Variational Laplace Autoencoders

Code available at
http://vision.snu.ac.kr/projects/VLAE

Yookoon Park Chris Dongjoo Kim Gunhee Kim

Seoul National University, Seoul, South Korea

1. Contributions

- The ***Laplace approximation*** of the posterior to improve the training of latent deep generative models with:
  1. Full-covariance Gaussian posterior
  2. Direct covariance computation from the generative network behavior
- A novel posterior inference exploiting local linearity of ReLU networks
- *Variational Laplace Autoencoders*, a generalized framework for training latent deep generative models

2. Background

2.1. Variational Autoencoders

- Generative network
  $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\boldsymbol{g}_\theta(\mathbf{z}), \sigma^2\mathbf{I}), p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$
- **Amortized inference** of $p_\theta(\mathbf{z}|\mathbf{x})$
  $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}_\phi^2(\mathbf{x})))$
- Objective*:* the *ELBO*
  $$\mathcal{L}(\mathbf{x}) = \mathbb{E}_q[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})] = \log p_\theta(\mathbf{x}) - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x}))$$

2.2. Challenges of Amortized Inference

1) Limited ***expressiveness*** of $q_\phi(\mathbf{z}|\mathbf{x})$
   - The fully-factorized assumption
   - E.g. normalizing flows (Rezende & Mohamed, 2015; Kingma et al., 2016)
2) The ***amortization error***
   - The error due to dynamic inference (Cremer et al., 2018)
   - E.g. gradient-based refinements (Kim et al., 2018; Marino et al., 2018; Krishnan et al., 2018)

3. Posterior Inference based on Local Linearity

3.1. Probabilistic PCA (Tipping & Bishop, 1999)

- A linear Gaussian model
  $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{W}\mathbf{z} + \mathbf{b}, \sigma^2\mathbf{I}),\ p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$
- The ***exact*** posterior is
  $$p_\theta(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\frac{1}{\sigma^2}\boldsymbol{\Sigma}\mathbf{W}^{\mathrm{T}}(\mathbf{x} - \mathbf{b}), \boldsymbol{\Sigma}\right)$$
  $$\text{where } \boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2}\mathbf{W}^{\mathrm{T}}\mathbf{W} + \mathbf{I}\right)^{-1}$$

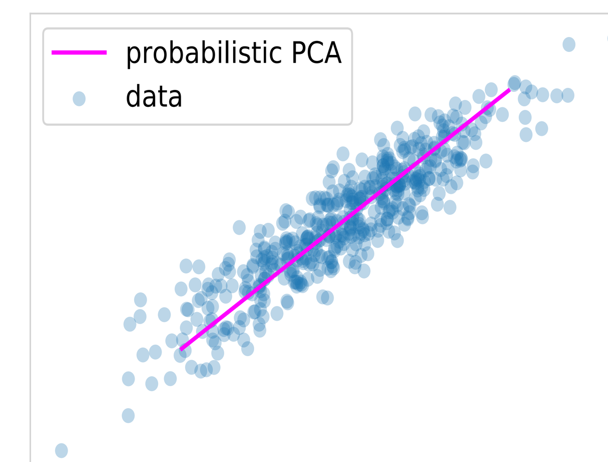


Fig 1. 1 dim p-PCA on 2 dim data

3.2. Piece-wise Linear Networks

- ReLU networks are *piece-wise linear* (Pascanu et al., 2014; Montufar et al., 2014)
  $$\boldsymbol{g}_\theta(\mathbf{z}) \approx \mathbf{W}_\mathbf{z}\,\mathbf{z} + \mathbf{b}_\mathbf{z}$$
- ***Locally*** equivalent to the *probabilistic PCA*
  $$p_\theta(\mathbf{x}|\mathbf{z}) \approx \mathcal{N}(\mathbf{W}_\mathbf{z}\,\mathbf{z} + \mathbf{b}_\mathbf{z}, \sigma^2\mathbf{I})$$

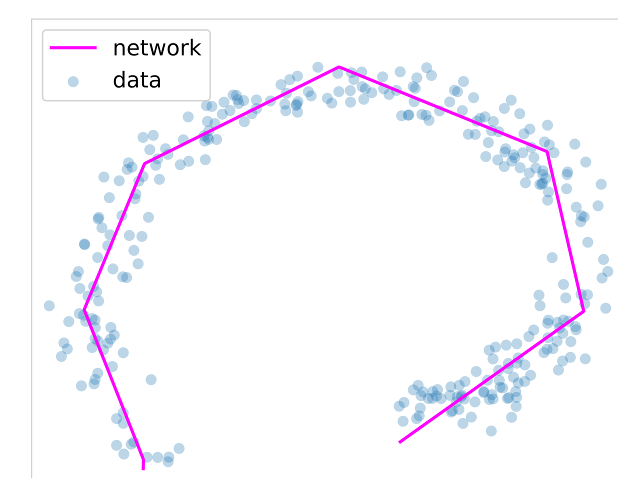


Fig 2. 1 dim ReLU VAE on 2 dim data

3.3 Posterior Inference for ReLU networks

1. Iteratively search for the ***posterior mode*** $\boldsymbol{\mu}$ for T steps
   - Solve under the linear assumption $\boldsymbol{g}_\theta(\boldsymbol{\mu}_t) \approx \mathbf{W}_t\,\boldsymbol{\mu}_t + \mathbf{b}_t$
   $$\boldsymbol{\mu}_{t+1} = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}\mathbf{W}_t^{\mathrm{T}}\mathbf{W}_t + \mathbf{I}\right)^{-1}\mathbf{W}_t^{\mathrm{T}}(\mathbf{x} - \mathbf{b})$$
2. Approximate the posterior using $p_\theta(\mathbf{x}|\mathbf{z}) \approx \mathcal{N}(\mathbf{W}_\mu\,\mathbf{z} + \mathbf{b}_\mu, \sigma^2\mathbf{I})$
   $$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ where } \boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2}\mathbf{W}_\mu^{\mathrm{T}}\mathbf{W}_\mu + \mathbf{I}\right)^{-1}$$

4. Variational Laplace Autoencoders

1. Search for 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*** of the posterior at $\boldsymbol{\mu}$:
   $$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ \text{ where } \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Lambda} = -\nabla_\mathbf{z}^2\log p(\mathbf{x}, \mathbf{z})|_{\mathbf{z}=\boldsymbol{\mu}}$$
3. Evaluate the ELBO using $q(\mathbf{z}|\mathbf{x})$ and train the model

5. Results

- **Baselines**: (1) VAE (2) Semi-Amortized (SA) VAE (Kim et al., 2018) (3) Householder Flows (HF) (Tomczak & Welling, 2016) (4) Inverse Autoregressive Flows (IAF) (Kingma et al., 2016)
- **Network**: fully-connected + ReLU activation

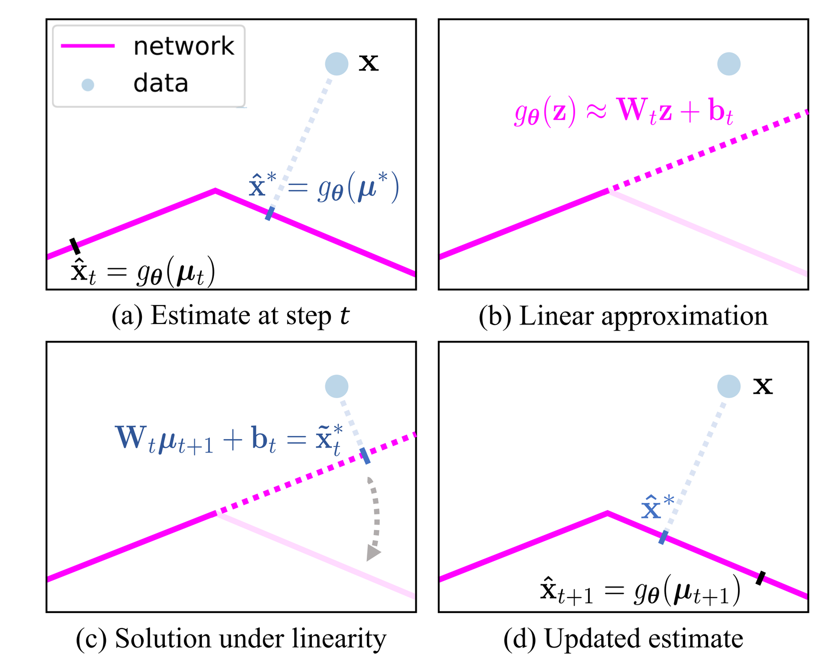


Fig 3. Mode searching (data space)

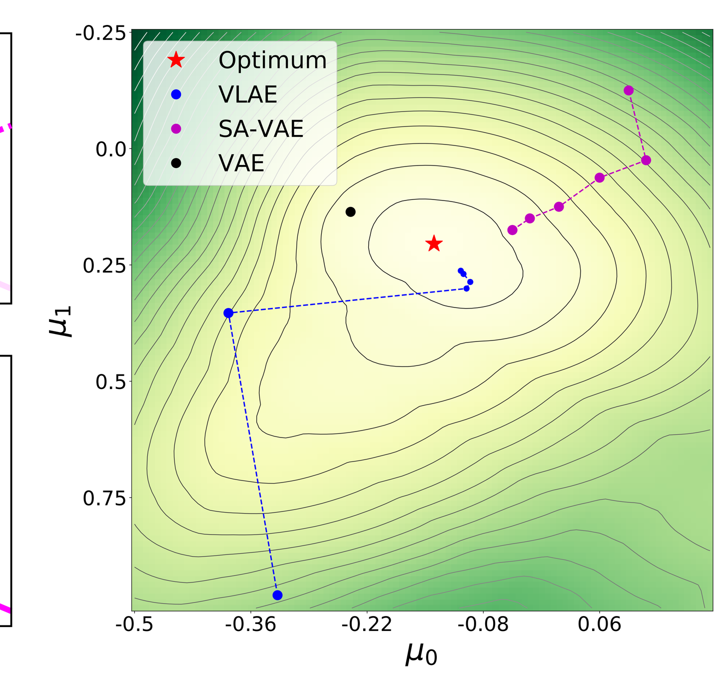


Fig 4. Mode searching (parameter space)

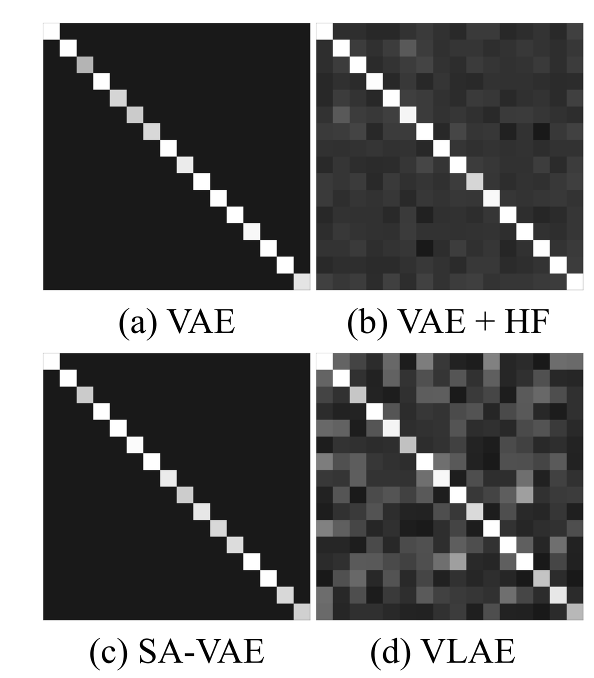


Fig 5. Posterior covariance matrices

Fig 6. Reconstruction images

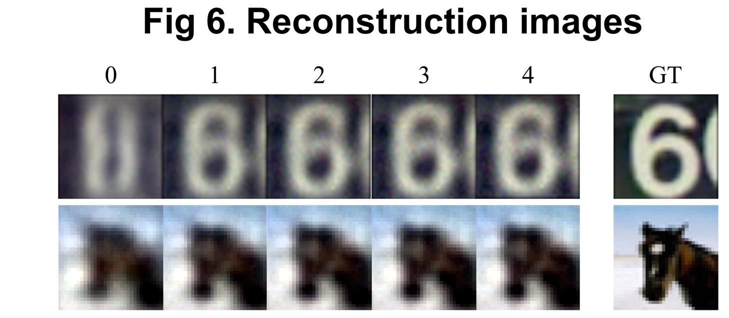
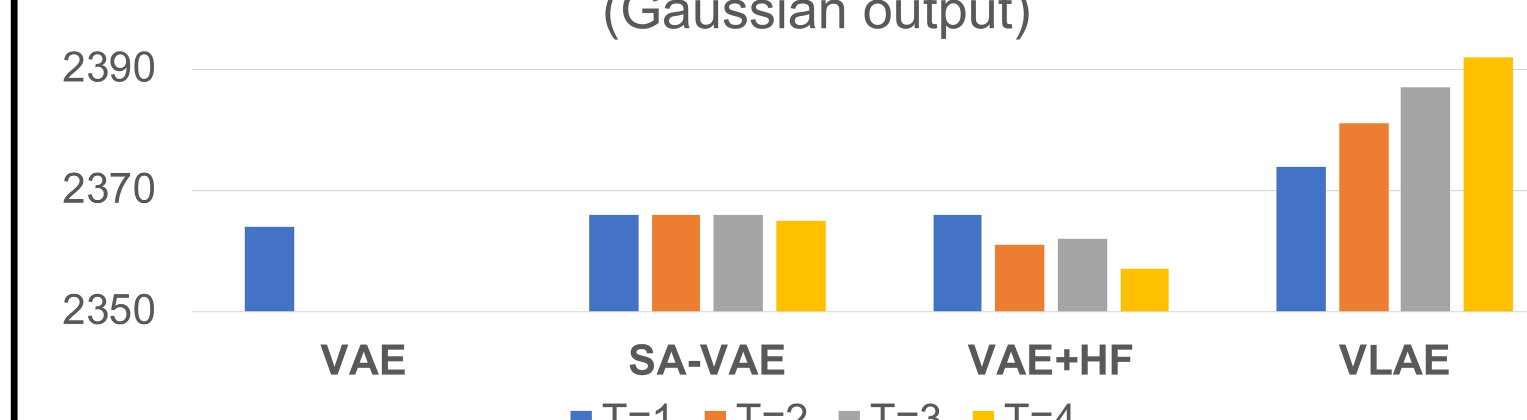


Fig 7. Reconstruction error vs. update steps

Log-likelihood Results on CIFAR10
(Gaussian output)



Log-likelihood Results on Binarized MNIST
(Bernoulli output)