Generalizing Hamiltonian Monte Carlo with Neural Networks

Daniel Levy[1], Matthew D. Hoffman[2], Jascha Sohl-Dickstein[3]
[1]Stanford University [2]Google Inc. [3]Google Brain

Motivation

- Sampling from analytically described distributions is ubiquitous in Machine Learning and many fields.
- Markov Chain Monte Carlo (MCMC) Methods promise a solution to this problem.
- However, they often need carefully handcrafted proposals or suffer from poor convergence and mixing.
- Neural networks have had great success at modeling highly-complex distributions.
- **Can we train MCMC kernels, parameterized by deep neural networks, that converge and mix quickly to their target distribution?**

Background

- **MCMC METHODS:** construct a sequence of correlated samples that converges in distribution to the target.

$$X_0 \sim \pi_0 \quad X_{t+1} \sim K(\cdot|X_t)$$

- If K is irreducible, aperiodic and admits p the target as a fixed point, then X_t converges in distribution to p.
- Last requirement is enforced through a *Metropolis-Hastings (MH) accept/reject step.*
- Issues: strong asymptotic guarantees but trade-off between low acceptance and slow exploration.
- **HAMILTONIAN MONTE CARLO**: for continuous distributions. *$p(x) = 1/Z \exp(U(x))$*. Extends the space with a momentum variable *v*. Proposes a new state by integrating with conservative dynamics (*U* potential energy, *$1/2\ v^T v$* kinetic energy).
- Asymptotic guarantees remain through an MH step.
- Leapfrog discretization limits integration error.

$$v^{\frac{1}{2}} = v - \frac{\epsilon}{2}\partial_x U(x); \quad x' = x + \epsilon v^{\frac{1}{2}}; \quad v' = v - \frac{\epsilon}{2}\partial_x U(x')$$

- **Still fails in a number of simple settings:**
  - Can take arbitrarily long to traverse ill-conditioned Gaussians
  - Cannot traverse low density zones
  - Cannot mix between energy levels.

Diagnostic Distributions

- **ILL-CONDITIONED GAUSSIAN:** Gaussian with diagonal covariance log-spaced between 10^{-2} and 10^2.

Can L2HMC learn a Diagonal Inertia Tensor?

- **STRONGLY-CORRELATED GAUSSIAN:** Rotated ICG.

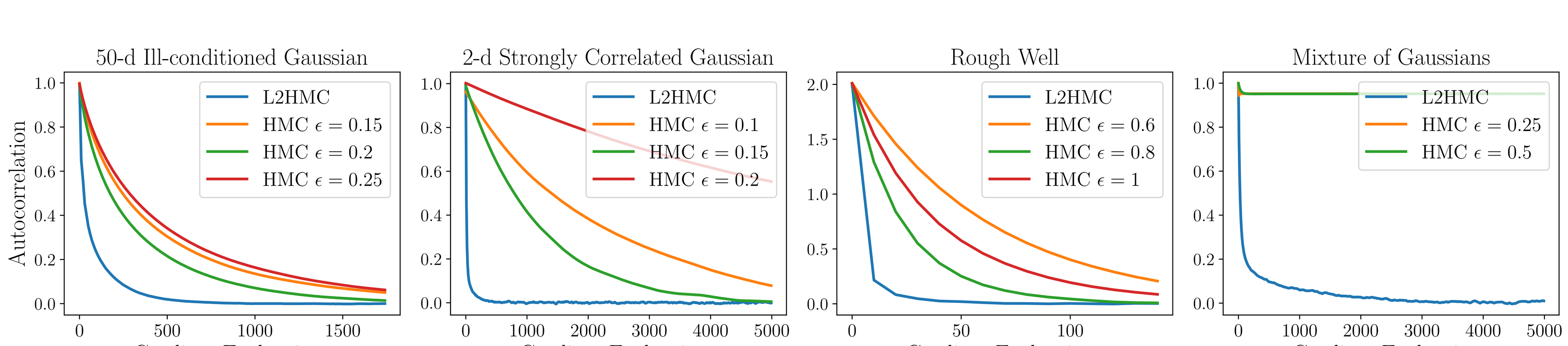
Can L2HMC approximate non-diagonal inertia tensors?

- **MIXTURE OF GAUSSIAN**: Mixture of isotropic Gaussians with variance $\sigma^2 = 0.1$ separated by 10 standard devs.

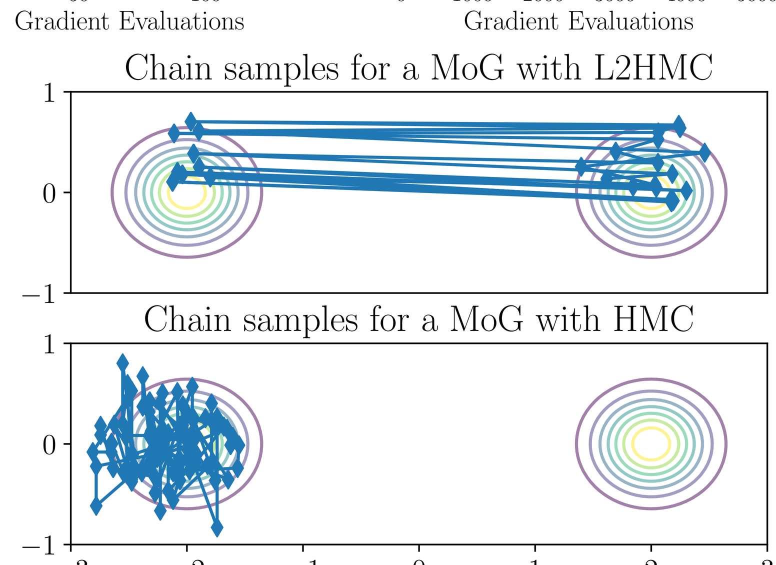
Can L2HMC learn within trajectory tempering?

- **ROUGH WELL:** Isotropic Gaussian, with low amplitude but high frequency noise.

Can L2HMC partially ignore the energy?



Distribution	ESS-HMC	ESS-L2HMC	Ratio
50-d ICG	7.30×10^{-2}	1.65×10^{-2}	4.4
2-d Rough Well	6.25×10^{-1}	1.16×10^{-1}	5.4
2-d SCG	2.32×10^{-1}	4.69×10^{-3}	49.5
2-d MoG	3.24×10^{-2}	$<< 2.61 \times 10^{-4}$	>> 124



Augmenting HMC

- To conserve exact sampling guarantees, the operator has to remain invertible, with a tractable determinant of its Jacobian.
- In HMC, this is enabled via *shear transformations* (each update updates linearly a subset of the variables by an amount determined by the complementary subset).
- **Can we define expressive operators with triangular Jacobian (shear composed with scaling)?**

$$\begin{aligned}
v' &= v \odot \exp(\frac{\epsilon}{2} S_v(\zeta_1)) - \frac{\epsilon}{2}(\partial_x U(x) \odot \exp(\epsilon Q_v(\zeta_1)) + T_v(\zeta_1)) \\
x' &= x_{\bar{m}^t} + m^t \odot [x \odot \exp(\epsilon S_x(\zeta_2)) + \epsilon(v' \odot \exp(\epsilon Q_x(\zeta_2)) + T_x(\zeta_2))] \\
x'' &= x'_{m^t} + \bar{m}^t \odot [x' \odot \exp(\epsilon S_x(\zeta_3)) + \epsilon(v' \odot \exp(\epsilon Q_x(\zeta_3)) + T_x(\zeta_3))] \\
v'' &= v' \odot \exp(\frac{\epsilon}{2} S_v(\zeta_4)) - \frac{\epsilon}{2}(\partial_x U(x'') \odot \exp(\epsilon Q_v(\zeta_4)) + T_v(\zeta_4))
\end{aligned}$$

$$\zeta_1 = (x, \partial_x U(x), t) \quad \zeta_2 = (x_{\bar{m}^t}, v, t) \quad \zeta_3 = (x'_{m^t}, v, t)\zeta_4 = (x'', \partial_x U(x''), t)$$

- This is a generalization of HMC as it is **non-volume preserving**, with **learnable parameters,** and **reduces to HMC** for Q=S=T=0.

Loss and Training Procedure

- Need a criterion to optimize Q, S and T.
- **Proxy for mixing**: variation on the Expected Square Jump Distance, the *reciprocal loss,* to encourage mixing across the entire state space.
- **Notations**: *$\xi = (x, v, d)$* initial state, *$\xi' = (x', v', d')$* proposed state, *$A(\xi'|\xi)$* acceptance probability, *q* initial distribution over the extended state space.

$$\ell_\lambda(\xi, \xi', A(\xi'|\xi)) = \frac{\lambda^2}{||x - x'||_2^2 A(\xi'|\xi)} - \frac{||x - x'||_2^2 A(\xi'|\xi)}{\lambda^2}$$

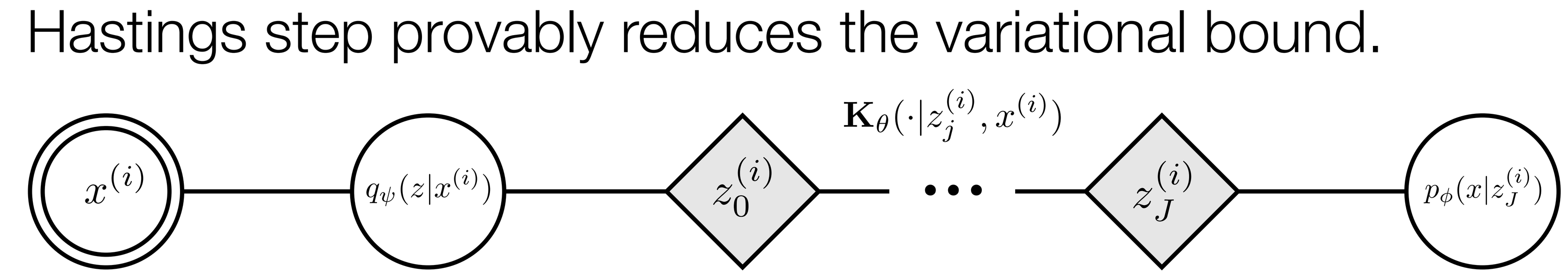
- **Loss:**

$$\mathcal{L}(\theta) = \mathbb{E}_{p(\xi)}\left[\ell_\lambda(\xi, \xi', A(\xi'|\xi))\right] + \mathbb{E}_{q(\xi)}\left[\ell_\lambda(\xi, \xi', A(\xi'|\xi))\right]$$

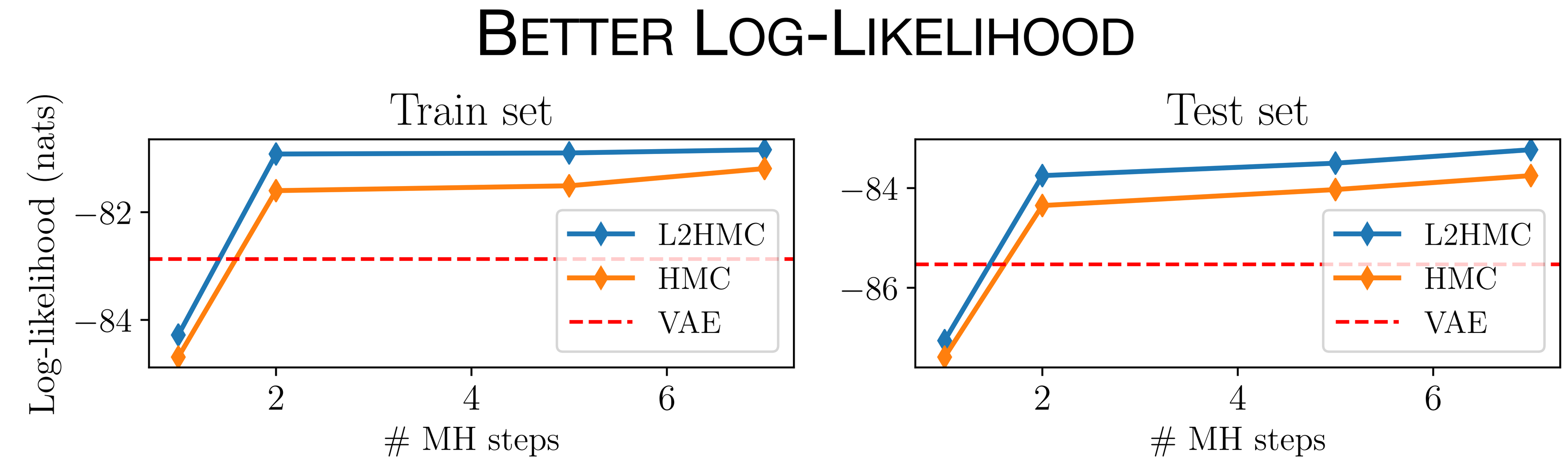
- Can be efficiently batched and requires only one forward/backward of the networks.

Exact Training of Generative Models

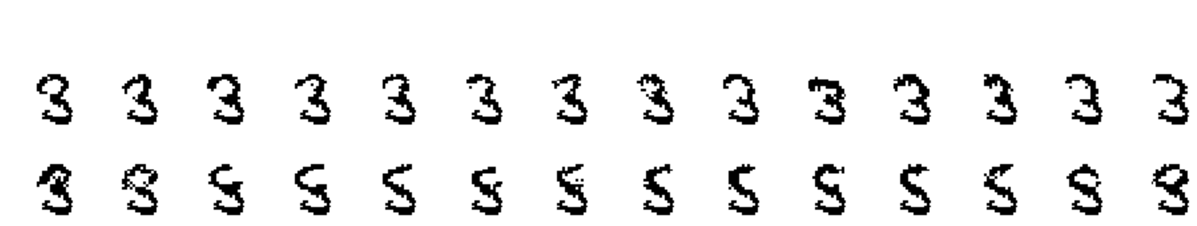
- **VAE**: training of latent variable generative models using a (variational) approximate posterior.
- **L2HMC-DGLM**: train a parametric sampler to perform efficient posterior sampling to obtain "more exact" posterior samples.
- Can start from approximate posterior — each Metropolis-Hastings step provably reduces the variational bound.



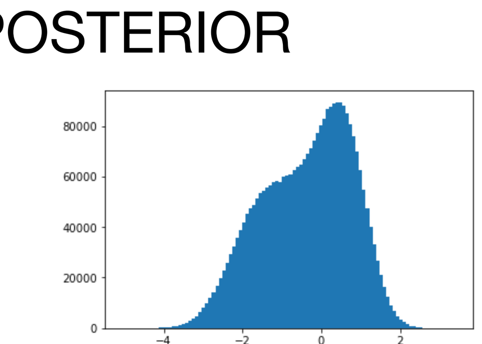
BETTER LOG-LIKELIHOOD



ENABLES MORE EXPRESSIVE POSTERIOR



Block-Gibbs inpainting of the top-half of a digit.
Top: using q_ψ, unimodal. **Bottom:** using L2HMC, multi-modal.



Projected posterior samples (PCA) exhibits Non-gaussianity