Theoretical Convergence Guarantees for Variational Autoencoders

Sobihan Surendran, Antoine Godichon-Baggioni, Sylvain Le Corff

Journées de Statistique (JdS), Marseille, 2025

Motivation

- **Advances in Generative Models:** Variational Autoencoders (VAE), Generative Adversarial Networks (GAN), and Diffusion Models.

Motivation

- **Advances in Generative Models:** Variational Autoencoders (VAE), Generative Adversarial Networks (GAN), and Diffusion Models.
- **Why VAE? Strengths and Relevance**

Motivation

- **Advances in Generative Models:** Variational Autoencoders (VAE), Generative Adversarial Networks (GAN), and Diffusion Models.
- **Why VAE? Strengths and Relevance**
  - **Structured Latent Space:** Encourages interpretable and disentangled representations.
  - **Sample Efficiency:** Performs well in **low sample size** scenarios (e.g., medical imaging).
  - **Latent Diffusion: VAE-based diffusion models** achieves state-of-the-art results in image generation.

Motivation

- **Advances in Generative Models:** Variational Autoencoders (VAE), Generative Adversarial Networks (GAN), and Diffusion Models.
- **Why VAE? Strengths and Relevance**
  - **Structured Latent Space:** Encourages interpretable and disentangled representations.
  - **Sample Efficiency:** Performs well in **low sample size** scenarios (e.g., medical imaging).
  - **Latent Diffusion: VAE-based diffusion models** achieves state-of-the-art results in image generation.
- **Theoretical Understanding:** Prior work has primarily focused on generalization bounds, ELBO approximations, and posterior collapse. However, the **Optimization in VAE remains underexplored**.

Table of Contents

Introduction: Variational Autoencoders

We consider **generative models** of the form:

$$p_\theta(x) = \int p_\theta(z)\, p_\theta(x|z) \mathrm{d}z \ ,$$

where x is an **observation** and z a **latent variable**.

Introduction: Variational Autoencoders

We consider **generative models** of the form:

$$p_\theta(x) = \int p_\theta(z)\, p_\theta(x|z)\mathrm{d}z \;,$$

where x is an **observation** and z a **latent variable**. The marginal log-likelihood is given by:

$$\log p_\theta(x) = \log \mathbb{E}_{p_\theta(\cdot|x)}\left[\frac{p_\theta(x,z)}{p_\theta(z|x)}\right] \gtrsim \underbrace{\mathbb{E}_{q_\phi(\cdot|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right]}_{\textbf{Evidence Lower Bound (ELBO)}} =: \mathcal{L}(\theta,\phi;x) \;,$$

where $q_\phi(z|x)$ is the **variational distribution**.

Gradient Computation.

- For $\nabla_\theta \mathcal{L}(\theta, \phi; x)$: easy to compute.

Optimization in Variational Autoencoders

Gradient Computation.

- For $\nabla_{\theta}\mathcal{L}(\theta, \phi; x)$: easy to compute.
- For $\nabla_{\phi}\mathcal{L}(\theta, \phi; x)$: more challenging; two main methods:
  - **Score function estimator:** general, but high variance.
  - **Pathwise estimator:** reparameterization trick, lower variance.

Optimization in Variational Autoencoders

Gradient Computation.

- For $\nabla_\theta \mathcal{L}(\theta, \phi; x)$: easy to compute.
- For $\nabla_\phi \mathcal{L}(\theta, \phi; x)$: more challenging; two main methods:
  - **Score function estimator:** general, but high variance.
  - **Pathwise estimator:** reparameterization trick, lower variance.

Consider the **Stochastic Gradient Descent (SGD)** update:

$$(\theta_{k+1}, \phi_{k+1}) = (\theta_k, \phi_k) + \gamma_{k+1} \widehat{\nabla}_{\theta,\phi} \mathcal{L}(\theta_k, \phi_k; \mathcal{D}_{k+1}) \ , \qquad (1)$$

- $\widehat{\nabla}_{\theta,\phi} \mathcal{L}(\theta_k, \phi_k; \mathcal{D}_{k+1})$ denotes an estimator of the gradient,
- $\mathcal{D}_{k+1}$ is the mini-batch of data used at iteration $k+1$,
- $\gamma_k > 0$ is the learning rate.

Deep Gaussian VAE: Setting

The **Deep Gaussian VAE** consists of a decoder and an encoder such that:

$$p_{\theta}(x|z) = \mathcal{N}(x; G_{\theta}(z), c^2 \mathrm{I}_{d_x}) \ ,$$
$$q_{\phi}(z|x) = \mathcal{N}(z; \mu_{\phi}(x), \Sigma_{\phi}(x)) \ .$$

Deep Gaussian VAE: Setting

The **Deep Gaussian VAE** consists of a decoder and an encoder such that:

$$p_{\theta}(x|z) = \mathcal{N}(x; G_{\theta}(z), c^2 \mathrm{I}_{d_x}) \ ,$$
$$q_{\phi}(z|x) = \mathcal{N}(z; \mu_{\phi}(x), \Sigma_{\phi}(x)) \ .$$

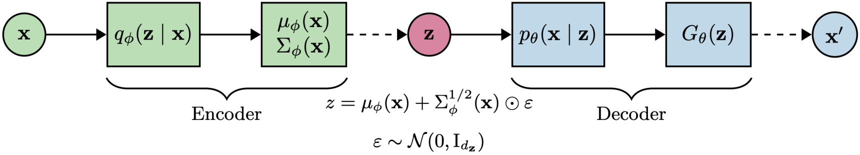

Figure: Architecture of a VAE using multivariate Gaussian distributions.

Convergence Analysis for Deep Gaussian VAE

Consider a Neural Network with the assumptions:

(i) For all $\phi \in \Phi$, $\lambda_{\min}(\Sigma_\phi(x)) \geq c_\Sigma$ and all activation functions are Lipschitz continuous and smooth.

(ii) There exists a constant a such that $\|\theta\|_\infty + \|\phi\|_\infty \leq a$ for all $\theta \in \Theta$ and $\phi \in \Phi$.

Convergence Analysis for Deep Gaussian VAE

Consider a Neural Network with the assumptions:

(i) For all $\phi \in \Phi$, $\lambda_{\min}(\Sigma_\phi(x)) \geq c_\Sigma$ and all activation functions are Lipschitz continuous and smooth.

(ii) There exists a constant a such that $\|\theta\|_\infty + \|\phi\|_\infty \leq a$ for all $\theta \in \Theta$ and $\phi \in \Phi$.

Convergence Analysis

Let $(\theta_n, \phi_n) \in \Theta \times \Phi$ be the n-th iterate of Adam, with $\gamma_n = C_\gamma n^{-1/2}$, $C_\gamma > 0$, and $\beta_1 < \sqrt{\beta_2} < 1$. For all $n \geq 1$, let $R \sim \mathcal{U}(\{0, \ldots, n\})$. Then,

$$\mathbb{E}\left[\left\|\nabla_{\theta,\phi}\mathcal{L}\left(\theta_R, \phi_R\right)\right\|^2\right] = \mathcal{O}\left(\frac{\mathcal{L}^*}{\sqrt{n}} + Na^{2(N-1)}\frac{d^* \log n}{(1-\beta_1)\sqrt{n}}\right),$$

where $\mathcal{L}^* = \mathcal{L}\left(\theta^*, \phi^*\right) - \mathcal{L}\left(\theta_0, \phi_0\right)$, $d^* = d_\theta + d_\phi$ is the dimension of the parameters, and N is the number of layers in the encoder and decoder.

Illustration of Our Convergence Rate

- **Generalized Soft-Clipping** (Lipschitz, smooth, and bounded between s_1 and s_2):

$$f(x) = \frac{1}{s} \log \left(\frac{1 + \mathrm{e}^{s(x-s_1)}}{1 + \mathrm{e}^{s(x-s_2)}} \right) + s_1 \ .$$

Illustration of Our Convergence Rate

- **Generalized Soft-Clipping** (Lipschitz, smooth, and bounded between s_1 and s_2):

$$f(x) = \frac{1}{s} \log \left(\frac{1 + e^{s(x - s_1)}}{1 + e^{s(x - s_2)}} \right) + s_1 \, .$$

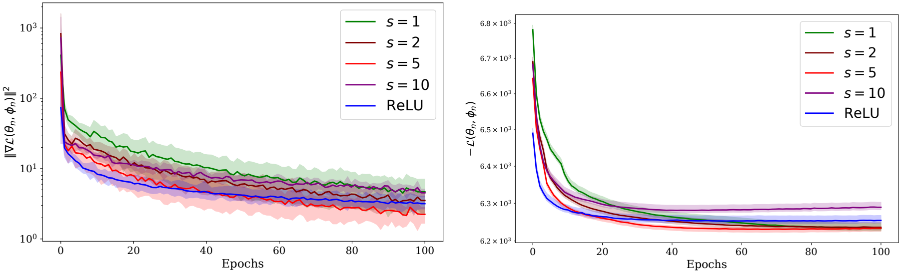

Figure: Squared norm of gradients and Negative ELBO on CelebA for VAE trained with Adam.

IWAE: Convergence Results

Objective: Obtain a **tighter ELBO** by using multiple importance weighted samples:

$$\log p_\theta(x) \geq \underbrace{\mathbb{E}_{q_\phi^{\otimes K}(\cdot|x)}\left[\log \frac{1}{K}\sum_{\ell=1}^{K} \frac{p_\theta(x, z^{(\ell)})}{q_\phi(z^{(\ell)}|x)}\right]}_{\textbf{IWAE}} \geq \underbrace{\mathbb{E}_{q_\phi(\cdot|x)}\left[\log \frac{p_\theta(x, z)}{q_\phi(z|x)}\right]}_{\textbf{VAE}} .$$

IWAE: Convergence Results

Objective: Obtain a **tighter ELBO** by using multiple importance weighted samples:

$$\log p_\theta(x) \geq \underbrace{\mathbb{E}_{q_\phi^{\otimes K}(\cdot|x)}\left[\log \frac{1}{K}\sum_{\ell=1}^{K}\frac{p_\theta(x, z^{(\ell)})}{q_\phi(z^{(\ell)}|x)}\right]}_{\textbf{IWAE}} \geq \underbrace{\mathbb{E}_{q_\phi(\cdot|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right]}_{\textbf{VAE}} .$$

Convergence Analysis (Informal)

Under similar assumptions to those for VAE, we have:

$$\mathbb{E}\left[\left\|\nabla_{\theta,\phi}\mathcal{L}_K^{\mathsf{IWAE}}(\theta_R, \phi_R)\right\|^2\right] = \mathcal{O}\left(\frac{\mathcal{L}_K^*}{\sqrt{n}} + d^* \frac{\log n}{BK\sqrt{n}}\right) ,$$

where B is the batch size and K is the number of variational samples.

Illustration of Our Convergence Rate in IWAE

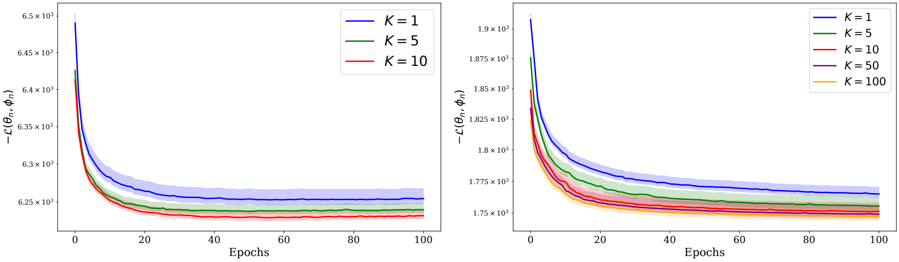

Figure: Negative ELBO in IWAE on CelebA and CIFAR-100 trained with Adam.

Illustration of Our Convergence Rate in IWAE

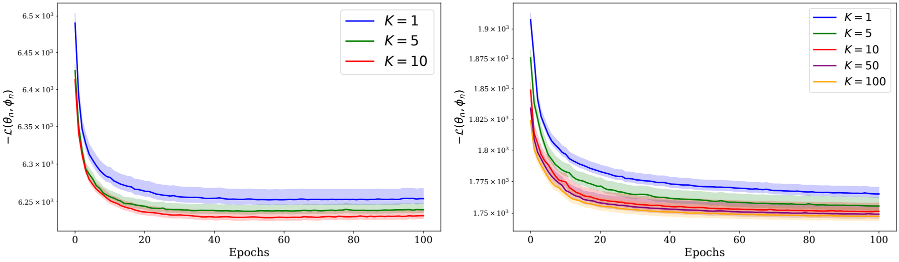

Figure: Negative ELBO in IWAE on CelebA and CIFAR-100 trained with Adam.

Link with Signal-to-Noise Ratio (SNR**)** [Rainforth et al. 2018].

SNR: expected gradient magnitude scaled by its standard deviation.

$$\mathrm{SNR}(\theta) = \sqrt{BK} \quad \mathrm{SNR}(\phi) = \sqrt{B/K}$$

Illustration of Our Convergence Rate in IWAE

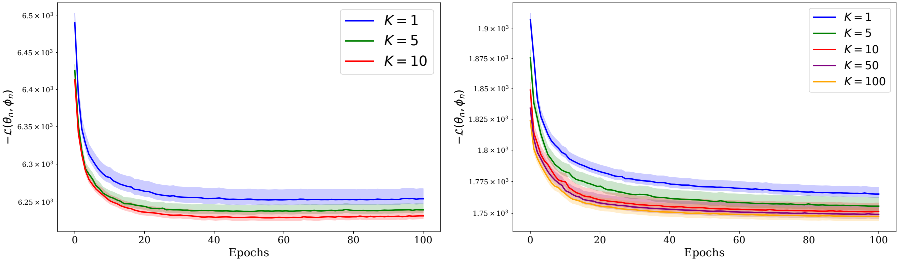

Figure: Negative ELBO in IWAE on CelebA and CIFAR-100 trained with Adam.

Link with Signal-to-Noise Ratio (SNR**)** [Rainforth et al. 2018].

SNR: expected gradient magnitude scaled by its standard deviation.

$$\mathrm{SNR}(\theta) = \sqrt{BK} \quad \mathrm{SNR}(\phi) = \sqrt{B/K}$$

$\Rightarrow$ **Gradually increase** K until a fixed threshold is reached.

$\Rightarrow$ Use **Rényi IWAE** [Daudel et al. 2023] with $\mathrm{SNR}(\theta, \phi) = \sqrt{BK}$.

Extension to Variational Inference

- Variational Inference is typically formulated as:

$$\phi^* \in \operatorname*{argmin}_{\phi \in \Phi} \mathsf{KL}(q_\phi \,\|\, p(\cdot|x))$$

where $q_\phi(z|x)$ is the variational distribution.

Extension to Variational Inference

- Variational Inference is typically formulated as:

$$\phi^* \in \underset{\phi \in \Phi}{\operatorname{argmin}} \mathsf{KL}(q_\phi \,\|\, p(\cdot|x)) \iff \phi^* \in \underset{\phi \in \Phi}{\operatorname{argmax}} \mathbb{E}_{q_\phi(\cdot|x)} \left[\log \frac{p(x,z)}{q_\phi(z|x)} \right]$$

where $q_\phi(z|x)$ is the variational distribution.

- Variational Inference is typically formulated as:

$$\phi^* \in \underset{\phi \in \Phi}{\operatorname{argmin}} \, \mathsf{KL}(q_\phi \,\|\, p(\cdot|x)) \iff \phi^* \in \underset{\phi \in \Phi}{\operatorname{argmax}} \, \mathbb{E}_{q_\phi(\cdot|x)} \left[\log \frac{p(x,z)}{q_\phi(z|x)} \right]$$

where $q_\phi(z|x)$ is the variational distribution.

Structural Assumptions in Prior Convergence Results.

Reference	Non-Concavity of $\log p$	Beyond Location-Scale Family for q_ϕ	Parameterization Type
Kim et al. 2024	✗	✗	Linear
Domke et al. 2023	✓	✗	Linear
Kim et al. 2023	✓	✗	Non-linear (scale)
Ours	✓	✓	**Non-linear**

Location-Scale Family: Distributions obtained by shifting and scaling a fixed base distribution, i.e., $Y = \mu + \sigma W$ with location μ and scale $\sigma > 0$.

Take Home Messages

- ☞ A convergence rate of $\mathcal{O}(n^{-1/2}\log n)$ for **VAE** with **SGD** and **Adam**, illustrated using the **Deep Gaussian VAE**.
- ☞ Increasing K in **IWAE** without tuning other parameters leads to vanishing SNR and poor gradient estimates for ϕ, hindering the learning of θ.
- ☞ New convergence results for **Variational Inference**, beyond location-scale families and linear parameterizations.

References

Rainforth, Tom, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh (2018). "Tighter variational bounds are not necessarily better". In: *International Conference on Machine Learning*. PMLR, pp. 4277–4285.

Daudel, Kamélia, Joe Benton, Yuyang Shi, and Arnaud Doucet (2023). "Alpha-divergence variational inference meets importance weighted auto-encoders: Methodology and asymptotics". In: *Journal of Machine Learning Research* 24.243, pp. 1–83.

Kim, Kyurae, Yian Ma, and Jacob Gardner (2024). "Linear Convergence of Black-Box Variational Inference: Should We Stick the Landing?" In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 235–243.

Domke, Justin, Robert Gower, and Guillaume Garrigos (2023). "Provable convergence guarantees for black-box variational inference". In: *Advances in Neural Information Processing Systems*. Vol. 36.

Kim, Kyurae, Jisu Oh, Kaiwen Wu, Yian Ma, and Jacob Gardner (2023). "On the convergence of black-box variational inference". In: *Advances in Neural Information Processing Systems*. Vol. 36.

Thank you for your attention!

SCAN ME

Find the full paper here (Accepted at AISTATS 2025)