Optimization for variational Monte Carlo with neural quantum states

MSG Seminar: Machine Learning in Science at NYU

Michael Lindsey (CIMS)
Joint work with Robert Webber (CIMS)

April 8, 2021
Setting and idea of VMC

- Mathematically, we want to compute

$$E_0 = \min_{\psi \in \mathcal{H}} \frac{\psi^* \mathcal{H} \psi}{\psi^* \psi},$$

where $\mathcal{H}$ is a Hermitian operator, $\psi = \psi(x) = \psi(x_1, \ldots, x_N)$ is a high-dimensional function (the wavefunction).

- Parametrize $\psi = \psi_\theta$, minimize

$$E(\theta) := \frac{\psi_\theta^* \mathcal{H} \psi_\theta}{\psi_\theta^* \psi_\theta}$$

- **Assumptions:** given $\theta, x$...
  - we can query $\psi_\theta$ at $x$, i.e., evaluate $\psi_\theta(x)$, ‘efficiently’
  - we can query $\mathcal{H} \psi_\theta$ efficiently, i.e., evaluate $[\mathcal{H} \psi_\theta](x)$, ‘efficiently’
Quantum many-body problems

- **Application:** determining ground-state energy / wavefunction of quantum many-body system
- Recent neural network-based approaches
  - **Quantum spin systems:** complex RBM [Carleo and Troyer 2017]
  - **Electronic structure:** neural network backflow [Luo and Clark 2019], FermiNet [Pfau et al 2020], PauliNet [Hermann et al 2020]
- Methodology below general to *any* setting for VMC
  - But experiments will be on quantum spin systems, where $x_1, \ldots, x_N \in \{\pm 1\}$
Energy evaluation by sampling

- Expand

\[
E(\theta) = \sum_x \frac{\psi_\theta(x)|\mathcal{H}\psi_\theta(x)|}{\sum_x |\psi_\theta(x)|^2}.
\]

\[
= \frac{\sum_x |\psi_\theta(x)|^2 \psi_\theta(x)}{\sum_x |\psi_\theta(x)|^2}
\]

- Then

\[
E(\theta) = \mathbb{E}_{x \sim \rho_\theta}[E_{loc}(x; \theta)]
\]

can be evaluated by sampling, where

\[
\rho_\theta(x) = \frac{|\psi_\theta(x)|^2}{\sum_{x'} |\psi_\theta(x')|^2}, \quad E_{loc}(x; \theta) := \frac{[\mathcal{H}\psi_\theta](x)}{\psi_\theta(x)}
\]

- Estimator satisfies zero variance property: if \(\psi_\theta\) is an eigenvector, then \(E_{loc}(\cdot; \theta) \equiv E(\theta)\)
Gradient evaluation

- Can’t just autograd the formula for $E(\theta)$!
- Compute analytically, then evaluate by sampling
- Obtain (omitting some dependence on $\theta$ for clarity)

$$g_i := \frac{\partial E}{\partial \theta_i} = \frac{\psi_i^* \mathcal{H} \psi}{\psi^* \psi},$$

where $\psi_i(s) = \frac{\partial \psi}{\partial \theta_i}(s) - \frac{\langle \psi, \frac{\partial \psi}{\partial \theta_i} \rangle}{\langle \psi, \psi \rangle} \psi(s)$ and $\mathcal{H} := \mathcal{H} - E(\theta)$

- $g$ (like $E$) can be estimated by sampling from $\rho_\theta$
  - Estimator satisfies zero variance property again

- Typically in VMC, one goes beyond first-order methods via the ‘linear method’
  - Solves Rayleigh-Ritz problem on tangent space to parametric manifold $\{\psi_\theta : \theta \in \mathbb{R}^n\}$
  - Requires modification to succeed, not yet successful in practice for NN-based ansatzes
Rayleigh-Gauss-Newton approach

- **New (but related) idea:** first-order Hessian approximation
- Can compute $\nabla^2 E(\theta) = H(\theta) + J(\theta)$, where

  $$H_{ij} = \frac{\psi_i^*\overline{H}\psi_j}{\psi^*\psi}, \quad J_{ij} = \frac{\psi_{ij}^*\overline{H}\psi}{\psi^*\psi},$$

  where $\psi_{ij}(s)$ depends on second derivatives w.r.t. $\theta$

- Notice that $J = 0$ if $\psi = \psi(\theta)$ is an eigenvector
- Hence approximate $\nabla^2 E \approx H$

- Analogy to Gauss-Newton (GN) method for nonlinear least squares
  - However, the setting is different due to the Rayleigh quotient objective
  - Hence we use the term *Rayleigh-Gauss-Newton* (RGN)
Need for stabilization

- Tempted to update $\theta \leftarrow \theta - H^{-1}g$
- However, away from the optimizer, $H \approx \nabla^2 E$ is inaccurate
- Analogous to Levenberg-Marquadt approach for GN, can consider the update

\[ \theta \leftarrow \theta - (H + \varepsilon^{-1})^{-1}g \]

- Closer to gradient descent with step size $\varepsilon$ when $\varepsilon > 0$ is small
- Can increase $\varepsilon$ as we get closer to the optimizer

- However, gradient descent privileges an *unnatural* metric on parameter space...
Natural gradients

- **Background:** stochastic reconfiguration (SR), also known as quantum natural gradient descent
  - cf., natural gradient for generative models in ML
- But first...what is gradient descent?
- Observe
  \[
  \theta - \varepsilon \nabla E(\theta) = \arg\min_{\theta' \in \mathbb{R}^n} \left\{ E(\theta) + \langle \nabla E(\theta), \theta' - \theta \rangle + \frac{1}{2\varepsilon} |\theta' - \theta|^2 \right\}
  \]
- Penalty \( d(\theta, \theta')^2 = |\theta' - \theta|^2 \) is **unnatural**
- Ideally replace with
  \[
  d_{FS}(\theta, \theta') = \angle \left( \frac{\psi_{\theta'}}{||\psi_{\theta'}||}, \frac{\psi_{\theta}}{||\psi_{\theta}||} \right)
  \]
- Instead expand
  \[
  d_{FS}(\theta, \theta')^2 \approx (\theta' - \theta)^* S(\theta' - \theta),
  \]
  where \( S = S(\theta) \) is PSD, can be evaluated by sampling
- Modified update:
  \[
  \theta - \varepsilon S^{-1} \nabla E(\theta)
  \]
Natural RGN

- **Idea:** integrate natural metric into RGN framework
- Consider the update

\[ \theta \leftarrow \text{argmin}_{\theta' \in \mathbb{R}^n} \left\{ E(\theta) + \langle g, \theta' - \theta \rangle + \frac{1}{2} (\theta' - \theta)^*(H + \varepsilon^{-1}S)(\theta' - \theta) \right\} \]

- Concretely,

\[ \theta \leftarrow \theta - (H - \varepsilon^{-1}S)^{-1}g, \]

where \(H, S, g\) all evaluated by sampling

- Can take \(\varepsilon\) larger as we approach the optimizer and \(H\) approaches the true Hessian
Figure: Comparison of optimization methods with ‘brute-force’ deterministic evaluation of $H, S, g$. (10-site 1D transverse-field Ising model, complex RBM ansatz [Carleo and Troyer 2017].)
Results: stochastic evaluation

Figure: 100-site 1D transverse-field Ising model, complex RBM ansatz.
## Concluding perspectives

<table>
<thead>
<tr>
<th>VMC</th>
<th>Deep learning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Optimal value</strong></td>
<td>Matters exclusively</td>
</tr>
<tr>
<td><strong>Digits of accuracy</strong></td>
<td>Can obtain many digits, indeed often require them</td>
</tr>
</tbody>
</table>

- Areas for further exploration:
  - Importance sampling for $\rho_\theta \propto |\psi_\theta|^2$
  - Matrix-free and/or compression approaches for huge parametrizations
  - *Beyond ground state*: excited states (ongoing work with R. Webber), dynamical properties (ongoing work with H. Zhang and J. Weare), ...