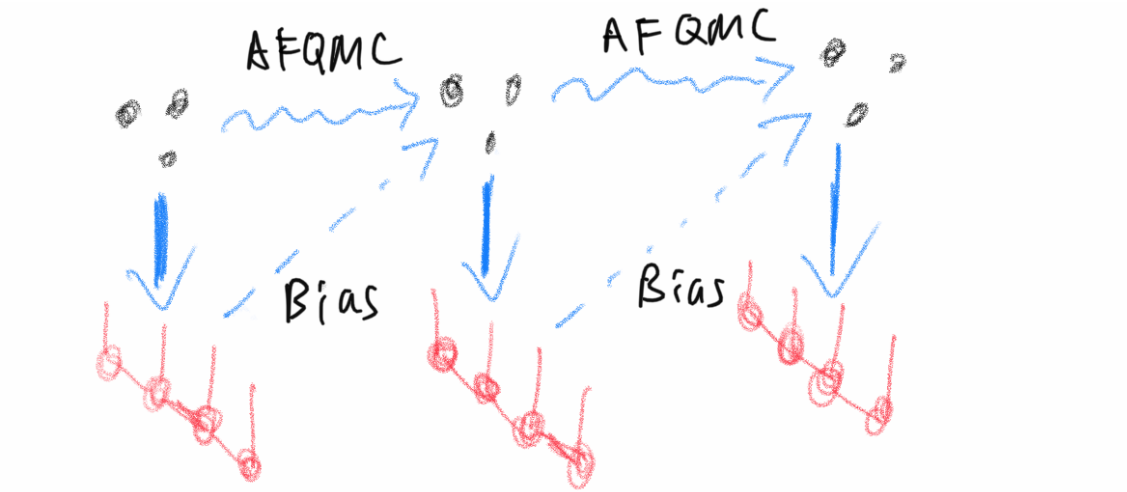
- If $\Psi_{\mathrm{tr}} = \Psi_0$:

$$\begin{aligned}\langle \Psi_{\mathrm{tr}}, H\tilde{B}^{(n)}(y)\Phi^{(n)}\rangle &= \langle H\Psi_{\mathrm{tr}}, \tilde{B}^{(n)}(y)\Phi^{(n)}\rangle \\ &= E_0\langle \Psi_{\mathrm{tr}}, \tilde{B}^{(n)}(y)\Phi^{(n)}\rangle = E_0\langle \Psi_{\mathrm{tr}}, \exp(-\Delta\tau H)\Phi^{(n)}\rangle\end{aligned}$$

- Zero variance: $\frac{\langle \Psi_{\mathrm{tr}}, H\tilde{B}^{(n)}(y)\Phi^{(n)}\rangle}{\langle \Psi_{\mathrm{tr}}, \tilde{B}^{(n)}(y)\Phi^{(n)}\rangle} = E_0$

Re-anchoring quantum Monte-Carlo with tensor-train sketching[1]:

- From current walkers $\{\Phi_k^{(n)}\}_{k=1}^N$, estimate a new Ψ_{tr} (in terms of tensor-train).
- Use new Ψ_{tr} to importance sample next episode of random walks. Hope improved Ψ_{tr} gives smaller energy variance and bias.
- Iterate back and forth.

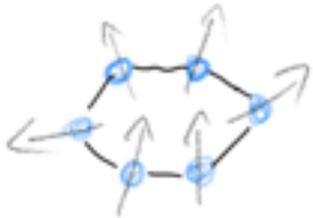


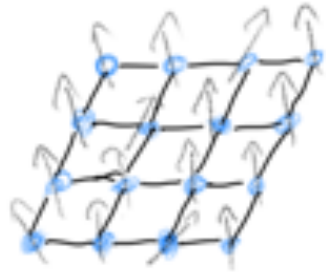
.

[1] Yu-Zhang-K. arxiv 2411.07194

Results I:

- 1D and 2D transverse-field Ising model.

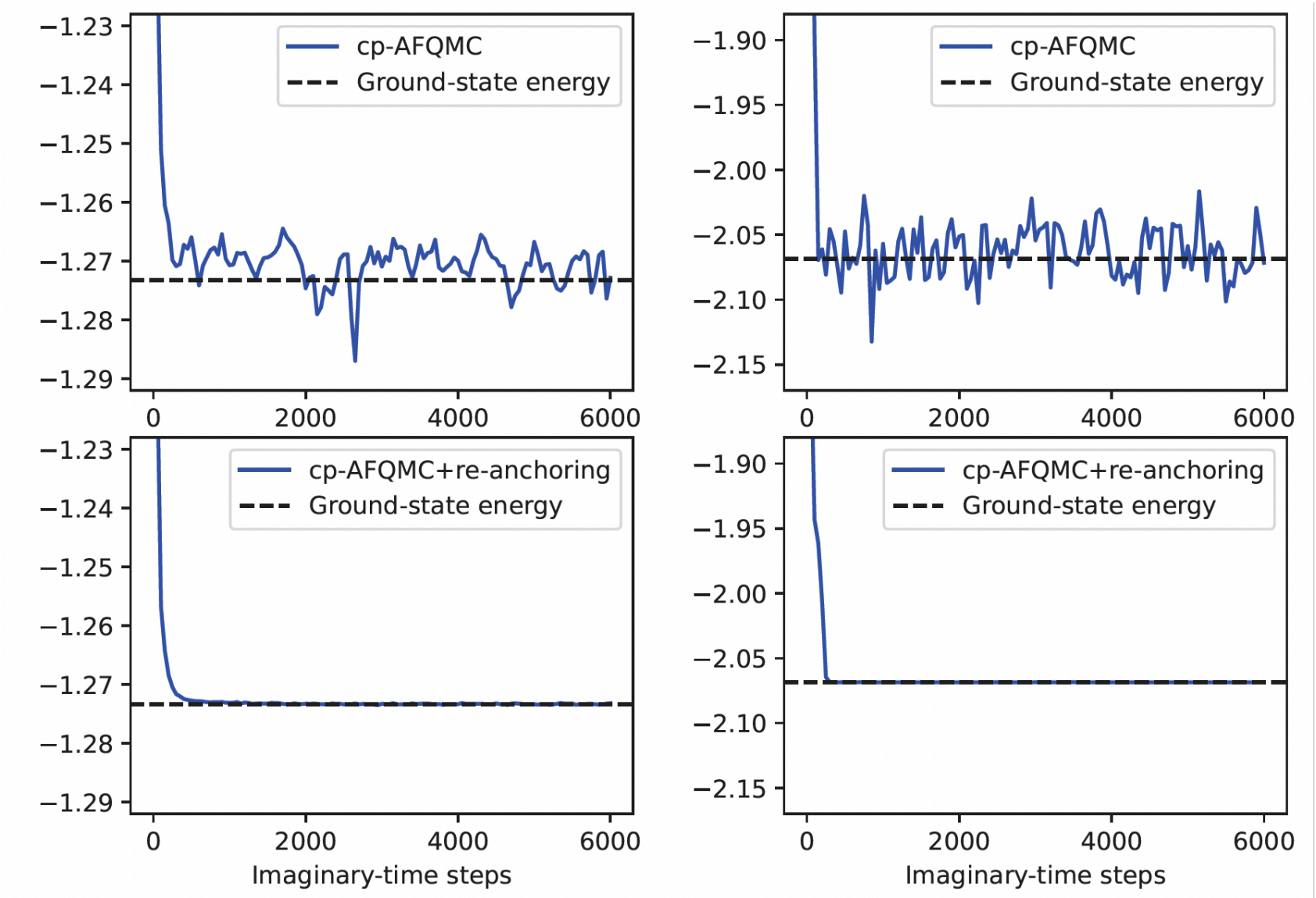




- Walkers number $4000, 8000$, 12000, 12000. Relative energy error:

	cp-AFQMC	cp-AFQMC with re-anchoring
32 spins 1D	$(+1.35 \pm 0.31) \times 10^{-3}$	$(+0.44 \pm 2.43) \times 10^{-5}$
64 spins 1D	$(+1.88 \pm 0.37) \times 10^{-3}$	$(+0.77 \pm 1.07) \times 10^{-5}$
96 spins 1D	$(+2.49 \pm 0.33) \times 10^{-3}$	$(-4.95 \pm 8.94) \times 10^{-6}$
4×16 spins	$(+4.18 \pm 1.42) \times 10^{-3}$	$(-0.66 \pm 1.96) \times 10^{-6}$

Results II:



96 1D spins

4x16 2D spins

Results III: Advantages over function approximation methods

- Energy matches direct minimization over rank-800 tensor train.
- We use only rank-4 tensor train as guide!

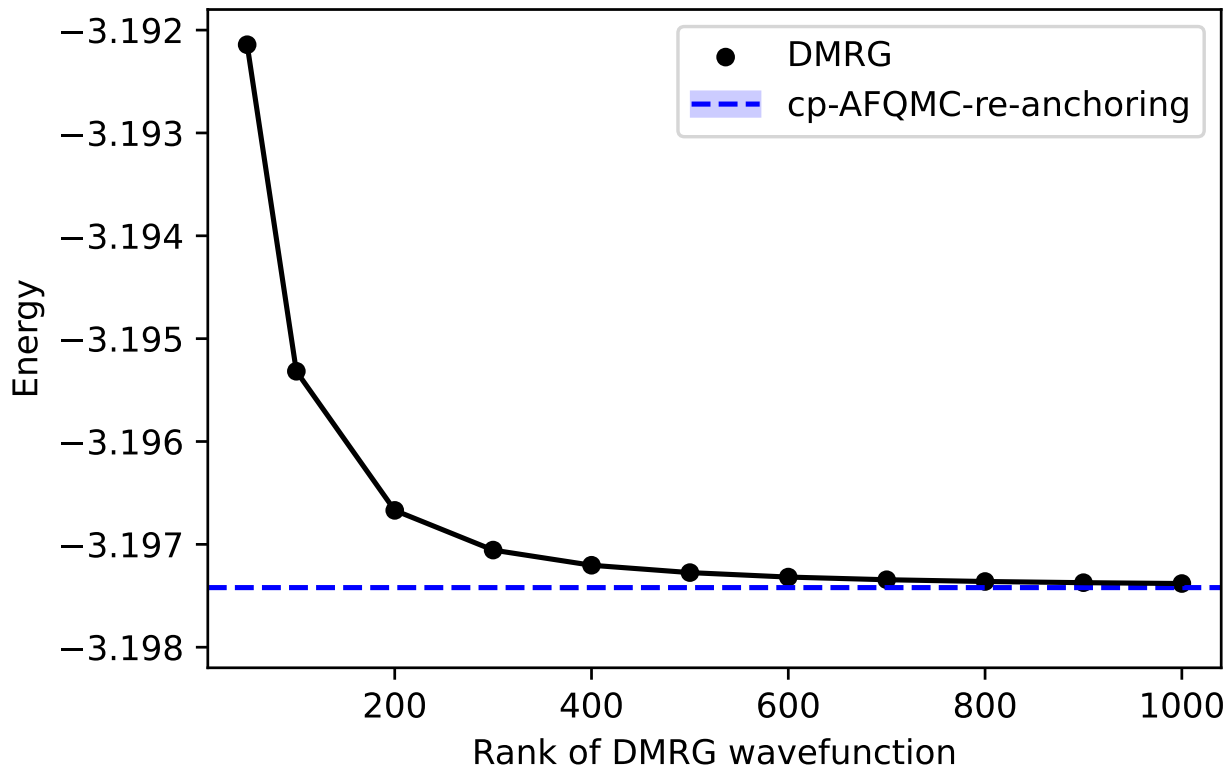


Figure: 8×8 model with magnetic field $g = 3.0$.

Computational cost:

- Applying $\tilde{B}^{(n)}(y)\Phi_k^{(n)}$ is $O(d)$ (tensor product structure). With N walkers, $O(Nd)$.

- Estimating a tensor-train Ψ_{tr} from walkers $\{\Phi_k^{(n)}\}_{k=1}^N$ is $O(Nd)$.
  - Use tensor-train (TT) sketching[1].
  - Almost no extra cost on top of Monte Carlo!

[1] Hur-Hoskins-Lindsey-Stoudenmire-K. ACHA 2023

Tensor-train sketching I

- Given $\Phi_k^{(n)} \in \mathbb{C}^{m^d}$. A tensor $\underbrace{m \times \cdots \times m}_{d \text{ times}}$

$$\hat{\Phi} := \sum_{k=1}^{N} \Phi_k^{(n)} \overset{N\to\infty}{\longrightarrow} \Phi^*$$

- $\hat{\Phi}$: empirical wavefunction. $\Phi^\star$: Ground truth wavefunction.
- Want to estimate $\Phi^\star$ from $\hat{\Phi}$.
- Crucial assumption: $\Phi^\star$ is low-rank:
  - Think of $\Phi^\star := \Phi^\star(x_1, \ldots x_d)$ (function of d-variables).
  - The matrix $\Phi^\star(x_{1:k}; x_{k+1:d}) \in \mathbb{C}^{m^k \times m^{k-d}}$ with row/col indexed by $x_{1:k}/x_{k+1:d}$ is rank-r
  - Low correlation/entanglement between $x_{1:k}$ and $x_{k+1:d}$.

Tensor-train sketching II

- Equivalent to $\Phi^\star$ being a tensor train (TT):

$$\Phi^\star(x_1, \ldots, x_d) = G_1^\star(x_1, :) G_2^*(:, x_2, :) \cdots G_{d-1}^\star(:, x_{d-1}, :) G_d^\star(:, x_d)$$

  - $G_1^\star \in \mathbb{C}^{m \times r}, G_i^\star \in \mathbb{C}^{r \times m \times r}, G_d^\star \in \mathbb{C}^{r \times m}$.
  - $r = 1$: separable state.

- Generalize low-rank matrices.

- Want to determine d cores $G_i^\star$'s in $O(d)$ time.

Matrix example: $d = 2$ case

- If $\Phi^\star \in \mathbb{C}^{m \times m}$ rank-r.
- Randomized linear algebra: $\text{Range}(\Phi^\star) = \text{Range}(\Phi^\star T)$ for some chosen sketch matrix T of size $m \times r$.
- Low-rank decomposition of $\Phi^\star = G_1^\star G_2^\star$ by *sketching* ($G_1^\star$ $m \times r, G_2^\star$ $r \times m$):

$$\begin{aligned} G_1^\star &= \Phi^\star T & G_1^\star &= \Phi^\star T \\ G_1^\star G_2^\star &= \Phi^\star & \implies (\Phi^\star T) G_2^\star &= \Phi^\star \end{aligned}$$

- Over-determined. Use a second sketch S $m \times r$:

$$\begin{aligned} G_1^\star &= \Phi^\star T && \textbf{(Range finding)} \\ (S^* \Phi^\star T) G_2^* &= S^* \Phi^\star && \textbf{(Interpolate)} \end{aligned}$$

- Finally with empirical distribution $\hat{\Phi}$:

$$\begin{aligned} G_1 &= \hat{\Phi} T && \textbf{(Range finding)} \\ (S^* \hat{\Phi} T) G_2 &= S^* \hat{\Phi} && \textbf{(Interpolate)} \end{aligned}$$

- Just solving two linear system. $O(r^3 + mr^2)$. Not the dominating cost.

Matrix example: Computational cost

- Forming equation is expensive.
- Naively:
  - Form $\hat{\Phi} = \sum_{k=1}^{N} \Phi_k^{(n)}$. Cost $O(Nm^2)$.
  - Sketch $\hat{\Phi}T$. Cost $O(m^2 r)$.
- However $\hat{\Phi} = \sum_{k=1}^{N} a_k b_k^*$ (each sample is a rank-1 outer product).
- $(a_k b_k^*)T$ is $O(mr)$
- Total cost $O(Nmr)$.

High-d case

- Parallelly solve a system of equations

$$\begin{aligned} G_1(x_1,:) &= \hat{B}_1(x_1,:) \\ \hat{A}_1 G_2(:,x_2,:) &= \hat{B}_2(:,x_2,:) \\ &\vdots \\ \hat{A}_{d-1} G_d(:,x_d) &= \hat{B}_d(:,x_d) \end{aligned}$$

  - $\hat{B}_k$ $r \times m \times r$. $\hat{A}_{k-1}$ $r \times r$.

- $\hat{A}_{k-1}, \hat{B}_k$'s formed with $O(dNmr)$ complexity.

Computational scaling:

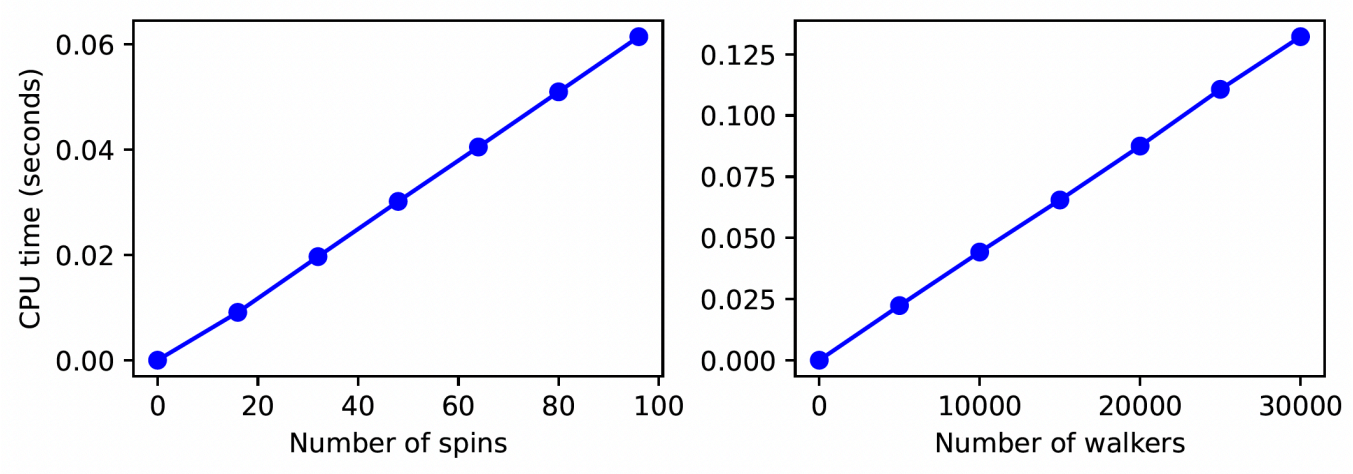


Figure: Left: different number of spins with 2000 walkers; right: different number of walkers for 16 spins.

Comparisons with "Ground truth" wavefunction:

- Proposed method provides more that an energy estimate.
- Fitted trial wavefunction $\Psi_{\text{tr}} \approx \Psi_0$
- Overlap $|\langle \Psi_{\text{tr}}, \Psi_0 \rangle|$:

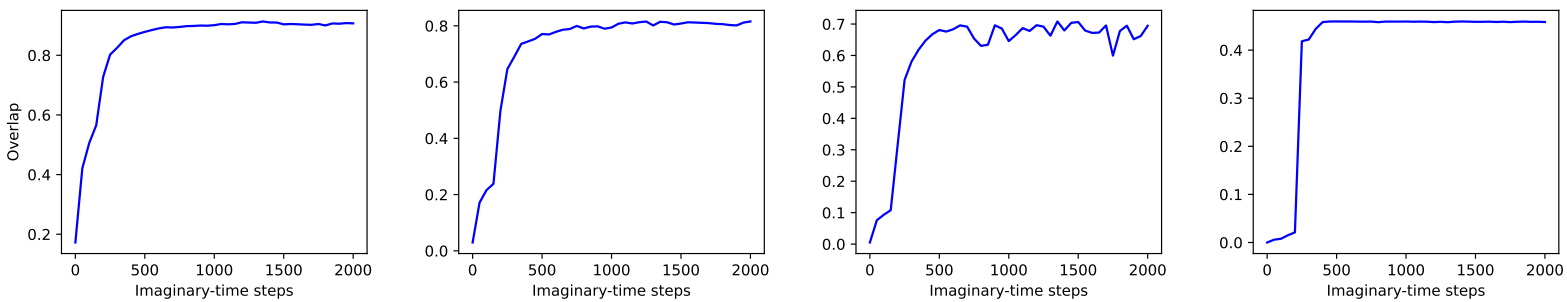
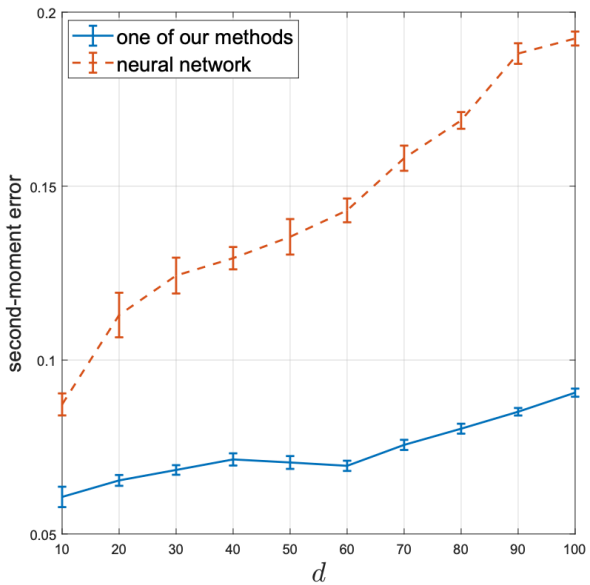


Figure: The overlap between TT trial wavefunction and the ground-state wavefunction. 32 spins in 1D; 64 spins in 1D; 96 spins in 1D; 4×16 spins.

- "Ground truth" from a density matrix renormalization group calculation. (Could be wrong!)

Guarantees:

- Can say things rigorously in density estimation setting
- Hur-Hoskins-Lindsey-Stoudenmire-K. 23: Guarantees on learning Markovian distribution.
- Peng-Yang-K.-Wang 24: Tensor density estimator by Convolution-Deconvolution. Unified framework for more general densities.



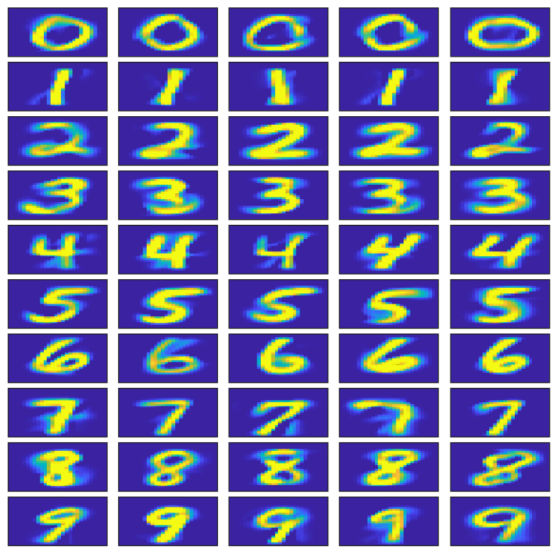


Figure: (Left) Comparison with neural-network for some ground truth density. (Right) Generated MNIST data

Conclusions:

- Combine quantum Monte-Carlo and wavefunction methods.
- Best of both worlds?
- Future work: electronic problems.
- Acknowledgements: Supported by DMS-2111563, DMS-2339439, DE-SC0022232, Sloan foundation.
- Thank you!