Flow-based generative models for MCMC in lattice field theory

Michael S. Albergo, Gurtej Kanwar, Phiala E. Shanahan
Center for Theoretical Physics, MIT

1 [Albergo, GK, Shanahan 1904.12072]
Machine learning for lattice theories

Quantum field theories

Real-world lattices
Machine learning for lattice theories

Quantum field theories

Real-world lattices

lattice theories
Machine learning for lattice theories

- Quantum field theories
- Real-world lattices

lattice theories

numerical methods

- Spectrum
- Structure
- Thermodynamics
- ...

Page 4
Machine learning for lattice theories

- Quantum field theories
- Real-world lattices

lattice theories

numerical methods

- Spectrum
- Structure
- Thermodynamics
- ...

hard to reach continuum limit / critical point in some theories
Machine learning for lattice theories

Quantum field theories

Real-world lattices

lattice theories

numerical methods

+ ML

- Spectrum
- Structure
- Thermodynamics
- ...

hard to reach continuum limit / critical point in some theories
Computational approach to lattice theories

- Markov Chain Monte Carlo allows estimating path integrals / partition functions

approximately distributed ~ $p(\phi)$
Computational approach to lattice theories

- Markov Chain Monte Carlo allows estimating path integrals / partition functions
  - Need to wait for "burn-in period"
  - Need to take many steps before drawing independent samples

- Burn-in and correlations both related to Markov chain "autocorrelation time"
  → smaller autocorrelation time means less computational cost!

\[
\tau^\text{int}_O = \frac{1}{2} + \lim_{\tau_{\text{max}} \to \infty} \sum_{\tau=1}^{\tau_{\text{max}}} \frac{\rho_O(\tau)}{\rho_O(0)}
\]
Critical slowing down

- As parameters in the action approach criticality, for Markov chains using local updates, autocorrelation time diverges.

- Fitting $\tau^{\text{int}}$ to power law behavior gives dynamical critical exponents:

  $$\tau^{\text{int}} = \alpha_{\mathcal{O}} L^{\nu_{\mathcal{O}}}$$

- Smaller dynamical critical exponent = cheaper, closer approach to criticality.
Critical slowing down

- As parameters in the action approach criticality, for Markov chains using local updates, **autocorrelation time diverges**

- Fitting $\tau^{\text{int}}$ to power law behavior gives dynamical critical exponents

$$\tau^{\text{int}}_\mathcal{O} = \alpha_\mathcal{O} L^z_\mathcal{O}$$

- Smaller dynamical critical exponent = cheaper, closer approach to criticality

CSD also affects (naive simulation of) simpler models:
- $\mathbb{C}P^{N-1}$ [Flynn, et al. 1504.06292]
- $O(N)$ [Frick, et al. PRL 63, 2613]
- $\phi^4$ [Vierhaus doi:10.18452/14138]
- ...

Wilson loop

topological charge

continuum limit

$Q^2_5$

$W_{0.5 \text{ fm}, 0.5 \text{ fm}}$
Sampling lattice configs $\approx$ generating images

- **likely** (log prob = 22)
- **likely** (log prob = 5)
- **unlikely** (log prob = -6107)

[Karras, Lane, Aila / NVIDIA 1812.04948]
Sampling lattice configs vs. generating images

✓ Probability density can be computed for a given sample (up to normalization)

\[ p(\ldots) = \frac{e^{-S(\ldots)}}{Z} \]

✓ Physics distributions have many symmetries

✗ For lattice field theories, \(10^9\) to \(10^{12}\) variables per config

✗ Often few, e.g. \(O(1000)\), samples available (fewer than \# vars!)
  ○ Hard to use ML training paradigms that rely on existing samples from distribution
Flow-based generative models

Using a change-of-variables, produce a distribution approximating what you want.

[Rezende & Mohamed 1505.05770]
Flow-based generative models

Using a change-of-variables, produce a distribution approximating what you want.

\[
\tilde{p}_f(\phi) = \left| \det \frac{\partial f^{-1}(z)}{\partial z} \right|^{-1} r(z)
\]

Easily sampled

Approximates desired dist.

[Rezende & Mohamed 1505.05770]
Flow-based generative models

We chose real non-volume preserving (real NVP) flows for our work.

[Dinh et al. 1605.08803]
Flow-based generative models

We chose real non-volume preserving (real NVP) flows for our work.

\[ \tilde{p}_f(\phi) = \left| \det \frac{\partial f^{-1}(z)}{\partial z} \right|^{-1} r(z) \]

[Dinh et al. 1605.08803]
Training by minimizing a loss function

- Desired distribution is known up to normalization: \( p(\phi) = e^{-S(\phi)} / Z \)

- KL divergence \( D_{KL} \geq 0 \) measures "distance" between distributions

  "badness" of approximating \( Q \) by \( P \)

  "badness" of approximating \( P \) by \( Q \)

- For our application, train to minimize shifted KL divergence

\[ L(\tilde{p}_f) := D_{KL}(\tilde{p}_f||p) - \log Z \]

[Shibuya, "Demystifying KL Divergence"]

[Zhang, E, Wang 1809.10188]
Making things exact via MCMC

- Borrow idea from standard approach to lattice physics: Markov Chain Monte Carlo (MCMC)
- Use generative model for proposals in a Metropolis-Hastings step

\[ A(\phi^{(i-1)}, \phi') = \min \left( 1, \frac{\tilde{p}(\phi^{(i-1)}) p(\phi')}{p(\phi^{(i-1)}) \tilde{p}(\phi')} \right) \]

Proposal independent of previous sample
Overview of algorithm

Parameterize flow using Real NVP coupling layers

Each layer contains arbitrary neural nets
Overview of algorithm

Parameterize flow using Real NVP coupling layers

Each layer contains arbitrary neural nets

Training step

Draw samples from model

Compute loss function

Gradient descent

Desired accuracy?

Save trained model
Overview of algorithm

Parameterize flow using Real NVP coupling layers

Training step

- Draw samples from model
- Compute loss function
- Gradient descent

Markov chain using samples from model

Desired accuracy?

Save trained model

Each layer contains arbitrary neural nets

Generating samples is "embarrassingly parallel"
Toy model: scalar $\phi^4$ lattice field theory

- One real number $\phi(x) \in (-\infty, \infty)$ per lattice site $x$ (2D lattice)

- Action: relativistic scalar with quartic coupling

$$ S(\phi) = \sum_x \left( \sum_y \frac{1}{2} \phi(x) \Box(x, y) \phi(y) + \frac{1}{2} m^2 \phi(x)^2 + \lambda \phi(x)^4 \right) $$
Toy model: **scalar $\phi^4$ lattice field theory**

- One real number $\phi(x) \in (-\infty, \infty)$ per lattice site $x$ (2D lattice)

- Action: relativistic scalar with quartic coupling

$$S(\phi) = \sum_x \left( \sum_y \frac{1}{2} \phi(x) \Box(x, y) \phi(y) + \frac{1}{2} m^2 \phi(x)^2 + \lambda \phi(x)^4 \right)$$

- 5 lattice sizes $L^2 = \{6^2, 8^2, 10^2, 12^2, 14^2\}$ with bare parameters tuned to follow a line of constant physics (symmetric phase)

<table>
<thead>
<tr>
<th></th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
<th>E5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>$m^2$</td>
<td>$-4$</td>
<td>$-4$</td>
<td>$-4$</td>
<td>$-4$</td>
<td>$-4$</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>$6.975$</td>
<td>$6.008$</td>
<td>$5.550$</td>
<td>$5.276$</td>
<td>$5.113$</td>
</tr>
<tr>
<td>$m_p L$</td>
<td>$3.96(3)$</td>
<td>$3.97(5)$</td>
<td>$4.00(4)$</td>
<td>$3.96(5)$</td>
<td>$4.03(6)$</td>
</tr>
</tbody>
</table>

- **HMC** and **local Metropolis** compared against our **ML method**
Comparing observables (1)

$$G_c(x) = \frac{1}{V} \sum_y \langle \phi(y) \phi(y + x) \rangle$$

Green's functions...

correlation falls off with separation in both directions on periodic lattice
Comparing observables (2)

\[ m_p = -\partial_t \log \langle \tilde{G}_c(0, t) \rangle \]

Green's functions...

... and pole masses agree.

effective pole mass plateaus to true pole mass
Comparing observables (3)

Ising energy and two-point susceptibility agree.

\[ E = \frac{1}{d} \sum_{1 \leq \mu \leq d} G_c(\hat{\mu}) \]

nearest neighbor response grows with shrinking lattice spacing

susceptibility (total lattice response to an impulse) diverges in the continuum limit
Critical slowing down

Dynamical critical exponents

\[ z_\mathcal{O} = 1.4 \text{ to } 2.7 \]

\[ z_\mathcal{O} = 0.8 \text{ to } 1.4 \]

by spending time training up-front, autocorrelations are fixed during sampling

Dynamical critical exponents compatible with zero
Towards gauge (and other) theories

- Real NVP only directly works on fields taking real values $\phi(x) \in (-\infty, \infty)$

- What about fields taking values in compact domains (gauge theories, O(N) models, etc.)?
  - Stereographic projection coupled with standard methods may work [Gemici, Rezende, Mohammed 1611.02304]

- What about discrete models (Ising, Potts, etc.)?
  - Some recent ideas emerging [Ziegler & Rush 1901.10548]
Better choices for neural networks

- Our $\phi^4$ results use fully-connected neural networks, but Real NVP authors suggest convolutions, and hierarchical structure
  - Translational invariance, improved scaling
  - Preliminary results for $\phi^4$ indicates that this works!

- Convolutions also make scaling physical volume easy

![Images showing training results and transfer learning](image)
Towards higher dimensions

- Costs scale up, but no theoretical obstacle
- Preliminary: 3D $\phi^4$ easily accessible, (solvable) memory bottleneck for 4D

30% acc, no hyperparameter tuning required

Samples generated for $\phi^4$ theory with $V=8^3$, $m^2=-6.0$, $\lambda=14.590$
mL ~ 4, matching CSD investigation of [Vierhaus, Thesis, doi:10.18452/14138]
Machine learning for lattice theories

- Quantum field theories
- Real-world lattices

lattice theories

numerical methods

- Spectrum
- Structure
- Thermodynamics
- ...

Real-world lattices

Quantum field theories
Machine learning for lattice theories

- Quantum field theories
- Real-world lattices

lattice theories

numerical methods

- Spectrum
- Structure
- Thermodynamics
- ...

Cost

1/a
Machine learning for lattice theories

- Spectrum
- Structure
- Thermodynamics
- ...

Quantum field theories

Real-world lattices

lattice theories

numerical methods

+ ML

Cost vs. 1/a
Backup slides
Image generation via ML

1. **Likelihood free methods:** [Goodfellow et al. 1406.2661]
   E.g. Generative Adversarial Networks (GANs)
   - ✗ Needs many real samples
   - ✗ No associated likelihood for each produced sample

2. **Autoencoding:** [Kingma & Welling 1312.6114]
   E.g. Variational Auto-Encoders (VAEs)
   - ✔ Good for human interpretability
   - ✗ Same issues as GANs

3. **Flow-based:** [Rezende & Mohamed 1505.05770]
   E.g. Normalizing flows
   - ✔ Exactly known likelihood for each sample
   - ✔ Can be trained with samples from itself
Image generation via ML

1. **Likelihood free methods:**
   - E.g. Generative Adversarial Networks (GANs)
     - ✗ Needs many real samples
     - ✗ No associated likelihood for each produced sample
   
   [Goodfellow et al. 1406.2661]

2. **Autoencoding:**
   - E.g. Variational Auto-Encoders (VAEs)
     - ✔ Good for human interpretability
     - ✗ Same issues as GANs
   
   [Kingma & Welling 1312.6114]

3. **Flow-based:**
   - E.g. Normalizing flows
     - ✔ Exactly known likelihood for each sample
     - ✔ Can be trained with samples from itself
   
   [Rezende & Mohamed 1505.05770]

   [Goodfellow et al. 1406.2661]

   [Kingma & Welling 1312.6114]

   [Shen & Liu 1612.05363]
Real NVP coupling layer

- Affine transformation of half the variables: scaling by exp(s), translation by t
- s and t are neural networks depending on untransformed variables only
- Simple inverse and Jacobian
Desired distribution is known up to normalization:

\[ p(\phi) = e^{-S(\phi)}/Z \]

For our application, train to minimize shifted KL divergence

\[
L(\tilde{p}_f) := D_{KL}(\tilde{p}_f||p) - \log Z \\
= \int \prod \limits_j d\phi_j \tilde{p}_f(\phi) \left( \log \tilde{p}_f(\phi) + S(\phi) \right)
\]

This loss allows self-training: sampling with respect to model distribution \( \tilde{p}_f(\phi) \) to estimate loss.
ML model for scalar lattice field theory

- Prior distribution chosen to be uncorrelated Gaussian, i.e. for each site $x$,
  \[
  \phi(x) \sim \mathcal{N}(0, 1)
  \]

- Real NVP model:
  - 8-12 Real NVP coupling layers
  - Alternating checkerboard pattern for variable split
  - 2-6 fully connected layers with 100-1024 hidden units

- Trained using shifted KL loss with Adam optimizer
  - Target fixed acceptance rate in Metropolis-Hastings MCMC
Samples from ML model vs standard algorithms

<table>
<thead>
<tr>
<th>Method</th>
<th>$10^2$</th>
<th>$12^2$</th>
<th>$14^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML 50%</td>
<td><img src="image1.png" alt="Images" /></td>
<td><img src="image2.png" alt="Images" /></td>
<td><img src="image3.png" alt="Images" /></td>
</tr>
<tr>
<td>Local</td>
<td><img src="image4.png" alt="Images" /></td>
<td><img src="image5.png" alt="Images" /></td>
<td><img src="image6.png" alt="Images" /></td>
</tr>
<tr>
<td>HMC</td>
<td><img src="image7.png" alt="Images" /></td>
<td><img src="image8.png" alt="Images" /></td>
<td><img src="image9.png" alt="Images" /></td>
</tr>
</tbody>
</table>

By eye, ML model produces varied samples and correlations at the right scale.
Physical limits of scalar $\phi^4$ lattice field theory

$$S(\phi) = \sum_x \left( \sum_y \frac{1}{2} \phi(x) \Delta(x,y) \phi(y) + \frac{1}{2} m^2 \phi(x)^2 + \lambda \phi(x)^4 \right)$$

- Ising: $\lambda \to \infty$, $m^2 / \lambda < 0$
- Gaussian: $\lambda \to 0$