Theoretical Convergence Guarantees for Variational Autoencoders

Sobihan Surendran[1,2], Antoine Godichon-Baggioni[1], and Sylvain Le Corff[1]

[1]Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, Paris, France

[2]LOPF, Califrais' Machine Learning Lab, Paris, France

SCAN ME

Goal

☞ Provide a non-asymptotic analysis of **Variational Autoencoders** and **Importance Weighted Autoencoders** with **Stochastic Gradient Descent**.
☞ Establish theoretical guarantees and illustrate the results using **Deep Gaussian VAE**.
☞ Extend the analysis to **Black Box Variational Inference**.

Introduction

We consider generative models of the form $p_\theta(x,z) = p_\theta(z)\,p_\theta(x|z)$, where x is an observation and z a latent variable. The marginal log-likelihood is given by:

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

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

The Pathwise Gradient.

- ▶ Reparameterization trick [1]: $z = g(\varepsilon,\phi)$, where $\varepsilon \sim p_\varepsilon$ (known distribution).
- ▶ Pathwise gradient of the ELBO:

$$\nabla_\phi \mathcal{L}(\theta,\phi;x) = \mathbb{E}_{p_\varepsilon}\left[\nabla_z \log w_{\theta,\phi}(x,z)\cdot \nabla_\phi g(\varepsilon,\phi)\right] - \mathbb{E}_{p_\varepsilon}\left[\nabla_\phi \log q_\phi(g(\varepsilon,\phi)\mid x)\right]\ ,$$

where $w_{\theta,\phi}(x,z) = p_\theta(x,z)/q_\phi(z|x)$ the unnormalized importance weights.

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

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

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

Smoothness of $\mathcal{L}$ + Gradient variance bound $\Rightarrow \mathbb{E}\left[\left\|\nabla \mathcal{L}(\theta_n,\phi_n)\right\|^2\right] = \mathcal{O}(n^{-1/2}\log n)$.

Deep Gaussian VAE

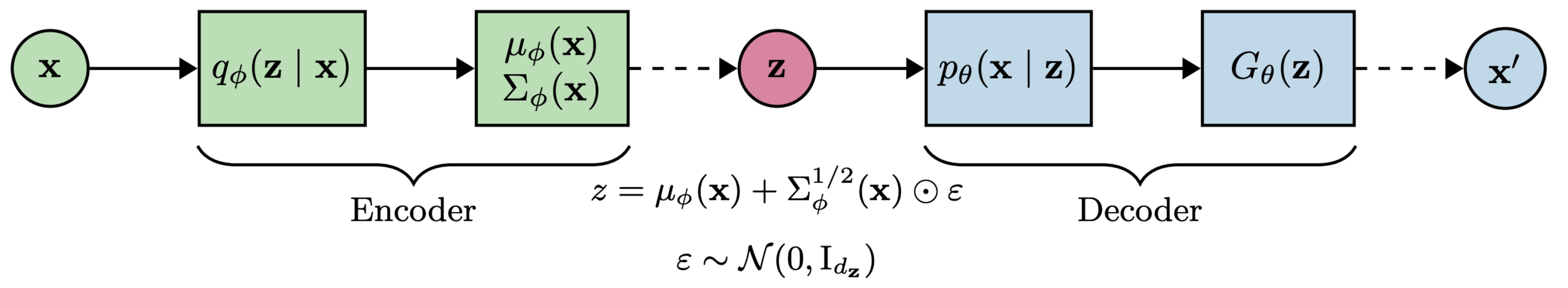

Figure: Illustration of the Architecture of a VAE using multivariate Gaussian distributions.

Convergence Analysis for Deep Gaussian VAE

Consider a neural network with the assumptions:

(i) $\|G_\theta(z)\| \leq C_G$, $\|\mu_\phi(x)\| \leq C_\mu$, $\lambda_{\min}(\Sigma_\phi(x)) \geq c_\Sigma$, 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 $\phi\in\Phi$.

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 $R\in\{0,\ldots,n\}$ be a uniformly distributed random variable. Then,

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

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

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

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

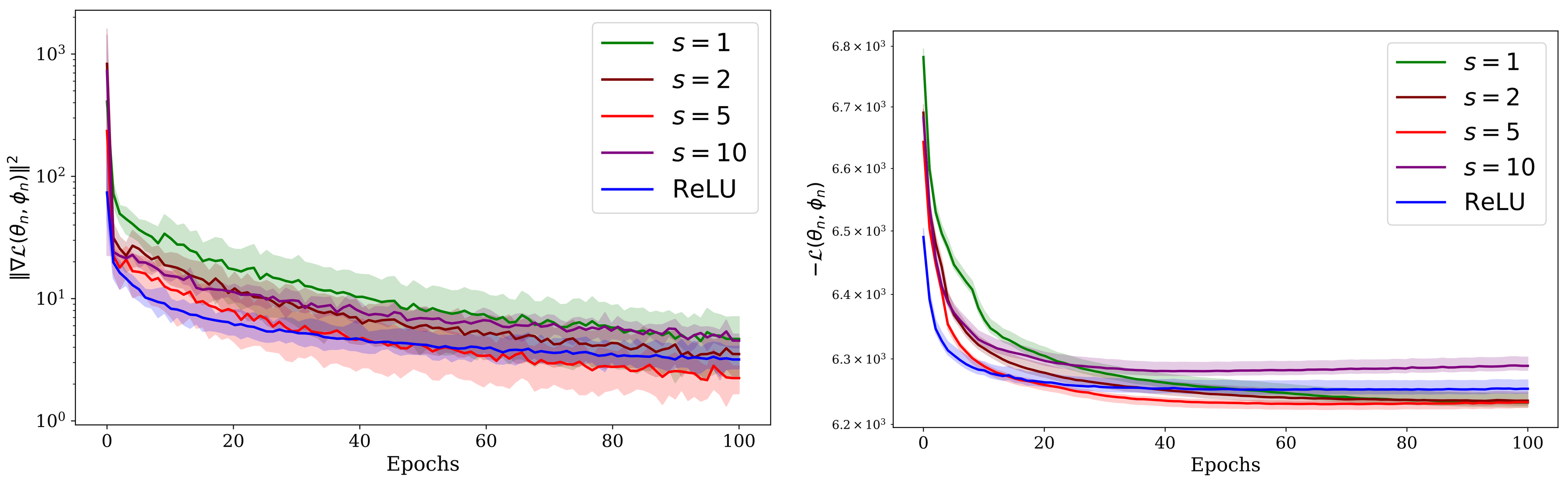

Figure: Squared norm of gradients and Negative ELBO on CelebA for VAE trained with Adam.

Importance Weighted Autoencoder

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

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

Convergence Analysis for IWAE

Assuming the same conditions as for VAE and mild regularity on weights $w_{\theta,\phi}$, we have:

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

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

Link with Signal-to-Noise Ratio (SNR) **[2].**
SNR: expected gradient magnitude scaled by its standard deviation.

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

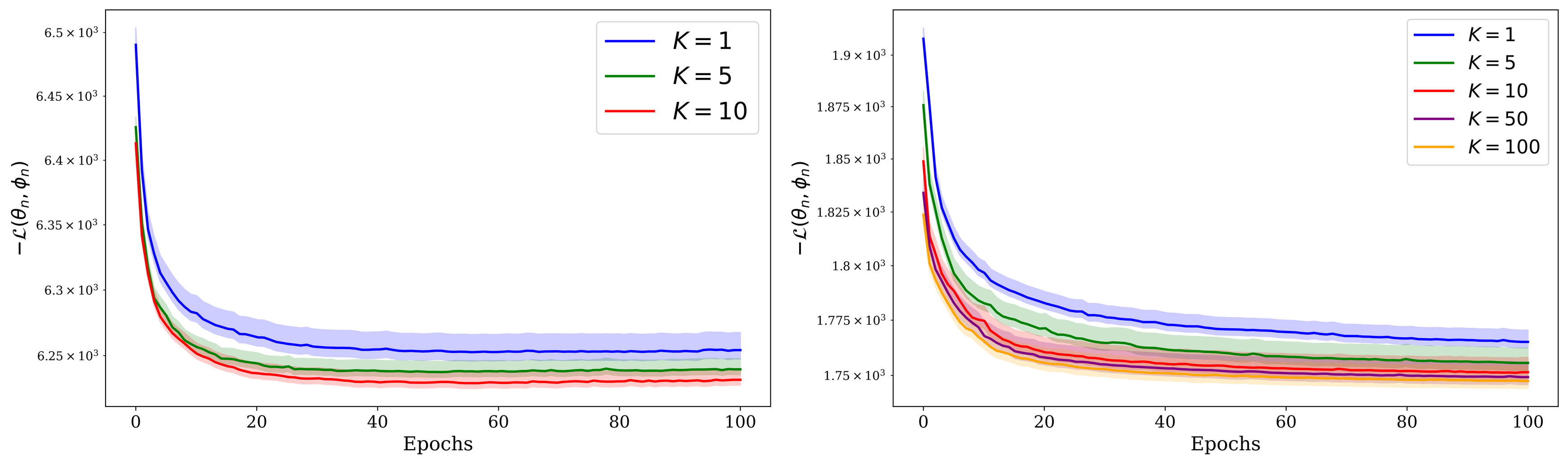

Figure: Negative ELBO in IWAE on CelebA (on the left) and CIFAR-100 (on the right) trained with Adam.

$\Rightarrow$ **gradually increase** K until a threshold, or use **Rényi IWAE** with $\text{SNR}(\theta,\phi) = \sqrt{BK}$.

Extension to Black Box Variational Inference

Black Box Variational Inference (BBVI) is typically formulated as:

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

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. [3]	✗	✗	Linear
Domke et al. [4]	✓	✗	Linear
Kim et al. [5]	✓	✗	Non-linear (scale)
Ours	✓	✓	**Non-linear**

Conclusion

☞ A convergence rate of $\mathcal{O}(n^{-1/2}\log n)$ for **VAE** with **SGD** and **Adam**.
☞ Illustration of the results using the **Deep Gaussian VAE**, that supports our theoretical claims, with similar empirical results for standard VAE with ReLU.
☞ 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 **BBVI**, beyond location-scale families and linear parameterizations.

References

[1] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *International Conference on Learning Representations*. 2014.

[2] Tom Rainforth et al. "Tighter variational bounds are not necessarily better". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 4277–4285.

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

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

[5] Kyurae Kim et al. "On the convergence of black-box variational inference". In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023.