On the Anatomy of MCMC-Based Maximum Likelihood Learning of Energy-Based Models

Erik Nijkamp*, Mitch Hill*, Tian Han, Song-Chun Zhu, and Ying Nian Wu

Motivation

The Energy-Based Model is a flexible and powerful tool for representing emergent behavior in complex systems. Recent works investigate image modeling using the Gibbs potential

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp\{-U(x;\theta)\}$$

where $U(x;\theta)$ is ConvNet with weights θ and scalar output. These works achieve realistic image synthesis with short-run MCMC sampling, but long-run MCMC samples are oversaturated and unrealistic, indicating the models do not correctly assign probability mass in the image space.

[5] [3] [1]

Contributions

- Identification of two axes which characterize each ML parameter update: 1) energy difference of positive and negative samples, and 2) MCMC convergence or non-convergence
- The first ConvNet potentials with realistic steady-state samples
- The first ConvNet potentials trained using ML with purely noise-initialized MCMC, yielding a non-invertible flow capable of realistic and diverse generation from noise
- Mapping the macroscopic structure of image potentials for unsupervised clustering

Maximum Likelihood (ML)

In the unsupervised ML framework, one learns a parameter θ such that $p_\theta(x)$ approximates the data distribution $q(x)$ by minimizing the KL-Divergence between q and p_θ:

$$D_{KL}(q \| p_\theta) = \log Z(\theta) + E_q[U(X;\theta)].$$

Using $\nabla \log Z(\theta) = -E_{p_\theta}[\nabla_\theta U(X;\theta)]$, the gradient for learning θ is:

$$\nabla \mathcal{L}(\theta) = \nabla E_q[U(X;\theta)] - E_{p_\theta}[\nabla_\theta U(X;\theta)]$$

$$\approx \nabla_\theta \left(\frac{1}{n}\sum_{i=1}^{n} U(X_i^+;\theta) - \frac{1}{m}\sum_{i=1}^{m} U(X_i^-;\theta) \right)$$

where the positive samples $\{X_i^+\}_{i=1}^n$ are i.i.d. from the data density q, and negative samples $\{X_i^-\}_{i=1}^m$ are i.i.d. from p_θ. In practice, $\{X_i^+\}_{i=1}^n$ are a batch of training images and $\{X_i^-\}_{i=1}^m$ are obtained after L steps of Langevin sampling:

$$X_{\ell+1} = X_\ell - \frac{\varepsilon^2}{2}\nabla_X U(X_\ell;\theta) + \varepsilon Z_\ell, \; Z_\ell \sim \mathrm{N}(0, I).$$

MCMC samples can be initialized from the data samples, persistent chains, or noise.

Diagnostics of Maximum Likelihood Learning

The practical behavior of ML learning is governed by the short-run density $s_t(x)$ of samples X_i^- given the finite-step MCMC sampler used at training step t, and the computational loss

$$d_{s_t}(\theta) = E_q[U(X;\theta)] - E_{s_t}[U(X;\theta)].$$

Two outcomes occur for each update on the parameter path $\{\theta_t\}_{t=1}^{T+1}$:

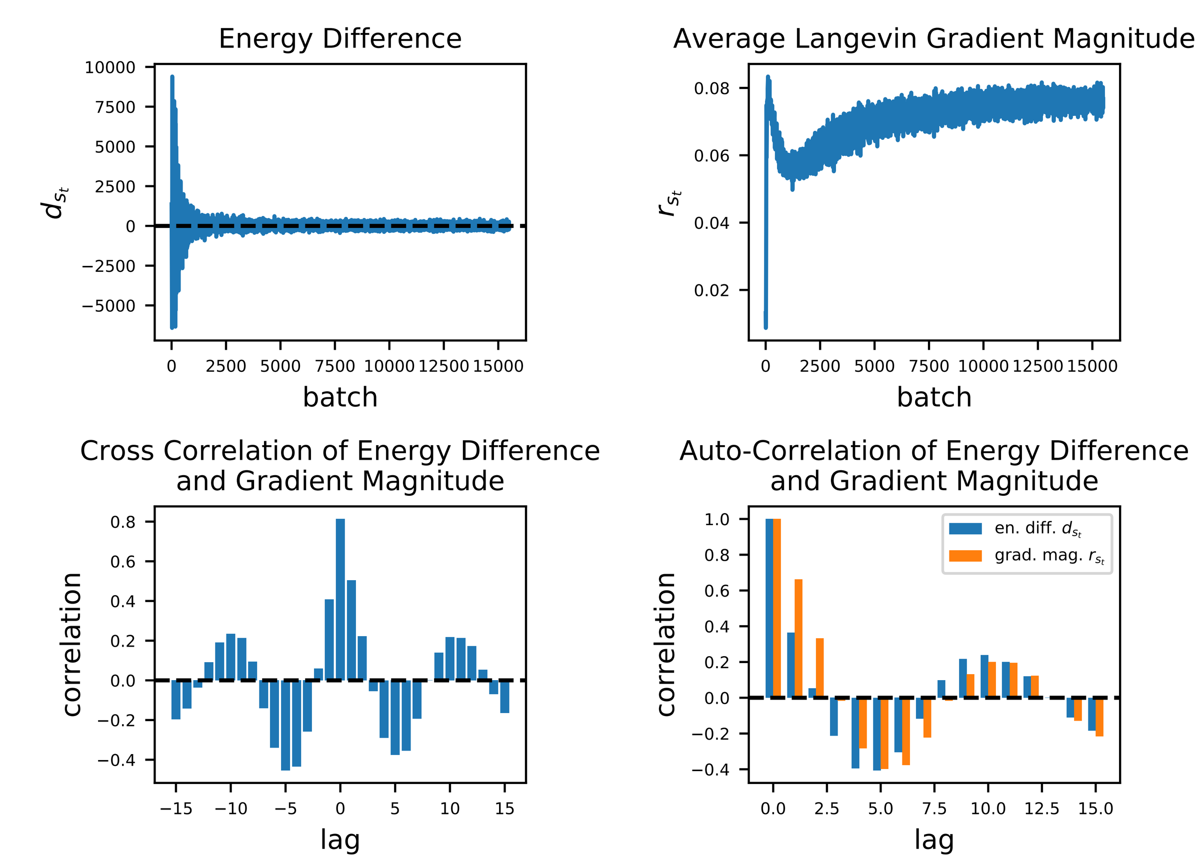
1. $d_{s_t}(\theta_t) < 0$ or $d_{s_t}(\theta_t) > 0$
2. $s_t \approx p_{\theta_t}$ (MCMC convergence) or $s_t \not\approx p_{\theta_t}$ (MCMC non-convergence) .

For stable convergent and non-convergent learning, $\{d_{s_t}(\theta_t)\}_{t=1}^{T+1}$ is distributed symmetrically around 0. The average Langevin gradient

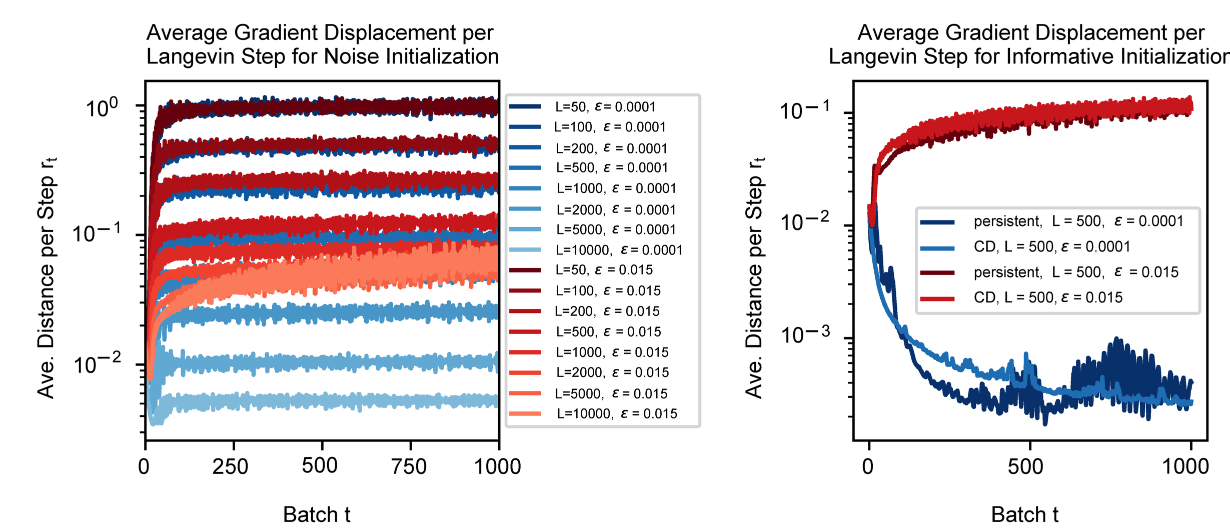
$$r_t = \frac{\varepsilon^2}{2} E_{w_t}\left[\frac{1}{L+1}\sum_{\ell=0}^{L} \left\| \nabla_Y U(Y_t^{(\ell)};\theta_t) \right\|_2 \right],$$

where w_t is the joint density a Langevin chain, converges to a constant value. For convergent learning r_t balances with ε, while for non-convergent learning r_t balances with the average distance between the samples from the initial MCMC distribution and samples from q.

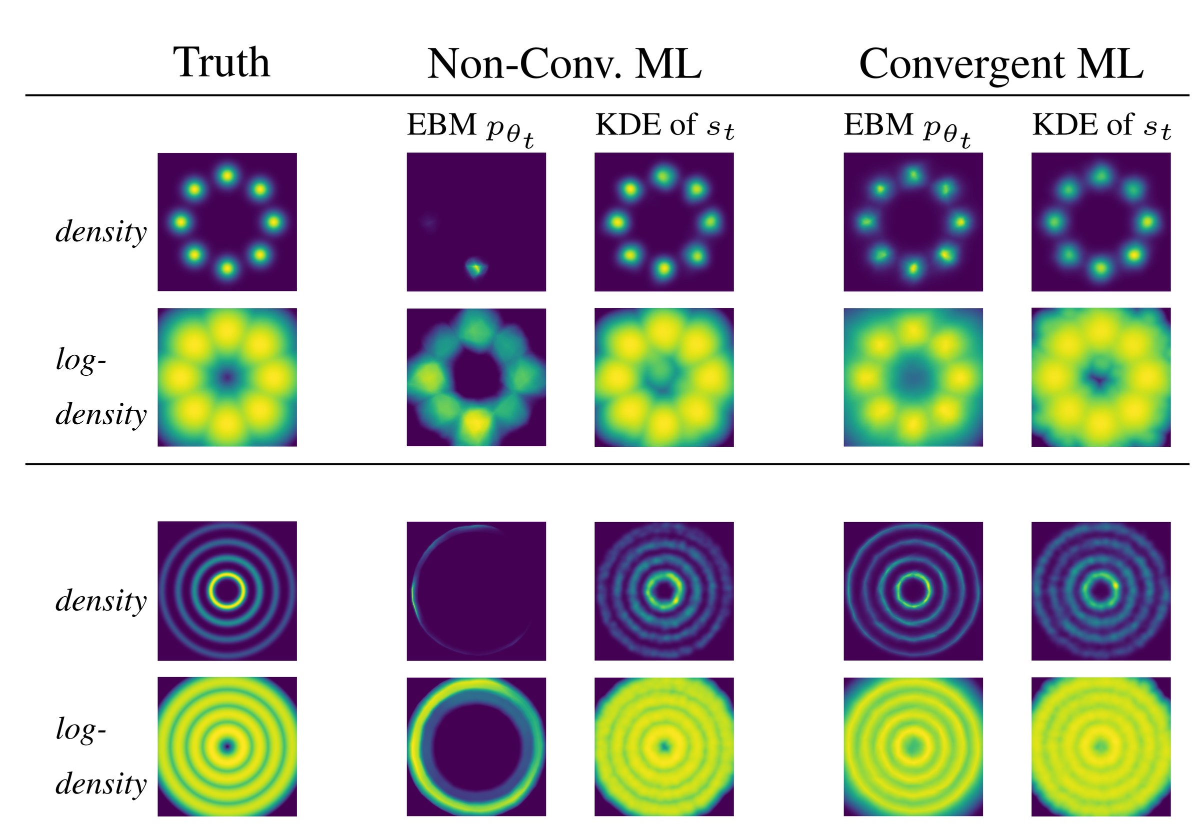
Diagnostics for Single Implementation



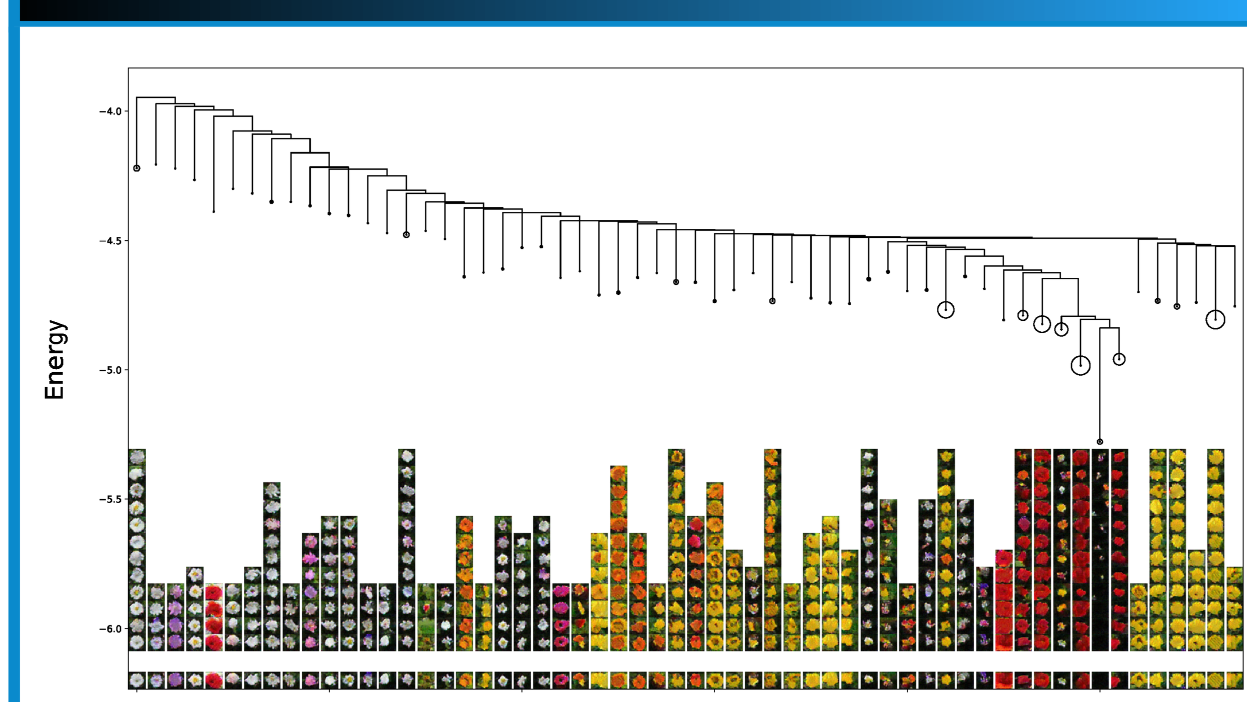
Langevin Gradient Across Implementations



2D Toy Experiments



Energy Landscape Mapping



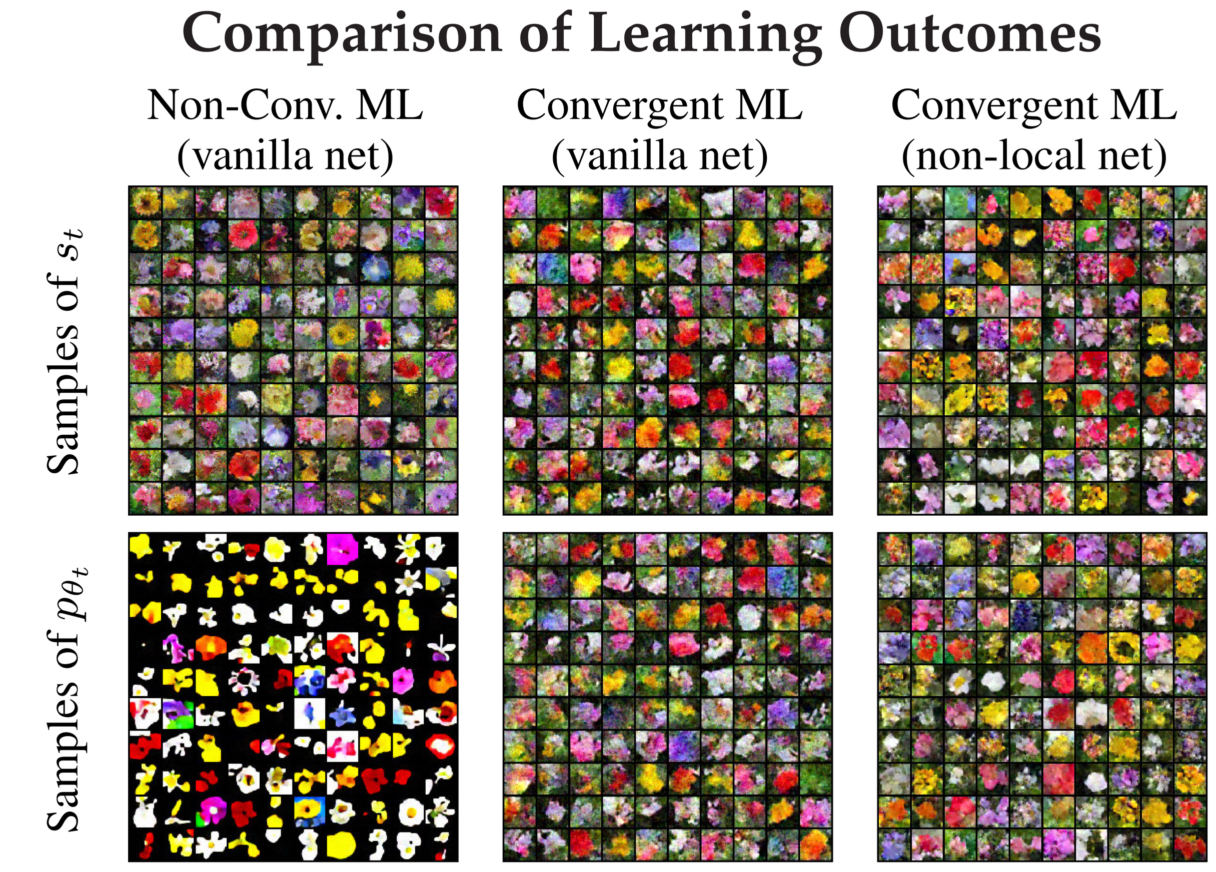
Metastable structures discovered via magnetized diffusion on a flower image potential. See [2].

Code

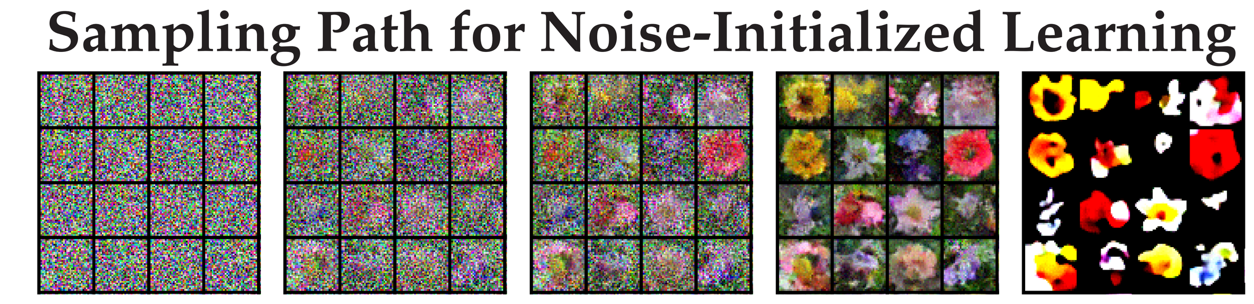
```
https://github.com/point0bar1/ebm-anatomy
```

Image Experiments

Comparison of Learning Outcomes



Sampling Path for Noise-Initialized Learning



0 25 50 100 500

MCMC Steps

References

[1] Y. Du and I. Mordatch. Implicit Generation and Generalization in Energy-Based Models. In *NeurIPS '19*

[2] M. Hill, E. Nijkamp, S.C. Zhu. Building a Telescope to Look Into High-Dimensional Image Spaces. In *QAM '19*

[3] K. Lee, W. Xu, F. Fan, Z. Tu. Wasserstein Introspective Neural Networks. In *CVPR '18*

[4] E. Nijkamp, M. Hill, S.C. Zhu, Y.N. Wu. Learning Non-Convergent Non-Persistent Short-Run MCMC Toward Energy-Based Model. In *NeurIPS '19*

[5] J. Xie, Y. Lu, S.C. Zhu, Y.N. Wu. A Theory of Generative ConvNet. In *ICML '16*