Langevin Autoencoders for Learning Deep Latent Variable Models

Shohei Taniguchi, Yusuke Iwasawa, Wataru Kumagai, Yutaka Matsuo

東京大学
THE UNIVERSITY OF TOKYO

Deep Latent Variable Models

Definition

$\mathbf{x}$: observation

$\mathbf{z}$: latent variable

$\boldsymbol{\theta}$: model parameter

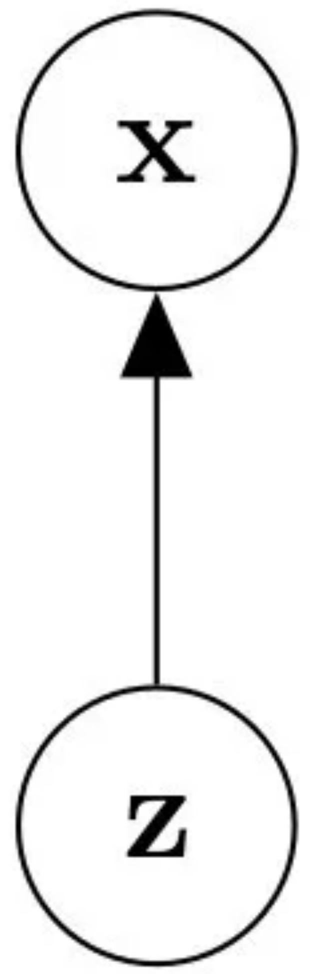
$$p(\mathbf{x};\boldsymbol{\theta}) = \int p(\mathbf{x} \mid z;\boldsymbol{\theta})p(z)dz$$

- $p(\mathbf{x} \mid z;\boldsymbol{\theta})$ is typically constructed using a deep neural network $f_{\mathbf{x}|\mathbf{z}}$

e.g.,

$$p\left(\mathbf{z}\right) = \mathcal{N}\left(\mathbf{z};\mathbf{0},I\right)$$

$$p\left(\mathbf{x} \mid z;\boldsymbol{\theta}\right) = \mathcal{N}\left(\mathbf{x};f_{\mathbf{x}|\mathbf{z}}\left(z;\boldsymbol{\theta}\right),\sigma^2 I\right)$$



Deep Latent Variable Models

Training via Maximum Likelihood

$\mathbf{x}$: observation

$\mathbf{z}$: latent variable

θ : model parameter

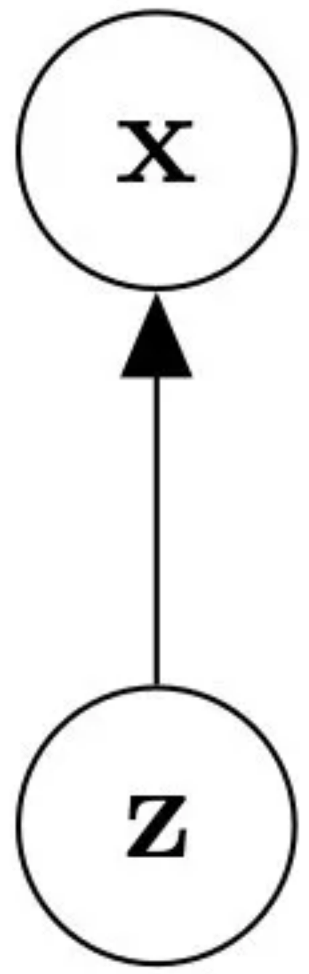
$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\hat{p}_{\text{data}}(\boldsymbol{x})}[\log p(\boldsymbol{x};\boldsymbol{\theta})] \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{p\left(\boldsymbol{z}^{(i)} \mid \boldsymbol{x}^{(i)};\boldsymbol{\theta}\right)} \left[\nabla_{\boldsymbol{\theta}} \log p\left(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}; \boldsymbol{\theta}\right) \right]$$

- $\hat{p}_{\text{data}}$: Empirical distribution defined by training set

  n: Minibatch size

- Due to the intractability of the posterior $p\left(\boldsymbol{z} \mid \boldsymbol{x};\boldsymbol{\theta}\right)$,

  approximation is needed



Langevin Dynamics

$\mathbf{x}$: **observation**
$\mathbf{z}$: **latent variable**
$\boldsymbol{\theta}$: **model parameter**

- Langevin dynamics (LD) is an MCMC based on the following Langevin equation

$$dz = -\nabla_z U\left(\boldsymbol{x}, z; \boldsymbol{\theta}\right) dt + \sqrt{2} dB$$

$$U\left(\boldsymbol{x}, z; \boldsymbol{\theta}\right) = -\log p\left(\boldsymbol{x}, z; \boldsymbol{\theta}\right)$$

- This stochastic differential equation has the posterior as a stationary distribution

Langevin Dynamics

$\mathbf{x}$: **observation**

$\mathbf{z}$: **latent variable**

$\boldsymbol{\theta}$: **model parameter**

η : **stepsize**

- By simulating the dynamics, samples asymptotically approach to the posterior

$$\text{for } t = 1, \ldots, T$$

$$\boldsymbol{z}_{t+1} \sim \mathcal{N}\left(\boldsymbol{z}_{t+1}; \boldsymbol{z}_t - \eta \nabla_{\boldsymbol{z}_t} U\left(\boldsymbol{x}, \boldsymbol{z}_t; \boldsymbol{\theta}\right), 2\eta \boldsymbol{I}\right)$$

- Optionally, Metropolis-Hastings rejection steps can be added for calibrating discretization error

Langevin Dynamics

Pros

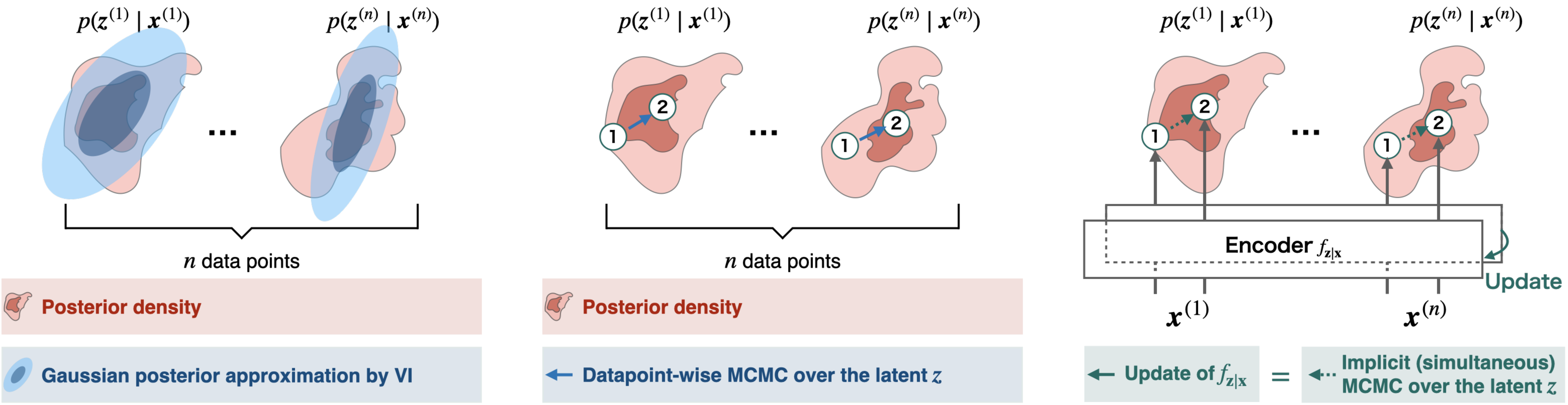
- Samples are asymptotically unbiased
  - High approximation power

Cons

- Datapoint-wise sampling is costly
  - Need to run MCMC independently for all minibatch data $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$

Method

Amortized Langevin Dynamics
Overview



Variational Inference

Langevin Dynamics

Amortized Langevin Dynamics (ours)

MCMC on the latent space is replaced by MCMC on the encoder's parameter space

Amortized Langevin Dynamics
Formulation

$\mathbf{x}$: observation
$\mathbf{z}$: latent variable
$\boldsymbol{\theta}$: model parameter
$f_{\mathbf{z}|\mathbf{x}}$: encoder

- Consider an SDE on the parameter of an encoder $f_{\mathbf{z}|\mathbf{x}}$ that maps $\mathbf{x}$ into $\mathbf{z}$

$$d\boldsymbol{\phi} = -\nabla_{\boldsymbol{\phi}} V\left(\boldsymbol{\phi}\right) dt + \sqrt{2} dB$$

$$V\left(\boldsymbol{\phi}\right) = \sum_{i=1}^{n} U\left(\boldsymbol{x}^{(i)}, f_{\mathbf{z}|\mathbf{x}}\left(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}\right); \boldsymbol{\theta}\right)$$

Amortized Langevin Dynamics

Algorithm

$\mathbf{x}$: observation

$\mathbf{z}$: latent variable

$\boldsymbol{\theta}$: model parameter

$f_{\mathbf{z}|\mathbf{x}}$: encoder

- By simulating the SDE, samples of $\boldsymbol{\phi}$ are collected
- MCMC over the latents is implicitly performed by collecting the mapping by $f_{\mathbf{z}|\mathbf{x}}$

for $t = 1, \ldots, T$

$$\boldsymbol{\phi}_{t+1} \sim \mathcal{N}\left(\boldsymbol{\phi}_{t+1}; \boldsymbol{\phi}_t - \eta \nabla_{\boldsymbol{\phi}} V\left(\boldsymbol{\phi}\right), 2\eta \boldsymbol{I}\right)$$

for $i = 1, \ldots, n$

$$z_{t+1}^{(i)} = f_{\mathbf{z}|\mathbf{x}}\left(\boldsymbol{x}^{(i)}; \boldsymbol{\phi}_{t+1}\right)$$

Amortized Langevin Dynamics

Theoretical Analysis

$\mathbf{x}$: observation

$\mathbf{z}$: latent variable

$\boldsymbol{\theta}$: model parameter

$f_{\mathbf{z}|\mathbf{x}}$: encoder

- When the following conditions are satisfied, ALD has the true posterior as a stationary distribution

  1. Encoder takes the form of $f_{\mathbf{z}|\mathbf{x}}(\boldsymbol{x}; \boldsymbol{\Phi}) = \boldsymbol{\Phi} g(\boldsymbol{x})$

  2. Rank of $\boldsymbol{G}$ is n, where $\boldsymbol{G}$ is a matrix with $g\left(\boldsymbol{x}^{(i)}\right)$ in row $\boldsymbol{G}_{i,:}$

- **ALD is valid as MCMC under mild assumptions**

Amortized Langevin Dynamics

Remarks

$\mathbf{x}$: observation
$\mathbf{z}$: latent variable
$\boldsymbol{\theta}$: model parameter
$f_{\mathbf{z}|\mathbf{x}}$: encoder

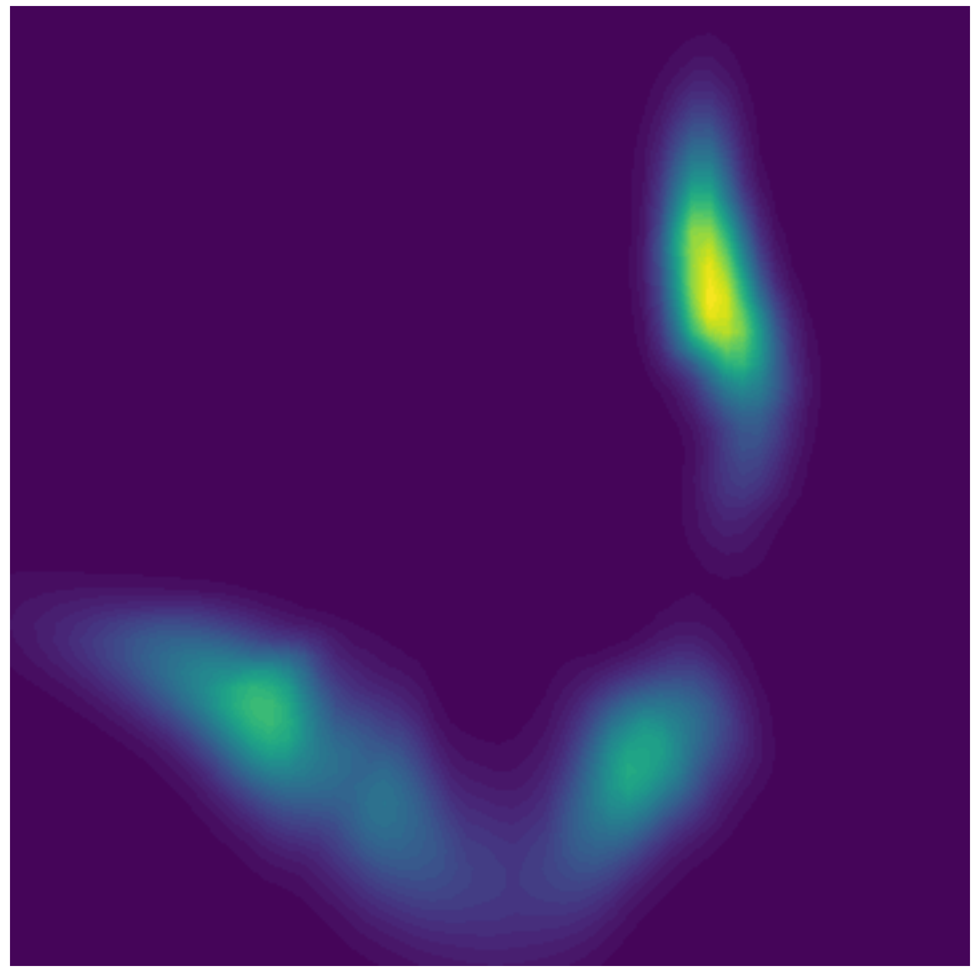
1. ALD completely removes datapoint-wise iterations
2. ALD is valid as an MCMC algorithm
3. Encoder may accelerate the convergence of MCMC

- We name the learning algorithm of DLVMs using ALD the ***Langevin autoencoder***

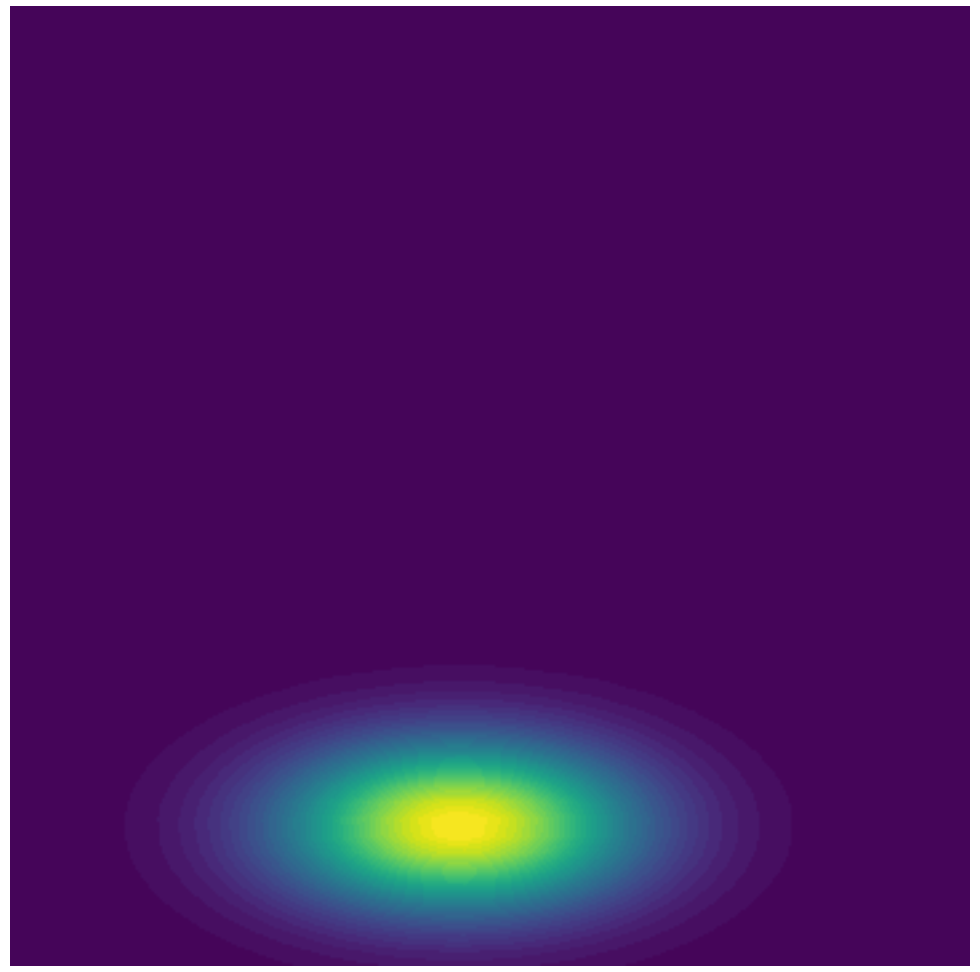
Experiments

Toy Example

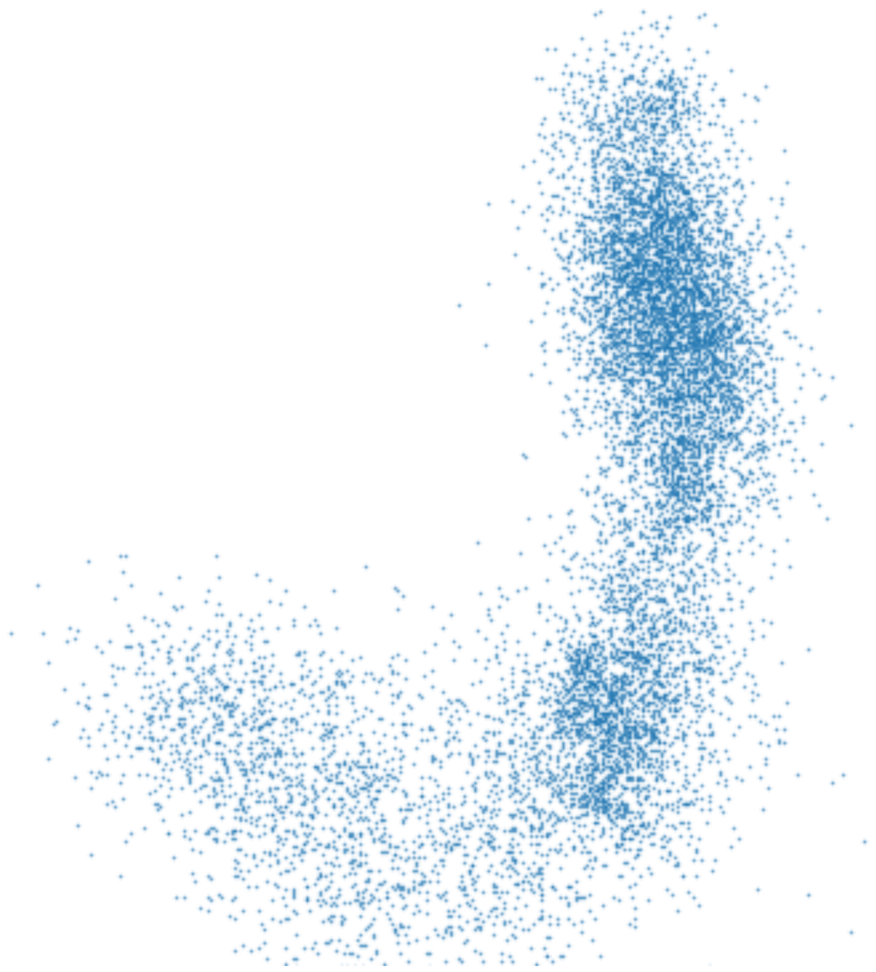
Visualization of Posterior Approximation



Ground Truth



Variational Inference



ALD (ours)

Our ALD can accurately capture the multimodality

Experiments

Image Generation

Test Likelihood Comparison

	MNIST	SVHN	CIFAR-10	CelebA
VAE	1.189 ± 0.002	4.442 ± 0.003	4.820 ± 0.005	4.671 ± 0.001
VAE-flow	1.183 ± 0.001	4.454 ± 0.016	4.828 ± 0.005	4.667 ± 0.005
Hoffman [2017]	1.189 ± 0.002	4.440 ± 0.007	4.831 ± 0.005	4.662 ± 0.011
LAE (ours)	$\mathbf{1.177} \pm 0.001$	$\mathbf{4.412} \pm 0.002$	$\mathbf{4.773} \pm 0.003$	$\mathbf{4.636} \pm 0.003$

Our LAE consistently outperforms VAE and existing LD-based method

Future Works

- Extend to more sophisticated MCMC (e.g., Hamiltonian Monte Carlo)
- Completely remove the bias of gradient estimation using unbiased MCMC method

Conclusion

- To train a deep latent variable model (DLVM), we need to approximate intractable posterior distribution
- Langevin dynamics (LD) can be a possible choice, but it is too slow due to datapoint-wise sampling process
- We propose **amortized Langevin dynamics** (ALD), which alleviate the problem by introducing an encoder
- ALD-based learning algorithm of DLVM named the **Langevin autoencoder** empirically outperforms existing methods (e.g., VAE and other LD-based method)