Randomized methods for quantum many-body problems: a mathematical primer

Robert J. Webber[1]

[1]Computing and Mathematical Sciences, California Institute of Technology

Institute for Pure & Applied Mathematics
March 9, 2022

Plan for talk

1. Introduction
2. Examples
   - The quantum spin problem
   - The electronic structure problem
3. FCIQMC
   - Deterministic power method
   - FCIQMC
   - FCIFRI
4. Variational Monte Carlo
   - Gradients and Hessians
   - Optimization methods
   - Deterministic analysis
   - Stochastic analysis

The difficulty of calculating the ground-state energy

Recent developments

- Ground-state energies are needed for chemistry.

Our motivation

The difficulty of calculating the ground-state energy

Recent developments

- Ground-state energies are needed for chemistry.
- The difficulty of finding the ground-state energy grows exponentially with number of electrons/spins.

Our motivation

The difficulty of calculating the ground-state energy

Recent developments

- Ground-state energies are needed for chemistry.
- The difficulty of finding the ground-state energy grows exponentially with number of electrons/spins.
- Quantum Monte Carlo (QMC) comes to the rescue, providing energies for
  1. Benzene (C_6H_6) with 42 interacting electrons
  2. Lattices with 100 spins, hence 2^{100} spin configurations

Our motivation

The difficulty of calculating the ground-state energy

Recent developments

- Ground-state energies are needed for chemistry.
- The difficulty of finding the ground-state energy grows exponentially with number of electrons/spins.
- Quantum Monte Carlo (QMC) comes to the rescue, providing energies for
  1. Benzene (C_6H_6) with 42 interacting electrons
  2. Lattices with 100 spins, hence 2^{100} spin configurations

Our motivation

- QMC can be extended even further, but first the sampling and optimization strategies it employs need to be improved.

The difficulty of calculating the ground-state energy

Recent developments

- Ground-state energies are needed for chemistry.
- The difficulty of finding the ground-state energy grows exponentially with number of electrons/spins.
- Quantum Monte Carlo (QMC) comes to the rescue, providing energies for
  1. Benzene (C_6H_6) with 42 interacting electrons
  2. Lattices with 100 spins, hence 2^{100} spin configurations

Our motivation

- QMC can be extended even further, but first the sampling and optimization strategies it employs need to be improved.
  1. Long convergence times.
  2. Loss of stability.
  3. Convergence to unreasonable solutions.

The difficulty of calculating the ground-state energy

Recent developments

- Ground-state energies are needed for chemistry.
- The difficulty of finding the ground-state energy grows exponentially with number of electrons/spins.
- Quantum Monte Carlo (QMC) comes to the rescue, providing energies for
  1. Benzene (C_6H_6) with 42 interacting electrons
  2. Lattices with 100 spins, hence 2^{100} spin configurations

Our motivation

- QMC can be extended even further, but first the sampling and optimization strategies it employs need to be improved.
  1. Long convergence times.
  2. Loss of stability.
  3. Convergence to unreasonable solutions.
- We need to understand QMC mathematically in order to improve it.

Extreme eigenvalue problem

In QMC applications, we need to solve discrete eigenvalue problems involving matrices with dimensions up to $10^{108} \times 10^{108}$.

Extreme eigenvalue problem

In QMC applications, we need to solve discrete eigenvalue problems involving matrices with dimensions up to $10^{108} \times 10^{108}$.

Even storing a single vector of this size would be unmanageable.

Extreme eigenvalue problem

In QMC applications, we need to solve discrete eigenvalue problems involving matrices with dimensions up to $10^{108} \times 10^{108}$.

Even storing a single vector of this size would be unmanageable.

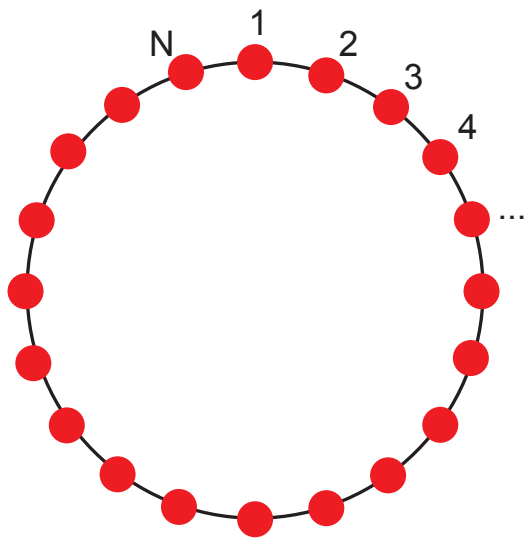
Question: How do such large matrices arise?

The quantum spin problem

Consider the *transverse-field Ising model* for spin-1/2 particles on a periodic 1-D lattice, specified by

$$\mathcal{H} = -\sum_{i\sim j} \sigma_i^z \sigma_j^z - h \sum_i \sigma_i^x$$

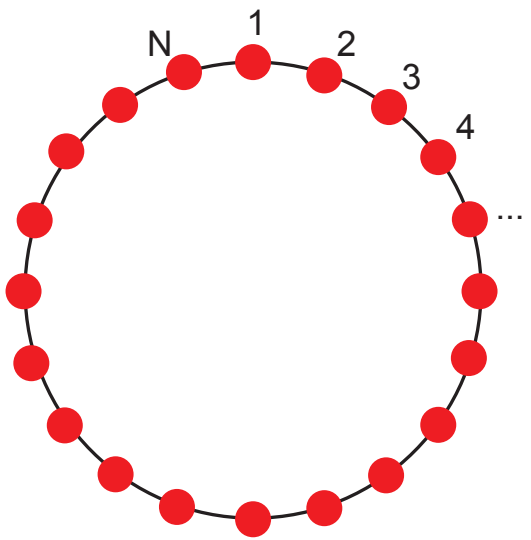
- σ_i^x and σ_i^z are Pauli operators for the i-th spin
- $i \sim j$ signifies that i and j are neighboring spins
- h is a real-valued parameter



The quantum spin problem

Consider the *transverse-field Ising model* for spin-1/2 particles on a periodic 1-D lattice, specified by

$$\mathcal{H} = -\sum_{i\sim j} \sigma_i^z \sigma_j^z - h \sum_i \sigma_i^x$$



- σ_i^x and σ_i^z are Pauli operators for the i-th spin
- $i \sim j$ signifies that i and j are neighboring spins
- h is a real-valued parameter

Question: How can we represent $\mathcal{H}$ as a matrix?

The quantum spin problem

Question: How can we represent $\mathcal{H}$ as a matrix?

The quantum spin problem

Question: How can we represent $\mathcal{H}$ as a matrix?

Consider a wavefunction

$$|\psi\rangle = |\psi_1\rangle \otimes |\psi_2\rangle \otimes \cdots \otimes |\psi_N\rangle,$$

which is a tensor product of $|+\rangle$ and $|-\rangle$ states on the individual spins.

The quantum spin problem

Question: How can we represent $\mathcal{H}$ as a matrix?

Consider a wavefunction

$$|\psi\rangle = |\psi_1\rangle \otimes |\psi_2\rangle \otimes \cdots \otimes |\psi_N\rangle,$$

which is a tensor product of $|+\rangle$ and $|-\rangle$ states on the individual spins.

Example 1.

$$|\psi\rangle = |+-+-\ldots+-\rangle$$

Example 2.

$$|\psi\rangle = |++--\ldots--\rangle$$

The quantum spin problem

Question: How can we represent $\mathcal{H}$ as a matrix?

Consider a wavefunction

$$|\psi\rangle = |\psi_1\rangle \otimes |\psi_2\rangle \otimes \cdots \otimes |\psi_N\rangle,$$

which is a tensor product of $|+\rangle$ and $|-\rangle$ states on the individual spins.

Example 1.

$$|\psi\rangle = |+-+-\ldots+-\rangle$$

Example 2.

$$|\psi\rangle = |++--\ldots--\rangle$$

$\implies$ The tensor product wavefunctions form a (complete) orthonormal basis of size 2^N.

The quantum spin problem

Question: How can we represent $\mathcal{H}$ as a matrix?

Introduce the raising and lowering operators

- σ_i^+, which raises $|\psi_i\rangle$ from $|-\rangle$ to $|+\rangle$,
- σ_i^-, which lowers $|\psi_i\rangle$ from $|+\rangle$ to $|-\rangle$,

and insist that

$$\sigma_i^+\sigma_i^+ = \sigma_i^-\sigma_i^- = 0.$$

The quantum spin problem

Question: How can we represent $\mathcal{H}$ as a matrix?

Introduce the raising and lowering operators

- σ_i^+, which raises $|\psi_i\rangle$ from $|-\rangle$ to $|+\rangle$,
- σ_i^-, which lowers $|\psi_i\rangle$ from $|+\rangle$ to $|-\rangle$,

and insist that

$$\sigma_i^+\sigma_i^+ = \sigma_i^-\sigma_i^- = 0.$$

$\implies$ $\mathcal{H}$ has the matrix representation

$$\mathcal{H} = -\sum_{i\sim j}\sigma_i^z\sigma_j^z - h\sum_i \sigma_i^x$$

$$= -\sum_{i\sim j}\sigma_i^z\sigma_j^z - h\sum_i(\sigma_i^+ + \sigma_i^-)\,.$$

The quantum spin problem

To see the sparsity pattern of $\mathcal{H}$, operate on the left and right with basis vectors:

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = \left\langle \psi^1 \right| \underbrace{- \sum_{i \sim j} \sigma_i^z \sigma_j^z}_{\text{null excitation}} \underbrace{- h \sum_i (\sigma_i^+ + \sigma_i^-)}_{\text{single excitation}} \left| \psi^2 \right\rangle$$

The quantum spin problem

To see the sparsity pattern of $\mathcal{H}$, operate on the left and right with basis vectors:

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = \left\langle \psi^1 \middle| \underbrace{- \sum_{i \sim j} \sigma_i^z \sigma_j^z}_{\text{null excitation}} \underbrace{- h \sum_i (\sigma_i^+ + \sigma_i^-)}_{\text{single excitation}} \middle| \psi^2 \right\rangle$$

– If ψ^1 and ψ^2 differ by 0 spins,

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = \langle \psi^1 | - \sum\nolimits_{i \sim j} \sigma_i^z \sigma_j^z | \psi^2 \rangle.$$

The quantum spin problem

To see the sparsity pattern of $\mathcal{H}$, operate on the left and right with basis vectors:

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = \left\langle \psi^1 \middle| \underbrace{- \sum_{i \sim j} \sigma_i^z \sigma_j^z}_{\text{null excitation}} \underbrace{- h \sum_i (\sigma_i^+ + \sigma_i^-)}_{\text{single excitation}} \middle| \psi^2 \right\rangle$$

- If ψ^1 and ψ^2 differ by 0 spins,

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = \langle \psi^1 | - \sum\nolimits_{i \sim j} \sigma_i^z \sigma_j^z | \psi^2 \rangle.$$

- If ψ^1 and ψ^2 differ by 1 spin,

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = -h.$$

The quantum spin problem

To see the sparsity pattern of $\mathcal{H}$, operate on the left and right with basis vectors:

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = \left\langle \psi^1 \middle| \underbrace{- \sum_{i \sim j} \sigma_i^z \sigma_j^z}_{\text{null excitation}} \underbrace{- h \sum_i (\sigma_i^+ + \sigma_i^-)}_{\text{single excitation}} \middle| \psi^2 \right\rangle$$

- If ψ^1 and ψ^2 differ by 0 spins,

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = \langle \psi^1 | - \sum\nolimits_{i \sim j} \sigma_i^z \sigma_j^z | \psi^2 \rangle.$$

- If ψ^1 and ψ^2 differ by 1 spin,

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = -h.$$

- If ψ^1 and ψ^2 differ by 2+ spins,

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = 0.$$

The quantum spin problem

To see the sparsity pattern of $\mathcal{H}$, operate on the left and right with basis vectors:

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = \left\langle \psi^1 \middle| \underbrace{- \sum_{i \sim j} \sigma_i^z \sigma_j^z}_{\text{null excitation}} \underbrace{- h \sum_i (\sigma_i^+ + \sigma_i^-)}_{\text{single excitation}} \middle| \psi^2 \right\rangle$$

- If ψ^1 and ψ^2 differ by 0 spins,

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = \langle \psi^1 | - \sum\nolimits_{i \sim j} \sigma_i^z \sigma_j^z | \psi^2 \rangle.$$

- If ψ^1 and ψ^2 differ by 1 spin,

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = -h.$$

- If ψ^1 and ψ^2 differ by 2+ spins,

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = 0.$$

$\Longrightarrow$ The matrix $\mathcal{H}$ has $N + 1$ nonzero entries per column.

The quantum spin problem

2 spins

$$\begin{pmatrix} -2 & -h & -h & \\ -h & 2 & & -h \\ -h & & 2 & -h \\ & -h & -h & -2 \end{pmatrix}$$

The quantum spin problem

2 spins

$$\begin{pmatrix} -2 & -h & -h & \\ -h & 2 & & -h \\ -h & & 2 & -h \\ & -h & -h & -2 \end{pmatrix}$$

3 spins

$$\begin{pmatrix} -3 & -h & -h & & -h & & & \\ -h & 1 & & -h & & -h & & \\ -h & & 1 & -h & & & -h & . \\ & -h & -h & 1 & & & & -h \\ -h & & & & 1 & -h & -h & \\ & -h & & & -h & 1 & & -h \\ & & -h & & -h & & 1 & -h \\ & & & -h & & -h & -h & -3 \end{pmatrix}$$

The quantum spin problem

3 spins

2 spins

$$\begin{pmatrix} -2 & -h & -h & \\ -h & 2 & & -h \\ -h & & 2 & -h \\ & -h & -h & -2 \end{pmatrix}$$

$$\begin{pmatrix} -3 & -h & -h & & -h & & & \\ -h & 1 & & -h & & -h & & \\ -h & & 1 & -h & & & -h & . \\ & -h & -h & 1 & & & & -h \\ -h & & & & 1 & -h & -h & \\ & -h & & & -h & 1 & & -h \\ & & -h & & -h & & 1 & -h \\ & & & -h & & -h & -h & -3 \end{pmatrix}$$

For a system of N spins, the matrix has dimensions $2^N \times 2^N$,

The quantum spin problem

3 spins

2 spins

$$\begin{pmatrix} -2 & -h & -h & \\ -h & 2 & & -h \\ -h & & 2 & -h \\ & -h & -h & -2 \end{pmatrix}$$

$$\begin{pmatrix} -3 & -h & -h & & -h & & & \\ -h & 1 & & -h & & -h & & \\ -h & & 1 & -h & & & -h & . \\ & -h & -h & 1 & & & & -h \\ -h & & & & 1 & -h & -h & \\ & -h & & & -h & 1 & & -h \\ & & -h & & -h & & 1 & -h \\ & & & -h & & -h & -h & -3 \end{pmatrix}$$

For a system of N spins, the matrix has dimensions $2^N \times 2^N$,

but there are just $N + 1$ nonzero entries per column.

The electronic structure problem

Consider the Born-Oppenheimer model for N electrons orbiting a field of fixed nuclei, specified by

$$\mathcal{H} = -\frac{1}{2}\sum_i \nabla^2_{\boldsymbol{r}_i} - \sum_{i,A} \frac{Z_A}{|\boldsymbol{r}_i - \boldsymbol{R}_A|} + \sum_{j>i} \frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|}.$$

- $\boldsymbol{R}_A$ is the location of nucleus A.
- Z_A is the atomic number of A.
- $\boldsymbol{r}_i$ is the location of electron i.

The electronic structure problem

Consider the Born-Oppenheimer model for N electrons orbiting a field of fixed nuclei, specified by

$$\mathcal{H} = -\frac{1}{2}\sum_i \nabla^2_{\boldsymbol{r}_i} - \sum_{i,A} \frac{Z_A}{|\boldsymbol{r}_i - \boldsymbol{R}_A|} + \sum_{j>i} \frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|}.$$

- $\boldsymbol{R}_A$ is the location of nucleus A.
- Z_A is the atomic number of A.
- $\boldsymbol{r}_i$ is the location of electron i.

Allow each electron to occupy a spin up or down state

$$|\omega_i\rangle = |\uparrow\rangle \qquad \text{or} \qquad |\omega_i\rangle = |\downarrow\rangle,$$

and insist on the antisymmetry property

$$\psi(\cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_j, \cdots) = -\psi(\cdots, \boldsymbol{x}_j, \cdots, \boldsymbol{x}_i, \cdots),$$

where $|\boldsymbol{x}_i\rangle = |\boldsymbol{r}_i, \omega_i\rangle$ denotes the position and spin of electron i.

The electronic structure problem

Question: How can we represent $\mathcal{H}$ as a matrix?

The electronic structure problem

Question: How can we represent $\mathcal{H}$ as a matrix?

Consider an orthonormal set of single-electron spatial orbitals

$$|\phi_1\rangle, |\phi_2\rangle, \ldots, |\phi_M\rangle$$

and the associated spin orbitals

$$\begin{aligned}
|\chi_1\rangle &= |\phi_1\rangle \otimes |\uparrow\rangle, \\
|\chi_2\rangle &= |\phi_1\rangle \otimes |\downarrow\rangle, \\
&\ldots, \\
|\chi_{2M}\rangle &= |\phi_M\rangle \otimes |\downarrow\rangle.
\end{aligned}$$

The electronic structure problem

Question: How can we represent $\mathcal{H}$ as a matrix?

Next introduce the basis of Slater determinants

$$|ij\cdots k\rangle = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_i(\boldsymbol{x}_1) & \chi_j(\boldsymbol{x}_1) & \cdots & \chi_k(\boldsymbol{x}_1) \\ \chi_i(\boldsymbol{x}_2) & \chi_j(\boldsymbol{x}_2) & \cdots & \chi_k(\boldsymbol{x}_2) \\ \vdots & \vdots & & \vdots \\ \chi_i(\boldsymbol{x}_N) & \chi_j(\boldsymbol{x}_N) & \cdots & \chi_k(\boldsymbol{x}_N) \end{vmatrix}$$

The electronic structure problem

Question: How can we represent $\mathcal{H}$ as a matrix?

Next introduce the basis of Slater determinants

$$|ij\cdots k\rangle = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_i(\boldsymbol{x}_1) & \chi_j(\boldsymbol{x}_1) & \cdots & \chi_k(\boldsymbol{x}_1) \\ \chi_i(\boldsymbol{x}_2) & \chi_j(\boldsymbol{x}_2) & \cdots & \chi_k(\boldsymbol{x}_2) \\ \vdots & \vdots & & \vdots \\ \chi_i(\boldsymbol{x}_N) & \chi_j(\boldsymbol{x}_N) & \cdots & \chi_k(\boldsymbol{x}_N) \end{vmatrix}$$

Slater determinants satisfy the antisymmetry property

$$|\cdots i\cdots j\cdots\rangle = -|\cdots j\cdots i\cdots\rangle.$$

The electronic structure problem

Question: How can we represent $\mathcal{H}$ as a matrix?

Next introduce the basis of Slater determinants

$$|ij\cdots k\rangle = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_i(\boldsymbol{x}_1) & \chi_j(\boldsymbol{x}_1) & \cdots & \chi_k(\boldsymbol{x}_1) \\ \chi_i(\boldsymbol{x}_2) & \chi_j(\boldsymbol{x}_2) & \cdots & \chi_k(\boldsymbol{x}_2) \\ \vdots & \vdots & & \vdots \\ \chi_i(\boldsymbol{x}_N) & \chi_j(\boldsymbol{x}_N) & \cdots & \chi_k(\boldsymbol{x}_N) \end{vmatrix}$$

Slater determinants satisfy the antisymmetry property

$$|\cdots i\cdots j\cdots\rangle = -|\cdots j\cdots i\cdots\rangle.$$

$\Longrightarrow$ Slater determinants form an (incomplete) orthonormal basis of size $\binom{2M}{N}$.

The electronic structure problem

Question: How can we represent $\mathcal{H}$ as a matrix?

Introduce the raising and lowering operators

- a_i^+, which satisfies $a_i^+|j\cdots k\rangle = |ij\cdots k\rangle$ if i is not yet occupied,
- a_i^-, which satisfies $a_i^+|ij\cdots k\rangle = |j\cdots k\rangle$,

and insist that

$$a_i^+ a_i^+ = a_i^- a_i^- = 0.$$

The electronic structure problem

Question: How can we represent $\mathcal{H}$ as a matrix?

Introduce the raising and lowering operators

- a_i^+, which satisfies $a_i^+|j\cdots k\rangle = |ij\cdots k\rangle$ if i is not yet occupied,
- a_i^-, which satisfies $a_i^+|ij\cdots k\rangle = |j\cdots k\rangle$,

and insist that

$$a_i^+ a_i^+ = a_i^- a_i^- = 0.$$

$\Longrightarrow$ In the basis of Slater determinants, $\mathcal{H}$ has the matrix representation

$$\mathcal{H} = \sum_{i,j=1}^{2M} \langle i|h|j\rangle a_i^+ a_j + \frac{1}{2}\sum_{i,j,k,l=1}^{2M} \langle ij|kl\rangle a_i^+ a_j^+ a_l a_k$$

where $\langle i|h|j\rangle$ and $\langle ij|kl\rangle$ are integrals over one or two electron positions.

The electronic structure problem

To see the sparsity pattern of $\mathcal{H}$, operate on the left and right with basis vectors:

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = \left\langle \psi^1 \middle| \underbrace{\sum_{i,j=1}^{2M} \langle i|h|j\rangle a_i^+ a_j}_{\text{null / single excitation}} + \frac{1}{2} \underbrace{\sum_{i,j,k,l=1}^{2M} \langle ij|kl\rangle a_i^+ a_j^+ a_l a_k}_{\text{null / single / double excitation}} \middle| \psi^2 \right\rangle$$

The electronic structure problem

To see the sparsity pattern of $\mathcal{H}$, operate on the left and right with basis vectors:

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = \left\langle \psi^1 \middle| \underbrace{\sum_{i,j=1}^{2M} \langle i|h|j\rangle a_i^+ a_j}_{\text{null / single excitation}} + \frac{1}{2} \underbrace{\sum_{i,j,k,l=1}^{2M} \langle ij|kl\rangle a_i^+ a_j^+ a_l a_k}_{\text{null / single / double excitation}} \middle| \psi^2 \right\rangle$$

– Double excitations lead to

$$\langle ij\cdots|\mathcal{H}|kl\cdots\rangle = \langle ij|kl\rangle - \langle ij|lk\rangle.$$

The electronic structure problem

To see the sparsity pattern of $\mathcal{H}$, operate on the left and right with basis vectors:

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = \left\langle \psi^1 \middle| \underbrace{\sum_{i,j=1}^{2M} \langle i|h|j\rangle a_i^+ a_j}_{\text{null / single excitation}} + \frac{1}{2} \underbrace{\sum_{i,j,k,l=1}^{2M} \langle ij|kl\rangle a_i^+ a_j^+ a_l a_k}_{\text{null / single / double excitation}} \middle| \psi^2 \right\rangle$$

- Double excitations lead to

$$\langle ij\cdots|\mathcal{H}|kl\cdots\rangle = \langle ij|kl\rangle - \langle ij|lk\rangle.$$

- Single excitations lead to

$$\langle ik_1k_2\cdots|\mathcal{H}|jk_1k_2\cdots\rangle = \langle i|h|j\rangle + \sum_{n=1}^{N-1} [\langle ik_n|jk_n\rangle - \langle ik_n|k_nj\rangle]$$

The electronic structure problem

To see the sparsity pattern of $\mathcal{H}$, operate on the left and right with basis vectors:

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = \left\langle \psi^1 \middle| \underbrace{\sum_{i,j=1}^{2M} \langle i|h|j\rangle a_i^+ a_j}_{\text{null / single excitation}} + \frac{1}{2} \underbrace{\sum_{i,j,k,l=1}^{2M} \langle ij|kl\rangle a_i^+ a_j^+ a_l a_k}_{\text{null / single / double excitation}} \middle| \psi^2 \right\rangle$$

- Double excitations lead to

$$\langle ij\cdots|\mathcal{H}|kl\cdots\rangle = \langle ij|kl\rangle - \langle ij|lk\rangle.$$

- Single excitations lead to

$$\langle ik_1k_2\cdots|\mathcal{H}|jk_1k_2\cdots\rangle = \langle i|h|j\rangle + \sum_{n=1}^{N-1}[\langle ik_n|jk_n\rangle - \langle ik_n|k_nj\rangle]$$

- Null excitations lead to

$$\langle i_1i_2\cdots|\mathcal{H}|i_1i_2\cdots\rangle = \sum_{n=1}^{N}\langle i_n|h|i_n\rangle + \sum_{m,n=1}^{N}[\langle i_mi_n|i_mi_n\rangle - \langle i_mi_n|i_ni_m\rangle].$$

The electronic structure problem

To see the sparsity pattern of $\mathcal{H}$, operate on the left and right with basis vectors:

$$\langle \psi^1 | \mathcal{H} | \psi^2 \rangle = \left\langle \psi^1 \middle| \underbrace{\sum_{i,j=1}^{2M} \langle i|h|j\rangle a_i^+ a_j}_{\text{null / single excitation}} + \frac{1}{2} \underbrace{\sum_{i,j,k,l=1}^{2M} \langle ij|kl\rangle a_i^+ a_j^+ a_l a_k}_{\text{null / single / double excitation}} \middle| \psi^2 \right\rangle$$

- Double excitations lead to

$$\langle ij\cdots|\mathcal{H}|kl\cdots\rangle = \langle ij|kl\rangle - \langle ij|lk\rangle.$$

- Single excitations lead to

$$\langle ik_1k_2\cdots|\mathcal{H}|jk_1k_2\cdots\rangle = \langle i|h|j\rangle + \sum_{n=1}^{N-1}[\langle ik_n|jk_n\rangle - \langle ik_n|k_nj\rangle]$$

- Null excitations lead to

$$\langle i_1i_2\cdots|\mathcal{H}|i_1i_2\cdots\rangle = \sum_{n=1}^{N}\langle i_n|h|i_n\rangle + \sum_{m,n=1}^{N}[\langle i_mi_n|i_mi_n\rangle - \langle i_mi_n|i_ni_m\rangle].$$

$\Longrightarrow$ The matrix $\mathcal{H}$ has $\mathcal{O}(N^2M^2)$ nonzero entries per column.

Traditional methods

For small matrices, there are iterative eigenvalue solvers, e.g., power method, subspace iteration, Lanczos algorithm, Jacobi-Davidson method.

Traditional methods

For small matrices, there are iterative eigenvalue solvers, e.g., power method, subspace iteration, Lanczos algorithm, Jacobi-Davidson method.

The *power method* is the simplest method for finding $\lambda_{\max}(\boldsymbol{A})$:

Traditional methods

For small matrices, there are iterative eigenvalue solvers, e.g., power method, subspace iteration, Lanczos algorithm, Jacobi-Davidson method.

The *power method* is the simplest method for finding $\lambda_{\max}(\boldsymbol{A})$:

Power method

1. $\boldsymbol{x}^{t+1} = \boldsymbol{A}\boldsymbol{x}^{t}$
2. $\boldsymbol{x}^{t+1} = \boldsymbol{x}^{t+1} / \|\boldsymbol{x}^{t+1}\|$

Traditional methods

For small matrices, there are iterative eigenvalue solvers, e.g., power method, subspace iteration, Lanczos algorithm, Jacobi-Davidson method.

The *power method* is the simplest method for finding $\lambda_{\max}(\boldsymbol{A})$:

Power method

1. $\boldsymbol{x}^{t+1} = \boldsymbol{A}\boldsymbol{x}^t$
2. $\boldsymbol{x}^{t+1} = \boldsymbol{x}^{t+1}/\|\boldsymbol{x}^{t+1}\|$

- Historical eigenvalue estimator

$$\hat{\lambda}^t = \frac{\langle \boldsymbol{x}^0, \boldsymbol{A}\boldsymbol{x}^t \rangle}{\langle \boldsymbol{x}^0, \boldsymbol{x}^t \rangle}.$$

- Modern eigenvalue estimator

$$\hat{\lambda}^t = \frac{\langle \boldsymbol{x}^t, \boldsymbol{A}\boldsymbol{x}^t \rangle}{\langle \boldsymbol{x}^t, \boldsymbol{x}^t \rangle}.$$

Convergence of power method

Convergence of power method

Consider a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with largest-magnitude eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \cdots$ and eigenvectors $\boldsymbol{v}^1, \boldsymbol{v}^2, \ldots$, and set

$$R = \left|\frac{\lambda_2}{\lambda_1}\right|, \qquad \theta = \angle(\boldsymbol{x}^0, \boldsymbol{v}^1) = \arccos\left(\frac{\langle \boldsymbol{x}^0, \boldsymbol{v}^1 \rangle}{\|\boldsymbol{x}^0\| \|\boldsymbol{v}^1\|}\right).$$

1. Eigenvector estimates $\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots$ satisfy

$$\tan \angle(\boldsymbol{x}^t, \boldsymbol{v}^1) \leq R^t \tan \theta.$$

2. Historical or modern eigenvalue estimates $\hat{\lambda}^0, \hat{\lambda}^1, \ldots$ satisfy

$$\left|\frac{\hat{\lambda}^t - \lambda_1}{\lambda_1}\right| \leq \frac{2R^t \tan^2 \theta}{1 - R^t \tan^2 \theta} \qquad \text{or} \qquad \left|\frac{\hat{\lambda}^t - \lambda_1}{\lambda_1}\right| \leq 2R^{2t} \tan^2 \theta.$$

Convergence of power method

Convergence of power method

Consider a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with largest-magnitude eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \cdots$ and eigenvectors $\boldsymbol{v}^1, \boldsymbol{v}^2, \ldots$, and set

$$R = \left|\frac{\lambda_2}{\lambda_1}\right|, \qquad \theta = \angle(\boldsymbol{x}^0, \boldsymbol{v}^1) = \arccos\left|\frac{\langle \boldsymbol{x}^0, \boldsymbol{v}^1 \rangle}{\|\boldsymbol{x}^0\| \|\boldsymbol{v}^1\|}\right|.$$

1. Eigenvector estimates $\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots$ satisfy

$$\tan \angle(\boldsymbol{x}^t, \boldsymbol{v}^1) \leq R^t \tan \theta.$$

2. Historical or modern eigenvalue estimates $\hat{\lambda}^0, \hat{\lambda}^1, \ldots$ satisfy

$$\left|\frac{\hat{\lambda}^t - \lambda_1}{\lambda_1}\right| \leq \frac{2\,R^t \tan^2 \theta}{1 - R^t \tan^2 \theta} \quad \text{or} \quad \left|\frac{\hat{\lambda}^t - \lambda_1}{\lambda_1}\right| \leq 2\,R^{2t} \tan^2 \theta.$$

Limitations of power method

We can estimate the ground state for small systems by applying the power method to $\boldsymbol{A} = \boldsymbol{I} - \epsilon \boldsymbol{H}$ for small enough $\epsilon > 0$,

Limitations of power method

We can estimate the ground state for small systems by applying the power method to $\boldsymbol{A} = \boldsymbol{I} - \epsilon \boldsymbol{H}$ for small enough $\epsilon > 0$,

but this leads to increasingly dense vectors at each iteration.

Limitations of power method

We can estimate the ground state for small systems by applying the power method to $\boldsymbol{A} = \boldsymbol{I} - \epsilon \boldsymbol{H}$ for small enough $\epsilon > 0$,

but this leads to increasingly dense vectors at each iteration.

Question: How can we adapt the power method to larger matrices?

Full Configuration Interaction Quantum Monte Carlo

Introduce random walkers $\xi_t^1, \ldots, \xi_t^{N_t} \in \{1, 2, \ldots, d\}$ with positive or negative signs $S_t^1, \ldots, S_t^{N_t} \in \{+1, -1\}$.

Full Configuration Interaction Quantum Monte Carlo

Introduce random walkers $\xi_t^1, \ldots, \xi_t^{N_t} \in \{1, 2, \ldots, d\}$ with positive or negative signs $S_t^1, \ldots, S_t^{N_t} \in \{+1, -1\}$.

---

1. **Birth:** For each walker $\xi_t^i = j$:
   a) With probability p_{jk}, spawn N_{jk} particles at a new location $k \neq j$, where

$$N_{jk} = \lfloor |\boldsymbol{A}_{jk}|/p_{jk} + U \rfloor, \qquad U \sim \text{Unif}(0, 1).$$

   b) Assign the walkers the sign S_t^i if $\boldsymbol{A}_{jk} > 0$ or $-S_t^j$ if $\boldsymbol{A}_{jk} < 0$.

Full Configuration Interaction Quantum Monte Carlo

Introduce random walkers $\xi_t^1, \ldots, \xi_t^{N_t} \in \{1, 2, \ldots, d\}$ with positive or negative signs $S_t^1, \ldots, S_t^{N_t} \in \{+1, -1\}$.

---

1. **Birth:** For each walker $\xi_t^i = j$:
   a) With probability p_{jk}, spawn N_{jk} particles at a new location $k \neq j$, where

$$N_{jk} = \lfloor |\boldsymbol{A}_{jk}|/p_{jk} + U \rfloor, \qquad U \sim \text{Unif}(0, 1).$$

   b) Assign the walkers the sign S_t^i if $\boldsymbol{A}_{jk} > 0$ or $-S_t^j$ if $\boldsymbol{A}_{jk} < 0$.
2. **Death:** For each walker $\xi_t^i = j$:
   a) Replace ξ_t^i with N_{jj} walkers, where

$$N_{jj} = \lfloor |\boldsymbol{A}_{jj}| + U \rfloor, \qquad U \sim \text{Unif}(0, 1).$$

   b) Assign the walkers the sign S_t^i if $\boldsymbol{A}_{jj} > 0$ or $-S_t^i$ if $\boldsymbol{A}_{jj} < 0$.

Full Configuration Interaction Quantum Monte Carlo

Introduce random walkers $\xi_t^1, \ldots, \xi_t^{N_t} \in \{1, 2, \ldots, d\}$ with positive or negative signs $S_t^1, \ldots, S_t^{N_t} \in \{+1, -1\}$.

---

1. **Birth:** For each walker $\xi_t^i = j$:
   a) With probability p_{jk}, spawn N_{jk} particles at a new location $k \neq j$, where

$$N_{jk} = \lfloor |\boldsymbol{A}_{jk}|/p_{jk} + U \rfloor, \qquad U \sim \text{Unif}(0, 1).$$

   b) Assign the walkers the sign S_t^i if $\boldsymbol{A}_{jk} > 0$ or $-S_t^j$ if $\boldsymbol{A}_{jk} < 0$.
2. **Death:** For each walker $\xi_t^i = j$:
   a) Replace ξ_t^i with N_{jj} walkers, where

$$N_{jj} = \lfloor |\boldsymbol{A}_{jj}| + U \rfloor, \qquad U \sim \text{Unif}(0, 1).$$

   b) Assign the walkers the sign S_t^i if $\boldsymbol{A}_{jj} > 0$ or $-S_t^i$ if $\boldsymbol{A}_{jj} < 0$.
3. **Annihilation:** Cancel walkers on the same site with opposite signs.

Full Configuration Interaction Quantum Monte Carlo

FCIQMC approximates $\boldsymbol{x}^{t+1} = \boldsymbol{A}\boldsymbol{x}^{t}$ in the sense that

$$\mathrm{E}[\mathbf{X}^{t+1}] = \boldsymbol{A}\boldsymbol{X}^{t}, \qquad \text{where} \qquad \begin{cases} \boldsymbol{X}^{t} = \sum_{i=1}^{N_t} S_t^i \delta_{\xi_t^i} \\ \boldsymbol{X}^{t+1} = \sum_{i=1}^{N_{t+1}} S_{t+1}^i \delta_{\xi_{t+1}^i} \end{cases}.$$

Full Configuration Interaction Quantum Monte Carlo

FCIQMC approximates $\boldsymbol{x}^{t+1} = \boldsymbol{A}\boldsymbol{x}^{t}$ in the sense that

$$\mathrm{E}[\mathbf{X}^{t+1}] = \boldsymbol{A}\boldsymbol{X}^{t}, \qquad \text{where} \qquad \begin{cases} \boldsymbol{X}^{t} = \sum_{i=1}^{N_t} S_t^i \delta_{\xi_t^i} \\ \boldsymbol{X}^{t+1} = \sum_{i=1}^{N_{t+1}} S_{t+1}^i \delta_{\xi_{t+1}^i} \end{cases}.$$

– The **birth** step perform multiplication by $\boldsymbol{A}_{\text{off-diag}}$.

Full Configuration Interaction Quantum Monte Carlo

FCIQMC approximates $\boldsymbol{x}^{t+1} = \boldsymbol{A}\boldsymbol{x}^{t}$ in the sense that

$$\mathrm{E}[\mathbf{X}^{t+1}] = \boldsymbol{A}\boldsymbol{X}^{t}, \qquad \text{where} \qquad \begin{cases} \boldsymbol{X}^{t} = \sum_{i=1}^{N_t} S_t^i \delta_{\xi_t^i} \\ \boldsymbol{X}^{t+1} = \sum_{i=1}^{N_{t+1}} S_{t+1}^i \delta_{\xi_{t+1}^i} \end{cases}.$$

- The **birth** step perform multiplication by $\boldsymbol{A}_{\text{off-diag}}$.
- The **death** step performs multiplication by $\boldsymbol{A}_{\text{diag}}$.

Full Configuration Interaction Quantum Monte Carlo

FCIQMC approximates $\boldsymbol{x}^{t+1} = \boldsymbol{A}\boldsymbol{x}^{t}$ in the sense that

$$\mathrm{E}[\mathbf{X}^{t+1}] = \boldsymbol{A}\boldsymbol{X}^{t}, \qquad \text{where} \qquad \begin{cases} \boldsymbol{X}^{t} = \sum_{i=1}^{N_t} S_t^i \delta_{\xi_t^i} \\ \boldsymbol{X}^{t+1} = \sum_{i=1}^{N_{t+1}} S_{t+1}^i \delta_{\xi_{t+1}^i} \end{cases}.$$

- The **birth** step perform multiplication by $\boldsymbol{A}_{\text{off-diag}}$.
- The **death** step performs multiplication by $\boldsymbol{A}_{\text{diag}}$.
- The **annihilation** step performs addition $\boldsymbol{A} = \boldsymbol{A}_{\text{diag}} + \boldsymbol{A}_{\text{off-diag}}$.

Full Configuration Interaction Quantum Monte Carlo

FCIQMC approximates $\boldsymbol{x}^{t+1} = \boldsymbol{A}\boldsymbol{x}^t$ in the sense that

$$\mathrm{E}[\mathbf{X}^{t+1}] = \boldsymbol{A}\boldsymbol{X}^t, \qquad \text{where} \qquad \begin{cases} \boldsymbol{X}^t = \sum_{i=1}^{N_t} S_t^i \delta_{\xi_t^i} \\ \boldsymbol{X}^{t+1} = \sum_{i=1}^{N_{t+1}} S_{t+1}^i \delta_{\xi_{t+1}^i} \end{cases}.$$

- The **birth** step perform multiplication by $\boldsymbol{A}_{\text{off-diag}}$.
- The **death** step performs multiplication by $\boldsymbol{A}_{\text{diag}}$.
- The **annihilation** step performs addition $\boldsymbol{A} = \boldsymbol{A}_{\text{diag}} + \boldsymbol{A}_{\text{off-diag}}$.
- To stabilize the walker population, FCIQMC adapts the shift, i.e.,

$$\boldsymbol{A} = \boldsymbol{I} - \epsilon(\boldsymbol{H} - \delta_t \boldsymbol{I}), \qquad \delta_t = \delta_{t-1} - .01 \log\left(\frac{N_t}{N_{t-1}}\right).$$

Convergence of FCIQMC

Question: Does FCIQMC converge?

Convergence of FCIQMC

Question: Does FCIQMC converge?

---

1. The system $((\xi_t^i)_{1\leq i\leq N_t}, (S_t^i)_{1\leq i\leq N_t}, \delta_t)$ is Markovian.

Convergence of FCIQMC

Question: Does FCIQMC converge?

---

1. The system $((\xi_t^i)_{1\leq i\leq N_t}, (S_t^i)_{1\leq i\leq N_t}, \delta_t)$ is Markovian.
2. We anticipate the system converges to a stationary measure μ.

Convergence of FCIQMC

Question: Does FCIQMC converge?

---

1. The system $((\xi_t^i)_{1\leq i\leq N_t}, (S_t^i)_{1\leq i\leq N_t}, \delta_t)$ is Markovian.
2. We anticipate the system converges to a stationary measure μ.
3. Moreover, we anticipate

$$\mathrm{E}_\mu[\boldsymbol{X}^0] = \mathrm{E}_\mu\left[\sum_{i=1}^{N_0} S_0^i \delta_{X_0^i}\right]$$

lies close to the ground state.

Convergence of FCIQMC

Question: Does FCIQMC converge?

---

1. The system $((\xi_t^i)_{1\le i\le N_t}, (S_t^i)_{1\le i\le N_t}, \delta_t)$ is Markovian.
2. We anticipate the system converges to a stationary measure μ.
3. Moreover, we anticipate

$$\mathrm{E}_\mu[\boldsymbol{X}^0] = \mathrm{E}_\mu\left[\sum_{i=1}^{N_0} S_0^i \delta_{X_0^i}\right]$$

   lies close to the ground state.
4. Then, it makes sense to estimate the ground-state energy using

$$\hat{\lambda}^t = \frac{\sum_{s=t_{\min}}^{t} \langle \boldsymbol{X}^0, \boldsymbol{A}\boldsymbol{X}^s \rangle}{\sum_{s=t_{\min}}^{t} \langle \boldsymbol{X}^0, \boldsymbol{X}^s \rangle},$$

   which is a ratio of **convergent** averages of a Markov chain, with convergence rate $\sim t^{-1/2}$.

Full Configuration Interaction Fast Randomized Iteration

Question: Why not approximate $\boldsymbol{x}^{t+1} = \boldsymbol{A}\boldsymbol{x}^t$ more directly?

Full Configuration Interaction Fast Randomized Iteration

Question: Why not approximate $\boldsymbol{x}^{t+1} = \boldsymbol{A}\boldsymbol{x}^t$ more directly?

Introduce a random compression operator $\boldsymbol{\Phi} : \mathbb{R}^d \to \mathbb{R}^d$ such that

(a) $\boldsymbol{\Phi}(\boldsymbol{x})$ has at most m nonzero entries.

(b) $\mathrm{E}\boldsymbol{\Phi}(\boldsymbol{x}) = \boldsymbol{x}$.

Full Configuration Interaction Fast Randomized Iteration

Question: Why not approximate $\boldsymbol{x}^{t+1} = \boldsymbol{A}\boldsymbol{x}^t$ more directly?

Introduce a random compression operator $\boldsymbol{\Phi} : \mathbb{R}^d \to \mathbb{R}^d$ such that

(a) $\boldsymbol{\Phi}(\boldsymbol{x})$ has at most m nonzero entries.

(b) $\mathrm{E}\boldsymbol{\Phi}(\boldsymbol{x}) = \boldsymbol{x}$.

FCIFRI (Lim & Weare, 2017)

1. $\boldsymbol{X}^{t+1} = \boldsymbol{A}\boldsymbol{\Phi}(\boldsymbol{X}^t)$
2. $\boldsymbol{X}^{t+1} = \boldsymbol{X}^{t+1}/\|\boldsymbol{X}^{t+1}\|_1$

Full Configuration Interaction Fast Randomized Iteration

Question: Why not approximate $\boldsymbol{x}^{t+1} = \boldsymbol{A}\boldsymbol{x}^t$ more directly?

Introduce a random compression operator $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ such that

(a) $\Phi(\boldsymbol{x})$ has at most m nonzero entries.

(b) $\mathrm{E}\Phi(\boldsymbol{x}) = \boldsymbol{x}$.

FCIFRI (Lim & Weare, 2017)

1. $\boldsymbol{X}^{t+1} = \boldsymbol{A}\Phi(\boldsymbol{X}^t)$
2. $\boldsymbol{X}^{t+1} = \boldsymbol{X}^{t+1}/\|\boldsymbol{X}^{t+1}\|_1$

Again, we approximate the dominant eigenvalue of $\boldsymbol{A}$ using

$$\hat{\lambda}^t = \frac{\sum_{s=t_{\min}}^{t} \langle \boldsymbol{X}^0, \boldsymbol{A}\boldsymbol{X}^s \rangle}{\sum_{s=t_{\min}}^{t} \langle \boldsymbol{X}^0, \boldsymbol{X}^s \rangle}.$$

and anticipate a $\sim t^{-1/2}$ convergence rate.

Full Configuration Interaction Fast Randomized Iteration

Question: Why not approximate $\boldsymbol{x}^{t+1} = \boldsymbol{A}\boldsymbol{x}^t$ more directly?

Introduce a random compression operator $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ such that

(a) $\Phi(\boldsymbol{x})$ has at most m nonzero entries.

(b) $\mathrm{E}\Phi(\boldsymbol{x}) = \boldsymbol{x}$.

FCIFRI (Lim & Weare, 2017)

1. $\boldsymbol{X}^{t+1} = \boldsymbol{A}\Phi(\boldsymbol{X}^t)$
2. $\boldsymbol{X}^{t+1} = \boldsymbol{X}^{t+1}/\|\boldsymbol{X}^{t+1}\|_1$

Again, we approximate the dominant eigenvalue of $\boldsymbol{A}$ using

$$\hat{\lambda}^t = \frac{\sum_{s=t_{\min}}^{t} \langle \boldsymbol{X}^0, \boldsymbol{A}\boldsymbol{X}^s \rangle}{\sum_{s=t_{\min}}^{t} \langle \boldsymbol{X}^0, \boldsymbol{X}^s \rangle}.$$

and anticipate a $\sim t^{-1/2}$ convergence rate.

Question: How should we choose the compression operator Φ?

Full Configuration Interaction Fast Randomized Iteration

Proposition (optimal compression)

For any $\boldsymbol{x} \in \mathbb{R}^d$, let α be a permutation of $\{1, \ldots, d\}$ such that

$$|\boldsymbol{x}_{\alpha(1)}| \geq |\boldsymbol{x}_{\alpha(2)}| \geq \cdots \geq |\boldsymbol{x}_{\alpha(d)}|.$$

Then, the compression operator $\boldsymbol{\Phi}$ that minimizes $\mathrm{E}\|\boldsymbol{\Phi}(\boldsymbol{x}) - \boldsymbol{x}\|^2$ is characterized as follows:

Full Configuration Interaction Fast Randomized Iteration

Proposition (optimal compression)

For any $\boldsymbol{x} \in \mathbb{R}^d$, let α be a permutation of $\{1, \ldots, d\}$ such that

$$|\boldsymbol{x}_{\alpha(1)}| \geq |\boldsymbol{x}_{\alpha(2)}| \geq \cdots \geq |\boldsymbol{x}_{\alpha(d)}|.$$

Then, the compression operator $\boldsymbol{\Phi}$ that minimizes $\mathrm{E}\|\boldsymbol{\Phi}(\boldsymbol{x}) - \boldsymbol{x}\|^2$ is characterized as follows:

(a) The largest-magnitude entries are preserved exactly, i.e.,

$$\boldsymbol{\Phi}(\boldsymbol{x})_{\alpha(i)} = \boldsymbol{x}_{\alpha(i)}, \qquad i = 1, \ldots, k.$$

Full Configuration Interaction Fast Randomized Iteration

Proposition (optimal compression)

For any $\boldsymbol{x} \in \mathbb{R}^d$, let α be a permutation of $\{1, \ldots, d\}$ such that

$$|\boldsymbol{x}_{\alpha(1)}| \geq |\boldsymbol{x}_{\alpha(2)}| \geq \cdots \geq |\boldsymbol{x}_{\alpha(d)}|.$$

Then, the compression operator $\boldsymbol{\Phi}$ that minimizes $\mathrm{E}\|\boldsymbol{\Phi}(\boldsymbol{x}) - \boldsymbol{x}\|^2$ is characterized as follows:

(a) The largest-magnitude entries are preserved exactly, i.e.,

$$\boldsymbol{\Phi}(\boldsymbol{x})_{\alpha(i)} = \boldsymbol{x}_{\alpha(i)}, \qquad i = 1, \ldots, k.$$

(b) The smallest-magnitude entries are randomly perturbed, i.e., $\boldsymbol{\Phi}(\boldsymbol{x})_i = \boldsymbol{x}_i / \boldsymbol{p}_i$ with probability $\boldsymbol{p}_i$ and $\boldsymbol{\Phi}(\boldsymbol{x})_i = 0$ otherwise, where

$$\boldsymbol{p}_i = \frac{(m-k)|\boldsymbol{x}_i|}{\sum_{j=k+1}^{d} |\boldsymbol{x}_{\alpha(j)}|}.$$

Full Configuration Interaction Fast Randomized Iteration

Proposition (optimal compression)

For any $\boldsymbol{x} \in \mathbb{R}^d$, let α be a permutation of $\{1, \ldots, d\}$ such that

$$|\boldsymbol{x}_{\alpha(1)}| \geq |\boldsymbol{x}_{\alpha(2)}| \geq \cdots \geq |\boldsymbol{x}_{\alpha(d)}|.$$

Then, the compression operator $\boldsymbol{\Phi}$ that minimizes $\mathrm{E}\|\boldsymbol{\Phi}(\boldsymbol{x}) - \boldsymbol{x}\|^2$ is characterized as follows:

(a) The largest-magnitude entries are preserved exactly, i.e.,

$$\boldsymbol{\Phi}(\boldsymbol{x})_{\alpha(i)} = \boldsymbol{x}_{\alpha(i)}, \qquad i = 1, \ldots, k.$$

(b) The smallest-magnitude entries are randomly perturbed, i.e., $\boldsymbol{\Phi}(\boldsymbol{x})_i = \boldsymbol{x}_i / \boldsymbol{p}_i$ with probability $\boldsymbol{p}_i$ and $\boldsymbol{\Phi}(\boldsymbol{x})_i = 0$ otherwise, where

$$\boldsymbol{p}_i = \frac{(m-k)|\boldsymbol{x}_i|}{\sum_{j=k+1}^{d} |\boldsymbol{x}_{\alpha(j)}|}.$$

(c) The number k is as small as possible, while ensuring $\boldsymbol{p}_{\alpha(k+1)} \leq 1$.

Empirical tests

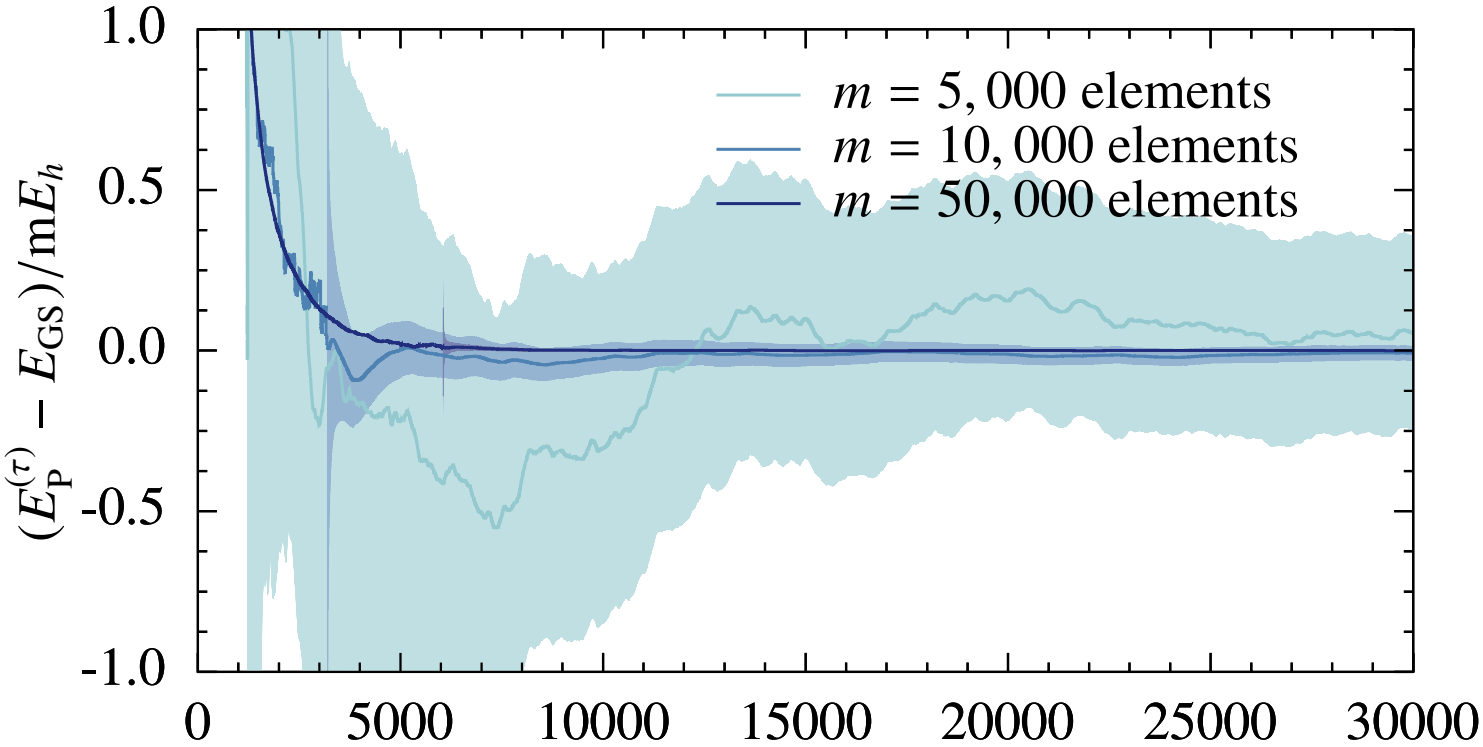


Figure: FCIFRI ground-state energy estimates for Neon (8e, 22o).

Limitations

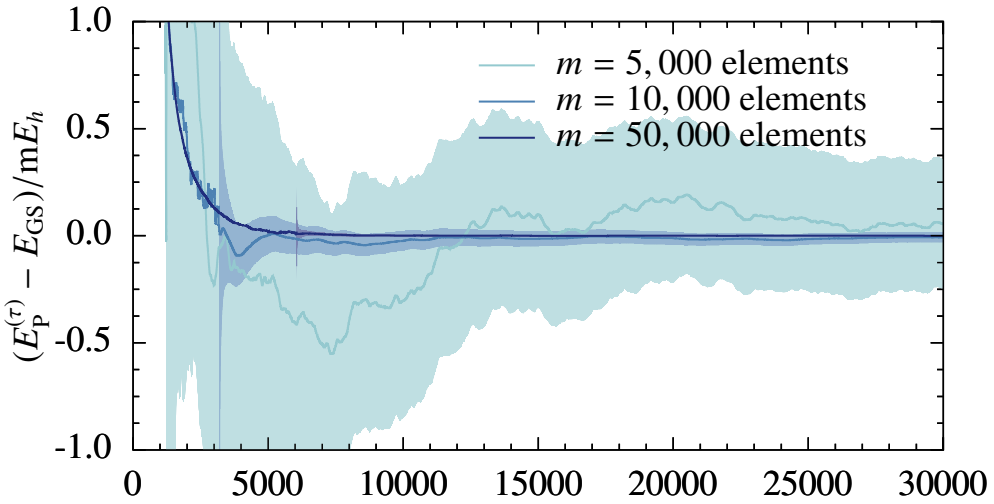


Figure: FCIFRI ground-state energy estimates for Neon (8e, 22o).

Limitations

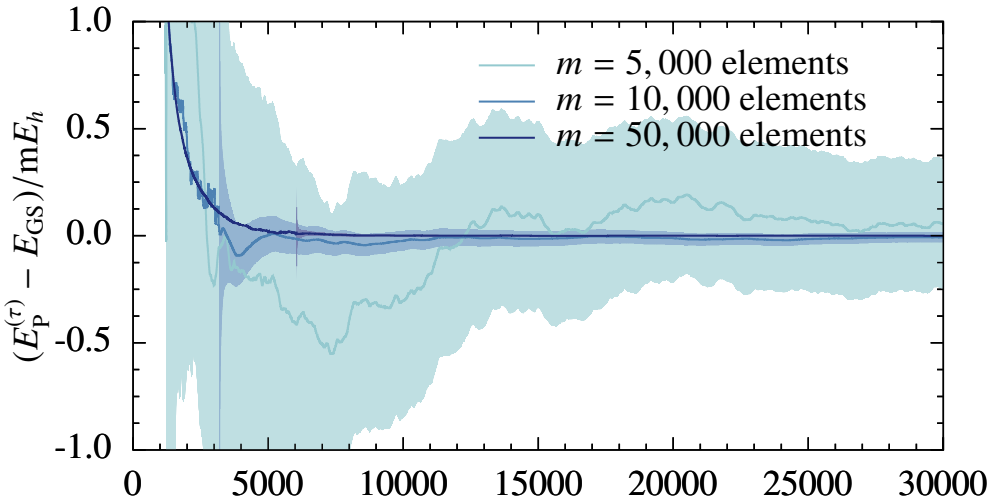


Figure: FCIFRI ground-state energy estimates for Neon (8e, 22o).

1. FCIFRI is more accurate than FCIQMC but FCIQMC is cheaper than FCIFRI. In practice, we combine them.

Limitations

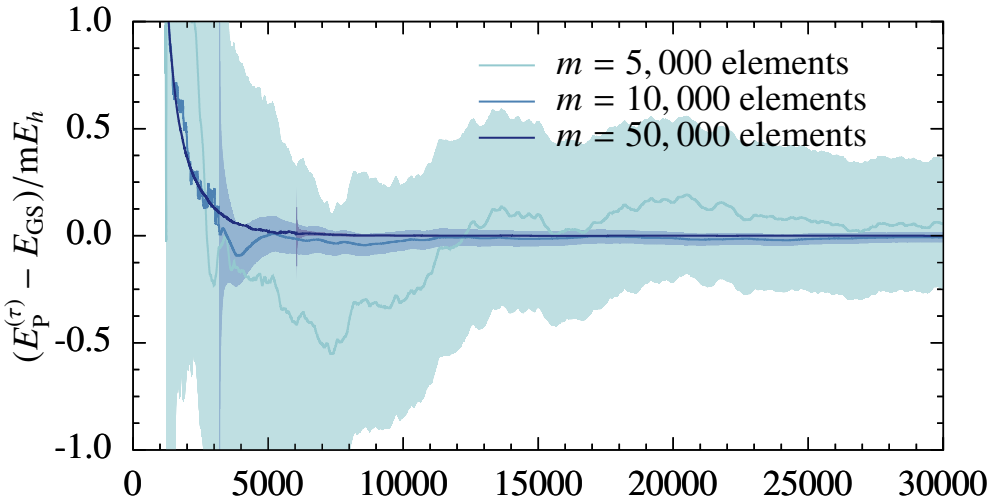


Figure: FCIFRI ground-state energy estimates for Neon (8e, 22o).

1. FCIFRI is more accurate than FCIQMC but FCIQMC is cheaper than FCIFRI. In practice, we combine them.
2. There is not a mathematical understanding of when these methods succeed versus fail (i.e., what matrix properties).

Introducing variational Monte Carlo

The ground state: The ground-state wavefunction ψ solves a Hermitian eigenvalue problem $\lambda\psi = \mathcal{H}\psi$ with λ as small as possible.

Introducing variational Monte Carlo

The ground state: The ground-state wavefunction ψ solves a Hermitian eigenvalue problem $\lambda\psi = \mathcal{H}\psi$ with λ as small as possible.
The variational principle: The ground-state wavefunction ψ minimizes

$$\mathcal{E}[\psi] = \frac{\langle \psi, \mathcal{H}\psi \rangle}{\langle \psi, \psi \rangle}.$$

Introducing variational Monte Carlo

The ground state: The ground-state wavefunction ψ solves a Hermitian eigenvalue problem $\lambda\psi = \mathcal{H}\psi$ with λ as small as possible.
The variational principle: The ground-state wavefunction ψ minimizes

$$\mathcal{E}[\psi] = \frac{\langle \psi, \mathcal{H}\psi \rangle}{\langle \psi, \psi \rangle}.$$

Variational Monte Carlo: VMC minimizes $\mathcal{E}[\psi]$ over a wavefunction class $\psi_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ is a vector of real- or complex-valued parameters.

Introducing variational Monte Carlo

The ground state: The ground-state wavefunction ψ solves a Hermitian eigenvalue problem $\lambda\psi = \mathcal{H}\psi$ with λ as small as possible.
The variational principle: The ground-state wavefunction ψ minimizes

$$\mathcal{E}[\psi] = \frac{\langle \psi, \mathcal{H}\psi \rangle}{\langle \psi, \psi \rangle}.$$

Variational Monte Carlo: VMC minimizes $\mathcal{E}[\psi]$ over a wavefunction class $\psi_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ is a vector of real- or complex-valued parameters.

Variational Monte Carlo (VMC)

1. Draw samples from the wavefunction density

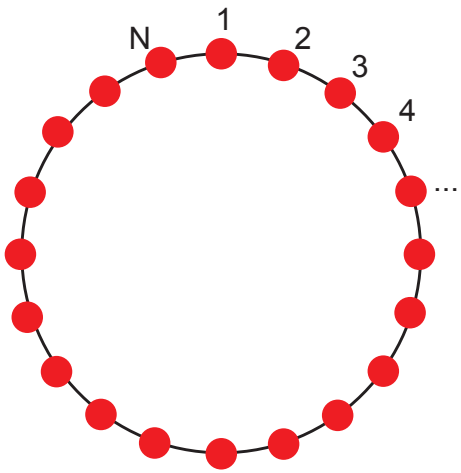
$$\rho_{\boldsymbol{\theta}}(\boldsymbol{x}) \propto |\psi_{\boldsymbol{\theta}}(\boldsymbol{x})|^2.$$

2. Use the random samples to estimate the energy gradient $\boldsymbol{g} = \nabla_{\boldsymbol{\theta}} \mathcal{E}[\psi_{\boldsymbol{\theta}}]$ and potentially other quantities for the optimization.
3. Update the $\boldsymbol{\theta}$ parameters to reduce the energy.

Introducing variational Monte Carlo

Main example: We will apply VMC to the transverse-field Ising model for spin-1/2 particles on a periodic 1-D lattice, specified by

$$\mathcal{H} = -\sum_{i\sim j} \sigma_i^z \sigma_j^z - h \sum_i \sigma_i^x$$

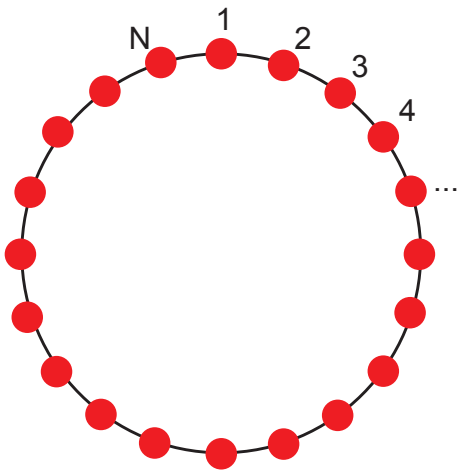


- σ_i^x and σ_i^z are Pauli operators
- $i \sim j$ signifies that i and j are neighbors
- h is a real-valued parameter

Introducing variational Monte Carlo

Main example: We will apply VMC to the transverse-field Ising model for spin-1/2 particles on a periodic 1-D lattice, specified by

$$\mathcal{H} = -\sum_{i\sim j} \sigma_i^z \sigma_j^z - h \sum_i \sigma_i^x$$



- σ_i^x and σ_i^z are Pauli operators
- $i \sim j$ signifies that i and j are neighbors
- h is a real-valued parameter

Ansatz: ψ is a restricted Bolzmann machine (RBM) ansatz

$$\psi_{\boldsymbol{w},\boldsymbol{b}}(\boldsymbol{\sigma}) = \prod_{i=1}^{\alpha} \prod_{\mathcal{T}} \cosh\left(\sum_j \boldsymbol{w}_{ij} \left(\mathcal{T}\boldsymbol{\sigma}\right)_j + \boldsymbol{b}_i \right).$$

where $\mathcal{T}$ ranges over translation operators on the lattice, and $\boldsymbol{w}$ and $\boldsymbol{b}$ are vectors of complex-valued parameters, called *weights* and *biases*.

Introducing variational Monte Carlo

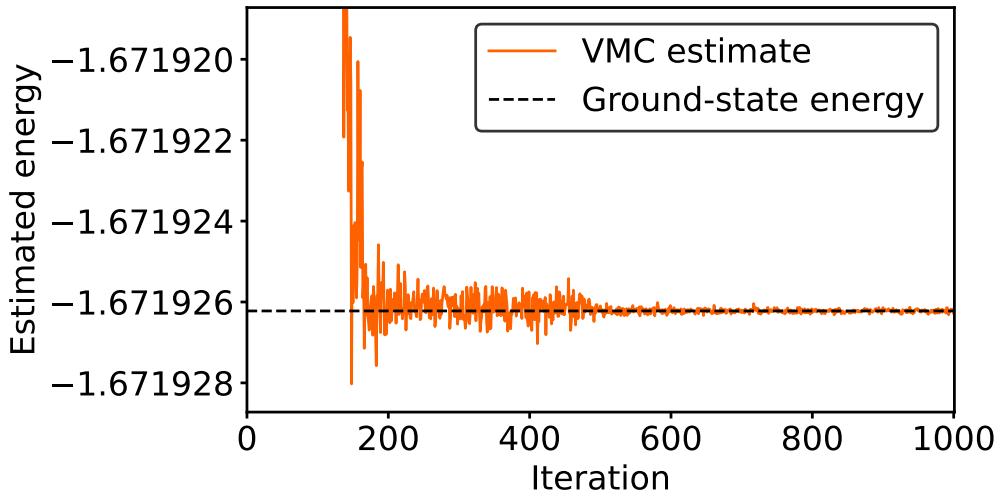


Figure: VMC ground-state energy estimates for a 200×1 Ising model with a transverse magnetic field ($h = 1.5$).

Introducing variational Monte Carlo

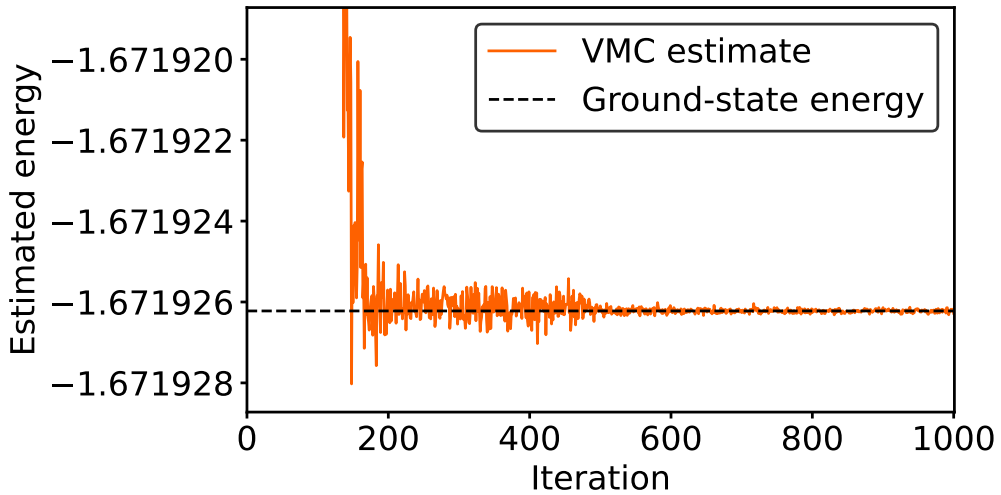


Figure: VMC ground-state energy estimates for a 200×1 Ising model with a transverse magnetic field ($h = 1.5$).

$\rightarrow$ VMC estimated energies decrease and improve over time.

Introducing variational Monte Carlo

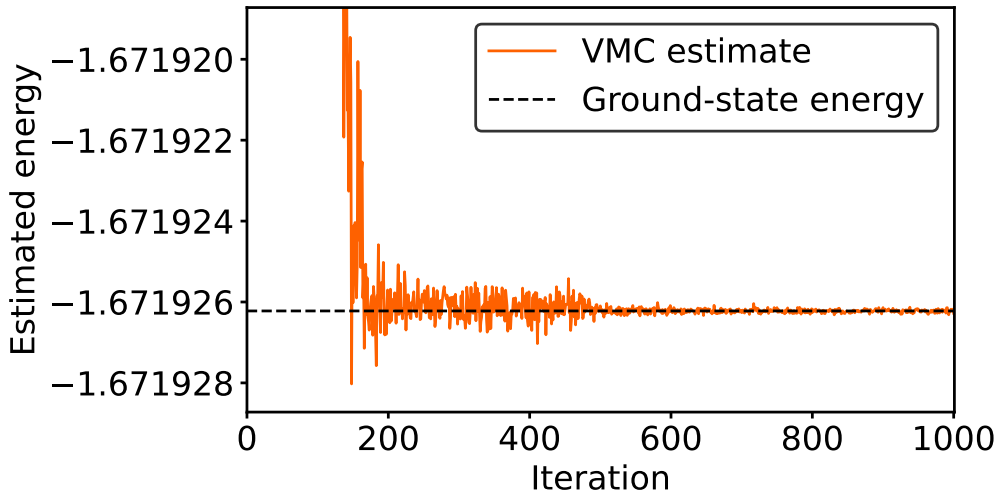


Figure: VMC ground-state energy estimates for a 200×1 Ising model with a transverse magnetic field ($h = 1.5$).

$\rightarrow$ VMC estimated energies decrease and improve over time.

$\rightarrow$ The variance in the energies also decreases. This is because of the *vanishing variance principle*, a special feature of VMC (proved later).

Introducing variational Monte Carlo

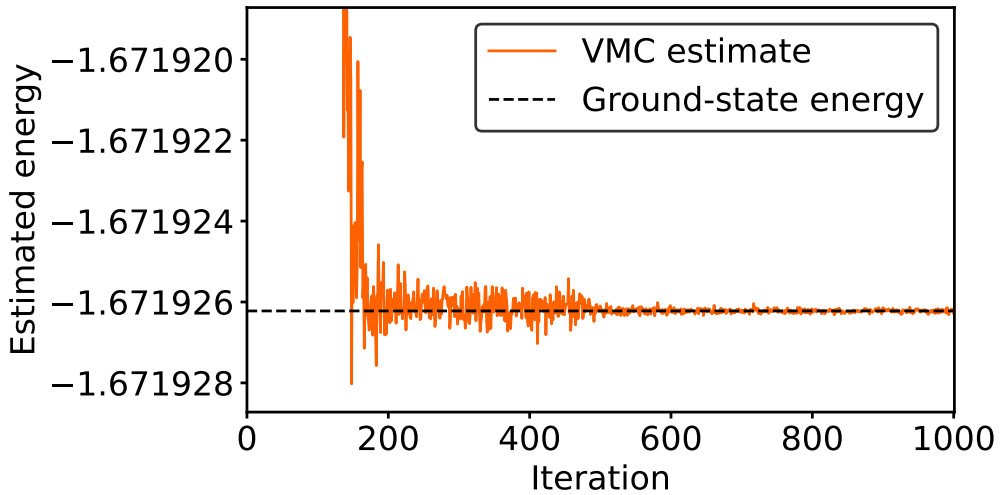


Figure: VMC ground-state energy estimates for a 200×1 Ising model with a transverse magnetic field ($h = 1.5$).

$\rightarrow$ VMC estimated energies decrease and improve over time.

$\rightarrow$ The variance in the energies also decreases. This is because of the *vanishing variance principle*, a special feature of VMC (proved later).

Question: How precisely do we update the $\boldsymbol{\theta}$ parameters?

Gradients and Hessians

1. Fix a vector of parameters $\boldsymbol{\theta}$ and consider a small update $\boldsymbol{\theta} + \boldsymbol{\delta}$. The resulting (intermediate-normalized) wavefunction is

$$\widehat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}} = \frac{\langle \psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}} \rangle}{\langle \psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}+\boldsymbol{\delta}} \rangle} \psi_{\boldsymbol{\theta}+\boldsymbol{\delta}}.$$

Gradients and Hessians

1. Fix a vector of parameters $\boldsymbol{\theta}$ and consider a small update $\boldsymbol{\theta} + \boldsymbol{\delta}$. The resulting (intermediate-normalized) wavefunction is

$$\widehat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}} = \frac{\langle \psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}} \rangle}{\langle \psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}+\boldsymbol{\delta}} \rangle} \psi_{\boldsymbol{\theta}+\boldsymbol{\delta}}.$$

2. Apply a Taylor series expansion

$$\widehat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}} = \widehat{\psi} + \sum\nolimits_i \boldsymbol{\delta}_i \widehat{\psi}_i + \frac{1}{2} \sum\nolimits_{ij} \boldsymbol{\delta}_i \boldsymbol{\delta}_j \widehat{\psi}_{ij} + \mathcal{O}(|\boldsymbol{\delta}|^3),$$

where

$$\begin{aligned}
\widehat{\psi} &= \widehat{\psi}_{\boldsymbol{\theta}} = \psi_{\boldsymbol{\theta}}, \\
\widehat{\psi}_i &= \partial_{\boldsymbol{\theta}_i} \widehat{\psi}_{\boldsymbol{\theta}} = \partial_{\boldsymbol{\theta}_i} \psi_{\boldsymbol{\theta}} - \frac{\langle \psi_{\boldsymbol{\theta}}, \partial_{\boldsymbol{\theta}_i} \psi_{\boldsymbol{\theta}} \rangle}{\langle \psi_{\boldsymbol{\theta}}, \psi_{\boldsymbol{\theta}} \rangle} \psi_{\boldsymbol{\theta}}, \\
\widehat{\psi}_{ij} &= \partial^2_{\boldsymbol{\theta}_i \boldsymbol{\theta}_j} \widehat{\psi}_{\boldsymbol{\theta}} = \dots.
\end{aligned}$$

Gradients and Hessians

3. Use the Taylor series expansion for $\hat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}$ to calculate

$$\underbrace{\mathcal{E}\left[\widehat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}\right] - \mathcal{E}\left[\widehat{\psi}_{\boldsymbol{\theta}}\right]}_{\text{energy difference}} = \underbrace{\boldsymbol{\delta}^* \boldsymbol{g} + \boldsymbol{g}^* \boldsymbol{\delta}}_{\text{gradient terms}} + \underbrace{\boldsymbol{\delta}^* \boldsymbol{H} \boldsymbol{\delta} + \Re(\boldsymbol{\delta}^T \overline{\boldsymbol{J}}\, \boldsymbol{\delta})}_{\text{Hessian terms}} + \mathcal{O}(|\boldsymbol{\delta}|^3),$$

$$\boldsymbol{g}_i = \frac{\left\langle \widehat{\psi}_i, \widehat{\mathcal{H}}\, \widehat{\psi} \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle}, \quad \boldsymbol{H}_{ij} = \frac{\left\langle \widehat{\psi}_i, \widehat{\mathcal{H}}\, \widehat{\psi}_j \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle}, \quad \boldsymbol{J}_{ij} = \frac{\left\langle \widehat{\psi}_{ij}, \widehat{\mathcal{H}}\, \widehat{\psi} \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle},$$

where $\widehat{\mathcal{H}} = \mathcal{H} - \mathcal{E}\left[\widehat{\psi}\right]$.

Gradients and Hessians

3. Use the Taylor series expansion for $\hat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}$ to calculate

$$\underbrace{\mathcal{E}\left[\widehat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}\right] - \mathcal{E}\left[\widehat{\psi}_{\boldsymbol{\theta}}\right]}_{\text{energy difference}} = \underbrace{\boldsymbol{\delta}^* \boldsymbol{g} + \boldsymbol{g}^* \boldsymbol{\delta}}_{\text{gradient terms}} + \underbrace{\boldsymbol{\delta}^* \boldsymbol{H} \boldsymbol{\delta} + \Re(\boldsymbol{\delta}^T \overline{\boldsymbol{J}}\, \boldsymbol{\delta})}_{\text{Hessian terms}} + \mathcal{O}(|\boldsymbol{\delta}|^3),$$

$$\boldsymbol{g}_i = \frac{\left\langle \widehat{\psi}_i, \widehat{\mathcal{H}}\, \widehat{\psi} \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle}, \quad \boldsymbol{H}_{ij} = \frac{\left\langle \widehat{\psi}_i, \widehat{\mathcal{H}}\, \widehat{\psi}_j \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle}, \quad \boldsymbol{J}_{ij} = \frac{\left\langle \widehat{\psi}_{ij}, \widehat{\mathcal{H}}\, \widehat{\psi} \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle},$$

where $\widehat{\mathcal{H}} = \mathcal{H} - \mathcal{E}\left[\widehat{\psi}\right]$.

Three main takeaways:

Gradients and Hessians

3. Use the Taylor series expansion for $\hat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}$ to calculate

$$\underbrace{\mathcal{E}\left[\widehat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}\right] - \mathcal{E}\left[\widehat{\psi}_{\boldsymbol{\theta}}\right]}_{\text{energy difference}} = \underbrace{\boldsymbol{\delta}^* \boldsymbol{g} + \boldsymbol{g}^* \boldsymbol{\delta}}_{\text{gradient terms}} + \underbrace{\boldsymbol{\delta}^* \boldsymbol{H} \boldsymbol{\delta} + \Re(\boldsymbol{\delta}^T \overline{\boldsymbol{J}}\, \boldsymbol{\delta})}_{\text{Hessian terms}} + \mathcal{O}(|\boldsymbol{\delta}|^3),$$

$$\boldsymbol{g}_i = \frac{\left\langle \widehat{\psi}_i, \widehat{\mathcal{H}}\, \widehat{\psi} \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle}, \quad \boldsymbol{H}_{ij} = \frac{\left\langle \widehat{\psi}_i, \widehat{\mathcal{H}}\, \widehat{\psi}_j \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle}, \quad \boldsymbol{J}_{ij} = \frac{\left\langle \widehat{\psi}_{ij}, \widehat{\mathcal{H}}\, \widehat{\psi} \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle},$$

where $\widehat{\mathcal{H}} = \mathcal{H} - \mathcal{E}\left[\widehat{\psi}\right]$.

Three main takeaways:

1. For real $\boldsymbol{\theta}$, the gradient is $2\boldsymbol{g}$ and the Hessian is $2\boldsymbol{H} + 2\boldsymbol{J}$.

Gradients and Hessians

3. Use the Taylor series expansion for $\hat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}$ to calculate

$$\underbrace{\mathcal{E}\left[\widehat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}\right] - \mathcal{E}\left[\widehat{\psi}_{\boldsymbol{\theta}}\right]}_{\text{energy difference}} = \underbrace{\boldsymbol{\delta}^* \boldsymbol{g} + \boldsymbol{g}^* \boldsymbol{\delta}}_{\text{gradient terms}} + \underbrace{\boldsymbol{\delta}^* \boldsymbol{H} \boldsymbol{\delta} + \Re(\boldsymbol{\delta}^T \overline{\boldsymbol{J}}\, \boldsymbol{\delta})}_{\text{Hessian terms}} + \mathcal{O}(|\boldsymbol{\delta}|^3),$$

$$\boldsymbol{g}_i = \frac{\left\langle \widehat{\psi}_i, \widehat{\mathcal{H}}\, \widehat{\psi} \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle}, \quad \boldsymbol{H}_{ij} = \frac{\left\langle \widehat{\psi}_i, \widehat{\mathcal{H}}\, \widehat{\psi}_j \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle}, \quad \boldsymbol{J}_{ij} = \frac{\left\langle \widehat{\psi}_{ij}, \widehat{\mathcal{H}}\, \widehat{\psi} \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle},$$

where $\widehat{\mathcal{H}} = \mathcal{H} - \mathcal{E}\left[\widehat{\psi}\right]$.

Three main takeaways:

1. For real $\boldsymbol{\theta}$, the gradient is $2\boldsymbol{g}$ and the Hessian is $2\boldsymbol{H} + 2\boldsymbol{J}$.
2. For complex $\boldsymbol{\theta}$, the Wirtinger gradient is $\begin{pmatrix} \boldsymbol{g} \\ \overline{\boldsymbol{g}} \end{pmatrix}$, and the Wirtinger Hessian is $\begin{pmatrix} \boldsymbol{H} & \boldsymbol{J} \\ \overline{\boldsymbol{J}} & \overline{\boldsymbol{H}} \end{pmatrix}$.

Gradients and Hessians

3. Use the Taylor series expansion for $\hat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}$ to calculate

$$\underbrace{\mathcal{E}\left[\widehat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}\right] - \mathcal{E}\left[\widehat{\psi}_{\boldsymbol{\theta}}\right]}_{\text{energy difference}} = \underbrace{\boldsymbol{\delta}^* \boldsymbol{g} + \boldsymbol{g}^* \boldsymbol{\delta}}_{\text{gradient terms}} + \underbrace{\boldsymbol{\delta}^* \boldsymbol{H} \boldsymbol{\delta} + \Re(\boldsymbol{\delta}^T \overline{\boldsymbol{J}}\, \boldsymbol{\delta})}_{\text{Hessian terms}} + \mathcal{O}(|\boldsymbol{\delta}|^3),$$

$$\boldsymbol{g}_i = \frac{\left\langle \widehat{\psi}_i, \widehat{\mathcal{H}}\,\widehat{\psi} \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle}, \quad \boldsymbol{H}_{ij} = \frac{\left\langle \widehat{\psi}_i, \widehat{\mathcal{H}}\,\widehat{\psi}_j \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle}, \quad \boldsymbol{J}_{ij} = \frac{\left\langle \widehat{\psi}_{ij}, \widehat{\mathcal{H}}\,\widehat{\psi} \right\rangle}{\left\langle \widehat{\psi}, \widehat{\psi} \right\rangle},$$

where $\widehat{\mathcal{H}} = \mathcal{H} - \mathcal{E}\left[\widehat{\psi}\right]$.

Three main takeaways:

1. For real $\boldsymbol{\theta}$, the gradient is $2\boldsymbol{g}$ and the Hessian is $2\boldsymbol{H} + 2\boldsymbol{J}$.
2. For complex $\boldsymbol{\theta}$, the Wirtinger gradient is $\begin{pmatrix} \boldsymbol{g} \\ \overline{\boldsymbol{g}} \end{pmatrix}$, and the Wirtinger Hessian is $\begin{pmatrix} \boldsymbol{H} & \boldsymbol{J} \\ \overline{\boldsymbol{J}} & \overline{\boldsymbol{H}} \end{pmatrix}$.
3. $\boldsymbol{g} \to \boldsymbol{0}$ and $\boldsymbol{J} \to \boldsymbol{0}$ as $\hat{\psi}$ approaches any eigenstate of $\mathcal{H}$.

Gradients and Hessians

Proposition (ground state regularity)

The vector $\boldsymbol{g}$ and matrix $\boldsymbol{J}$ are bounded by

$$|\boldsymbol{g}_i| \leq \frac{\| \widehat{\psi}_i \|}{\| \widehat{\psi} \|} \min_{\lambda \in \mathbb{R}} \frac{\| (\mathcal{H} - \lambda) \widehat{\psi} \|}{\| \widehat{\psi} \|}, \qquad |\boldsymbol{J}_{ij}| \leq \frac{\| \widehat{\psi}_{ij} \|}{\| \widehat{\psi} \|} \min_{\lambda \in \mathbb{R}} \frac{\| (\mathcal{H} - \lambda) \widehat{\psi} \|}{\| \widehat{\psi} \|}.$$

Therefore, $\boldsymbol{g} \to \mathbf{0}$ and $\boldsymbol{J} \to \mathbf{0}$ as $\min_{\lambda \in \mathbb{R}} \| (\mathcal{H} - \lambda) \widehat{\psi} \| / \| \widehat{\psi} \| \to 0$, assuming uniformly bounded $\| \widehat{\psi}_i \| / \| \widehat{\psi} \|$ and $\| \widehat{\psi}_{ij} \| / \| \widehat{\psi} \|$ terms.

Gradients and Hessians

Proposition (ground state regularity)

The vector $\boldsymbol{g}$ and matrix $\boldsymbol{J}$ are bounded by

$$|\boldsymbol{g}_i| \leq \frac{\| \widehat{\psi}_i \|}{\| \widehat{\psi} \|} \min_{\lambda \in \mathbb{R}} \frac{\| (\mathcal{H} - \lambda) \widehat{\psi} \|}{\| \widehat{\psi} \|}, \qquad |\boldsymbol{J}_{ij}| \leq \frac{\| \widehat{\psi}_{ij} \|}{\| \widehat{\psi} \|} \min_{\lambda \in \mathbb{R}} \frac{\| (\mathcal{H} - \lambda) \widehat{\psi} \|}{\| \widehat{\psi} \|}.$$

Therefore, $\boldsymbol{g} \to \mathbf{0}$ and $\boldsymbol{J} \to \mathbf{0}$ as $\min_{\lambda \in \mathbb{R}} \| (\mathcal{H} - \lambda) \widehat{\psi} \| / \| \widehat{\psi} \| \to 0$, assuming uniformly bounded $\| \widehat{\psi}_i \| / \| \widehat{\psi} \|$ and $\| \widehat{\psi}_{ij} \| / \| \widehat{\psi} \|$ terms.

The Hessian structure has implications for VMC optimization.

Various optimization methods

Gradient descent methods: Choose $\boldsymbol{\delta}$ to minimize

$$\underbrace{\mathcal{E}_{\text{linear}}\left[\hat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}\right] - \mathcal{E}\left[\hat{\psi}_{\boldsymbol{\theta}}\right]}_{\text{linearized energy difference}} = \underbrace{\boldsymbol{\delta}^* \boldsymbol{g} + \boldsymbol{g}^* \boldsymbol{\delta}}_{\text{gradient}},$$

plus a penalization term that keeps the update small.

Various optimization methods

Gradient descent methods: Choose $\boldsymbol{\delta}$ to minimize

$$\underbrace{\mathcal{E}_{\text{linear}}\left[\hat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}\right] - \mathcal{E}\left[\hat{\psi}_{\boldsymbol{\theta}}\right]}_{\text{linearized energy difference}} = \underbrace{\boldsymbol{\delta}^* \boldsymbol{g} + \boldsymbol{g}^* \boldsymbol{\delta}}_{\text{gradient}},$$

plus a penalization term that keeps the update small.

---

1. The penalization term is either

$$\underbrace{\epsilon^{-1} |\boldsymbol{\delta}|^2}_{\text{GD}} \qquad \text{or} \qquad \underbrace{\epsilon^{-1} \angle\left(\hat{\psi}_{\boldsymbol{\theta}}, \hat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}\right)^2}_{\text{natural GD}},$$

   where $\epsilon > 0$ is a tunable parameter.

Various optimization methods

Gradient descent methods: Choose $\boldsymbol{\delta}$ to minimize

$$\underbrace{\mathcal{E}_{\text{linear}}\left[\hat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}\right] - \mathcal{E}\left[\hat{\psi}_{\boldsymbol{\theta}}\right]}_{\text{linearized energy difference}} = \underbrace{\boldsymbol{\delta}^* \boldsymbol{g} + \boldsymbol{g}^* \boldsymbol{\delta}}_{\text{gradient}},$$

plus a penalization term that keeps the update small.

---

1. The penalization term is either

$$\underbrace{\epsilon^{-1} |\boldsymbol{\delta}|^2}_{\text{GD}} \qquad \text{or} \qquad \underbrace{\epsilon^{-1} \angle\big(\hat{\psi}_{\boldsymbol{\theta}}, \hat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}} \big)^2}_{\text{natural GD}},$$

   where $\epsilon > 0$ is a tunable parameter.

2. The term $\angle\big(\hat{\psi}_{\boldsymbol{\theta}}, \hat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}} \big)^2$ is approximated using

$$\angle\big(\hat{\psi}_{\boldsymbol{\theta}}, \hat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}} \big)^2 = \boldsymbol{\delta}^* \boldsymbol{S} \boldsymbol{\delta} + \mathcal{O}\big(|\boldsymbol{\delta}|^3 \big), \qquad \boldsymbol{S}_{ij} = \frac{\big\langle \hat{\psi}_i, \hat{\psi}_j \big\rangle}{\big\langle \hat{\psi}, \hat{\psi} \big\rangle}.$$

   Natural GD ('stochastic reconfiguration') uses the penalization $\epsilon^{-1} \boldsymbol{\delta}^* (\boldsymbol{S} + \eta \boldsymbol{I}) \boldsymbol{\delta}$, where $\eta > 0$ helps keep the updates small.

Various optimization methods

GD and natural GD

Choose $\boldsymbol{\delta}$ to solve

$$\min_{\boldsymbol{\delta}} \left[\boldsymbol{\delta}^* \boldsymbol{g} + \boldsymbol{g}^* \boldsymbol{\delta} + \frac{\boldsymbol{\delta}^* \boldsymbol{R} \boldsymbol{\delta}}{\epsilon} \right],$$

where $\boldsymbol{R} = \boldsymbol{I}$ in GD and $\boldsymbol{R} = \boldsymbol{S} + \eta \boldsymbol{I}$ in natural GD. Equivalently, set

$$\boldsymbol{\delta} = -\epsilon \boldsymbol{R}^{-1} \boldsymbol{g}.$$

Various optimization methods

Rayleigh-Gauss-Newton method: Choose $\boldsymbol{\delta}$ to minimize

$$\underbrace{\mathcal{E}_{\text{quad}}\left[\widehat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}\right] - \mathcal{E}\left[\widehat{\psi}_{\boldsymbol{\theta}}\right]}_{\text{quadratic energy difference}} = \underbrace{\boldsymbol{\delta}^* \boldsymbol{g} + \boldsymbol{g}^* \boldsymbol{\delta}}_{\text{gradient}} + \underbrace{\boldsymbol{\delta}^* \boldsymbol{H} \boldsymbol{\delta}}_{\text{quasi-Hessian}},$$

plus a penalization term that keeps the update small.

Various optimization methods

Rayleigh-Gauss-Newton method: Choose $\boldsymbol{\delta}$ to minimize

$$\underbrace{\mathcal{E}_{\text{quad}}\left[\widehat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}\right] - \mathcal{E}\left[\widehat{\psi}_{\boldsymbol{\theta}}\right]}_{\text{quadratic energy difference}} = \underbrace{\boldsymbol{\delta}^* \boldsymbol{g} + \boldsymbol{g}^* \boldsymbol{\delta}}_{\text{gradient}} + \underbrace{\boldsymbol{\delta}^* \boldsymbol{H} \boldsymbol{\delta}}_{\text{quasi-Hessian}},$$

plus a penalization term that keeps the update small.

---

1. The missing Hessian term $\Re(\boldsymbol{\delta}^T \overline{\boldsymbol{J}}\, \boldsymbol{\delta})$ is small in practice and $\to \mathbf{0}$ as ψ approaches the true ground state.

Various optimization methods

Rayleigh-Gauss-Newton method: Choose $\boldsymbol{\delta}$ to minimize

$$\underbrace{\mathcal{E}_{\text{quad}}\left[\widehat{\psi}_{\boldsymbol{\theta}+\boldsymbol{\delta}}\right] - \mathcal{E}\left[\widehat{\psi}_{\boldsymbol{\theta}}\right]}_{\text{quadratic energy difference}} = \underbrace{\boldsymbol{\delta}^* \boldsymbol{g} + \boldsymbol{g}^* \boldsymbol{\delta}}_{\text{gradient}} + \underbrace{\boldsymbol{\delta}^* \boldsymbol{H} \boldsymbol{\delta}}_{\text{quasi-Hessian}},$$

plus a penalization term that keeps the update small.

---

1. The missing Hessian term $\Re(\boldsymbol{\delta}^T \overline{\boldsymbol{J}}\, \boldsymbol{\delta})$ is small in practice and $\to \mathbf{0}$ as ψ approaches the true ground state.
2. As a penalization term, we use $\epsilon^{-1} \boldsymbol{\delta}^* (\boldsymbol{S} + \eta \boldsymbol{I}) \boldsymbol{\delta}$.

Various optimization methods

RGN method (Webber & Lindsey, 2021)

Choose $\boldsymbol{\delta}$ to solve

$$\min_{\boldsymbol{\delta}} \left[\boldsymbol{\delta}^* \boldsymbol{g} + \boldsymbol{g}^* \boldsymbol{\delta} + \boldsymbol{\delta}^* \boldsymbol{H} \boldsymbol{\delta} + \frac{\boldsymbol{\delta}^* \boldsymbol{R} \boldsymbol{\delta}}{\epsilon} \right],$$

where $\boldsymbol{R} = \boldsymbol{S} + \eta \boldsymbol{I}$. Equivalently, set

$$\boldsymbol{\delta} = -\left(\boldsymbol{H} + \epsilon^{-1} \boldsymbol{R} \right)^{-1} \boldsymbol{g}.$$

Convergence rate analysis

Observation: GD, natural GD, and RGN can all be written as:

$$\boldsymbol{P}^i \left(\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i\right) = -\boldsymbol{g}\left(\boldsymbol{\theta}^i\right), \qquad i = 1, 2, \ldots$$

where $\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i$ is the parameter update, $-\boldsymbol{g}\left(\boldsymbol{\theta}^i\right)$ is the negative energy gradient, and $\boldsymbol{P}^i$ is the preconditioning matrix.

Convergence rate analysis

Observation: GD, natural GD, and RGN can all be written as:

$$\boldsymbol{P}^i \left(\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i\right) = -\boldsymbol{g}\left(\boldsymbol{\theta}^i\right), \qquad i = 1, 2, \ldots$$

where $\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i$ is the parameter update, $-\boldsymbol{g}\left(\boldsymbol{\theta}^i\right)$ is the negative energy gradient, and $\boldsymbol{P}^i$ is the preconditioning matrix.

Method	Preconditioner $\boldsymbol{P}$
Gradient descent	$\epsilon^{-1}\boldsymbol{I}$
Natural gradient descent	$\epsilon^{-1}\left(\boldsymbol{S} + \eta\boldsymbol{I}\right)$
Rayleigh-Gauss-Newton	$\boldsymbol{H} + \epsilon^{-1}\left(\boldsymbol{S} + \eta\boldsymbol{I}\right)$

Table: Different preconditioners for energy minimization.

Convergence rate analysis

Observation: GD, natural GD, and RGN can all be written as:

$$\boldsymbol{P}^i \left(\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i\right) = -\boldsymbol{g}\left(\boldsymbol{\theta}^i\right), \qquad i = 1, 2, \ldots$$

where $\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i$ is the parameter update, $-\boldsymbol{g}\left(\boldsymbol{\theta}^i\right)$ is the negative energy gradient, and $\boldsymbol{P}^i$ is the preconditioning matrix.

Method	Preconditioner $\boldsymbol{P}$
Gradient descent	$\epsilon^{-1}\boldsymbol{I}$
Natural gradient descent	$\epsilon^{-1}\left(\boldsymbol{S} + \eta\boldsymbol{I}\right)$
Rayleigh-Gauss-Newton	$\boldsymbol{H} + \epsilon^{-1}\left(\boldsymbol{S} + \eta\boldsymbol{I}\right)$

Table: Different preconditioners for energy minimization.

This motivates a convergence rate analysis for general parameter updates $\boldsymbol{P}^i \left(\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i\right) = -\boldsymbol{g}\left(\boldsymbol{\theta}^i\right)$

Convergence rate analysis

Proposition (convergence rate)

Consider the parameter updates $\boldsymbol{P}^i \left(\boldsymbol{\theta}^{i+1} - \boldsymbol{\theta}^i\right) = -\boldsymbol{g}\left(\boldsymbol{\theta}^i\right)$, where $(\boldsymbol{P}^1)^{-1}, (\boldsymbol{P}^2)^{-1}, \ldots$ are uniformly bounded. Assume $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots$ converges to a local energy minimizer $\boldsymbol{\theta}^*$, and the Hessian or Wirtinger Hessian is positive definite at $\boldsymbol{\theta}^*$. Then,

$$\limsup_{i\to\infty} \frac{\mathcal{E}\left[\psi_{\boldsymbol{\theta}^{i+1}}\right] - \mathcal{E}\left[\psi_{\boldsymbol{\theta}^*}\right]}{\mathcal{E}\left[\psi_{\boldsymbol{\theta}^i}\right] - \mathcal{E}\left[\psi_{\boldsymbol{\theta}^*}\right]} \leq \limsup_{i\to\infty} \left\| \boldsymbol{I} - (\boldsymbol{H}+\boldsymbol{J})^{\frac{1}{2}} \boldsymbol{P}_i^{-1} (\boldsymbol{H}+\boldsymbol{J})^{\frac{1}{2}} \right\|_2^2$$

or

$$\limsup_{i\to\infty} \frac{\mathcal{E}\left[\psi_{\boldsymbol{\theta}^{i+1}}\right] - \mathcal{E}\left[\psi_{\boldsymbol{\theta}^*}\right]}{\mathcal{E}\left[\psi_{\boldsymbol{\theta}^i}\right] - \mathcal{E}\left[\psi_{\boldsymbol{\theta}^*}\right]} \leq \limsup_{i\to\infty} \left\| \boldsymbol{I} - \begin{pmatrix} \boldsymbol{H} & \boldsymbol{J} \\ \boldsymbol{J} & \boldsymbol{H} \end{pmatrix}^{\frac{1}{2}} \begin{pmatrix} \boldsymbol{P}_i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_i \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{H} & \boldsymbol{J} \\ \boldsymbol{J} & \boldsymbol{H} \end{pmatrix}^{\frac{1}{2}} \right\|_2^2$$

in the real and complex cases, respectively, where $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{\theta}^*)$ and $\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{\theta}^*)$.

Convergence rate analysis

The convergence rate analysis ensures the following:

Convergence rate analysis

The convergence rate analysis ensures the following:

1. If RGN is applied with ϵ^i converging to infinity and $\psi_{\boldsymbol{\theta}^i}$ converging to the ground state, the energies converge *superlinearly*:

$$\limsup_{i \to \infty} \frac{\mathcal{E}\left[\psi_{\boldsymbol{\theta}^{i+1}}\right] - \mathcal{E}\left[\psi_{\boldsymbol{\theta}^*}\right]}{\mathcal{E}\left[\psi_{\boldsymbol{\theta}^i}\right] - \mathcal{E}\left[\psi_{\boldsymbol{\theta}^*}\right]} = 0.$$

Convergence rate analysis

The convergence rate analysis ensures the following:

1. If RGN is applied with ϵ^i converging to infinity and $\psi_{\boldsymbol{\theta}^i}$ converging to the ground state, the energies converge *superlinearly*:

$$\limsup_{i\to\infty} \frac{\mathcal{E}\left[\psi_{\boldsymbol{\theta}^{i+1}}\right] - \mathcal{E}\left[\psi_{\boldsymbol{\theta}^*}\right]}{\mathcal{E}\left[\psi_{\boldsymbol{\theta}^i}\right] - \mathcal{E}\left[\psi_{\boldsymbol{\theta}^*}\right]} = 0.$$

2. Other methods have a *slower linear* convergence rate, quantified by the mismatch between the preconditioner $\boldsymbol{P}^i$ and Hessian.

Empirical tests

We applied different VMC optimizers to the 10×1 TFI model, which is small enough that $\mathcal{E}$, $\boldsymbol{g}$, $\boldsymbol{S}$, and $\boldsymbol{H}$ can be computed by exact summation.

Empirical tests

We applied different VMC optimizers to the 10×1 TFI model, which is small enough that $\mathcal{E}$, $\boldsymbol{g}$, $\boldsymbol{S}$, and $\boldsymbol{H}$ can be computed by exact summation.

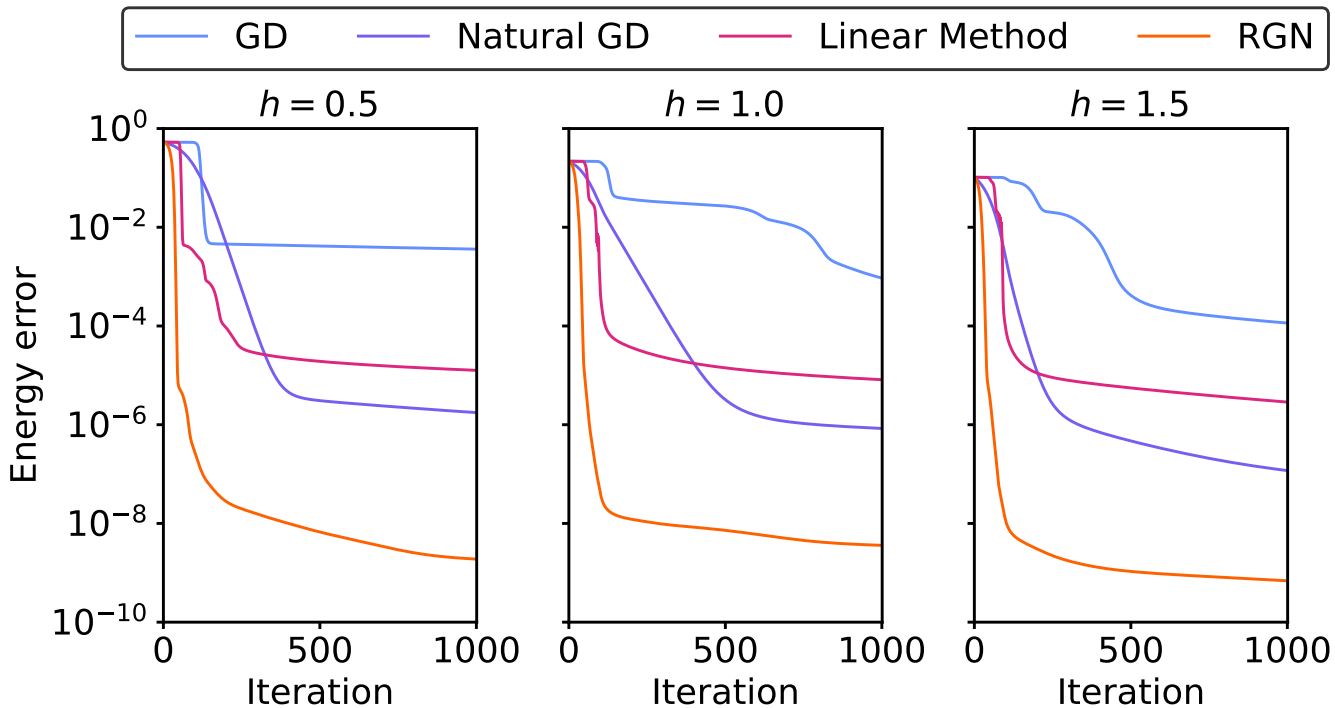


Figure: RGN energies converge quickly in ferromagnetic ($h = 0.5$, left), transitional ($h = 1.0$, center), and paramagnetic ($h = 1.5$, right) regimes.

Stochastic sampling

Question: How can we estimate $\mathcal{E}$, $\boldsymbol{g}$, $\boldsymbol{S}$, and $\boldsymbol{H}$ stochastically?

Stochastic sampling

Question: How can we estimate $\mathcal{E}$, $\boldsymbol{g}$, $\boldsymbol{S}$, and $\boldsymbol{H}$ stochastically?

---

1. Generate samples $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \ldots, \boldsymbol{\sigma}_T$ from the wavefunction density

$$\rho(\boldsymbol{\sigma}) = \frac{|\psi(\boldsymbol{\sigma})|^2}{\langle \psi, \psi \rangle}$$

using *Markov chain Monte Carlo*, i.e., propose moving from $\boldsymbol{\sigma}$ to $\boldsymbol{\sigma}'$ with probability $p(\boldsymbol{\sigma}, \boldsymbol{\sigma}')$ and accept the move with probability

$$\min\left\{ \frac{\rho(\boldsymbol{\sigma}')p(\boldsymbol{\sigma}', \boldsymbol{\sigma})}{\rho(\boldsymbol{\sigma})p(\boldsymbol{\sigma}, \boldsymbol{\sigma}')}, 1 \right\}$$

Otherwise, stay put.

Stochastic sampling

2. Approximate $\mathcal{E}$, $\boldsymbol{g}$, $\boldsymbol{S}$, and $\boldsymbol{H}$ using

$$\begin{aligned}
\hat{\mathcal{E}} &= \mathbb{E}_{\hat{\rho}}[E_L(\boldsymbol{\sigma})]\,, \\
\hat{\boldsymbol{g}}_i &= \mathrm{cov}_{\hat{\rho}}[\boldsymbol{\nu}_i(\boldsymbol{\sigma}), E_L(\boldsymbol{\sigma})]\,, \\
\hat{\boldsymbol{S}}_{ij} &= \mathrm{cov}_{\hat{\rho}}[\boldsymbol{\nu}_i(\boldsymbol{\sigma}), \boldsymbol{\nu}_j(\boldsymbol{\sigma})]\,, \\
\hat{\boldsymbol{H}}_{ij} &= \mathrm{cov}_{\hat{\rho}}[\boldsymbol{\nu}_i(\boldsymbol{\sigma}), E_{L,j}(\boldsymbol{\sigma})] - \hat{\boldsymbol{g}}_i\, \mathbb{E}_{\hat{\rho}}[\boldsymbol{\nu}_j(\boldsymbol{\sigma})] - \hat{\mathcal{E}}\hat{\boldsymbol{S}}_{ij}.
\end{aligned}$$

$\mathbb{E}_{\hat{\rho}}$ and $\mathrm{cov}_{\hat{\rho}}$ are expectations and covariances over the data, and

$$E_L(\boldsymbol{\sigma}) = \frac{\mathcal{H}\psi(\boldsymbol{\sigma})}{\psi(\boldsymbol{\sigma})}, \quad E_{L,i}(\boldsymbol{\sigma}) = \frac{\mathcal{H}\partial_{\boldsymbol{\theta}_i}\psi(\boldsymbol{\sigma})}{\psi(\boldsymbol{\sigma})}, \quad \boldsymbol{\nu}_i(\boldsymbol{\sigma}) = \frac{\partial_{\boldsymbol{\theta}_i}\psi(\boldsymbol{\sigma})}{\psi(\boldsymbol{\sigma})}.$$

Analysis of sampling

Proposition (sampling error)

Assume the MCMC sampler is geometrically ergodic with respect to ρ, and for some $\epsilon > 0$, $\mathbb{E}_\rho |E_L(\boldsymbol{\sigma})|^{4+\epsilon} < \infty$ and $\sup_i \mathbb{E}_\rho |\nu_i(\boldsymbol{\sigma})|^{4+\epsilon} < \infty$. Then, as $T \to \infty$,

$$\sqrt{T}\left(\hat{\mathcal{E}}_T - \mathcal{E}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, v^2\right), \qquad \sqrt{T}\left(\hat{\boldsymbol{g}}_T - \boldsymbol{g}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right),$$

where the asymptotic variances v^2 and $\boldsymbol{\Sigma}$ are given by

$$v^2 = \sum_{t=0}^{\infty} \mathrm{cov}_{\boldsymbol{\sigma}_0 \sim \rho}\left[E_L(\boldsymbol{\sigma}_0), E_L(\boldsymbol{\sigma}_t)\right] + \sum_{t=1}^{\infty} \mathrm{cov}_{\boldsymbol{\sigma}_0 \sim \rho}\left[E_L(\boldsymbol{\sigma}_t), E_L(\boldsymbol{\sigma}_0)\right],$$

$$\boldsymbol{\Sigma}_{ij} = \sum_{t=0}^{\infty} \mathrm{cov}_{\boldsymbol{\sigma}_0 \sim \rho}\left[\boldsymbol{g}_i'(\boldsymbol{\sigma_0}), \boldsymbol{g}_j'(\boldsymbol{\sigma}_t)\right] + \sum_{t=1}^{\infty} \mathrm{cov}_{\boldsymbol{\sigma}_0 \sim \rho}\left[\boldsymbol{g}_i'(\boldsymbol{\sigma_t}), \boldsymbol{g}_j'(\boldsymbol{\sigma}_0)\right],$$

and $\boldsymbol{g}'$ is defined as

$$\boldsymbol{g}'(\boldsymbol{\sigma}) = \overline{\left(\boldsymbol{\nu}(\boldsymbol{\sigma}) - \mathbb{E}_{\boldsymbol{\sigma}' \sim \rho}[\boldsymbol{\nu}\left(\boldsymbol{\sigma}'\right)]\right)}\left(E_L(\boldsymbol{\sigma}) - \mathcal{E}\right).$$

Analysis of sampling

The proposition has two main takeaways:

Analysis of sampling

The proposition has two main takeaways:

1. **The 'vanishing variance principle':** When ψ approaches an eigenstate, var $\hat{\mathcal{E}}_T$ approaches zero regardless of the MCMC sampler.

Analysis of sampling

The proposition has two main takeaways:

1. **The 'vanishing variance principle':** When ψ approaches an eigenstate, var $\hat{\mathcal{E}}_T$ approaches zero regardless of the MCMC sampler.
2. **Variance reduction:** To reduce the variance in $\hat{\mathcal{E}}_T$ and $\hat{\boldsymbol{g}}_T$, (a) increase the number of samples or (b) reduce time-correlations among samples.

Analysis of sampling

The proposition has two main takeaways:

1. **The 'vanishing variance principle':** When ψ approaches an eigenstate, var $\hat{\mathcal{E}}_T$ approaches zero regardless of the MCMC sampler.
2. **Variance reduction:** To reduce the variance in $\hat{\mathcal{E}}_T$ and $\hat{\boldsymbol{g}}_T$, (a) increase the number of samples or (b) reduce time-correlations among samples.

Variance reduction is essential for achieving the best VMC energies.

Empirical tests

1. We used an MCMC sampler with random proposals based on flipping a single spin.

Empirical tests

1. We used an MCMC sampler with random proposals based on flipping a single spin.
2. We combined 50 MCMC samplers per core across 48 cores and performed 4000 MCMC steps per parameter update.

Empirical tests

1. We used an MCMC sampler with random proposals based on flipping a single spin.
2. We combined 50 MCMC samplers per core across 48 cores and performed 4000 MCMC steps per parameter update.

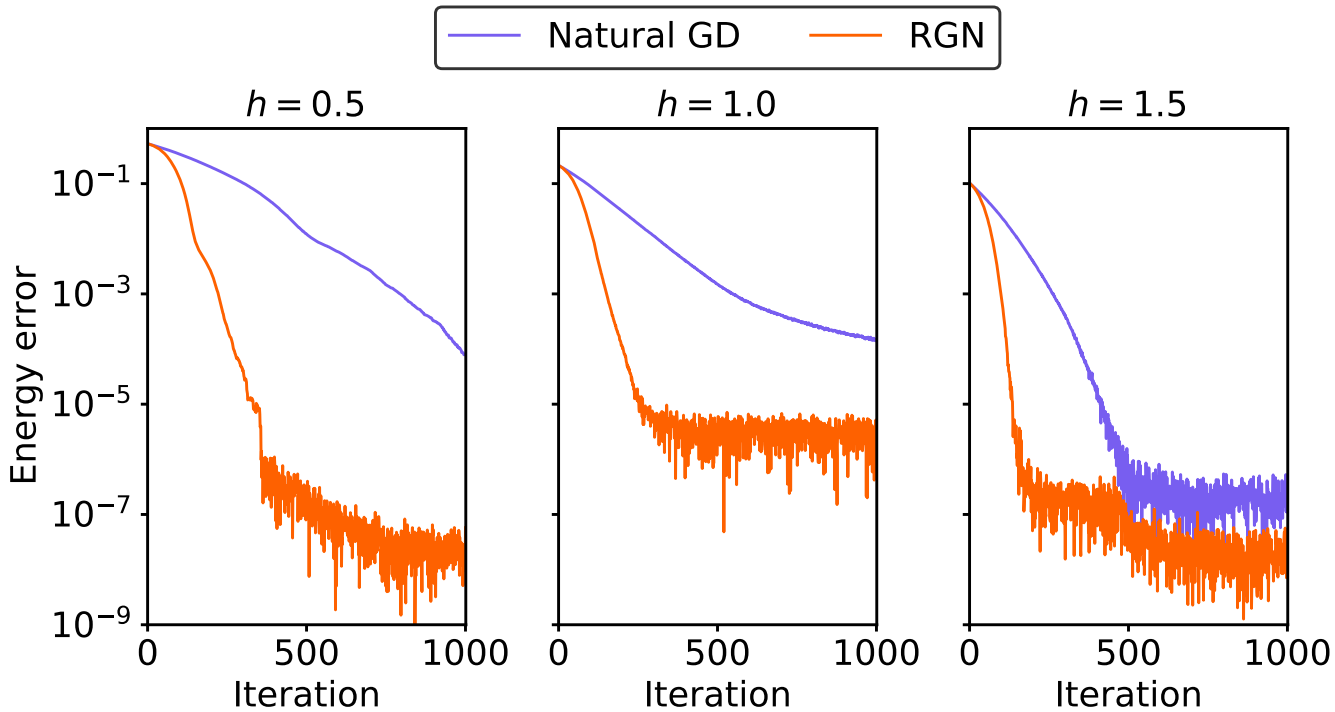


Figure: VMC ground-state energy estimates for TFI models on a 200×1 lattice.

Limitations

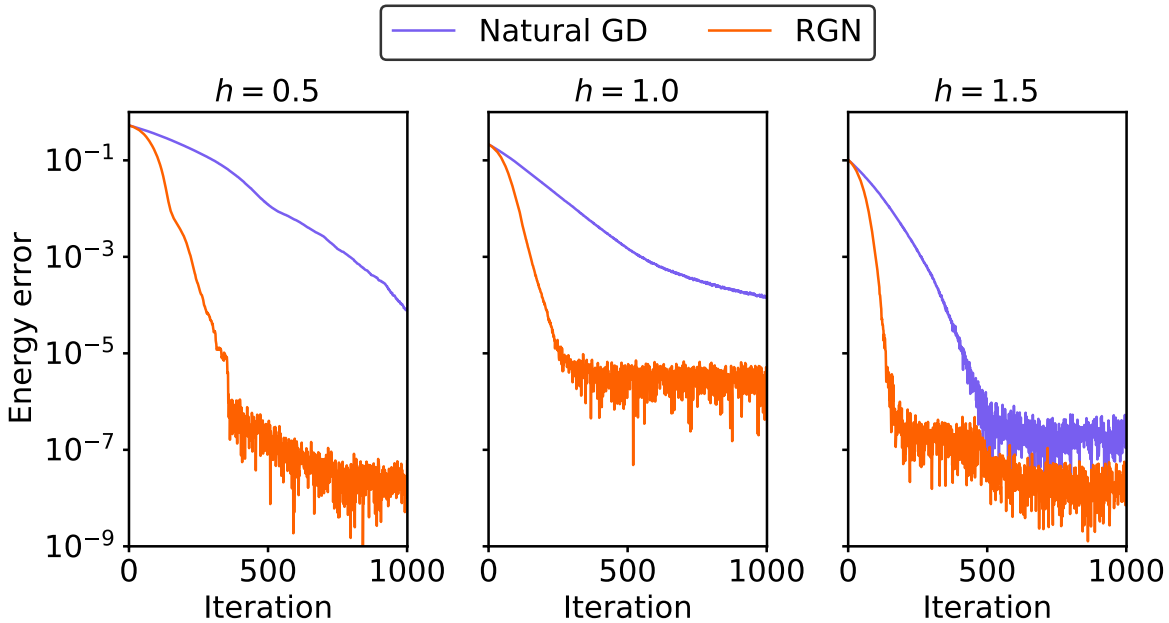


Figure: VMC ground-state energy estimates for TFI models on a 200×1 lattice.

Limitations

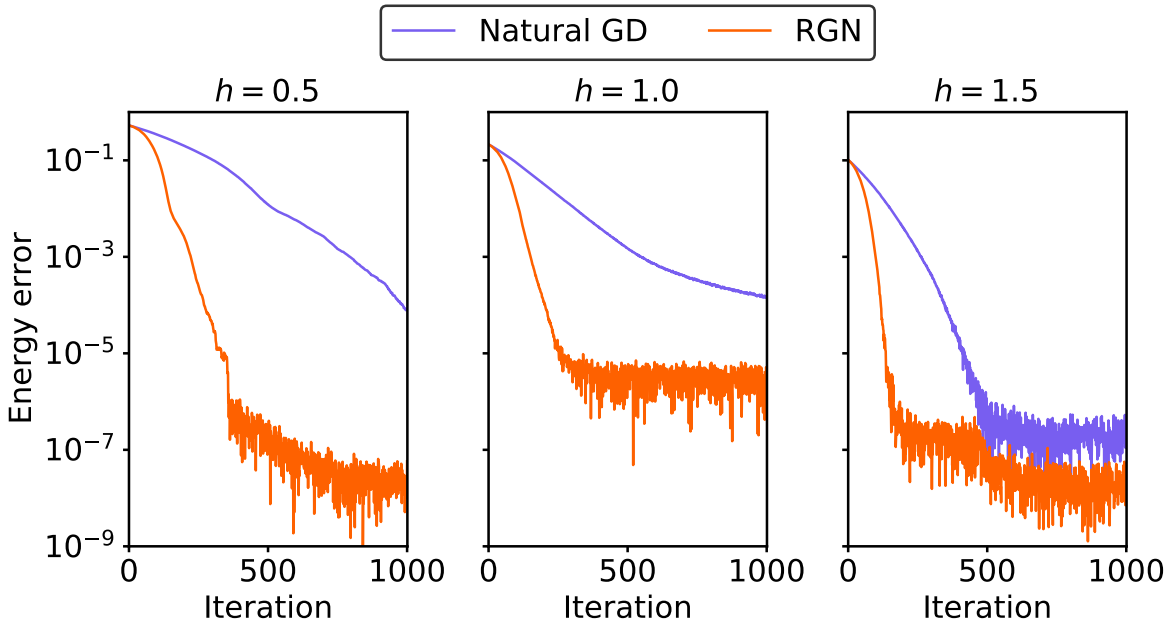


Figure: VMC ground-state energy estimates for TFI models on a 200×1 lattice.

1. In principle, VMC converges nicely using natural GD or especially RGN.

Limitations

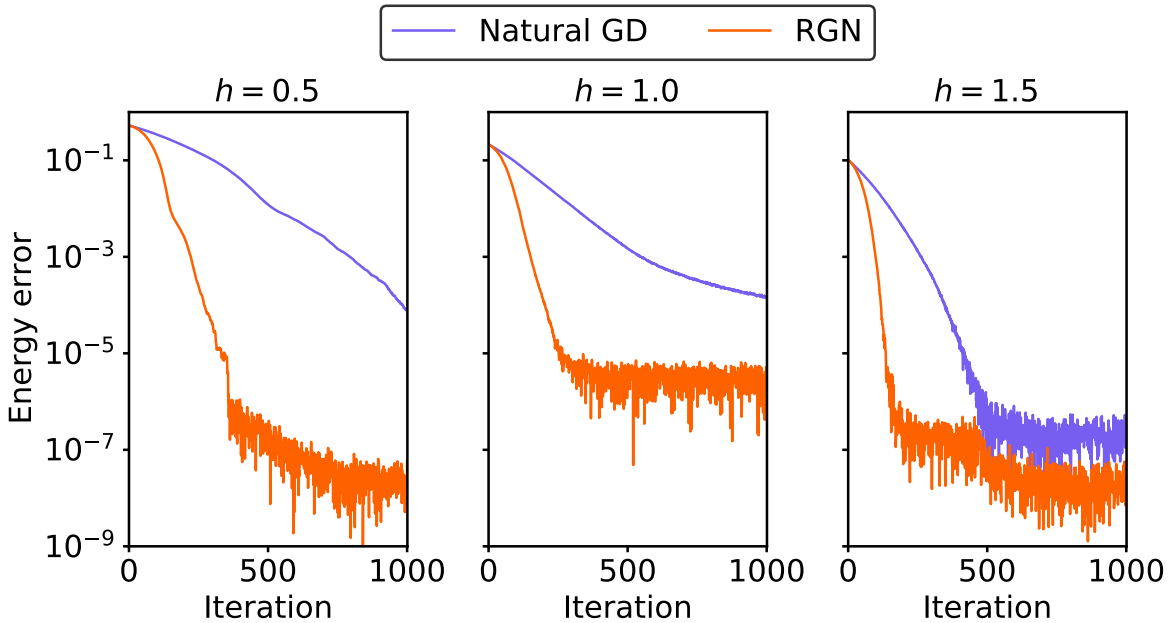


Figure: VMC ground-state energy estimates for TFI models on a 200×1 lattice.

1. In principle, VMC converges nicely using natural GD or especially RGN.
2. In practice, it is challenging to (a) choose an ansatz, (b) gather enough samples, and (c) perform the linear algebra calculations.