Variational Laplace Autoencoders

Yookoon Park, Chris Dongjoo Kim and Gunhee Kim
Vision and Learning Lab
Seoul National University, South Korea

Introduction

- Variational Autoencoders
- Two Challenges of Amortized Variational Inference
- Contributions

Variational Autoencoders (VAEs)

- Generative network θ

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\boldsymbol{g}_\theta(\mathbf{z}), \sigma^2\mathbf{I}),\ p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$

- Inference network ϕ: ***amortized inference*** of $p_\theta(\mathbf{z}|\mathbf{x})$

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}_\phi^2(\mathbf{x})))$$

- Networks jointly trained by maximizing the *Evidence Lower Bound (ELBO)*

$$\begin{aligned}\mathcal{L}(\mathbf{x}) = \mathbb{E}_q\left[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right] &= \log p_\theta(\mathbf{x}) - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x})) \\ &\leq \log p_\theta(\mathbf{x})\end{aligned}$$

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *ICLR*, 2014.
Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014

Two Challenges of Amortized Variational Inference

1. Enhancing the ***expressiveness*** of $q_\phi(\mathbf{z}|\mathbf{x})$
    - The full-factorized assumption is restrictive to capture complex posteriors
    - E.g. normalizing flows (Rezende & Mohamed, 2015; Kingma et al., 2016)
2. Reducing the ***amortization error*** of $q_\phi(\mathbf{z}|\mathbf{x})$
    - The error due to the inaccuracy of the inference network
    - E.g. gradient-based refinements of $q_\phi(\mathbf{z}|\mathbf{x})$ (Kim et al, 2018; Marino et al., 2018; Krishnan et al. 2018)

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. In *ICML*, 2015.
Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. In *NeurIPS*, 2016.
Kim, Y., Wiseman, S., Millter, A. C., Sontag, D., and Rush, A. M. Semi-amortized variational autoencoders. In *ICML*, 2018.
Marino, J., Yisong, Y., and Mandt, S. Iterative amortized inference. In *ICML*, 2018.
Krishnan, R. G., Liang, D., and Hoffman, M. D. On the challenges of learning with inference networks on sparse high-dimensional data. In *AISTAT*, 2018.

Contributions

- The ***Laplace approximation*** of the posterior to improve the training of latent deep generative models with:
  1. Enhanced **expressiveness** of full-covariance Gaussian posterior
  2. Reduced **amortization error** due to direct covariance computation from the generative network behavior

- A novel posterior inference exploiting local linearity of ReLU networks

Approach

- Posterior Inference using Local Linear Approximations
- Generalization: Variational Laplace Autoencoders

Observation 1: Probabilistic PCA

- A linear Gaussian model
  (Tipping & Bishop, 1999)

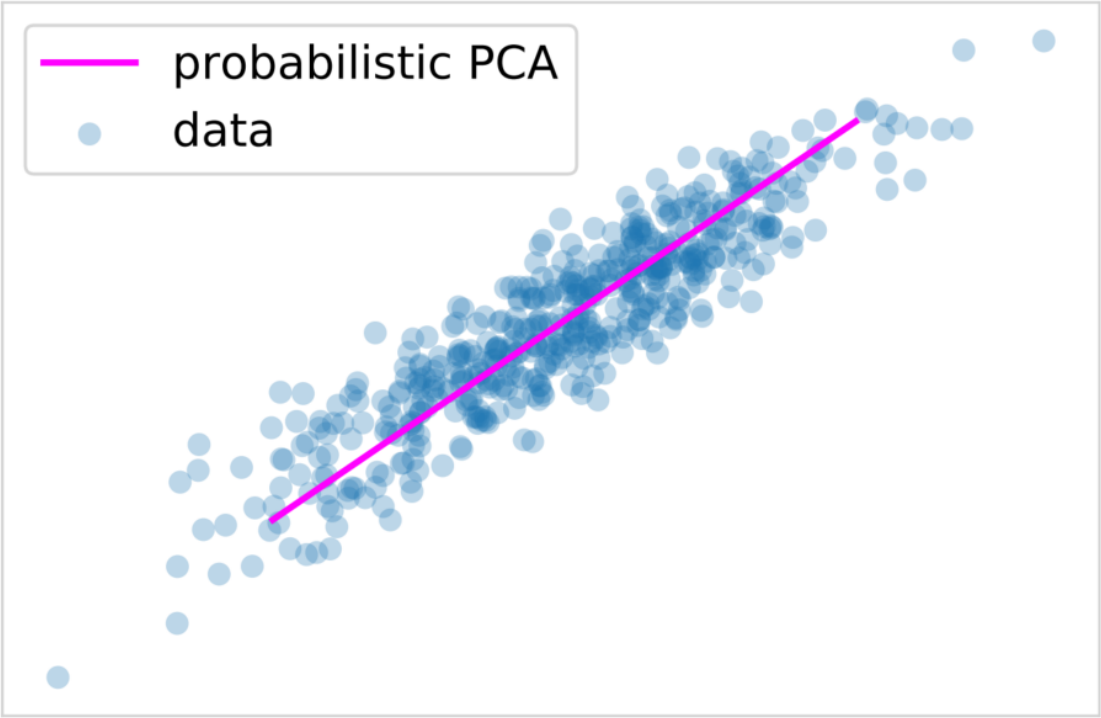
$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{W}\mathbf{z} + \mathbf{b}, \sigma^2\mathbf{I})$$

- The posterior distribution is ***exactly***

$$p_\theta(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\frac{1}{\sigma^2}\boldsymbol{\Sigma}\mathbf{W}^{\mathrm{T}}(\mathbf{x} - \mathbf{b}), \boldsymbol{\Sigma}\right)$$

$$\text{where } \boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2}\mathbf{W}^{\mathrm{T}}\mathbf{W} + \mathbf{I}\right)^{-1}$$



Toy example. 1-dim pPCA on 2-dim data

Tipping, M. E. and Bishop, C. M. Probabilistic Principal Component Analysis. *J. R. Statist. Soc. B*, 61(3):611–622, 1999.

Observation 2: Piece-wise Linear ReLU Networks

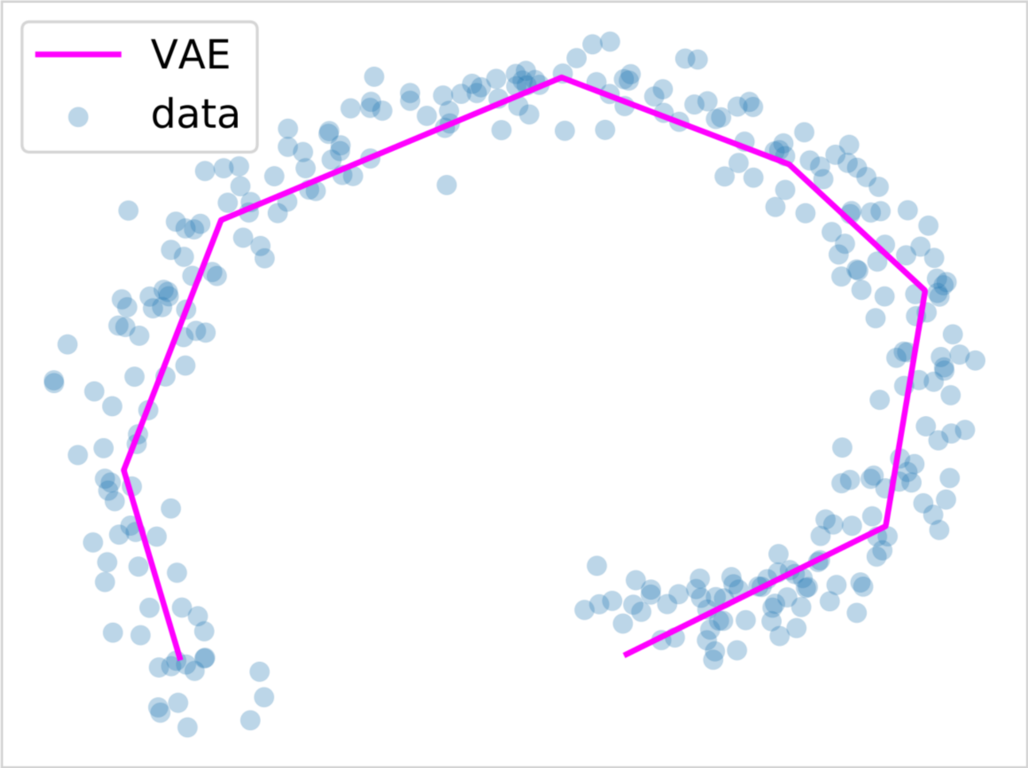
- ReLU networks are *piece-wise linear*

  (Pascanu et al., 2014; Montufar et al., 2014)

$$g_\theta(\mathbf{z}) \approx \mathbf{W_z}\,\mathbf{z} + \mathbf{b_z}$$

- Locally equivalent to *probabilistic PCA*

$$p_\theta(\mathbf{x}|\mathbf{z}) \approx \mathcal{N}(\mathbf{W_z}\,\mathbf{z} + \mathbf{b_z}, \sigma^2\mathbf{I})$$



Toy example. 1-dim ReLU VAE on 2-dim data

Pascanu, R., Montufar, G., and Bengio, Y. On the number of response regions of deep feedforward networks with piecewise linear activations. In *ICLR*, 2014.
Montufar, G., Pascanu, R., Cho, K., and Bengio, Y. On the number of linear regions of deep neural networks. In *NeurIPS*, 2014.

Posterior Inference using Local Linear Approximations

Linear models give exact posterior distribution

Observation 1

+

ReLU networks are locally linear

Observation 2

⇓

Posterior approximation based on the local linearity

Posterior Inference using Local Linear Approximations

1. Iteratively find the ***posterior mode*** $\boldsymbol{\mu}$ where the density is concentrated
    - Solve under the linear assumption $\boldsymbol{g}_\theta(\boldsymbol{\mu}_t) \approx \mathbf{W}_t\, \boldsymbol{\mu}_t + \mathbf{b}_t$

$$\boldsymbol{\mu}_{t+1} = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}\mathbf{W}_t^{\mathbf{T}}\mathbf{W}_t + \mathbf{I}\right)^{-1}\mathbf{W}_t^{\mathbf{T}}(\mathbf{x} - \mathbf{b})$$

    - Repeat for T steps

2. Posterior approximation using $p_\theta(\mathbf{x}|\mathbf{z}) \approx \mathcal{N}(\mathbf{W}_{\boldsymbol{\mu}}\, \mathbf{z} + \mathbf{b}_{\boldsymbol{\mu}}, \sigma^2\mathbf{I})$

$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{where } \boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2}\mathbf{W}_{\boldsymbol{\mu}}^{\mathbf{T}}\mathbf{W}_{\boldsymbol{\mu}} + \mathbf{I}\right)^{-1}$$

Generalization: Variational Laplace Autoencoders

1. Find the posterior mode s.t. $\nabla_{\mathbf{z}} \log p(\mathbf{x}, \mathbf{z})|_{\mathbf{z}=\boldsymbol{\mu}} = 0$
    - Initialize $\boldsymbol{\mu}_0$ using the inference network
    - Iteratively refine $\boldsymbol{\mu}_t$ (e.g. use gradient-descent)

2. The ***Laplace approximation*** defines the posterior as:

$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ where } \boldsymbol{\Sigma}^{-\mathbf{1}} = \boldsymbol{\Lambda} = -\nabla_{\mathbf{z}}^2 \log p(\mathbf{x}, \mathbf{z})|_{\mathbf{z}=\boldsymbol{\mu}}$$

3. Evaluate the ELBO using $q(\mathbf{z}|\mathbf{x})$ and train the model

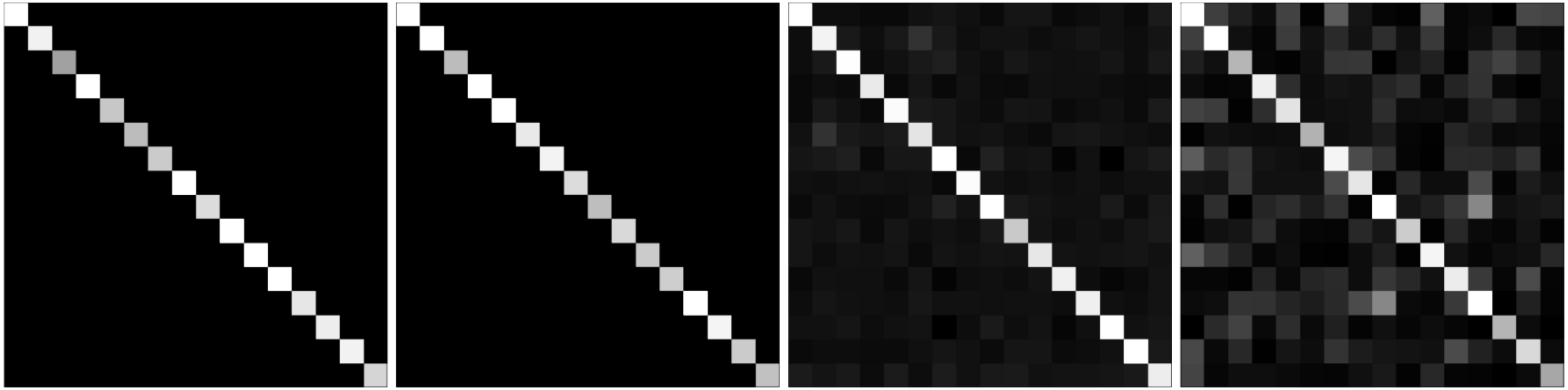
Results

- Posterior Covariance
- Log-likelihood Results

Experiments

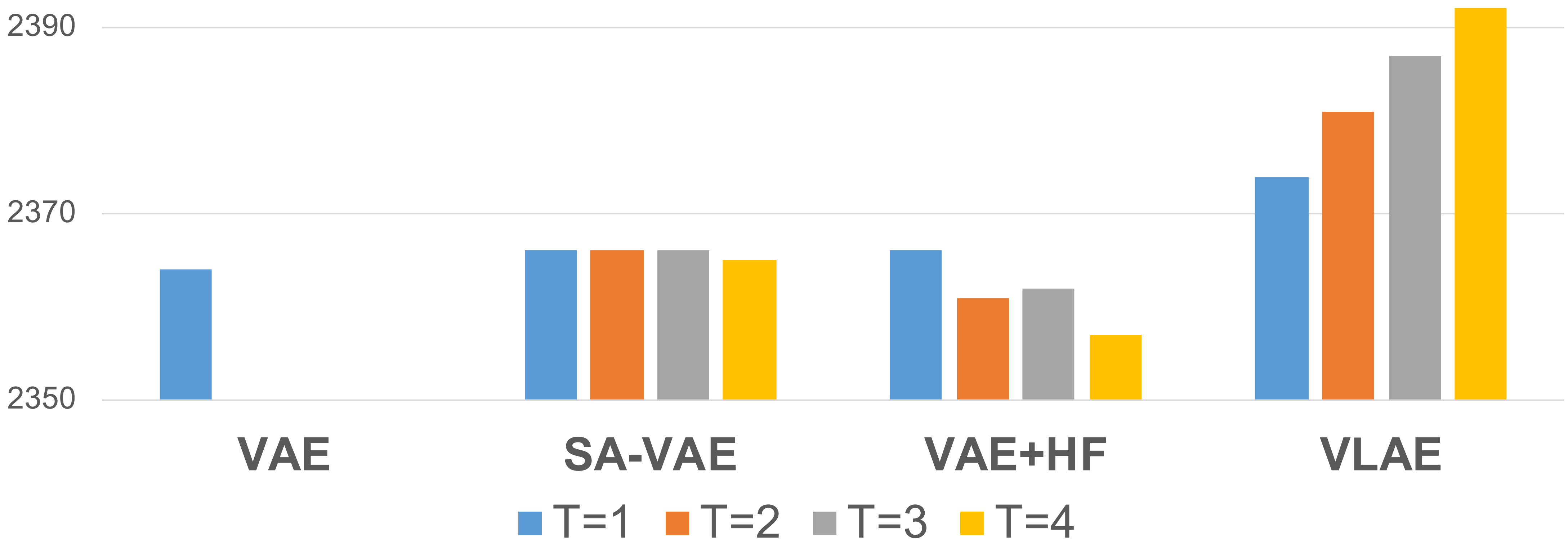
- Image datasets: MNIST, OMNIGLOT, Fashion MNIST, SVHN, CIFAR10
- Baselines
  - VAE
  - Semi-Amortized (SA) VAE (Kim et al, 2018)
  - VAE + Householder Flows (HF) (Tomczak & Welling, 2016)
  - Variational Laplace Autoencoder (VLAE)
- T=1, 2, 4, 8 (number of iterative updates or flows)

Posterior Covariance Matrices



(a) VAE (b) SA-VAE (c) VAE + HF (d) VLAE

Log-likelihood Results on CIFAR10



Thank you

Visit our poster session at **Pacific Ballroom #2**

Code available at : https://github.com/yookoon/VLAE