Theoretical Convergence Guarantees for Variational Autoencoders

Sobihan Surendran[1,2] **Antoine Godichon-Baggioni**[1] **Sylvain Le Corff**[1]
[1]Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, Paris, France
[2]LOPF, Califrais' Machine Learning Lab, Paris, France

Abstract

Variational Autoencoders (VAE) are popular generative models used to sample from complex data distributions. Despite their empirical success in various machine learning tasks, significant gaps remain in understanding their theoretical properties, particularly regarding convergence guarantees. This paper aims to bridge that gap by providing non-asymptotic convergence guarantees for VAE trained using both Stochastic Gradient Descent and Adam algorithms. We derive a convergence rate of $\mathcal{O}(\log n/\sqrt{n})$, where n is the number of iterations of the optimization algorithm, with explicit dependencies on the batch size, the number of variational samples, and other key hyperparameters. Our theoretical analysis applies to both Linear VAE and Deep Gaussian VAE, as well as several VAE variants, including β-VAE and IWAE. Additionally, we empirically illustrate the impact of hyperparameters on convergence, offering new insights into the theoretical understanding of VAE training.

1 INTRODUCTION

Probabilistic inference in generative models is a long-standing problem in particular for complex and high dimensional distributions. Many solutions to estimate the likelihood of data or to infer parameters are based on Importance Sampling, Sequential Monte Carlo (SMC) or Markov Chain Monte Carlo (MCMC)-based algorithms. Although their convergence properties have been analyzed in numerous research works, they face convergence problems and lead to slow training procedures in high-dimensional settings. Recently, many advances have been proposed in generative modeling, leading to the emergence of novel approaches such as Variational Autoencoders (VAE) (Kingma and Welling, 2014; Rezende et al., 2014; Ranganath et al., 2014), Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), and Probabilistic and Score-Based Diffusion Models (Ho et al., 2020; Song et al., 2021). These models offer alternative training paradigms that address many of the limitations associated with MCMC-based approaches. Among these, VAE are notable for their ability to learn low dimensional latent variables, while maintaining a clear probabilistic interpretation, enabling both scalable training and meaningful generative modeling.

VAE have been successfully applied in many different contexts such as text generation (Bowman et al., 2016), image generation (Vahdat and Kautz, 2020), image segmentation (Kohl et al., 2018), representation learning (Chen et al., 2017), music generation (Roberts et al., 2018), dimensionality reduction (Kaur et al., 2021), anomaly detection (Park et al., 2022) and state estimation and image reconstruction (Cohen et al., 2022). There are several popular variants of VAE, including IWAE (Burda et al., 2016), β-VAE (Higgins et al., 2017), VQ-VAE (Van Den Oord et al., 2017), and Conditional VAE (Sohn et al., 2015).

Theoretical properties of Variational Inference have only recently been analyzed. For instance, Chérief-Abdellatif et al. (2022) provides generalization bounds on the theoretical reconstruction error using PAC-Bayes theory, while Huggins et al. (2020) offers variational error bounds for posterior mean and uncertainty estimates. Mbacke et al. (2024) extend these analyses by offering statistical guarantees for reconstruction, generation, and regeneration. Furthermore, Tang and Yang (2021) highlights the importance of covariance matrices in Gaussian encoders and decoders, providing variational excess risk bounds for Empirical Bayes Variational Autoencoders (EBVAE) applied to Gaussian models. The authors provide an oracle result for the EBVAE estimator which is crucial to prove

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

the consistency of the estimator. Additionally, theoretical insights into posterior collapse in VAE are provided by (Razavi et al., 2019; Lucas et al., 2019; Wang and Ziyin, 2022), particularly in the context of Linear VAE. More recently, Domke et al. (2023); Kim et al. (2024, 2023) established convergence results for Black-Box Variational Inference (BBVI) under location-scale parameterization.

Most theoretical guarantees for Variational Inference procedures have been established for independent data. However, for sequential data, such as time series, Sequential VAE have been developed. Various model architectures for Sequential VAE have been proposed, including those by (Chung et al., 2015; Fraccaro et al., 2016; Marino et al., 2018; Kim et al., 2020; Campbell et al., 2021; Bayer et al., 2021). However, there are relatively few theoretical results dedicated to dependent or structured data. Notably, (Chagneux et al., 2024) and (Gassiat and Le Corff, 2024) derive variational excess risk bounds for general state space models.

Therefore, obtaining convergence guarantees and non-asymptotic convergence rates for VAE remains mainly an open problem. In this paper, we address this gap by providing non-asymptotic convergence guarantees for VAE, applicable to both independent and sequential data, when using standard Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951) and Adam (Kingma and Ba, 2015) optimization algorithms. More precisely, our contributions are summarized as follows.

- We first establish the smoothness of the expected ELBO which is a crucial step to obtain a convergence rate of $\mathcal{O}(\log n/\sqrt{n})$, where n is the number of iterations of the optimization algorithm, with explicit dependency on the batch size B and the number of variational samples K used at each gradient step (Theorem 3.2).
- We demonstrate that our results apply to both Linear VAE (Section 3.2) and Deep Gaussian VAE (Section 3.3), as well as to several variants, including β-VAE, IWAE, and Sequential VAE.
- We extend our analysis to BBVI (Section 3.5), deriving new convergence guarantees under weaker assumptions, notably without requiring location-scale parameterization.
- We illustrate our convergence results by analyzing empirically the impact of hyperparameters, particularly the regularization parameter β in β-VAE, the number of variational samples in IWAE, as well as the number of layers and activation functions.

2 NOTATION AND BACKGROUND

Notations. In the following, for all distribution μ (resp. probability density p) we write $\mathbb{E}_\mu$ (resp. $\mathbb{E}_p$) the expectation under μ (resp. under p). Given a measurable space $(\mathsf{X}, \mathcal{X})$, where $\mathcal{X}$ is a countably generated σ-algebra, let $\mathsf{M}(\mathsf{X})$ denote the set of all measurable functions defined on $(\mathsf{X}, \mathcal{X})$. We define $\mathcal{F}_{\mathrm{SL}}$ as the set of measurable functions that are both smooth and Lipschitz continuous, and $\mathcal{F}_{\mathrm{b}}$ as the set of bounded measurable functions. The Hadamard product of vectors u and v is denoted by $u \odot v$. For $d \geq 1$, let I_d denote the $d \times d$ identity matrix. The probability density function of the normal distribution with mean μ and covariance matrix Σ is denoted by $\mathcal{N}(\cdot; \mu, \Sigma)$. For probability measures P and Q defined on the same probability space, the Kullback-Leibler (KL) divergence is defined as $\mathrm{D}_{\mathrm{KL}}(Q\|P) = \mathbb{E}_Q[\log(\mathrm{d}Q/\mathrm{d}P)]$. Given a fully connected neural network with N layers, input z, parameters $\theta = \{(W_i, b_i)\}_{i=1}^N$, and activation functions $f = \{f_i\}_{i=1}^N$, the output of the neural network is written:

$$\mathsf{NN}(z; \theta, f, N) = f_N \left(\cdots f_1 \left(W_1 z + b_1 \right) \cdots \right) .$$

We define $\|\theta\|_\infty := \max\{\max_i \|W_i\|, \max_i \|b_i\|\}$, where $\|\cdot\|$ represents the Euclidean norm for vectors and the spectral norm for matrices.

Background. Let $\mathsf{X} \subseteq \mathbb{R}^{d_x}$ and $\mathsf{Z} \subseteq \mathbb{R}^{d_z}$ denote the data space and the latent space, respectively. We consider a dataset $\mathcal{D} = \{x_1, \ldots, x_B\}$ of independent copies of a random variable $x \in \mathsf{X}$ sampled from an unknown probability distribution π. In generative models, particularly those involving latent variables, a classical objective is to maximize the marginal likelihood of the observed data. This marginal likelihood for a given observation x is typically expressed as:

$$\log p_\theta(x) = \log \mathbb{E}_{p_\theta(\cdot|x)} \left[\frac{p_\theta(x, Z)}{p_\theta(Z|x)} \right] ,$$

where $(x, z) \mapsto p_\theta(x, z)$ is the joint likelihood of the observation $x \in \mathsf{X}$ and the latent variable $z \in \mathsf{Z}$. The model is composed of a conditional likelihood $(x, z) \mapsto p_\theta(x|z)$ from a parametric family indexed by $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ (e.g., the weights of a neural network) and of a prior $z \mapsto p_\theta(z)$ (e.g., a standard Gaussian density). Under some simple technical assumptions, by Fisher's identity, we have:

$$\nabla_\theta \log p_\theta(x) = \int \nabla_\theta \log p_\theta(x, z) p_\theta(z|x) \mathrm{d}z . \tag{1}$$

However, in most cases, the conditional density $z \mapsto p_\theta(z|x)$ can only be sampled from approximately using Markov Chain or Sequential Monte Carlo methods, see for instance (Neal, 1993; Andrieu et al., 2010;

Thin et al., 2021; Neklyudov and Welling, 2022). Variational Autoencoders introduce an additional parameter $\phi \in \Phi \subseteq \mathbb{R}^{d_\phi}$ and a family of variational distributions $(x, z) \mapsto q_\phi(z|x)$ to approximate the posterior distribution. Parameters are estimated by maximizing the Evidence Lower Bound (ELBO):

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(\cdot|x)}\left[\log \frac{p_\theta(x, Z)}{q_\phi(Z|x)}\right] =: \mathcal{L}(\theta, \phi; x) \ .$$

The ELBO can be further rewritten as:

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(\cdot|x)}\left[\log p_\theta(x|Z)\right] - \mathrm{D}_{\mathrm{KL}}\left(q_\phi(\cdot|x) \| p\right) \ .$$

The ELBO consists of two terms: (i) the reconstruction term, which quantifies the capability to accurately reconstruct the original input data from its latent representation, and (ii) the regularization term, expressed as the KL divergence, which encourages the latent space of the VAE to follow the prior distribution.

3 THEORETICAL PROPERTIES OF VAE

The expected Evidence Lower Bound over the data distribution π is defined as follows:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{\pi,\phi}\left[\log \frac{p_\theta(X, Z)}{q_\phi(Z|X)}\right] = \mathbb{E}_\pi\left[\mathcal{L}(\theta, \phi; X)\right] \ ,$$

where $\mathbb{E}_{\pi,\phi}$ denotes the expectation under the distribution $\pi(\mathrm{d}x)q_\phi(\mathrm{d}z|x)$. In order to optimize θ and ϕ, one needs to compute the gradients of the ELBO with respect to these parameters. Under classical regularity assumptions, the gradient with respect to θ is given by:

$$\nabla_\theta \mathcal{L}(\theta, \phi) = \mathbb{E}_{\pi,\phi}\left[\nabla_\theta \log p_\theta(X, Z)\right] \ .$$

Computing the gradient with respect to the variational parameters ϕ is more challenging since the inner expectation depends on q_ϕ. There are two common methods for computing this gradient.

The Pathwise Gradient. The reparametrization trick involves expressing the random variable z as a deterministic transform $z = g(\varepsilon, \phi)$, where ε is an auxiliary independent random variable drawn from a known distribution p_ε. Using this trick, the ELBO can be expressed as:

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{p_\varepsilon}\left[\log w_{\theta,\phi}(x, g(\varepsilon, \phi))\right] \ ,$$

where $w_{\theta,\phi}(x, z) = p_\theta(x, z)/q_\phi(z|x)$ the unnormalized importance weights and $\mathbb{E}_{p_\varepsilon}$ is the expectation under the law of ε when $\varepsilon \sim p_\varepsilon$. The pathwise gradient (Kingma and Welling, 2014; Rezende et al., 2014) of the ELBO is given by:

$$\begin{aligned}\nabla_\phi^{\mathsf{P}} \mathcal{L}(\theta, \phi; x) = \mathbb{E}_{p_\varepsilon}\left[\nabla_z \log w_{\theta,\phi}(x, z) \nabla_\phi g(\varepsilon, \phi)\right] \\ - \mathbb{E}_{p_\varepsilon}\left[\nabla_\phi \log q_\phi(g(\varepsilon, \phi)|x)\right] \ .\end{aligned}$$

The Score Function Gradient. Alternatively, the score function gradient, also known as the Reinforce gradient (Glynn, 1990; Williams, 1992; Paisley et al., 2012), can be utilized. Unlike the reparameterization trick, this method does not necessitate reparameterization and is applicable to a wider range of variational distributions. Proposition A.1 provides the form of the score function gradient with respect to ϕ:

$$\nabla_\phi^{\mathsf{S}} \mathcal{L}(\theta, \phi) = \mathbb{E}_{\pi,\phi}\left[\log \frac{p_\theta(X, Z)}{q_\phi(Z|X)} \nabla_\phi \log q_\phi(Z|X)\right] \ .$$

The gradient estimator of the ELBO for a given batch of observations $\{x_i\}_{i=1}^B$, where B is the batch size, with respect to θ and ϕ can then be computed using Monte Carlo sampling as follows:

$$\widehat{\nabla}_{\theta,\phi} \mathcal{L}(\theta, \phi; \{x_i\}_{i=1}^B) = \frac{1}{B}\sum_{i=1}^{B} \frac{1}{K}\sum_{\ell=1}^{K} \tilde{g}_{i,\ell} \ ,$$

where $\tilde{g}_{i,\ell}$ is either the gradient estimator via the score function $\tilde{g}_{i,\ell}^{\mathsf{S}}$ as defined in (5) or the pathwise gradient estimator $\tilde{g}_{i,\ell}^{\mathsf{P}}$ as described in (6).

The pathwise gradient estimator often yields lower-variance estimates than the score function estimator (Miller et al., 2017; Buchholz et al., 2018), but its variance can sometimes exceed that of the score function estimator, especially when the score function correlates with other components of the pathwise estimator. Several methods have been proposed to further reduce variance, such as the Rao-Blackwellization estimator (Ranganath et al., 2014), Control Variates (Liévin et al., 2020), Stop Gradient estimator (Roeder et al., 2017), Quasi-Monte Carlo VAE (Buchholz et al., 2018), and Multi-Level Monte Carlo estimator (Fujisawa and Sato, 2021; He et al., 2022). While our analysis focuses on the convergence rate of score function and pathwise gradient estimators, our convergence results also apply to most of these other methods.

3.1 Convergence Analysis of VAE in the General Setting

In this section, we derive the convergence rates of VAE with both the score function estimator and the pathwise gradient estimator. SGD is a widely used method for training statistical models based on deep architectures. It produces a sequence of parameter estimates as follows: $(\theta_0, \phi_0) \in \Theta \times \Phi$ and for all $k \in \mathbb{N}$,

$$(\theta_{k+1}, \phi_{k+1}) = (\theta_k, \phi_k) + \gamma_{k+1} \widehat{\nabla}_{\theta,\phi} \mathcal{L}(\theta_k, \phi_k; \mathcal{D}_{k+1}) \ , \tag{2}$$

where $\widehat{\nabla}_{\theta,\phi} \mathcal{L}(\theta_k, \phi_k; \mathcal{D}_{k+1})$ denotes an estimator of the gradient, defined in (4), $\mathcal{D}_{k+1}$ is the mini-batch of data used at iteration $k+1$ and for all $k \geq 1$, $\gamma_k > 0$ is the

learning rate. In recent years, several adaptive methods have been proposed, which leverage past gradients to avoid saddle points and handle ill-conditioned problems. Popular adaptive methods include Adagrad (Duchi et al., 2011), RMSProp (Tieleman et al., 2012), Adadelta (Zeiler, 2012), and Adam (Kingma and Ba, 2015). Recent studies have highlighted the superior performance of Adam, a method that iteratively updates the parameters θ and ϕ to effectively maximize the ELBO, as detailed in Algorithm 1.

Consider the following assumptions.

Assumption 1. *There exists $\alpha \in \mathsf{M}(\mathsf{X} \times \mathsf{Z})$ such that for all $\theta \in \Theta$, $\phi \in \Phi$, $x \in \mathsf{X}$ and $z \in \mathsf{Z}$,*

$$\max\{|\log p_\theta(x,z)|, |\log q_\phi(z|x)|\} \leq \alpha(x,z) .$$

Assumption 1 corresponds to bounding the logarithm of the joint probability density $p_\theta(x,z)$ and variational log density $q_\phi(z|x)$. In other words, the probability densities are bounded in log space, which in turn guarantees that the score remains bounded. Although this assumption is typically satisfied in models with a compact state space, compactness is not a necessary condition in this context. This assumption is analyzed in detail in Lemma D.4 for Gaussian distributions.

Assumption 2. *(i) Score Function: there exist M, L_1, and $L_2 \in \mathsf{M}(\mathsf{X} \times \mathsf{Z})$ such that for all $\theta, \theta' \in \Theta$, $\phi, \phi' \in \Phi$, $x \in \mathsf{X}$ and $z \in \mathsf{Z}$,*

$$\max\{\|\nabla_\theta \log p_\theta(x,z)\|, \|\nabla_\phi \log q_\phi(z|x)\|\} \leq M(x,z) ,$$

$$\|\nabla_\theta \log p_\theta(x,z) - \nabla_\theta \log p_{\theta'}(x,z)\| \leq L_1(x,z)\,\|\theta - \theta'\| ,$$

$$\|\nabla_\phi \log q_\phi(z|x) - \nabla_\phi \log q_{\phi'}(z|x)\| \leq L_2(x,z)\,\|\phi - \phi'\| .$$

(ii) Pathwise Gradient: there exist M, L_p, and $L_q \in \mathsf{M}(\mathsf{X} \times \mathsf{Z})$ such that for all $\theta, \theta' \in \Theta$, $\phi, \phi' \in \Phi$, $x \in \mathsf{X}$, $\varepsilon \in \mathsf{Z}$, writing $z = g(\varepsilon, \phi)$ and $z' = g(\varepsilon, \phi')$,

$$\begin{aligned}
&\max\{\|\nabla_{z,\theta} \log p_\theta(x,z)\|, \|\nabla_z \log q_\phi(z|x)\|\} \leq M(x,\varepsilon) ,\\
&\quad \|\nabla_{z,\theta} \log p_\theta(x,z) - \nabla_{z,\theta} \log p_{\theta'}(x,z')\| \\
&\qquad\qquad \leq L_p(x,\varepsilon)\,(\|\theta - \theta'\| + \|z - z'\|) ,\\
&\quad \|\nabla_z \log q_\phi(z|x) - \nabla_z \log q_{\phi'}(z'|x)\| \\
&\qquad\qquad \leq L_q(x,\varepsilon)\,(\|\phi - \phi'\| + \|z - z'\|) .
\end{aligned}$$

Assumption 2 is divided into two parts: (i) presents the regularity conditions needed for convergence when the gradient is computed via the score function, and (ii) specifies the assumptions required for convergence when using the pathwise gradient. Assumption 2(i) concerns the boundedness and Lipschitz continuity of the score functions associated with the distributions $p_\theta(x,z)$ and $q_\phi(z|x)$. This assumption is standard in the literature on Reinforcement Learning (RL) and Maximum Likelihood Estimation (MLE). In RL, assumptions about the boundedness and Lipschitz continuity of the score functions of the policy are commonly used to prove the convergence rates of policy gradient algorithms (Papini et al., 2018; Shen et al., 2019; Fallah et al., 2021; Liu et al., 2020) and actor-critic algorithms (Castro and Meir, 2010; Qiu et al., 2021). Importantly, the bounds on the score function in prior works are independent of state variables, requiring a compact state space. In contrast, our approach considers variable dependence, relaxing this assumption and allowing for unbounded state spaces. Furthermore, we impose additional assumptions on the score functions associated with the variational distribution. In MLE, these assumptions are used to establish the convergence of recursive MLE in Non-Linear State-Space Models (Tadić and Doucet, 2020). However, these assumptions are typically formulated in the original space rather than log space. Assumption 2(ii) is similar to Assumption 2(i), with the key difference being that it considers the gradient with respect to z instead of ϕ for the variational log density. Additionally, it accounts for the joint gradient with respect to both z and θ for the conditional decoder log density.

Assumption 3. *There exist M_g and $L_g \in \mathsf{M}(\mathsf{X} \times \mathsf{Z})$ such that for all $\theta, \theta' \in \Theta$, $\phi, \phi' \in \Phi$, $x \in \mathsf{X}$ and $\varepsilon \in \mathsf{Z}$,*

$$\|\nabla_\phi g(\varepsilon, \phi)\| \leq M_g(x,\varepsilon) ,$$

$$\|\nabla_\phi g(\varepsilon, \phi) - \nabla_\phi g(\varepsilon, \phi')\| \leq L_g(x,\varepsilon)\,\|\phi - \phi'\| .$$

Assumption 3 concerns the boundedness and smoothness of the reparameterization trick function g. It is important to note that under the boundedness of the gradients, the smoothness of g and the Lipchitz condition of $\nabla_z \log q_\phi(z|x)$ are equivalent to the Lipchitz condition of the score function associated with the variational density. We show that these assumptions hold in both Linear and Deep Gaussian VAE. Under these assumptions, we first establish the smoothness of the expected ELBO, in both the score function and pathwise gradient cases, which is a critical step to prove the convergence rate.

Proposition 3.1. *For all $\theta, \theta' \in \Theta$ and $\phi, \phi' \in \Phi$,*

(i) Score Function: under Assumptions 1 and 2(i),

$$\left\|\nabla^{\mathsf{S}}_{\theta,\phi}\mathcal{L}(\theta,\phi) - \nabla^{\mathsf{S}}_{\theta,\phi}\mathcal{L}(\theta',\phi')\right\| \leq L^{\mathsf{S}}\,\|(\theta,\phi) - (\theta',\phi')\| ,$$

(ii) Pathwise Gradient: under Assumptions 2(ii) and 3,

$$\|\nabla^{\mathsf{P}}_{\theta,\phi}\mathcal{L}(\theta,\phi) - \nabla^{\mathsf{P}}_{\theta,\phi}\mathcal{L}(\theta',\phi')\| \leq L^{\mathsf{P}}\|(\theta,\phi) - (\theta',\phi')\|,$$

where L^{S} and L^{P} are defined in Lemma B.1 and B.2 respectively.

We establish in Lemma D.6 that these smoothness constants are well-defined and finite in the Gaussian case. With these results established, we now derive the convergence rates for VAE using both gradient estimators.

Theorem 3.2. *Let $(\theta_n, \phi_n) \in \Theta \times \Phi$ be the n-th iterate of the recursion (2), where $\gamma_n = C_\gamma n^{-1/2}$ with $C_\gamma > 0$. Assume that for $m \in \{\mathsf{S}, \mathsf{P}\}$, $\sigma_m^2 = \sup_{\theta\in\Theta,\phi\in\Phi} \mathbb{E}_\pi[\mathbb{E}_{q_\phi(\cdot|X_i)}[\|\tilde{g}_{i,\ell}^m - \nabla_{\theta,\phi}^m \mathcal{L}(\theta,\phi)\|^2]] < +\infty$. For all $n \geq 1$, let $R \in \{0,\dots,n\}$ be a uniformly distributed random variable. Then,*

(i) Score Function: under Assumptions 1 and 2(i), and for $C_\gamma \leq 1/L^{\mathsf{S}}$,

$$\mathbb{E}\left[\left\|\nabla_{\theta,\phi}^{\mathsf{S}} \mathcal{L}\left(\theta_R, \phi_R\right)\right\|^2\right] \leq \frac{2\mathcal{L}^* + L^{\mathsf{S}} C_\gamma \sigma_{\mathsf{S}}^2 \log n/(BK)}{C_\gamma \sqrt{n}} ,$$

(ii) Pathwise Gradient: under Assumptions 2(ii) and 3, and for $C_\gamma \leq 1/L^{\mathsf{P}}$,

$$\mathbb{E}\left[\left\|\nabla_{\theta,\phi}^{\mathsf{P}} \mathcal{L}\left(\theta_R, \phi_R\right)\right\|^2\right] \leq \frac{2\mathcal{L}^* + L^{\mathsf{P}} C_\gamma \sigma_{\mathsf{P}}^2 \log n/(BK)}{C_\gamma \sqrt{n}} ,$$

where $\mathcal{L}^* = \mathcal{L}\left(\theta^*, \phi^*\right) - \mathcal{L}\left(\theta_0, \phi_0\right)$.

Theorem 3.2 provides the classical convergence rate of $\mathcal{O}\left(\log n/\sqrt{n}\right)$ for non-convex problems. This rate indicates that increasing the batch size B and the number of samples K from the variational distribution can improve convergence by reducing the second term. However, larger values of B and K also increase computational costs. Therefore, it is crucial to balance the convergence rate with computational efficiency by choosing appropriate values for B and K. In practice, it is common to use a larger batch size B while setting $K = 1$. Additionally, we observe that high variance results in slower convergence, making the pathwise gradient estimator more favorable than the score function estimator. Theorem 3.3 provides the convergence rate of Adam, as defined in Algorithm 1.

Theorem 3.3. *Let $(\theta_n, \phi_n) \in \Theta\times\Phi$ be the n-th iterate of the recursion in Algorithm 1, where $\gamma_n = C_\gamma n^{-1/2}$ with $C_\gamma > 0$. Suppose that $\beta_1 < \sqrt{\beta_2} < 1$ and that the assumptions of Theorem 3.2 hold. Then, for $m \in \{\mathsf{S}, \mathsf{P}\}$,*

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

where $\mathcal{L}^* = \mathcal{L}\left(\theta^*, \phi^*\right) - \mathcal{L}\left(\theta_0, \phi_0\right)$, $d^* = d_\theta + d_\phi$ is the total dimension of the parameters, and L^{S} and L^{P} are the smoothness constants for the score function and pathwise gradient cases, respectively.

Theorem 3.3 provides a convergence rate similar to that of SGD, $\mathcal{O}\left(\log n/\sqrt{n}\right)$ but with an additional factor of the total dimension d^*, reflecting the impact of the adaptive step sizes. In practice, Adam typically uses $\beta_1 = 0.9$ and $\beta_2 = 0.999$ (Kingma and Ba, 2015; Zaheer et al., 2018; Reddi et al., 2018), which satisfy the condition $\beta_1 < \sqrt{\beta_2} < 1$.

3.2 Linear Gaussian VAE

We consider the following VAE model: for all $x \in \mathsf{X}$ and $z \in \mathsf{Z}$,

$$\begin{aligned} p_\theta(x|z) &= \mathcal{N}(x; W_1 z + b_1, c^2 \mathrm{I}_{d_x}), \\ q_\phi(z|x) &= \mathcal{N}(z; W_2 x + b_2, D), \end{aligned} \tag{3}$$

where $\theta = (W_1, b_1, c^2) \in \mathbb{R}^{d_x \times d_z} \times \mathbb{R}^{d_x} \times \mathbb{R}_+^*$ and $\phi = (W_2, b_2, D) \in \mathbb{R}^{d_z \times d_x} \times \mathbb{R}^{d_z} \times \mathbb{R}^{d_z \times d_z}$. The matrix D is a diagonal covariance matrix and serves as an amortized variance for each input point. It is sufficient to achieve the global optimum of this model (Lucas et al., 2019). Conditionally on x, the output of the Linear VAE follows a Gaussian distribution with mean $W_1(W_2 x + b_2) + b_1$ and variance $W_1 D W_1^\top$.

While analytic solutions for deep latent models are generally not available, the Linear VAE provides analytic solutions for optimal parameters, allowing us to gain insights into various phenomena associated with VAE training. Proposition C.1 provides an analytical form of the expected ELBO for the Linear VAE, which can be used to analyze the convergence rate. The following corollary derives the convergence rate for the Linear VAE.

Corollary 3.4. *Consider the Linear Gaussian VAE defined in (3) with $\theta = (W_1, b_1)$ and $\phi = (W_2, b_2, D)$ and let $c_D > 0$ such that $\lambda_{\min}(D) \geq c_D$. Assume that the inputs have bounded second moments and there exists some constant a such that for all $\theta \in \Theta$ and $\phi \in \Phi$,*

$$\|\theta\|_\infty + \|\phi\|_\infty \leq a.$$

Let $(\theta_n, \phi_n) \in \Theta\times\Phi$ be the n-th iterate of the recursion in Algorithm 1, where $\gamma_n = C_\gamma n^{-1/2}$ with $C_\gamma > 0$. Assume that $\beta_1 < \sqrt{\beta_2} < 1$. For all $n \geq 1$, let $R \in \{0,\dots,n\}$ be a uniformly distributed random variable. Then,

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

where $\mathcal{L}^* = \mathcal{L}\left(\theta^*, \phi^*\right) - \mathcal{L}\left(\theta_0, \phi_0\right)$.

In Lucas et al. (2019), it is shown that the ELBO objective for a Linear VAE does not introduce any local maxima beyond the marginal loglikelihood. Corollary 3.4 indicates a convergence rate of $\mathcal{O}(\log n/\sqrt{n})$ with respect to the marginal loglikelihood.

3.3 Deep Gaussian VAE

We show that the assumptions of our main results hold for classical architectures of neural networks and discuss the choice of hyperparameters for this architecture. The deep Gaussian VAE consists of a decoder and an encoder such that $p_\theta(x|z) = \mathcal{N}(x; G_\theta(z), c^2 \mathrm{I}_{d_x})$ and $q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \Sigma_\phi(x))$. Consider the following neural network formulations for the encoder and decoder:

$$\mathcal{F}_G = \Big\{ G_\theta(z) = \mathsf{NN}(z; \theta, f, N_{dd}),\, \theta = \{W_\ell, b_\ell\}_{\ell=1}^{N_{dd}},\\ \sigma_\ell \in \mathcal{F}_{\mathrm{SL}}, \ell = 1, \dots, N_{dd},\ \|G_\theta(z)\| \le C_G \Big\},$$

and

$$\mathcal{F}_{\mu,\Sigma} = \Big\{ (\mu_\phi(x), \Sigma_\phi(x)) = \mathsf{NN}(x; \phi, f, N_{ed}),\\ \phi = \{W_\ell, b_\ell\}_{\ell=1}^{N_{ed}},\ f_\ell \in \mathcal{F}_{\mathrm{SL}}, \ell = 1, \dots, N_{ed},\\ \|\mu_\phi(x)\| \le C_\mu,\, \lambda_{\min}(\Sigma_\phi(x)) \ge c_\Sigma,\, \|\Sigma_\phi(x)\| \le C_\Sigma \Big\}.$$

In $\mathcal{F}_G$ and $\mathcal{F}_{\mu,\Sigma}$, f represents an activation function. Specifically, Proposition D.2 suggests that f can be chosen from a set that includes the sigmoid, Tanh, softplus (Glorot et al., 2011), or CELU (Clevert et al., 2016; Barron, 2017) defined in Appendix D.1.

A common approach to satisfy the condition on the minimum eigenvalue of the covariance matrix is to add a regularization term. However, in practice, since the encoder outputs the log variance, it is more efficient to impose a lower bound directly on the log variance. This ensures that the eigenvalues of the covariance matrix remain bounded from below, thus maintaining numerical stability. Furthermore, to ensure that $\|\mu_\phi(x)\| \le C_\mu$ and $\|\Sigma_\phi(x)\| \le C_\Sigma$, it is essential for the activation function in the final layer to be bounded. While sigmoid and tanh are commonly used for their bounded nature, they may not be suitable when linearity is required, as often needed for the encoder. An alternative is Hard Tanh (Tang and Yang, 2021), which clips the identity function, achieving both linearity and boundedness, though it lacks differentiability. To address this, we propose a generalized soft-clipping activation function, an extension of (Klimek and Perelstein, 2020), designed to be approximately linear within a specified interval. We also provide a smoothness analysis in Proposition 3.5.

Proposition 3.5. *For $s_1, s_2 \in \mathbb{R}$ with $s_1 \le s_2$ and $s \in \mathbb{R}_+^*$, the generalized soft-clipping activation function defined by*

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

is bounded between s_1 and s_2, and is Lipschitz continuous and smooth.

The generalized soft-clipping activation function is approximately linear within the interval (s_1, s_2). The parameter s plays a crucial role in determining the shape and sharpness of the transition between s_1 and s_2.

Theorem 3.6. *Let $G_\theta(z) \in \mathcal{F}_G$ and $(\mu_\phi(x), \Sigma_\phi(x)) \in \mathcal{F}_{\mu,\Sigma}$ for all $(x, z) \in \mathsf{X} \times \mathsf{Z}$. Assume that there exists $C_{rec} \in \mathsf{M}(\mathsf{X} \times \mathsf{Z})$ such that $\|x - G_\theta(z)\| \le C_{rec}(x, z)$ for all $\theta \in \Theta$ and $(x, z) \in \mathsf{X} \times \mathsf{Z}$. Assume also that the data distribution π has a finite fourth moment, and that there exists some constant a such that for all $\theta \in \Theta$ and $\phi \in \Phi$,*

$$\|\theta\|_\infty + \|\phi\|_\infty \le a\ .$$

Let $(\theta_n, \phi_n) \in \Theta \times \Phi$ be the n-th iterate of the recursion in Algorithm 1, where $\gamma_n = C_\gamma n^{-1/2}$ with $C_\gamma > 0$. Assume that $\beta_1 < \sqrt{\beta_2} < 1$. For all $n \ge 1$, let $R \in \{0, \dots, n\}$ be a uniformly distributed random variable. Then, for $m \in \{\mathsf{S}, \mathsf{P}\}$,

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

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 $C^{\mathsf{S}} = d_z^2 N_{max} a^{2(N_{max}-1)}$ and $C^{\mathsf{P}} = d_z N_{total} a^{2(N_{total}-1)}$. Here, $N_{max} = \max\{N_{ed}, N_{dd}\}$ denotes the maximum number of layers, while $N_{total} = N_{ed} + N_{dd}$ represents the total number of layers in the encoder and decoder.

Theorem 3.6 provides the convergence rate of $\mathcal{O}(\log n/\sqrt{n})$ for deep Gaussian VAE, which is commonly used in practice. In deep learning, it is standard practice to initialize weights using a distribution scaled by $\mathcal{O}(1/\sqrt{d})$, such as $\mathcal{N}(0, 1/\sqrt{d})$ or $\mathcal{U}(-1/\sqrt{d}, 1/\sqrt{d})$ (Glorot and Bengio, 2010; He et al., 2015; Li and Yuan, 2017). This initialization ensures that the spectral norm of the resulting weights matrix is typically $\mathcal{O}(1)$ (Rudelson and Vershynin, 2010). Consequently, assuming the compactness of the parameter space is well-justified. This assumption is also used to derive the excess risk of VAE (Tang and Yang, 2021).

Our choice of activation functions to achieve the convergence rate is reasonable and does not deviate significantly from commonly used activation functions. Although our results do not directly apply to the ReLU activation function, experimental results demonstrate that similar convergence rates can still be achieved using ReLU. Furthermore, since G_θ is bounded, the assumption regarding the reconstruction error C_{rec} can be easily verified if the inputs x are also bounded.

In our convergence rate analysis, the smoothness constant depends on the number of layers N and grows exponentially with N. Specifically, the leading term in the smoothness constant is of the form $N \times a^{2(N-1)}$. This growing exponential factor also appears in the

Lipschitz constant of neural networks, where the term is $\sqrt{N} \times a^{N-1}$ (Virmaux and Scaman, 2018; Combettes and Pesquet, 2020; Tang and Yang, 2021). As the number of hidden layers increases, both the smoothness constant and the total parameter dimension d^* grow, leading to a greater number of iterations required for convergence. Damm et al. (2023) shows that for Deep Gaussian VAE, the ELBO at stationary points is equal to the sum of the negative entropy of the prior distribution, the expected negative entropy of the observable distribution, and the average entropy of the variational distributions. Consequently, the ELBO in Deep Gaussian VAE converges to a sum of entropies at a rate of $\mathcal{O}(\log n/\sqrt{n})$.

3.4 Some Variants of VAE

3.4.1 β-VAE

β-VAE (Higgins et al., 2017) is a variant of VAE introducing a parameter β to control the trade-off between the reconstruction term and the regularization of the latent space. The ELBO for β-VAE is given by:

$$\mathcal{L}_\beta(\theta, \phi; x) = \mathbb{E}_{q_\phi(\cdot|x)} \left[\log p_\theta(x|Z)\right] - \beta D_{\mathrm{KL}}(q_\phi(\cdot|x)||p) \ ,$$

where the Lagrangian multiplier β is considered as a hyperparameter. The small values of β force decoders to use the latent variables, but this comes at the cost of a poor ELBO. The role of β invites a natural comparison to the parameter c^2 in the objective of the standard Gaussian VAE. Setting c^2 small in the standard VAE corresponds to setting β small in β-VAE. For a given β, one can find a corresponding c^2 (and a learning rate) such that the gradient updates to the network parameters are identical (Lucas et al., 2019).

Given that β plays a role analogous to $1/c^2$, when applying Theorems 3.3 and 3.6, we observe that if $\beta < \infty$, β-VAE with the same architecture as in Theorem 3.6, converges to a critical point of the expected ELBO at a rate of $\mathcal{O}(\log n/\sqrt{n})$. The smaller the value of β, the faster the convergence, due to the analogous role of β and $1/c^2$. Selecting a sufficiently small β leads to faster convergence and can help prevent posterior collapse (Wang and Ziyin, 2022).

3.4.2 IWAE

The Importance Weighted Autoencoder (IWAE) (Burda et al., 2016) is another extension of the VAE that incorporates importance weighting to obtain a tighter ELBO. The IWAE objective function is:

$$\mathcal{L}_K^{\mathsf{IS}}(\theta, \phi) = \mathbb{E}_\pi \left[\mathbb{E}_{q_\phi^{\otimes K}(\cdot|X)} \left[\log \frac{1}{K} \sum_{\ell=1}^{K} \frac{p_\theta(X, Z^{(\ell)})}{q_\phi(Z^{(\ell)}|X)} \right] \right] \ ,$$

where K corresponds to the number of samples drawn from the variational posterior distribution.

Assumption 4. *There exist $\alpha^-, \alpha^+ \in \mathsf{M}(\mathsf{X} \times \mathsf{Z})$ such that for all $\theta \in \Theta$, $\phi \in \Phi$, $x \in \mathsf{X}$, $\varepsilon, z \in \mathsf{Z}$ where $z = g(\varepsilon, \phi)$,*

$$\alpha^-(x, \varepsilon) \leq \max\{p_\theta(x, z), q_\phi(z|x)\} \leq \alpha^+(x, \varepsilon) \ .$$

Assumption 4 states the boundedness of both the joint probability density function and the variational density function. Given the existence of the reparametrization trick, Assumptions 1 and 4 are equivalent, and the bound is verified with $\alpha(x, z) = \max\{|\log \alpha^-(x, \varepsilon)|, |\log \alpha^+(x, \varepsilon)|\}$.

Theorem 3.7. *Let Assumptions 2(ii)-4 hold. Let $(\theta_n, \phi_n) \in \Theta \times \Phi$ be the n-th iterate of the recursion in Algorithm 1 where $\mathcal{L}$ is the IWAE objective, and $\gamma_n = C_\gamma n^{-1/2}$ with $C_\gamma > 0$. Assume that $\beta_1 < \sqrt{\beta_2} < 1$. For all $n \geq 1$, let $R \in \{0, \ldots, n\}$ be a uniformly distributed random variable. Then,*

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

where $d^* = d_\theta + d_\phi$, and L_K is as defined in (11).

We achieve a convergence rate similar to that of VAE and β-VAE. In particular, as K increases, the convergence rate improves and becomes nearly inversely proportional to K.

Link with Signal to Noise ratio. In Rainforth et al. (2018), the authors propose to measure the relative accuracy of the gradient estimates using the Signal to Noise ratio (SNR), i.e. the absolute value of the expected estimate of the gradient scaled by its standard deviation. They highlight that a low SNR is problematic, as it indicates that gradient estimates are dominated by noise. Our results align with (Rainforth et al., 2018, Theorem 1), which establishes that the SNR scales as $\sqrt{BK}$ for θ and $\sqrt{B/K}$ for ϕ, where B is the batch size. This means that increasing K independently of B might lead to vanishing SNR and poor gradient estimates for ϕ and motivate adaptive choices of K with respect to B.

3.5 Extension to Variational Inference

Black-Box Variational Inference (BBVI) is typically formulated as the maximization of the following objective function (Ranganath et al., 2014):

$$\mathcal{L}^{\mathsf{BBVI}}(\phi; x) = \mathbb{E}_{q_\phi(\cdot|x)} \left[\log p(x, z) - \log q_\phi(z|x)\right] \ ,$$

where q_ϕ is the variational distribution with parameter ϕ. As a special case of VAE, BBVI optimizes only ϕ, not θ. Existing convergence results for BBVI have been established in both convex (Domke, 2020; Kim et al., 2024) and non-convex settings (Domke et al.,

2023; Kim et al., 2023). These results typically rely on smoothness assumptions about the ELBO, which are often derived under linear parameterization or the location-scale family. The following corollary extends these results by removing the location-scale assumption, making them applicable to a broader class of reparameterization families.

Corollary 3.8. *Assume the following conditions hold. There exist M, M_g, L_g, L_p, and $L_q \in \mathsf{M}(\mathsf{X} \times \mathsf{Z})$ such that for all $\phi \in \Phi$, $x \in \mathsf{X}$ and $\varepsilon \in \mathsf{Z}$ with $z = g(\varepsilon, \phi)$,*

(i) $z \mapsto \log p(x, z)$ *is* $L_p(x, \varepsilon)$*-smooth.*

(ii) $z \mapsto \log q_\phi(z|x)$ *is* $M(x, \varepsilon)$*-Lipchitz and* $L_q(x, \varepsilon)$*-smooth.*

(iii) $\phi \mapsto g(\phi, \varepsilon)$ *is* $M_g(x, \varepsilon)$*-Lipchitz and* $L_g(x, \varepsilon)$*-smooth.*

Then, $\phi \mapsto \mathcal{L}^{\mathsf{BBVI}}(\phi)$ *is* L^{BBVI}*-smooth, where the smoothness constant* L^{BBVI} *is given by* (12).

Corollary 3.8 establishes the smoothness of the ELBO, which is crucial for achieving the convergence rate. Our assumptions are less restrictive than those in Domke (2020); Kim et al. (2023) and do not depend on a specific reparameterization trick, unlike prior works that often assume a location-scale parameterization. A detailed comparison is provided in Appendix F. This allows us to attain a convergence rate of $\mathcal{O}(\log n/\sqrt{n})$, as in Theorem 3.2.

4 EXPERIMENTS

In this section, we illustrate our theoretical results in the context of deep Gaussian VAE. The experiments were conducted using PyTorch (Paszke et al., 2017), and the source code can be found here[1].

Dataset and Model. We conduct our experiments on the CelebA dataset (Liu et al., 2018) and use a Convolutional Neural Network (CNN) architecture with Rectified Linear Unit (ReLU) and generalized soft-clipping activation functions for both the encoder and decoder networks. The latent space dimension is set to 100. We estimate the log-likelihood using the VAE, β-VAE, and IWAE models, all of which are trained for 100 epochs. Training is performed with Adam optimizer and learning rate decay defined as $\gamma_n = C_\gamma/\sqrt{n}$, where $C_\gamma = 0.001$. The momentum parameters are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the regularization parameter δ is fixed at 10^{-8}. Note that while all figures are plotted with respect to epochs, n here denotes the number of gradient updates. Additional details are provided in Appendix H.

For the first experiment, we illustrate our convergence results of the standard VAE with our choice of activation functions. Figure 1 shows the squared norm of the gradients $\|\nabla\mathcal{L}(\theta_n, \phi_n)\|^2$ and the Negative ELBO on the test dataset for both ReLU and the generalized soft-clipping activation function with various values of s. We observe a similar convergence rate for all values of s. However, selecting an appropriate value of s is crucial to achieve optimal convergence. Theoretically, smaller values of s should result in slower convergence rates, but in practice, choosing a very large s can introduce numerical instabilities due to the exponential term in the generalized soft-clipping function. Then, $s = 5$ appears to be a reasonable choice, balancing convergence rate and numerical stability.

Moreover, our choice of activation functions leads to improved convergence rates compared to the standard VAE with ReLU. While our theoretical analysis does not directly apply to ReLU, experimental results indicate that similar convergence performance can be achieved.

Figure 2 illustrates the squared norm of the gradients $\|\nabla\mathcal{L}(\theta_n, \phi_n)\|^2$ for both the β-VAE and IWAE models, evaluated across different values of β and K. The standard VAE is a special case of these models, cor-

[1] https://github.com/SobihanSurendran/VAE-Convergence-Guarantees

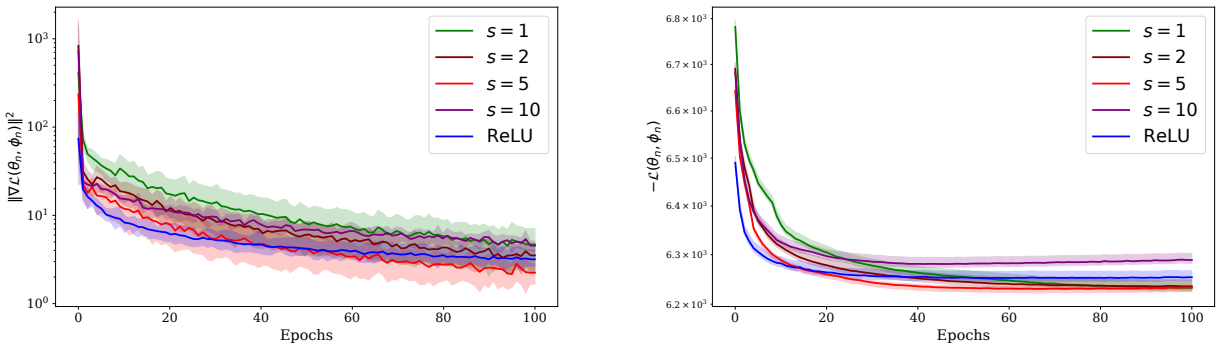

Figure 1: Squared norm of gradients and Negative ELBO on the test set of the CelebA for VAE trained with Adam and generalized soft-clipping activation function. Bold lines represent the mean over 5 independent runs.

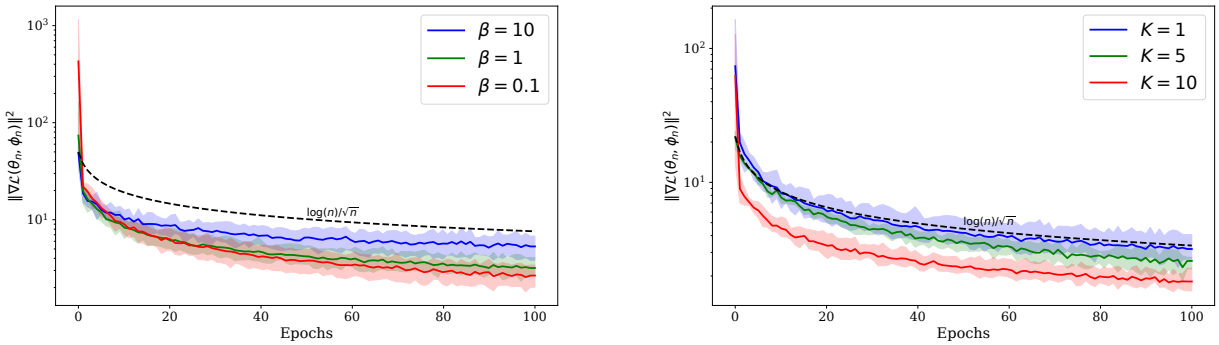

Figure 2: $\|\nabla\mathcal{L}(\theta_n, \phi_n)\|^2$ in β-VAE (on the left) and IWAE (on the right) trained with Adam. Bold lines represent the mean over 5 independent runs. The dashed curves correspond to the expected convergence rate $\mathcal{O}(\log n/\sqrt{n})$.

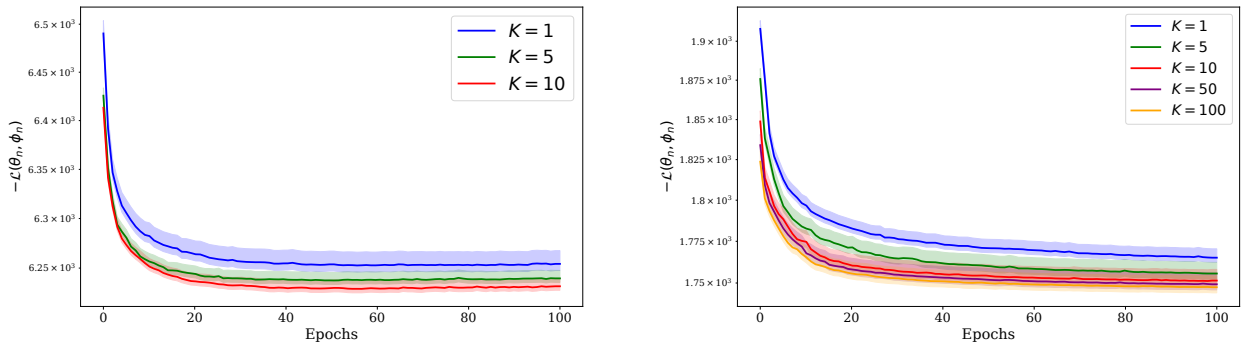

Figure 3: Negative ELBO in IWAE on the test set of the CelebA (on the left) and CIFAR-100 (on the right) trained with Adam. Bold lines represent the mean over 5 independent runs.

responding to $\beta = 1$ in the β-VAE and $K = 1$ in the IWAE. As expected, we observe that for the β-VAE, smaller values of β lead to faster convergence. Similarly, for the IWAE, increasing the value of K results in faster convergence. However, beyond a certain threshold, neither the gradient norm nor the objective value improves significantly (Figure 3), instead incurring unnecessary computational cost. This behavior aligns with the earlier discussion on the Signal to Noise Ratio, see Section 3.4.2.

While the objective is for the gradient norm to converge to zero, it should not do so too quickly. If the gradient becomes too small as K increases, the training procedure is prone to yield poor results regarding ϕ, thereby limiting improvements in θ. This highlights the need for careful selection of K. It is also crucial to consider the other hyperparameters, as they impact the convergence rate. When adjusting K, it is important to adjust the other parameters accordingly to optimize the convergence rate. One approach is to gradually increase K until a suitable threshold is reached (Surendran et al., 2024). Alternatively, Variational Rényi IWAE can be used, ensuring the SNR scales as $\sqrt{BK}$ for both θ and ϕ (Daudel et al., 2023).

5 DISCUSSION

This paper provides a non-asymptotic convergence analysis of VAE trained using both SGD and Adam algorithms. We derive a convergence rate of $\mathcal{O}(\log n/\sqrt{n})$, applicable to Linear VAE, Deep Gaussian VAE, and several VAE variants. Our analysis indicates that smaller values of β in β-VAE and the large values of K in IWAE lead to faster convergence rates. However, increasing K independently of the batch size B can lead to vanishing SNR and poor gradient estimates for ϕ, thereby hindering the learning of θ.

For Deep Gaussian VAE, we introduce a generalized soft-clipping activation function that supports our theoretical claims and yields improved convergence rates compared to standard VAE using ReLU. Although our analysis does not directly address ReLU, empirical results suggest that similar convergence rates can still be achieved. A promising direction for future work is to explore alternative distributions for the encoder and decoder, along with different deep architectures. Additionally, extending our results to Variational Rényi IWAE would be a valuable direction for future work.

Acknowledgements

The PhD of Sobihan Surendran was funded by the Paris Region PhD Fellowship Program of Région Ile-de-France. We would like to thank SCAI (Sorbonne Center for Artificial Intelligence) for providing the computing clusters. We also express our gratitude to the reviewers for their insightful comments and suggestions, which have helped improve this paper.

References

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(3):269–342, 2010.

J. T. Barron. Continuously differentiable exponential linear units. *arXiv preprint arXiv:1704.07483*, 2017.

J. Bayer, M. Soelch, A. Mirchev, B. Kayalibay, and P. van der Smagt. Mind the gap when conditioning amortised inference in sequential latent-variable models. In *International Conference on Learning Representations*, 2021.

S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, 2016.

A. Buchholz, F. Wenzel, and S. Mandt. Quasi-Monte Carlo variational inference. In *International Conference on Machine Learning*, pages 668–677. PMLR, 2018.

Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

A. Campbell, Y. Shi, T. Rainforth, and A. Doucet. Online variational filtering and parameter learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 18633–18645, 2021.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

G. Cardoso, Y. J. El Idrissi, S. Le Corff, E. Moulines, and J. Olsson. State and parameter learning with PaRIS particle Gibbs. In *International Conference on Machine Learning*, pages 3625–3675. PMLR, 2023.

D. D. Castro and R. Meir. A convergent online single time scale actor critic algorithm. *Journal of Machine Learning Research*, 11:367–410, 2010.

M. Chagneux, É. Gassiat, P. Gloaguen, and S. Le Corff. Additive smoothing error in backward variational inference for general state-space models. *Journal of Machine Learning Research*, 25(28):1–33, 2024.

V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21 (2):572–596, 2011.

X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. Variational lossy autoencoder. In *International Conference on Learning Representations*, 2017.

B.-E. Chérief-Abdellatif, Y. Shi, A. Doucet, and B. Guedj. On PAC-Bayesian reconstruction guarantees for VAEs. In *International Conference on Artificial Intelligence and Statistics*, pages 3066–3079. PMLR, 2022.

P. Chigansky and R. Liptser. Stability of nonlinear filters in nonmixing case. *The Annals of Applied Probability*, 14(4):2038–2056, 2004.

J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In *International Conference on Learning Representations*, 2016.

M. Cohen, G. Quispe, S. Le Corff, C. Ollion, and E. Moulines. Diffusion bridges vector quantized variational autoencoders. In *International Conference on Machine Learning*, 2022.

P. L. Combettes and J.-C. Pesquet. Lipschitz certificates for layered network structures driven by averaged activation operators. *SIAM Journal on Mathematics of Data Science*, 2(2):529–557, 2020.

B. Dai, Y. Wang, J. Aston, G. Hua, and D. Wipf. Connections with robust pca and the role of emergent sparsity in variational autoencoder models. *Journal of Machine Learning Research*, 19(41):1–42, 2018.

S. Damm, D. Forster, D. Velychko, Z. Dai, A. Fischer, and J. Lücke. The elbo of variational autoencoders converges to a sum of entropies. In *International Conference on Artificial Intelligence and Statistics*, pages 3931–3960. PMLR, 2023.

K. Daudel, J. Benton, Y. Shi, and A. Doucet. Alpha-divergence variational inference meets importance weighted auto-encoders: Methodology and asymptotics. *Journal of Machine Learning Research*, 24 (243):1–83, 2023.

J. Domke. Provable smoothness guarantees for black-box variational inference. In *International Con-*

ference on Machine Learning, pages 2587–2596. PMLR, 2020.

J. Domke, R. Gower, and G. Garrigos. Provable convergence guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

R. Douc, G. Fort, E. Moulines, and P. Priouret. Forgetting the initial distribution for hidden Markov models. *Stochastic processes and their applications*, 119(4):1235–1256, 2009.

R. Douc, A. Garivier, E. Moulines, and J. Olsson. Sequential Monte Carlo smoothing for general state space hidden Markov models. *Annals of Applied Probability*, 21(6):2109–2145, 2011.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

A. Fallah, K. Georgiev, A. Mokhtari, and A. Ozdaglar. On the convergence theory of debiased model-agnostic meta-reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 3096–3107, 2021.

M. Fraccaro, S. K. Sønderby, U. Paquet, and O. Winther. Sequential neural models with stochastic layers. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

M. Fujisawa and I. Sato. Multilevel Monte Carlo variational inference. *Journal of Machine Learning Research*, 22(278):1–44, 2021.

E. Gassiat and S. Le Corff. Variational excess risk bound for general state space models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

P. Gloaguen, S. Le Corff, and J. Olsson. A pseudo-marginal sequential Monte Carlo online smoothing algorithm. *Bernoulli*, 28(4):2606–2633, 2022.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

P. W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Z. He, Z. Xu, and X. Wang. Unbiased mlmc-based variational bayes for likelihood-free inference. *SIAM Journal on Scientific Computing*, 44 (4):A1884–A1910, 2022.

I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

J. Huggins, M. Kasprzak, T. Campbell, and T. Broderick. Validated variational inference via practical posterior error bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 1792–1802. PMLR, 2020.

N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin. On particle methods for parameter estimation in state-space models. *Statistical Science*, 30(3):328–351, 2015.

M. Karl, M. Soelch, J. Bayer, and P. Van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. In *International Conference on Learning Representations*, 2017.

D. Kaur, S. N. Islam, and M. A. Mahmud. A variational autoencoder-based dimensionality reduction technique for generation forecasting in cyber-physical smart grids. In *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6. IEEE, 2021.

G.-H. Kim, Y. Jang, H. Yang, and K.-E. Kim. Variational inference for sequential data with future likelihood estimates. In *International Conference on Machine Learning*, pages 5296–5305. PMLR, 2020.

K. Kim, J. Oh, K. Wu, Y. Ma, and J. Gardner. On the convergence of black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

K. Kim, Y. Ma, and J. Gardner. Linear convergence of black-box variational inference: Should we stick the landing? In *International Conference on Artificial Intelligence and Statistics*, pages 235–243. PMLR, 2024.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

M. Klimek and M. Perelstein. Neural network-based approach to phase space integration. *SciPost Physics*, 9(4):053, 2020.

S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, and O. Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

R. Krishnan, U. Shalit, and D. Sontag. Structured inference networks for nonlinear state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

R. G. Krishnan, U. Shalit, and D. Sontag. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

Y. Li and Y. Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

V. Liévin, A. Dittadi, A. Christensen, and O. Winther. Optimal variance control of the score-function gradient estimator for importance-weighted bounds. In *Advances in Neural Information Processing Systems*, volume 33, pages 16591–16602, 2020.

Y. Liu, K. Zhang, T. Basar, and W. Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *Advances in Neural Information Processing Systems*, volume 33, pages 7624–7636, 2020.

Z. Liu, P. Luo, X. Wang, and X. Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

J. Lucas, G. Tucker, R. B. Grosse, and M. Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

J. Marino, M. Cvitkovic, and Y. Yue. A general method for amortizing variational filtering. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

S. D. Mbacke, F. Clerc, and P. Germain. Statistical guarantees for variational autoencoders using pac-bayesian theory. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

A. Miller, N. Foti, A. D'Amour, and R. P. Adams. Reducing reparameterization gradient variance. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993.

K. Neklyudov and M. Welling. Orbital MCMC. In *International Conference on Artificial Intelligence and Statistics*, pages 5790–5814. PMLR, 2022.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

J. Olsson and J. Westerborn. Efficient particle-based online smoothing in general hidden Markov models: the PaRIS algorithm. *Bernoulli*, 23(3):1951–1996, 2017.

J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. SIAM, 2000.

J. Paisley, D. Blei, and M. Jordan. Variational bayesian inference with stochastic search. In *International Conference on Machine Learning*, pages 1367–1374. PMLR, 2012.

M. Papini, D. Binaghi, G. Canonaco, M. Pirotta, and M. Restelli. Stochastic variance-reduced policy gradient. In *International Conference on Machine Learning*, pages 4026–4035. PMLR, 2018.

S. Park, G. Adosoglou, and P. M. Pardalos. Interpreting rate-distortion of variational autoencoder and using model uncertainty for anomaly detection. *Annals of Mathematics and Artificial Intelligence*, 90 (7):735–752, 2022.

A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

S. Qiu, Z. Yang, J. Ye, and Z. Wang. On finite-time convergence of actor-critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2(2):652–664, 2021.

T. Rainforth, A. Kosiorek, T. A. Le, C. Maddison, M. Igl, F. Wood, and Y. W. Teh. Tighter variational

bounds are not necessarily better. In *International Conference on Machine Learning*, pages 4277–4285. PMLR, 2018.

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *International Conference on Artificial Intelligence and Statistics*, pages 814–822. PMLR, 2014.

A. Razavi, A. v. d. Oord, B. Poole, and O. Vinyals. Preventing posterior collapse with delta-vaes. In *International Conference on Learning Representations*, 2019.

S. J. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

J. Regier, M. I. Jordan, and J. McAuliffe. Fast black-box variational inference through stochastic trust-region optimization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286. PMLR, 2014.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck. A hierarchical latent vector model for learning long-term structure in music. In *International Conference on Machine Learning*, pages 4364–4373. PMLR, 2018.

G. Roeder, Y. Wu, and D. K. Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.

Z. Shen, A. Ribeiro, H. Hassani, H. Qian, and C. Mi. Hessian aided policy gradient. In *International Conference on Machine Learning*, pages 5729–5738. PMLR, 2019.

N. Shi and D. Li. RMSprop converges with proper hyperparameter. In *International Conference on Learning Representations*, 2021.

K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

S. Surendran, A. Fermanian, A. Godichon-Baggioni, and S. Le Corff. Non-asymptotic analysis of biased adaptive stochastic approximation. In *Advances in Neural Information Processing Systems*, volume 37, pages 12897–12943, 2024.

V. Z. Tadić and A. Doucet. Asymptotic properties of recursive particle maximum likelihood estimation. *IEEE Transactions on Information Theory*, 67(3): 1825–1848, 2020.

R. Tang and Y. Yang. On empirical bayes variational autoencoder: An excess risk bound. In *Conference on Learning Theory*, pages 4068–4125. PMLR, 2021.

A. Thin, Y. Janati El Idrissi, S. Le Corff, C. Ollion, E. Moulines, A. Doucet, A. Durmus, and C. X. Robert. NEO: Non equilibrium sampling on the orbits of a deterministic transform. In *Advances in Neural Information Processing Systems*, volume 34, pages 17060–17071, 2021.

T. Tieleman, G. Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2):26–31, 2012.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61 (3):611–622, 1999.

A. Vahdat and J. Kautz. Nvae: A deep hierarchical variational autoencoder. In *Advances in Neural Information Processing Systems*, volume 33, pages 19667–19679, 2020.

A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

A. Virmaux and K. Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Z. Wang and L. Ziyin. Posterior collapse of a linear latent variable model. In *Advances in Neural Information Processing Systems*, volume 35, pages 37537–37548, 2022.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

M. Zaheer, S. Reddi, D. Sachan, S. Kale, and S. Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Theoretical Convergence Guarantees for Variational Autoencoders: Supplementary Materials

Table of Contents

Notations. Given vectors $v = \left[v_1, v_2, \ldots, v_p\right]^\top$ and $u = \left[u_1, u_2, \ldots, u_d\right]^\top$, where u is a function of v, the derivative of u with respect to the vector v, denoted by $\nabla_v u$, is a matrix of size (d, p), and it is defined as follows:

$$\nabla_v u := \frac{\partial u}{\partial v}^\top ,$$

so that for all $1 \leq i \leq d$, $1 \leq j \leq p$, $(\nabla_v u)_{ij} = \partial u_i / \partial v_j$. For all $v \in \mathbb{R}^d$, we use $\mathrm{Diag}(v)$ to denote the diagonal matrix with diagonal given by v. For all $A \in \mathbb{R}^{d \times d}$, $\mathrm{diag}(A)$ is the vector obtained with the diagonal elements of A. The Hadamard product of vectors u and v is denoted by $u \odot v$, and $v^2 = v \odot v$ represents the elementwise product of v with itself. For $d \geq 1$, let I_d denote the $d \times d$ identity matrix and $\mathbf{1}$ be the vector in $\mathbb{R}^d$ whose entries are all equal to 1. Let $u \in \mathbb{R}^d$ be a vector and $A \in \mathbb{R}^{d \times p}$ be a matrix with columns $A_1, \ldots, A_p \in \mathbb{R}^d$, The element-wise product of u with the matrix A, denoted by $x \cdot A$, is defined as:

$$u \cdot A = \left[u \odot A_1, \ldots, u \odot A_p\right] ,$$

where $\odot$ denotes the Hadamard (element-wise) product. For matrices $W_1, \cdots, W_N$ where $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$, the product $\prod_{j=1}^N W_j$ is defined as:

$$\prod_{j=1}^{N} W_j = W_N W_{N-1} \cdots W_1 ,$$

which is generally not equal to $W_1 W_2 \cdots W_N$. The notation $\det(\cdot)$ denotes the determinant of a matrix, and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix.

A ADDITIONAL METHODOLOGICAL DETAILS

The conditional likelihood $p_\theta(x|z)$, referred to as the decoder distribution, is generally defined as a Gaussian distribution for real-valued data or a Bernoulli distribution for binary data. Specifically, for the Gaussian decoder where $p_\theta(x|z) = \mathcal{N}(x; G_\theta(z), \Gamma_\theta(z))$, the reconstruction loss simplifies to the Mean Squared Error if Γ_θ is assumed to be the identity matrix. For a Bernoulli decoder, this corresponds to the binary cross-entropy loss. The prior over the latent variables is typically chosen to be an isotropic multivariate Gaussian. The encoder distribution is also commonly modeled as a Gaussian.

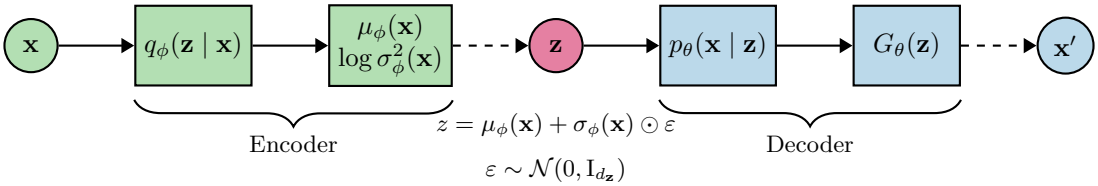

Figure 4: Illustration of the Architecture of a VAE with the Multivariate Gaussian.

Figure 4 displays an illustration of a VAE, where both the encoder and the decoder are represented as Gaussian distributions, and the prior over the latent variables is an isotropic multivariate Gaussian, as discussed in Section 3.3. It is important to note that the encoder outputs the logarithm of the variance instead of the variance to ensure that the variance is always positive, avoiding the need for explicit constraints on the output. If $z \mapsto q_\phi(z|x)$ is a multivariate Gaussian density with a diagonal covariance structure, the reparameterization trick can be expressed as:

$$z = g(\varepsilon, \phi) = \mu_\phi(x) + \sigma_\phi(x) \odot \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \mathrm{I}_{d_z}) .$$

This technique allows the Gaussian VAE to generate latent representations by sampling through the reparameterization trick, thus enabling gradient-based optimization with the pathwise gradient estimator.

Derivation of the Score Function Gradient. Proposition A.1 provides the form of the score function gradient of the expected ELBO with respect to ϕ.

Proposition A.1. *For all $\theta \in \Theta$, $\phi \in \Phi$, we have:*

$$\nabla^{\mathsf{S}}_\phi \mathcal{L}(\theta, \phi) = \mathbb{E}_\pi \left[\mathbb{E}_{q_\phi(\cdot|X)} \left[\log \frac{p_\theta(X, Z)}{q_\phi(Z|X)} \nabla_\phi \log q_\phi(Z|X) \right] \right] .$$

Proof. For a given observation $x \in \mathsf{X}$, the score function gradient of the ELBO with respect to ϕ is given by:

$$\begin{aligned}\nabla_\phi \mathcal{L}(\theta, \phi; x) &= \nabla_\phi \mathbb{E}_{q_\phi(\cdot|x)} \left[\log p_\theta(x, Z) - \log q_\phi(Z|x)\right] \\ &= \nabla_\phi \int (\log p_\theta(x, z) - \log q_\phi(z|x)) q_\phi(z|x) \, \mathrm{d}z \\ &= \int \nabla_\phi \left[(\log p_\theta(x, z) - \log q_\phi(z|x)) q_\phi(z|x)\right] \, \mathrm{d}z \\ &= \mathbb{E}_{q_\phi(\cdot|x)} \left[\nabla_\phi \log q_\phi(Z|x)(\log p_\theta(x, Z) - \log q_\phi(Z|x))\right] - \mathbb{E}_{q_\phi(\cdot|x)} \left[\nabla_\phi \log q_\phi(Z|x)\right] .\end{aligned}$$

Using the fact that $\mathbb{E}_{q_\phi(\cdot|x)} \left[\nabla_\phi \log q_\phi(Z|x)\right] = 0$ due to the regularity conditions on $q_\phi(z|x)$ yields

$$\begin{aligned}\nabla_\phi \mathcal{L}(\theta, \phi; x) &= \mathbb{E}_{q_\phi(\cdot|x)} \left[\nabla_\phi \log q_\phi(Z|x)(\log p_\theta(x, Z) - \log q_\phi(Z|x))\right] \\ &= \mathbb{E}_{q_\phi(\cdot|x)} \left[\log \frac{p_\theta(x, Z)}{q_\phi(Z|x)} \nabla_\phi \log q_\phi(Z|x)\right] .\end{aligned}$$

□

Score Function and Pathwise Gradient Estimator. The gradient estimator of the ELBO for a given batch of observations $\{x_i\}_{i=1}^B$, where B is the batch size, can be computed using Monte Carlo sampling as follows:

$$\widehat{\nabla}_{\theta,\phi} \mathcal{L}(\theta, \phi; \{x_i\}_{i=1}^B) = \frac{1}{B} \sum_{i=1}^{B} \frac{1}{K} \sum_{\ell=1}^{K} \tilde{g}_{i,\ell} \, , \tag{4}$$

where K denotes the number of samples drawn from the latent space, and $\tilde{g}_{i,\ell}$ is either the gradient estimator via the score function $\tilde{g}_{i,\ell}^{\mathsf{SF}}$ or the pathwise gradient estimator $\tilde{g}_{i,\ell}^{\mathsf{PF}}$. For the score function gradient estimator, the expression is given by:

$$\tilde{g}_{i,\ell}^{\mathsf{S}} = \left(\nabla_\theta \log \frac{p_\theta(x_i, z_i^{(\ell)})}{q_\phi(z_i^{(\ell)}|x_i)}, \log \frac{p_\theta(x_i, z_i^{(\ell)})}{q_\phi(z_i^{(\ell)}|x_i)} \nabla_\phi \log \frac{p_\theta(x_i, z_i^{(\ell)})}{q_\phi(z_i^{(\ell)}|x_i)}\right)^\top , \tag{5}$$

where, for all $1 \le i \le B$ and $1 \le \ell \le K$, $z_i^{(\ell)}$ are independent samples from $q_\phi(\cdot|x_i)$. The pathwise gradient estimator is given by:

$$\tilde{g}_{i,\ell}^{\mathsf{P}} = \left(\nabla_\theta \log \frac{p_\theta(x_i, g(\varepsilon_i^{(\ell)}, \phi))}{q_\phi(g(\varepsilon_i^{(\ell)}, \phi)|x_i)}, \nabla_z \log \frac{p_\theta(x_i, g(\varepsilon_i^{(\ell)}, \phi))}{q_\phi(g(\varepsilon_i^{(\ell)}, \phi)|x_i)} \nabla_\phi g(\varepsilon_i^{(\ell)}, \phi) - \nabla_\phi \log q_\phi(g(\varepsilon_i^{(\ell)}, \phi)|x_i)\right)^\top , \tag{6}$$

where, for all $1 \le i \le B$ and $1 \le \ell \le K$, $\varepsilon_i^{(\ell)}$ are independent samples from p_ε.

Algorithm 1 Adam Algorithm for ELBO Maximization

Input: Initial points θ_0, ϕ_0, maximum number of iterations n, step sizes $\{\gamma_k\}_{k\ge1}$, momentum parameters $\beta_1, \beta_2 \in [0, 1)$, regularization parameter $\delta \ge 0$ and batch size B.
Set $m_0 = 0$ and $v_0 = 0$.
for $k = 0$ to $n - 1$ **do**
  Sample a mini-batch $\{x_i\}_{i=1}^B$.
  Compute the stochastic gradient $g_{k+1} = \widehat{\nabla}_{\theta,\phi} \mathcal{L}(\theta_k, \phi_k; \{x_i\}_{i=1}^B)$ using (4).
  $m_{k+1} = \beta_1 m_k + (1 - \beta_1) g_{k+1}$.
  $v_{k+1} = \beta_2 v_k + (1 - \beta_2) g_{k+1} \odot g_{k+1}$.
  $(\theta_{k+1}, \phi_{k+1}) = (\theta_k, \phi_k) + \gamma_{k+1} \big[v_{k+1} + \delta\big]^{-1/2} m_{k+1}$.
end for
Output: $(\theta_k, \phi_k)_{0 \le k \le n}$.

B CONVERGENCE PROOFS

B.1 Proof of Proposition 3.1

We divide the proof into two lemmas where Lemma B.1 establishes the smoothness condition when the gradient is computed using the score function, while Lemma B.2 presents the smoothness condition for the pathwise gradient.

Lemma B.1. *Let Assumptions 1 and 2(i) hold. For all $\theta, \theta' \in \Theta$ and $\phi, \phi' \in \Phi$,*

$$\left\|\nabla^{\mathsf{S}}_{\theta,\phi}\mathcal{L}(\theta,\phi) - \nabla^{\mathsf{S}}_{\theta,\phi}\mathcal{L}(\theta',\phi')\right\| \leq L^{\mathsf{S}} \left\|(\theta,\phi) - (\theta',\phi')\right\| \ ,$$

where $L^{\mathsf{S}} = \sup_{\phi\in\Phi} \mathbb{E}_{\pi,\phi}\left[L_1(x,z) + 2\alpha(x,z)L_2(x,z) + 4M(x,z)^2 + 4\alpha(x,z)M(x,z)^2\right]$.

Proof. First, for all $\theta, \theta' \in \Theta$, $\phi, \phi' \in \Phi$, we have:

$$\begin{aligned}\left\|\nabla^{\mathsf{S}}_{\theta,\phi}\mathcal{L}(\theta,\phi) - \nabla^{\mathsf{S}}_{\theta,\phi}\mathcal{L}(\theta',\phi')\right\| &= \left\|\left(\nabla_\theta\mathcal{L}(\theta,\phi) - \nabla_\theta\mathcal{L}(\theta',\phi'), \nabla^{\mathsf{S}}_\phi\mathcal{L}(\theta,\phi) - \nabla^{\mathsf{S}}_\phi\mathcal{L}(\theta',\phi')\right)\right\| \\ &\leq \left\|\nabla_\theta\mathcal{L}(\theta,\phi) - \nabla_\theta\mathcal{L}(\theta',\phi')\right\| + \left\|\nabla^{\mathsf{S}}_\phi\mathcal{L}(\theta,\phi) - \nabla^{\mathsf{S}}_\phi\mathcal{L}(\theta',\phi')\right\| \ . \end{aligned} \tag{7}$$

Lipschitz condition of $\nabla_\theta\mathcal{L}(\theta,\phi)$.

$$\left\|\nabla_\theta\mathcal{L}(\theta,\phi) - \nabla_\theta\mathcal{L}(\theta',\phi')\right\| \leq \left\|\nabla_\theta\mathcal{L}(\theta,\phi) - \nabla_\theta\mathcal{L}(\theta',\phi)\right\| + \left\|\nabla_\theta\mathcal{L}(\theta',\phi) - \nabla_\theta\mathcal{L}(\theta',\phi')\right\|$$

Now, we bound each of these terms individually. By Assumption 2(i), for all $\theta, \theta' \in \Theta$, $\phi \in \Phi$,

$$\begin{aligned}\left\|\nabla_\theta\mathcal{L}(\theta,\phi) - \nabla_\theta\mathcal{L}(\theta',\phi)\right\| &= \left\|\mathbb{E}_{\pi,\phi}\left[\nabla_\theta \log p_\theta(x,z) - \nabla_\theta \log p_{\theta'}(x,z)\right]\right\| \\ &\leq \mathbb{E}_{\pi,\phi}\left[\left\|\nabla_\theta \log p_\theta(x,z) - \nabla_\theta \log p_{\theta'}(x,z)\right\|\right] \\ &\leq L^1_{dd}\left\|\theta - \theta'\right\| \ , \end{aligned}$$

where $L^1_{dd} = \mathbb{E}_{\pi,\phi}\left[L_1(x,z)\right]$. For the second term, we have:

$$\begin{aligned}\left\|\nabla_\theta\mathcal{L}(\theta',\phi) - \nabla_\theta\mathcal{L}(\theta',\phi')\right\| &= \left\|\mathbb{E}_{\pi,\phi}\left[\nabla_\theta \log p_{\theta'}(x,z)\right] - \mathbb{E}_{\pi,\phi'}\left[\nabla_\theta \log p_{\theta'}(x,z)\right]\right\| \\ &\leq \mathbb{E}_\pi\left[\left\|\int \nabla_\theta \log p_{\theta'}(x,z)(q_\phi(z|x) - q_{\phi'}(z|x))\,\mathrm{d}z\right\|\right] \\ &\leq \mathbb{E}_\pi\left[\int \left\|\nabla_\theta \log p_{\theta'}(x,z)(q_\phi(z|x) - q_{\phi'}(z|x))\right\|\,\mathrm{d}z\right] \\ &\leq \mathbb{E}_\pi\left[\int M(x,z)\left|q_\phi(z|x) - q_{\phi'}(z|x)\right|\,\mathrm{d}z\right] \ , \end{aligned}$$

where the inequality follows from Assumption 2(i). Then, using Assumption 2(i) and that for all $x \geq 1$, $x - 1 \leq x\log x \leq |x\log x|$,

$$\begin{aligned}\left\|\nabla_\theta\mathcal{L}(\theta',\phi) - \nabla_\theta\mathcal{L}(\theta',\phi')\right\| &\leq \mathbb{E}_\pi\left[\int M(x,z)\left(\frac{q_\phi(z|x)}{q_{\phi'}(z|x)} - 1\right)1_{q_\phi(z|x)\geq q_{\phi'}(z|x)}q_{\phi'}(z|x)\,\mathrm{d}z\right] \\ &\quad + \mathbb{E}_\pi\left[\int M(x,z)\left(\frac{q_{\phi'}(z|x)}{q_\phi(z|x)} - 1\right)1_{q_{\phi'}(z|x)>q_\phi(z|x)}q_\phi(z|x)\,\mathrm{d}z\right] \\ &\leq \mathbb{E}_\pi\left[\int M(x,z)\left|\frac{q_\phi(z|x)}{q_{\phi'}(z|x)}\log\frac{q_\phi(z|x)}{q_{\phi'}(z|x)}\right|q_{\phi'}(z|x)\,\mathrm{d}z\right] \\ &\quad + \mathbb{E}_\pi\left[\int M(x,z)\left|\frac{q_{\phi'}(z|x)}{q_\phi(z|x)}\log\frac{q_{\phi'}(z|x)}{q_\phi(z|x)}\right|q_\phi(z|x)\,\mathrm{d}z\right] \\ &\leq \mathbb{E}_\pi\left[\int M(x,z)\left|\log\frac{q_\phi(z|x)}{q_{\phi'}(z|x)}\right|q_\phi(z|x)\,\mathrm{d}z + \int M(x,z)\left|\log\frac{q_{\phi'}(z|x)}{q_\phi(z|x)}\right|q_{\phi'}(z|x)\,\mathrm{d}z\right] \\ &\leq \mathbb{E}_{\pi,\phi}\left[M(x,z)\left|\log q_\phi(z|x) - \log q_{\phi'}(z|x)\right|\right] + \mathbb{E}_{\pi,\phi'}\left[M(x,z)\left|\log q_\phi(z|x) - \log q_{\phi'}(z|x)\right|\right] \\ &\leq L^2_{dd}\left\|\phi - \phi'\right\| \ , \end{aligned}$$

where $L^2_{dd} = \mathbb{E}_{\pi,\phi}\left[M(x,z)^2\right] + \mathbb{E}_{\pi,\phi'}\left[M(x,z)^2\right]$. This completes the proof for the first term in (7).

Lipschitz condition of $\nabla^{\mathsf{S}}_{\phi}\mathcal{L}(\theta,\phi)$. This case can be treated similarly to the Lipschitz condition with respect to θ. For all $\theta,\theta' \in \Theta$, $\phi,\phi' \in \Phi$,

$$\left\|\nabla^{\mathsf{S}}_{\phi}\mathcal{L}(\theta,\phi) - \nabla^{\mathsf{S}}_{\phi}\mathcal{L}(\theta',\phi')\right\| \leq \left\|\nabla^{\mathsf{S}}_{\phi}\mathcal{L}(\theta,\phi) - \nabla^{\mathsf{S}}_{\phi}\mathcal{L}(\theta',\phi)\right\| + \left\|\nabla^{\mathsf{S}}_{\phi}\mathcal{L}(\theta',\phi) - \nabla^{\mathsf{S}}_{\phi}\mathcal{L}(\theta',\phi')\right\|$$

For all $\theta,\theta' \in \Theta$, $\phi \in \Phi$, using Assumption 2(i), we have:

$$\begin{aligned}
\left\|\nabla^{\mathsf{S}}_{\phi}\mathcal{L}(\theta,\phi) - \nabla^{\mathsf{S}}_{\phi}\mathcal{L}(\theta',\phi)\right\| &= \left\|\mathbb{E}_{\pi,\phi}\left[\log\frac{p_\theta(x,z)}{q_\phi(z|x)}\nabla_\phi \log q_\phi(z|x) - \log\frac{p_{\theta'}(x,z)}{q_\phi(z|x)}\nabla_\phi \log q_\phi(z|x)\right]\right\| \\
&\leq \mathbb{E}_{\pi,\phi}\left[\left\|\log\frac{p_\theta(x,z)}{q_\phi(z|x)}\nabla_\phi \log q_\phi(z|x) - \log\frac{p_{\theta'}(x,z)}{q_\phi(z|x)}\nabla_\phi \log q_\phi(z|x)\right\|\right] \\
&\leq \mathbb{E}_{\pi,\phi}\left[M(x,z)\left\|\log p_\theta(x,z) - \log p_{\theta'}(x,z)\right\|\right] \\
&\leq \mathbb{E}_{\pi,\phi}\left[M(x,z)^2\left\|\theta-\theta'\right\|\right] \\
&\leq L^1_{ed}\left\|\theta-\theta'\right\| \ ,
\end{aligned}$$

where $L^1_{ed} = \mathbb{E}_{\pi,\phi}\left[M(x,z)^2\right]$. On the other hand, for all $\theta' \in \Theta$, $\phi,\phi' \in \Phi$,

$$\begin{aligned}
\left\|\nabla^{\mathsf{S}}_{\phi}\mathcal{L}(\theta',\phi) - \nabla^{\mathsf{S}}_{\phi}\mathcal{L}(\theta',\phi')\right\| &= \left\|\mathbb{E}_{\pi,\phi}\left[\log\frac{p_{\theta'}(x,z)}{q_\phi(z|x)}\nabla_\phi \log q_\phi(z|x)\right] - \mathbb{E}_{\pi,\phi'}\left[\log\frac{p_{\theta'}(x,z)}{q_{\phi'}(z|x)}\nabla_\phi \log q_{\phi'}(z|x)\right]\right\| \\
&\leq A_1 + A_2 \ ,
\end{aligned}$$

where

$$\begin{aligned}
A_1 &= \left\|\mathbb{E}_{\pi,\phi}\left[\log\frac{p_{\theta'}(x,z)}{q_\phi(z|x)}\nabla_\phi \log q_\phi(z|x)\right] - \mathbb{E}_{\pi,\phi}\left[\log\frac{p_{\theta'}(x,z)}{q_{\phi'}(z|x)}\nabla_\phi \log q_{\phi'}(z|x)\right]\right\|, \\
A_2 &= \left\|\mathbb{E}_{\pi,\phi}\left[\log\frac{p_{\theta'}(x,z)}{q_{\phi'}(z|x)}\nabla_\phi \log q_{\phi'}(z|x)\right] - \mathbb{E}_{\pi,\phi'}\left[\log\frac{p_{\theta'}(x,z)}{q_{\phi'}(z|x)}\nabla_\phi \log q_{\phi'}(z|x)\right]\right\|.
\end{aligned}$$

By decomposing again A_1, we get that

$$\begin{aligned}
A_1 &\leq \mathbb{E}_{\pi,\phi}\left[\left\|\log\frac{p_{\theta'}(x,z)}{q_\phi(z|x)}\nabla_\phi \log q_\phi(z|x) - \log\frac{p_{\theta'}(x,z)}{q_{\phi'}(z|x)}\nabla_\phi \log q_{\phi'}(z|x)\right\|\right] \\
&\leq \mathbb{E}_{\pi,\phi}\left[\left\|\log p_{\theta'}(x,z)(\nabla_\phi \log q_\phi(z|x) - \nabla_\phi \log q_{\phi'}(z|x)\right\|\right] \\
&\quad + \mathbb{E}_{\pi,\phi}\left[\left\|\log q_\phi(z|x)\nabla_\phi \log q_\phi(z|x) - \log q_{\phi'}(z|x)\nabla_\phi \log q_{\phi'}(z|x)\right\|\right] \\
&\leq \mathbb{E}_{\pi,\phi}\left[\left\|\log p_{\theta'}(x,z)(\nabla_\phi \log q_\phi(z|x) - \nabla_\phi \log q_{\phi'}(z|x)\right\|\right] \\
&\quad + \mathbb{E}_{\pi,\phi}\left[\left|\log q_\phi(z|x)\right|\left\|\nabla_\phi \log q_\phi(z|x) - \nabla_\phi \log q_{\phi'}(z|x)\right\|\right] \\
&\quad + \mathbb{E}_{\pi,\phi}\left[\left\|\nabla_\phi \log q_{\phi'}(z|x)\right\|\left\|\log q_\phi(z|x) - \log q_{\phi'}(z|x)\right\|\right] \ .
\end{aligned}$$

Then, using the Mean Value Theorem along with Assumptions 1 and 2(i),

$$A_1 \leq 2\mathbb{E}_{\pi,\phi}\left[\alpha(x,z)L_2(x,z)\right]\left\|\phi-\phi'\right\| + \mathbb{E}_{\pi,\phi}\left[M(x,z)^2\right]\left\|\phi-\phi'\right\| \ .$$

For the term A_2, note that

$$\begin{aligned}
A_2 &\leq \mathbb{E}_{\pi}\left[\left\|\int \log\frac{p_{\theta'}(x,z)}{q_{\phi'}(z|x)}\nabla_\phi \log q_{\phi'}(z|x)(q_\phi(z|x) - q_{\phi'}(z|x))\,\mathrm{d}z\right\|\right] \\
&\leq \mathbb{E}_{\pi}\left[\int\left\|\log\frac{p_{\theta'}(x,z)}{q_{\phi'}(z|x)}\nabla_\phi \log q_{\phi'}(z|x)(q_\phi(z|x) - q_{\phi'}(z|x))\right\|\,\mathrm{d}z\right] \\
&\leq 2\mathbb{E}_{\pi}\left[\int \alpha(x,z)M(x,z)\left|q_\phi(z|x) - q_{\phi'}(z|x)\right|\,\mathrm{d}z\right],
\end{aligned}$$

where the inequality follows from Assumptions 1 and 2(i). Then, using Assumption 2(i) and that for all $x \geq 1$, $x - 1 \leq x \log x \leq |x \log x|$,

$$
\begin{aligned}
A_2 &\leq 2\mathbb{E}_\pi \left[\int \alpha(x,z) M(x,z) \left(\frac{q_\phi(z|x)}{q_{\phi'}(z|x)} - 1 \right) 1_{q_\phi(z|x) \leq q_{\phi'}(z|x)} q_{\phi'}(z|x) \, \mathrm{d}z \right] \\
&\quad + 2\mathbb{E}_\pi \left[\int \alpha(x,z) M(x,z) \left(\frac{q_{\phi'}(z|x)}{q_\phi(z|x)} - 1 \right) 1_{q_{\phi'}(z|x) > q_\phi(z|x)} q_\phi(z|x) \, \mathrm{d}z \right] \\
&\leq 2\mathbb{E}_\pi \left[\int \alpha(x,z) M(x,z) \left| \frac{q_\phi(z|x)}{q_{\phi'}(z|x)} \log \frac{q_\phi(z|x)}{q_{\phi'}(z|x)} \right| q_{\phi'}(z|x) \, \mathrm{d}z \right] \\
&\quad + 2\mathbb{E}_\pi \left[\int \alpha(x,z) M(x,z) \left| \frac{q_{\phi'}(z|x)}{q_\phi(z|x)} \log \frac{q_{\phi'}(z|x)}{q_\phi(z|x)} \right| q_\phi(z|x) \, \mathrm{d}z \right] \\
&\leq 2\mathbb{E}_\pi \left[\int \alpha(x,z) M(x,z) \left| \log \frac{q_\phi(z|x)}{q_{\phi'}(z|x)} \right| q_\phi(z|x) \, dz + \int \alpha(x,z) M(x,z) \left| \log \frac{q_{\phi'}(z|x)}{q_\phi(z|x)} \right| q_{\phi'}(z|x) \, \mathrm{d}z \right] \\
&\leq 2\mathbb{E}_{\pi,\phi} \left[\alpha(x,z) M(x,z) \left| \log q_\phi(z|x) - \log q_{\phi'}(z|x) \right| \right] + 2\mathbb{E}_{\pi,\phi'} \left[\alpha(x,z) M(x,z) \left| \log q_\phi(z|x) - \log q_{\phi'}(z|x) \right| \right] \\
&\leq 2 \left(\mathbb{E}_{\pi,\phi} \left[\alpha(x,z) M(x,z)^2 \right] + \mathbb{E}_{\pi,\phi'} \left[\alpha(x,z) M(x,z)^2 \right] \right) \left\| \phi - \phi' \right\| \, .
\end{aligned}
$$

By combining these two terms, we obtain:

$$
\left\| \nabla^{\mathsf{S}}_\phi \mathcal{L}(\theta, \phi) - \nabla^{\mathsf{S}}_\phi \mathcal{L}(\theta, \phi') \right\| \leq L_{ed} \left\| \phi - \phi' \right\| ,
$$

where $L^2_{ed} = \mathbb{E}_{\pi,\phi} \left[2\alpha(x,z) L_2(x,z) + M(x,z)^2 + 2\alpha(x,z) M(x,z)^2 \right] + 2\mathbb{E}_{\pi,\phi'} \left[\alpha(x,z) M(x,z)^2 \right]$ and concludes the argument for the second term in (7).

Then,

$$
\begin{aligned}
\left\| \nabla^{\mathsf{S}}_{\theta,\phi} \mathcal{L}(\theta, \phi) - \nabla^{\mathsf{S}}_{\theta,\phi} \mathcal{L}(\theta', \phi') \right\| &\leq \left(L^1_{dd} + L^1_{ed} \right) \left\| \theta - \theta' \right\| + \left(L^2_{dd} + L^2_{ed} \right) \left\| \phi - \phi' \right\| \, , \\
&\leq L^{\mathsf{S}} \left\| (\theta, \phi) - (\theta', \phi') \right\| \, ,
\end{aligned}
$$

where $L^{\mathsf{S}} = \sup_{\phi \in \Phi} \mathbb{E}_{\pi,\phi} \left[L_1(x,z) + 2\alpha(x,z) L_2(x,z) + 4M(x,z)^2 + 4\alpha(x,z) M(x,z)^2 \right]$. □

Lemma B.2. *Let Assumptions 2(ii) and 3 hold. For all $\theta, \theta' \in \Theta$ and $\phi, \phi' \in \Phi$,*

$$
\left\| \nabla^{\mathsf{P}}_{\theta,\phi} \mathcal{L}(\theta, \phi) - \nabla^{\mathsf{P}}_{\theta,\phi} \mathcal{L}(\theta', \phi') \right\| \leq L^{\mathsf{P}} \left\| (\theta, \phi) - (\theta', \phi') \right\| ,
$$

where $L^{\mathsf{P}} = \mathbb{E}_{\pi,p_\varepsilon} \left[L_p(x,\varepsilon) + M_g(x,\varepsilon)^2 \left(L_p(x,\varepsilon) + 2L_q(x,\varepsilon) \right) + 3L_g(x,\varepsilon) M(x,\varepsilon) + 2M_g(x,\varepsilon) L_q(x,\varepsilon) \right] + \mathbb{E}_{\pi,p_\varepsilon} \left[L_p(x,\varepsilon) M_g(x,\varepsilon) \right]$.

Proof. We proceed by dividing the proof into two cases, following the same structure as in the proof of Lemma B.1. We establish that the following inequalities hold: for all $\theta, \theta' \in \Theta$, $\phi, \phi' \in \Phi$,

$$
\begin{aligned}
\left\| \nabla_\theta \mathcal{L}(\theta, \phi) - \nabla_\theta \mathcal{L}(\theta', \phi') \right\| &\leq L^1_{dd} \left\| \theta - \theta' \right\| + L^2_{dd} \left\| \phi - \phi' \right\| \, , \\
\left\| \nabla^{\mathsf{P}}_\phi \mathcal{L}(\theta, \phi) - \nabla^{\mathsf{P}}_\phi \mathcal{L}(\theta', \phi') \right\| &\leq L^1_{ed} \left\| \theta - \theta' \right\| + L^2_{ed} \left\| \phi - \phi' \right\| \, ,
\end{aligned}
$$

where $L^1_{dd} = \mathbb{E}_{\pi,p_\varepsilon} \left[L_p(x,\varepsilon) \right]$, $L^2_{dd} = \mathbb{E}_{\pi,p_\varepsilon} \left[L_p(x,\varepsilon) M_g(x,\varepsilon) \right]$, $L^1_{ed} = \mathbb{E}_{\pi,p_\varepsilon} \left[M_g(x,\varepsilon) L_p(x,\varepsilon) \right]$ and $L^2_{ed} = \mathbb{E}_{\pi,p_\varepsilon} \left[M_g(x,\varepsilon)^2 \left(L_p(x,\varepsilon) + 2L_q(x,\varepsilon) \right) + 3L_g(x,\varepsilon) M(x,\varepsilon) + 2M_g(x,\varepsilon) L_q(x,\varepsilon) \right]$.

Lipschitz condition of $\nabla_\theta \mathcal{L}(\theta, \phi)$.

$$
\left\| \nabla_\theta \mathcal{L}(\theta, \phi) - \nabla_\theta \mathcal{L}(\theta', \phi') \right\| \leq \left\| \nabla_\theta \mathcal{L}(\theta, \phi) - \nabla_\theta \mathcal{L}(\theta', \phi) \right\| + \left\| \nabla_\theta \mathcal{L}(\theta', \phi) - \nabla_\theta \mathcal{L}(\theta', \phi') \right\| \tag{8}
$$

Now, we bound each of these terms individually. By Assumption 2(ii), for all $\theta, \theta' \in \Theta$, $\phi \in \Phi$,

$$
\begin{aligned}
\left\| \nabla_\theta \mathcal{L}(\theta, \phi) - \nabla_\theta \mathcal{L}(\theta', \phi) \right\| &= \left\| \mathbb{E}_{\pi,p_\varepsilon} \left[\nabla_\theta \log p_\theta(x, g(\varepsilon, \phi)) - \nabla_\theta \log p_{\theta'}(x, g(\varepsilon, \phi)) \right] \right\| \\
&\leq \mathbb{E}_{\pi,p_\varepsilon} \left[\left\| \nabla_\theta \log p_\theta(x, g(\varepsilon, \phi)) - \nabla_\theta \log p_{\theta'}(x, g(\varepsilon, \phi)) \right\| \right] \\
&\leq \mathbb{E}_{\pi,p_\varepsilon} \left[L_p(x,\varepsilon) \right] \left\| \theta - \theta' \right\| \, ,
\end{aligned}
$$

which concludes the bound for the first term in (8). For the second term, we have:

$$\begin{aligned}\|\nabla_\theta \mathcal{L}(\theta', \phi) - \nabla_\theta \mathcal{L}(\theta', \phi')\| &= \|\mathbb{E}_{\pi, p_\varepsilon}\left[\nabla_\theta \log p_{\theta'}(x, g(\varepsilon, \phi)) - \nabla_\theta \log p_{\theta'}(x, g(\varepsilon, \phi'))\right]\| \\ &\leq \mathbb{E}_{\pi, p_\varepsilon}\left[\|\nabla_\theta \log p_{\theta'}(x, g(\varepsilon, \phi)) - \nabla_\theta \log p_{\theta'}(x, g(\varepsilon, \phi'))\|\right] \ .\end{aligned}$$

Since for all $z, z' \in \mathsf{Z}$, $\|\nabla_\theta \log p_{\theta'}(x, z) - \nabla_\theta \log p_{\theta'}(x, z')\| \leq L_p(x, \varepsilon)\, \|z - z'\|$, it follows that:

$$\|\nabla_\theta \log p_{\theta'}(x, g(\varepsilon, \phi)) - \nabla_\theta \log p_{\theta'}(x, g(\varepsilon, \phi'))\| \leq L_p(x, \varepsilon)\, \|g(\varepsilon, \phi) - g(\varepsilon, \phi')\| \ .$$

Therefore,

$$\begin{aligned}\|\nabla_\theta \mathcal{L}(\theta', \phi) - \nabla_\theta \mathcal{L}(\theta', \phi')\| &\leq \mathbb{E}_{\pi, p_\varepsilon}\left[L_p(x, \varepsilon)\|g(\varepsilon, \phi) - g(\varepsilon, \phi')\|\right] \\ &\leq \mathbb{E}_{\pi, p_\varepsilon}\left[L_p(x, \varepsilon) M_g(x, \varepsilon)\right] \|\phi - \phi'\| \ ,\end{aligned}$$

which concludes the Lipschitz condition of $\nabla_\theta \mathcal{L}(\theta, \phi)$.

Lipschitz condition of $\nabla^{\mathsf{P}}_\phi \mathcal{L}(\theta, \phi)$. For all $\theta, \theta' \in \Theta$ and $\phi, \phi' \in \Phi$, we have the following inequality:

$$\left\|\nabla^{\mathsf{P}}_\phi \mathcal{L}(\theta, \phi) - \nabla^{\mathsf{P}}_\phi \mathcal{L}(\theta', \phi')\right\| \leq \left\|\nabla^{\mathsf{P}}_\phi \mathcal{L}(\theta, \phi) - \nabla^{\mathsf{P}}_\phi \mathcal{L}(\theta', \phi)\right\| + \left\|\nabla^{\mathsf{P}}_\phi \mathcal{L}(\theta', \phi) - \nabla^{\mathsf{P}}_\phi \mathcal{L}(\theta', \phi')\right\| \ . \tag{9}$$

We now handle each term separately. For the first term, we have:

$$\begin{aligned}\left\|\nabla^{\mathsf{P}}_\phi \mathcal{L}(\theta, \phi) - \nabla^{\mathsf{P}}_\phi \mathcal{L}(\theta', \phi)\right\| &\leq \mathbb{E}_{\pi, p_\varepsilon}\left[\|\nabla_\phi g(\varepsilon, \phi)\|\, \|\nabla_z \log p_\theta(x, g(\varepsilon, \phi)) - \nabla_z \log p_{\theta'}(x, g(\varepsilon, \phi))\|\right] \\ &\leq \mathbb{E}_{\pi, p_\varepsilon}\left[M_g(x, \varepsilon) L_p(x, \varepsilon)\right] \|\theta - \theta'\| \ ,\end{aligned}$$

which concludes the bound for the first term in (9). For the second term in (9), applying the triangle inequality, we have:

$$\left\|\nabla^{\mathsf{P}}_\phi \mathcal{L}(\theta', \phi) - \nabla^{\mathsf{P}}_\phi \mathcal{L}(\theta', \phi')\right\| \leq A_1 + A_2 + A_3 \ ,$$

where

$$\begin{aligned}A_1 &= \mathbb{E}_{\pi, p_\varepsilon}\left[\|\nabla_z \log p_{\theta'}(x, g(\varepsilon, \phi)\nabla_\phi g(\varepsilon, \phi) - \nabla_z \log p_{\theta'}(x, g(\varepsilon, \phi'))\nabla_\phi g(\varepsilon, \phi')\|\right] \ , \\ A_2 &= \mathbb{E}_{\pi, p_\varepsilon}\left[\|\nabla_z \log q_\phi(g(\varepsilon, \phi)|x)\nabla_\phi g(\varepsilon, \phi) - \nabla_z \log q_{\phi'}(g(\varepsilon, \phi')|x)\nabla_\phi g(\varepsilon, \phi')\|\right] \ , \\ A_3 &= \mathbb{E}_{\pi, p_\varepsilon}\left[\|\nabla_\phi \log q_\phi(z|x) - \nabla_\phi \log q_{\phi'}(z|x)\|\right] \ .\end{aligned}$$

For A_1, using the boundedness of the gradient with respect to z and the smoothness of g (Assumptions 2(ii) and 3), we obtain:

$$\begin{aligned}A_1 &= \mathbb{E}_{\pi, p_\varepsilon}\left[\|\nabla_z \log p_{\theta'}(x, g(\varepsilon, \phi))\nabla_\phi g(\varepsilon, \phi) - \nabla_z \log p_{\theta'}(x, g(\varepsilon, \phi'))\nabla_\phi g(\varepsilon, \phi')\|\right] \\ &\leq \mathbb{E}_{\pi, p_\varepsilon}\left[\|\nabla_z \log p_{\theta'}(x, g(\varepsilon, \phi))\nabla_\phi g(\varepsilon, \phi) - \nabla_z \log p_{\theta'}(x, g(\varepsilon, \phi'))\nabla_\phi g(\varepsilon, \phi)\|\right] \\ &\quad + \mathbb{E}_{\pi, p_\varepsilon}\left[\|\nabla_z \log p_{\theta'}(x, g(\varepsilon, \phi'))\nabla_\phi g(\varepsilon, \phi) - \nabla_z \log p_{\theta'}(x, g(\varepsilon, \phi'))\nabla_\phi g(\varepsilon, \phi')\|\right] \\ &\leq \mathbb{E}_{\pi, p_\varepsilon}\left[\|\nabla_\phi g(\varepsilon, \phi)\|\, \|\nabla_z \log p_{\theta'}(x, g(\varepsilon, \phi)) - \nabla_z \log p_{\theta'}(x, g(\varepsilon, \phi'))\|\right] \\ &\quad + \mathbb{E}_{\pi, p_\varepsilon}\left[\|\nabla_z \log p_{\theta'}(x, g(\varepsilon, \phi')\|\, \|\nabla_\phi g(\varepsilon, \phi) - \nabla_\phi g(\varepsilon, \phi')\|\right] \ .\end{aligned}$$

Since for all $z, z' \in \mathsf{Z}$, $\|\nabla_z \log p_{\theta'}(x, z) - \nabla_z \log p_{\theta'}(x, z')\| \leq L_p(x, \varepsilon)\, \|z - z'\|$, it follows that:

$$\|\nabla_z \log p_{\theta'}(x, g(\varepsilon, \phi)) - \nabla_z \log p_{\theta'}(x, g(\varepsilon, \phi'))\| \leq L_p(x, \varepsilon)\, \|g(\varepsilon, \phi) - g(\varepsilon, \phi')\| \ .$$

Therefore,

$$\begin{aligned}A_1 &\leq \mathbb{E}_{\pi, p_\varepsilon}\left[M_g(x, \varepsilon) L_p(x, \varepsilon)\, \|g(\varepsilon, \phi) - g(\varepsilon, \phi')\|\right] + \mathbb{E}_{\pi, p_\varepsilon}\left[M(x, \varepsilon)\, \|\nabla_\phi g(\varepsilon, \phi) - \nabla_\phi g(\varepsilon, \phi')\|\right] \\ &\leq \mathbb{E}_{\pi, p_\varepsilon}\left[M_g(x, \varepsilon)^2 L_p(x, \varepsilon) + M(x, \varepsilon) L_g(x, \varepsilon)\right] \|\phi - \phi'\| \ .\end{aligned}$$

For A_2, we have:

$$\begin{aligned}
A_2 &= \mathbb{E}_{\pi,p_\varepsilon}\left[\|\nabla_z \log q_\phi(g(\varepsilon,\phi)|x)\nabla_\phi g(\varepsilon,\phi) - \nabla_z \log q_{\phi'}(g(\varepsilon,\phi')|x)\nabla_\phi g(\varepsilon,\phi')\|\right] \\
&\leq \mathbb{E}_{\pi,p_\varepsilon}\left[\|\nabla_z \log q_\phi(g(\varepsilon,\phi)|x)\nabla_\phi g(\varepsilon,\phi) - \nabla_z \log q_{\phi'}(g(\varepsilon,\phi)|x)\nabla_\phi g(\varepsilon,\phi)\|\right] \\
&\quad + \mathbb{E}_{\pi,p_\varepsilon}\left[\|\nabla_z \log q_{\phi'}(g(\varepsilon,\phi)|x)\nabla_\phi g(\varepsilon,\phi) - \nabla_z \log q_{\phi'}(g(\varepsilon,\phi')|x)\nabla_\phi g(\varepsilon,\phi)\|\right] \\
&\quad + \mathbb{E}_{\pi,p_\varepsilon}\left[\|\nabla_z \log q_{\phi'}(g(\varepsilon,\phi')|x)\nabla_\phi g(\varepsilon,\phi) - \nabla_z \log q_{\phi'}(g(\varepsilon,\phi')|x)\nabla_\phi g(\varepsilon,\phi')\|\right] \\
&\leq \mathbb{E}_{\pi,p_\varepsilon}\left[\|\nabla_\phi g(\varepsilon,\phi)\|\,\|\nabla_z \log q_\phi(g(\varepsilon,\phi)|x) - \nabla_z \log q_{\phi'}(g(\varepsilon,\phi)|x)\|\right] \\
&\quad + \mathbb{E}_{\pi,p_\varepsilon}\left[\|\nabla_\phi g(\varepsilon,\phi)\|\,\|\nabla_z \log q_{\phi'}(g(\varepsilon,\phi)|x) - \nabla_z \log q_{\phi'}(g(\varepsilon,\phi')|x)\|\right] \\
&\quad + \mathbb{E}_{\pi,p_\varepsilon}\left[\|\nabla_z \log q_{\phi'}(g(\varepsilon,\phi')|x)\|\,\|\nabla_\phi g(\varepsilon,\phi) - \nabla_\phi g(\varepsilon,\phi')\|\right] .
\end{aligned}$$

Since for all $z, z' \in \mathsf{Z}$, $\|\nabla_z \log q_{\phi'}(z|x) - \nabla_z \log q_{\phi'}(z'|x)\| \leq L_q(x,\varepsilon)\,\|z - z'\|$, it follows that:

$$\|\nabla_z \log q_{\phi'}(g(\varepsilon,\phi)|x) - \nabla_z \log q_{\phi'}(g(\varepsilon,\phi')|x)\| \leq L_q(x,\varepsilon)\,\|g(\varepsilon,\phi) - g(\varepsilon,\phi')\| .$$

Therefore,

$$\begin{aligned}
A_2 &\leq \mathbb{E}_{\pi,p_\varepsilon}\left[M_g(x,\varepsilon)L_q(x,\varepsilon)\,\|\phi - \phi'\|\right] \\
&\quad + \mathbb{E}_{\pi,p_\varepsilon}\left[M_g(x,\varepsilon)L_q(x,\varepsilon)\,\|g(\varepsilon,\phi) - g(\varepsilon,\phi')\|\right] \\
&\quad + \mathbb{E}_{\pi,p_\varepsilon}\left[M_q(x,\varepsilon)\,\|\nabla_\phi g(\varepsilon,\phi) - \nabla_\phi g(\varepsilon,\phi')\|\right] \\
&\leq \mathbb{E}_{\pi,p_\varepsilon}\left[M_g(x,\varepsilon)L_q(x,\varepsilon) + M_g(x,\varepsilon)^2 L_q(x,\varepsilon) + M(x,\varepsilon)L_g(x,\varepsilon)\right]\|\phi - \phi'\| .
\end{aligned}$$

For A_3, following the same procedure as for the term A_2 we get:

$$\begin{aligned}
A_3 &= \mathbb{E}_{\pi,p_\varepsilon}\left[\|\nabla_\phi \log q_\phi(g(\varepsilon,\phi)|x) - \nabla_\phi \log q_{\phi'}(g(\varepsilon,\phi')|x)\|\right] \\
&= \mathbb{E}_{\pi,p_\varepsilon}\left[\|\nabla_z \log q_\phi(g(\varepsilon,\phi)|x)\nabla_\phi g(\varepsilon,\phi) - \nabla_z \log q_{\phi'}(g(\varepsilon,\phi')|x)\nabla_\phi g(\varepsilon,\phi')\|\right] \\
&\leq \mathbb{E}_{\pi,p_\varepsilon}\left[M_g(x,\varepsilon)L_q(x,\varepsilon) + M_g(x,\varepsilon)^2 L_q(x,\varepsilon) + M(x,\varepsilon)L_g(x,\varepsilon)\right]\|\phi - \phi'\| ,
\end{aligned}$$

which concludes the bound for the second term in (9), thereby completing the proof. □

B.2 Proof of Theorem 3.2

Proof. We address both the score function and pathwise gradient cases at the same time. Our proof is adapted from Theorem 2.1 in Ghadimi and Lan (2013), with modifications to account for the two variables and the dual randomness arising from both the data and the latent variables. Using the smoothness of $\mathcal{L}$ by Proposition 3.1 (Descent Lemma (Ortega and Rheinboldt, 2000; Nesterov, 2013) for maximization), for $m \in \{\mathsf{S}, \mathsf{P}\}$, we have:

$$\begin{aligned}
\mathcal{L}(\theta_{k+1},\phi_{k+1}) &\geq \mathcal{L}(\theta_k,\phi_k) + \left\langle \nabla^m_{\theta,\phi}\mathcal{L}(\theta_k,\phi_k), (\theta_{k+1},\phi_{k+1}) - (\theta_k,\phi_k)\right\rangle - \frac{L}{2}\|(\theta_{k+1},\phi_{k+1}) - (\theta_k,\phi_k)\|^2 \\
&\geq \mathcal{L}(\theta_k,\phi_k) + \gamma_{n+1}\left\langle \nabla^m_{\theta,\phi}\mathcal{L}(\theta_k,\phi_k), \widehat{\nabla}^m_{\theta,\phi}\mathcal{L}(\theta_k,\phi_k;\mathcal{D}_{k+1})\right\rangle - \frac{L}{2}\gamma^2_{k+1}\left\|\widehat{\nabla}^m_{\theta,\phi}\mathcal{L}(\theta_k,\phi_k;\mathcal{D}_{k+1})\right\|^2 ,
\end{aligned}$$

where $\mathcal{D}_{k+1}$ corresponds to the mini-batch of data used to compute the gradient estimator at iteration $k+1$. For all $k \geq 0$, let $\mathcal{F}_k = \sigma(\theta_0, \{\mathcal{D}_i, Z_i\}_{1\leq i\leq k})$, where $\mathcal{D}_i$ denotes the mini-batch of data used at iteration i, and Z_i represents all latent samples drawn at iteration i. Taking the conditional expectation with respect to the filtration $(\mathcal{F}_k)_{k\geq 0}$,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}(\theta_{k+1},\phi_{k+1}) \mid \mathcal{F}_k\right] &\geq \mathcal{L}(\theta_k,\phi_k) + \gamma_{k+1}\left\langle \nabla^m_{\theta,\phi}\mathcal{L}(\theta_k,\phi_k), \mathbb{E}\left[\widehat{\nabla}^m_{\theta,\phi}\mathcal{L}(\theta_k,\phi_k;\mathcal{D}_{k+1}) \mid \mathcal{F}_k\right]\right\rangle \\
&\quad - \frac{L\gamma^2_{k+1}}{2}\|\nabla^m_{\theta,\phi}\mathcal{L}(\theta_k,\phi_k)\|^2 - \frac{L\gamma^2_{k+1}}{2}\mathbb{E}\left[\|\widehat{\nabla}^m_{\theta,\phi}\mathcal{L}(\theta_k,\phi_k;\mathcal{D}_{k+1})) - \nabla^m_{\theta,\phi}\mathcal{L}(\theta_k,\phi_k)\|^2 \mid \mathcal{F}_k\right] .
\end{aligned}$$

Given the assumption on the variance of the gradient estimator, we obtain:

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\nabla}^m_{\theta,\phi}\mathcal{L}(\theta_k,\phi_k;\mathcal{D}_{k+1}) - \nabla^m_{\theta,\phi}\mathcal{L}(\theta_k,\phi_k)\right\|^2 \mid \mathcal{F}_k\right] &= \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B}\frac{1}{K}\sum_{\ell=1}^{K}\tilde{g}^m_{i,\ell} - \nabla^m_{\theta,\phi}\mathcal{L}(\theta_k,\phi_k)\right\|^2 \mid \mathcal{F}_k\right] \\
&\leq \frac{1}{BK}\mathbb{E}\left[\left\|\tilde{g}^m_{i,\ell} - \nabla^m_{\theta,\phi}\mathcal{L}(\theta_k,\phi_k)\right\|^2 \mid \mathcal{F}_k\right] \\
&\leq \frac{\sigma^2}{BK} .
\end{aligned}$$

Using this result, and noting that $\widehat{\nabla}^m_{\theta,\phi}\mathcal{L}(\theta,\phi;\mathcal{D}_{k+1})$ is an unbiased estimator of $\nabla^m_{\theta,\phi}\mathcal{L}(\theta,\phi)$, we get:

$$\mathbb{E}\left[\mathcal{L}\left(\theta_{k+1},\phi_{k+1}\right)\mid\mathcal{F}_k\right]\geq\mathcal{L}\left(\theta_k,\phi_k\right)+\gamma_{k+1}\|\nabla^m_{\theta,\phi}\mathcal{L}\left(\theta_k,\phi_k\right)\|^2-\frac{L\gamma_{k+1}^2}{2}\|\nabla^m_{\theta,\phi}\mathcal{L}\left(\theta_k,\phi_k\right)\|^2-\frac{L\gamma_{k+1}^2}{2}\frac{\sigma^2}{BK}\,.$$

Therefore,

$$\begin{aligned}\sum_{k=0}^{n}\left(\gamma_{k+1}-\frac{L\gamma_{k+1}^2}{2}\right)\mathbb{E}\left[\left\|\nabla^m_{\theta,\phi}\mathcal{L}\left(\theta_k,\phi_k\right)\right\|^2\right]&\leq\sum_{k=0}^{n}\mathbb{E}[\mathcal{L}\left(\theta_{k+1},\phi_{k+1}\right)-\mathcal{L}\left(\theta_k,\phi_k\right)]+\frac{L\sigma^2}{2BK}\sum_{k=0}^{n}\gamma_{k+1}^2\,,\\&\leq\mathbb{E}[\mathcal{L}\left(\theta_{n+1},\phi_{n+1}\right)]-\mathcal{L}\left(\theta_0,\phi_0\right)+\frac{L\sigma^2}{2BK}\sum_{k=0}^{n}\gamma_{k+1}^2\,.\end{aligned}$$

Consequently, by the definition of the discrete random variable R and choosing $\gamma_n=n^{-1/2}$,

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla^m_{\theta,\phi}\mathcal{L}\left(\theta_R,\phi_R\right)\right\|^2\right]&=\frac{1}{n}\sum_{k=0}^{n}\mathbb{E}\left[\left\|\nabla^m_{\theta,\phi}\mathcal{L}\left(\theta_k,\phi_k\right)\right\|^2\right]\\&\leq\sum_{k=0}^{n}\frac{\gamma_{k+1}}{\sum_{k=0}^{n}\gamma_{k+1}}\mathbb{E}\left[\left\|\nabla^m_{\theta,\phi}\mathcal{L}\left(\theta_k,\phi_k\right)\right\|^2\right]\\&\leq\frac{2\left(\mathbb{E}[\mathcal{L}\left(\theta_{n+1},\phi_{n+1}\right)]-\mathcal{L}\left(\theta_0,\phi_0\right)\right)+L\sigma^2\sum_{k=0}^{n}\gamma_{k+1}^2/(BK)}{\sqrt{n}}\,,\end{aligned}$$

which concludes the proof by noting that $\mathcal{L}\left(\theta_{n+1},\phi_{n+1}\right)\leq\mathcal{L}\left(\theta^*,\phi^*\right)$. □

B.3 Proof of Theorem 3.3

Proof. Since $\widehat{\nabla}_{\theta,\phi}\mathcal{L}\left(\theta,\phi;\mathcal{D}\right)$ is an unbiased estimator of the gradient of the expected ELBO, the proof is a direct consequence of (Shi and Li, 2021, Proposition 4.2) using the smoothness of the ELBO (Proposition 3.1). □

C LINEAR GAUSSIAN VAE

C.1 Analytic ELBO of the Linear VAE

While analytic solutions for deep latent models are generally not available, the Linear VAE provides analytic solutions for optimal parameters, allowing us to gain insights into various phenomena associated with VAE training. For instance, Dai et al. (2018) explore the connections between Linear VAE, probabilistic PCA (Tipping and Bishop, 1999), and robust PCA (Candès et al., 2011; Chandrasekaran et al., 2011), and analyze the local minima smoothing effects of VAE. Similarly, Lucas et al. (2019); Wang and Ziyin (2022) use Linear VAE to study the causes of posterior collapse Razavi et al. (2019). The following proposition provides the analytical form of the ELBO for the Linear VAE defined in (3), which is crucial for analyzing the convergence rate.

Proposition C.1. *The KL-divergence term and the reconstruction term can be expressed respectively for all* $x\in\mathsf{X}$ *as:*

$$KL(q_\phi(\cdot|x)\|p)=\frac{1}{2}\left(-\log\det D+\|W_2x+b_2\|^2+\operatorname{tr}(D)-d_z\right)\,,$$

$$\mathbb{E}_{q_\phi(\cdot|x)}\left[\log p_\theta(x|Z)\right]=\frac{1}{2c^2}\left[-\operatorname{tr}(W_1DW_1^\top)-\|W_1(W_2x+b_2)+b_1-x\|^2\right]-\frac{d_z}{2}\log 2\pi c^2\,.$$

Proof. First, we have:

$$\begin{aligned}\int q_\phi(z|x)\log p(z)\,\mathrm{d}z&=\mathbb{E}_{q_\phi(\cdot|x)}\left[-\frac{d_z}{2}\log 2\pi-\frac{1}{2}\|Z\|^2\right]\\&=-\frac{d_z}{2}\log 2\pi-\frac{1}{2}\sum_{i=1}^{d}\mathbb{E}_{q_\phi(\cdot|x)}\left[Z_i^2\right]\\&=-\frac{d_z}{2}\log 2\pi-\frac{1}{2}\left(\operatorname{tr}(D)+\|W_2x+b_2\|^2\right)\,.\end{aligned}$$

$$\int q_\phi(z|x)\log q_\phi(z|x)\,\mathrm{d}z = -\frac{d_z}{2}\log 2\pi - \frac{1}{2}\log\det D - \frac{1}{2}\mathbb{E}_{q_\phi(\cdot|x)}\left[(Z-W_2x-b_2)^\top D^{-1}(Z-W_2x-b_2)\right]$$
$$= -\frac{d_z}{2}\log 2\pi - \frac{1}{2}\log\det D - \frac{d_z}{2}\,.$$

By subtracting these two terms, we obtain the KL-divergence:

$$\mathrm{KL}(q_\phi(\cdot|x)||p) = \frac{1}{2}\left(-\log\det D + \|W_2x+b_2\|^2 + \mathrm{tr}(D) - d_z\right)\,.$$

For the reconstruction term, we have:

$$\begin{aligned}\mathbb{E}_{q_\phi(\cdot|x)}[\log p_\theta(x|Z)] &= -\frac{d_z}{2}\log 2\pi c^2 - \frac{1}{2c^2}\mathbb{E}_{q_\phi(\cdot|x)}\left[\|x-W_1Z-b_1\|^2\right]\\ &= -\frac{d_z}{2}\log 2\pi c^2 - \frac{1}{2c^2}\mathbb{E}_{q_\phi(\cdot|x)}\left[\|W_1Z\|^2 - 2(x-b_1)^\top W_1Z + \|x-b_1\|^2\right]\\ &= \frac{1}{2c^2}\left[-\mathrm{tr}(W_1DW_1^\top) - \|W_1(W_2x+b_2)\|^2 + 2(x-b_1)^\top W_1(W_2x+b_2) - \|x-b_1\|^2\right]\\ &\quad - \frac{d_z}{2}\log 2\pi c^2\,,\end{aligned}$$

where we used the fact that $W_1z \sim \mathcal{N}\left(W_1(W_2x+b_2), W_1DW_1^\top\right)$. □

C.2 Proof of Corollary 3.4

Proof. First, we compute the derivative of the ELBO with respect to each parameter of the encoder and decoder.

Derivatives of ELBO.

$$\nabla_{W_1}\mathcal{L}(\theta,\phi;x) = \frac{1}{c^2}\left((x-b_1)(W_2x+b_2)^\top - W_1D - W_1(W_2x+b_2)(W_2x+b_2)^\top\right)$$
$$\nabla_{W_2}\mathcal{L}(\theta,\phi;x) = \frac{1}{c^2}\left(W_1^\top(x-b_1)x^\top - W_1^\top W_1(W_2x+b_2)x^\top - c^2(W_2x+b_2)x^\top\right)$$
$$\nabla_{b_1}\mathcal{L}(\theta,\phi;x) = \frac{1}{c^2}\left(x-b_1-W_1(W_2x+b_2)\right)$$
$$\nabla_{b_2}\mathcal{L}(\theta,\phi;x) = \frac{1}{c^2}\left(W_1^\top(x-b_1) - W_1^\top W_1(W_2x+b_2) - c^2(W_2x+b_2)\right)$$
$$\nabla_D\mathcal{L}(\theta,\phi;x) = \frac{1}{2}\left(D^{-1} - \mathrm{I}_{d_z} - \frac{1}{c^2}\mathrm{diag}(W_1^\top W_1)\right)$$

Smoothness Property.

$$\left\|\nabla_{W_1}\mathcal{L}(\theta,\phi) - \nabla_{W_1'}\mathcal{L}(\theta,\phi)\right\| \le \frac{1}{c^2}\left(\|D\| + \left\|\mathbb{E}_\pi\left[(W_2x+b_2)(W_2x+b_2)^\top\right]\right\|\right)\|W_1-W_1'\|$$
$$\left\|\nabla_{W_2}\mathcal{L}(\theta,\phi) - \nabla_{W_2'}\mathcal{L}(\theta,\phi)\right\| \le \frac{1}{c^2}\left\|W_1^\top W_1 + c^2\mathrm{I}_{d_x}\right\|\left\|\mathbb{E}_\pi\left[xx^\top\right]\right\|\|W_2-W_2'\|$$
$$\left\|\nabla_{b_1}\mathcal{L}(\theta,\phi) - \nabla_{b_1'}\mathcal{L}(\theta,\phi)\right\| = \frac{1}{c^2}\|b_1-b_1'\|$$
$$\left\|\nabla_{b_2}\mathcal{L}(\theta,\phi) - \nabla_{b_2'}\mathcal{L}(\theta,\phi)\right\| \le \frac{1}{c^2}\left\|W_1^\top W_1 + c^2\mathrm{I}_{d_x}\right\|\|b_2-b_2'\|$$
$$\|\nabla_D\mathcal{L}(\theta,\phi) - \nabla_{D'}\mathcal{L}(\theta,\phi)\| = \frac{1}{2}\left\|D^{-1}-D'^{-1}\right\| \le \frac{1}{2}\left\|D^{-1}\right\|\left\|D'^{-1}\right\|\|D-D'\|$$

If $D = \mathrm{Diag}(\sigma_1^2,\dots,\sigma_{d_z}^2)$ as used in practice, the last inequality can be expressed as:

$$|\nabla_{\sigma^2}\mathcal{L}(\theta,\phi) - \nabla_{\sigma'^2}\mathcal{L}(\theta,\phi)| = \frac{1}{2\sigma^2\sigma'^2}\left|\sigma^2-\sigma'^2\right|.$$

Since the parameter space is compact, $\lambda_{\min}(D) \ge c_D$ for some $c_D > 0$, and the inputs have bounded second moments, it follows that the ELBO is smooth. The proof is completed using Theorem 3.3.

□

D DEEP GAUSSIAN VAE

D.1 Activation Functions in Deep Gaussian VAE

Definition D.1. The activation functions Sigmoid, Hyperbolic Tangent (Tanh), Softplus, and Continuously Differentiable Exponential Linear Units (CELU) are defined for all $x \in \mathbb{R}$ as follows:

$$\text{Sigmoid}(x) = \frac{1}{1+\mathrm{e}^{-x}}, \quad \text{Tanh}(x) = \frac{\mathrm{e}^{x}-\mathrm{e}^{-x}}{\mathrm{e}^{x}+\mathrm{e}^{-x}}, \quad \text{Softplus}(x) = \ln(1+\mathrm{e}^{x}) \ ,$$

and

$$\text{CELU}(x) = \begin{cases} x & \text{if } x > 0, \\ \alpha\left(\exp\left(\frac{x}{\alpha}\right)-1\right) & \text{if } x \leq 0. \end{cases}$$

Consider the following neural network formulations for the encoder and decoder:

$$\mathcal{F}_G = \Big\{ G_\theta : \mathbb{R}^{d_z} \to \mathbb{R}^{d_x} \, ; \, G_\theta(z) = \mathsf{NN}(z;\theta,f,N_{dd}),\, \theta = \{W_\ell, b_\ell\}_{\ell=1}^{N_{dd}} \in \Theta, \\ \sigma_\ell \in \mathcal{F}_{\text{SL}}, \ell = 1,\dots,N_{dd}, \text{ and } \|G_\theta(z)\| \leq C_G \Big\},$$

and

$$\mathcal{F}_{\mu,\Sigma} = \Big\{ (\mu_\phi,\Sigma_\phi) : \mathbb{R}^{d_x} \to \mathbb{R}^{d_z} \times \mathbb{R}^{d_z \times d_z} \, ; \, (\mu_\phi(x),\Sigma_\phi(x)) = \mathsf{NN}(x;\phi,f,N_{ed}),\, \phi = \{W_\ell, b_\ell\}_{\ell=1}^{N_{ed}} \in \Phi, \\ f_\ell \in \mathcal{F}_{\text{SL}}, \ell = 1,\dots,N_{ed},\, \|\mu_\phi(x)\| \leq C_\mu, \quad \lambda_{\min}(\Sigma_\phi(x)) \geq c_\Sigma, \text{ and } \|\Sigma_\phi(x)\| \leq C_\Sigma \Big\},$$

where $\mathcal{F}_{\text{SL}}$ denotes the set of functions that are both smooth and Lipschitz continuous and $\mathcal{F}_{\text{b}}$ denotes the set of bounded functions.

In $\mathcal{F}_G$ and $\mathcal{F}_{\mu,\Sigma}$, f represents an activation function, specifically one chosen from the set that includes the sigmoid, hyperbolic tangent (tanh), softplus (Glorot et al., 2011), or Continuously Differentiable Exponential Linear Units (CELU) (Clevert et al., 2016; Barron, 2017). These activation functions are crucial in neural network architectures as they ensure the assumptions made about the encoder and decoder distributions.

Proposition D.2. *The activation functions sigmoid, tanh, softplus, and CELU are Lipschitz continuous and smooth.*

Proof. We consider each activation function separately.

Sigmoid. The first and second derivatives of the sigmoid function are given for all x by:

$$f_1'(x) = f_1(x)(1-f_1(x)) \quad \text{and} \quad f_1''(x) = f_1(x)(1-f_1(x))(1-2f_1(x)) \ .$$

Since the sigmoid function is bounded by 1, it follows that for all x, $|f_1'(x)| \leq 1/4$ and $|f_1''(x)| \leq 1/4$. Therefore, the sigmoid activation function is Lipschitz continuous and smooth.

Tanh. The first and second derivatives of the hyperbolic tangent function are given for all x by:

$$f_2'(x) = 1 - f_2^2(x) \quad \text{and} \quad f_2''(x) = -2f_2(x)(1-f_2^2(x)) \ .$$

Since f_2 is bounded by 1, for all x, $|f_2'(x)| \leq 1$ and $|f_2''(x)| \leq 1$. Therefore, the Tanh activation function is Lipschitz continuous and smooth.

Softplus. The first and second derivatives of the softplus function are given for all x by:

$$f_3'(x) = f_1(x) \quad \text{and} \quad f_3''(x) = f_1(x)(1-f_1(x)) \ .$$

Since f_1 is bounded by 1, it follows that $|f_3'(x)| \leq 1$ and $|f_3''(x)| \leq 1/4$. Therefore, the softplus activation function is Lipschitz continuous and smooth.

CELU. The first and second derivatives of the CELU function are given for all x by:

$$f_4'(x) = \begin{cases} 1 & \text{if } x > 0, \\ \exp\left(\frac{x}{\alpha}\right) & \text{if } x \leq 0, \end{cases}$$

$$f_4''(x) = \begin{cases} 0 & \text{if } x > 0, \\ \frac{1}{\alpha}\exp\left(\frac{x}{\alpha}\right) & \text{if } x \leq 0. \end{cases}$$

Since $\mathrm{e}^x \leq 1$ for $x \leq 0$, it follows that $|f_4'(x)| \leq 1$ and $|f_4''(x)| \leq 1/\alpha$. Therefore, the CELU activation function is Lipschitz continuous and smooth. □

Proposition D.2 highlights that activation functions such as sigmoid, tanh, softplus, and CELU are suitable for use in the encoder and decoder of a network due to their Lipschitz continuity and smoothness.

D.2 Proof of Proposition 3.5

Proof. The first derivative of f is given by:

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

Given that $s_1 \leq s_2$ and that the sigmoid function is increasing, we have $f'(x) \geq 0$. Therefore, f is an increasing function. Additionally, since $f(x) \to s_1$ as $x \to -\infty$ and $f(x) \to s_2$ as $x \to +\infty$, it follows that f is bounded between s_1 and s_2. Furthermore, since $|f'(x)| \leq 1/s$, f is Lipchitz continuous. The second derivative of f is given by:

$$f''(x) = \frac{e^{s(x-s_1)}}{(1+e^{s(x-s_1)})^2} - \frac{e^{s(x-s_2)}}{(1+e^{s(x-s_2)})^2} .$$

Since $|f''(x)| \leq 1/2$, it follows that f is smooth. □

D.3 Proof of Theorem 3.6

First, the ELBO objective can be expressed for all $x \in \mathsf{X}$ as:

$$\mathcal{L}(\theta, \phi; x) = -\mathrm{KL}(q_\phi(\cdot|x)\|p) - \frac{1}{2c^2}\mathbb{E}_{q_\phi(\cdot|x)}\left[\|G_\theta(z) - x\|^2\right] - \frac{1}{2}\log(2\pi c^2).$$

We divide the proof into two parts: Section D.3.1 establishes the convergence rate using the score function gradient, and Section D.3.2 presents the convergence rate with the pathwise gradient.

D.3.1 Analysis with the score function

We first analyze the convergence rate using the score function in the Gaussian case.

Theorem D.3. *Let $c_0 > 0$, and consider*

$$\mathcal{F}_{dd} = \{(x,z) \mapsto p_\theta(x|z) = \mathcal{N}(x; G_\theta(z), c^2 \mathrm{I}_{d_x}) \,|\, G_\theta(z) \in \mathcal{F}_G, \theta \in \Theta \subseteq \mathbb{R}^{d_\theta}, c > c_0\} ,$$
$$\mathcal{F}_{ed} = \{(x,z) \mapsto q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \Sigma_\phi(x)) \,|\, (\mu_\phi(x), \Sigma_\phi(x)) \in \mathcal{F}_{\mu,\Sigma}, \phi \in \Phi \subseteq \mathbb{R}^{d_\phi}\} .$$

Assume that there exists $C_{rec} \in \mathsf{M}(\mathsf{X} \times \mathsf{Z})$ such that $\|x - G_\theta(z)\| \leq C_{rec}(x,z)$ for all $\theta \in \Theta$ and $(x,z) \in \mathsf{X} \times \mathsf{Z}$. Assume also that the data distribution π has a finite fourth moment, and that there exists some constant a such that for all $\theta \in \Theta$ and $\phi \in \Phi$,

$$\|\theta\|_\infty + \|\phi\|_\infty \leq a .$$

Let $(\theta_n, \phi_n) \in \Theta \times \Phi$ be the n-th iterate of the recursion in Algorithm 1, where $\gamma_n = C_\gamma n^{-1/2}$ with $C_\gamma > 0$. Assume that $\beta_1 < \sqrt{\beta_2} < 1$. For any $n \geq 1$, let $R \in \{0, \ldots, n\}$ be a uniformly distributed random variable. Then,

$$\mathbb{E}\left[\left\|\nabla\mathcal{L}\left(\theta_R, \phi_R\right)\right\|^2\right] = \mathcal{O}\left(\frac{\mathcal{L}^*}{\sqrt{n}} + d_z^2 \frac{N_{max} a^{2(N_{max}-1)}}{1-\beta_1}\frac{d^* \log n}{\sqrt{n}}\right) ,$$

where $\mathcal{L}^* = \mathcal{L}\left(\theta^*, \phi^*\right) - \mathcal{L}\left(\theta_0, \phi_0\right)$, $d^* = d_\theta + d_\phi$ is the total dimension of the parameters, $N_{max} = \max\{N_{ed}, N_{dd}\}$ represents the maximum number of layers in the model architecture.

We divide the proof into two lemmas. Lemma D.4 ensures Assumption 1, while Lemma D.5 guarantees Assumption 2(i).

Lemma D.4. *Let $c_0 > 0$, and consider*

$$\mathcal{F}_{dd} = \{(x,z) \mapsto p_\theta(x|z) = \mathcal{N}(x; G_\theta(z), c^2 \mathrm{I}_{d_x}) \,|\, G_\theta(z) \in \mathcal{F}_G, \theta \in \Theta \subseteq \mathbb{R}^{d_\theta}, c > c_0\} \,,$$
$$\mathcal{F}_{ed} = \{(x,z) \mapsto q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \Sigma_\phi(x)) \,|\, (\mu_\phi(x), \Sigma_\phi(x)) \in \mathcal{F}_{\mu,\Sigma}, \phi \in \Phi \subseteq \mathbb{R}^{d_\phi}\} \,.$$

Then, Assumption 1 is satisfied i.e. for all $\theta \in \Theta$, $\phi \in \Phi$, $x \in \mathsf{X}$ and $z \in \mathsf{Z}$,

$$\max\{|\log p_\theta(x,z)| \,, |\log q_\phi(z|x)|\} \leq \alpha(x,z) \,,$$

with

$$\alpha(x,z) = \max\left\{ \frac{d_z}{2} \log(2\pi C_\Sigma) + \frac{1}{c_\Sigma} \left(\|z\|^2 + C_\mu^2\right), \frac{d_z}{2} \log(2\pi c^2) + \frac{1}{c^2} \left(\|x\|^2 + C_G^2\right) \right\} \,.$$

Proof. For all $x \in \mathsf{X}$, the density function $z \mapsto q_\phi(z|x)$ of the Gaussian encoder is given by:

$$q_\phi(z|x) = \det(2\pi\Sigma_\phi(x))^{-1/2} \exp\left(-\frac{1}{2} (z - \mu_\phi(x))^\top \Sigma_\phi^{-1}(x)(z - \mu_\phi(x)) \right) \,,$$

where $\mu_\phi(x)$ and $\Sigma_\phi(x)$ are the mean and covariance matrix of the Gaussian distribution.

Bounding the Normalization Factor. The determinant of $\Sigma_\phi(x)$ can be expressed as:

$$\det(\Sigma_\phi(x)) = \prod_{i=1}^{d_z} \lambda_i(x),$$

where $\lambda_1(x), \lambda_2(x), \ldots, \lambda_{d_z}(x)$ are the eigenvalues of $\Sigma_\phi(x)$. Given the conditions $\lambda_{\min}(\Sigma_\phi(x)) \geq c_\Sigma$ and $\|\Sigma_\phi(x)\| \leq C_\Sigma$, it follows that:

$$c_\Sigma^{d_z} \leq \det(\Sigma_\phi(x)) \leq C_\Sigma^{d_z} \,.$$

Thus, the normalization factor is bounded by:

$$(2\pi C_\Sigma)^{-d_z/2} \leq \det(2\pi\Sigma_\phi(x))^{-1/2} \leq (2\pi c_\Sigma)^{-d_z/2} \,.$$

Bounding the Exponential Term. We have:

$$\exp\left(-\frac{1}{2c_\Sigma} \|z - \mu_\phi(x)\|^2 \right) \leq \exp\left(-\frac{1}{2} (z - \mu_\phi(x))^\top \Sigma_\phi^{-1}(x)(z - \mu_\phi(x)) \right) \leq \exp\left(-\frac{1}{2C_\Sigma} \|z - \mu_\phi(x)\|^2 \right) \,.$$

Since for all $x \in \mathsf{X}$ $\|\mu_\phi(x)\| \leq C_\mu$, for all $x \in \mathsf{X}$, and $z \in \mathsf{Z}$,

$$\|z - \mu_\phi(x)\|^2 \leq 2\|z\|^2 + 2\|\mu_\phi(x)\|^2 \leq 2\|z\|^2 + 2C_\mu^2 \,.$$

Similarly, we can bound from below using:

$$\|z - \mu_\phi(x)\|^2 \geq \frac{1}{2} \left(\|z\|^2 - \|\mu_\phi(x)\|^2 \right) \geq \frac{1}{2} \left(\|z\|^2 - C_\mu^2 \right) \,.$$

Therefore, we have uniform bounds on the exponential term that are independent of the encoder parameters ϕ:

$$\exp\left(-\frac{1}{c_\Sigma} \left(\|z\|^2 + C_\mu^2 \right) \right) \leq \exp\left(-\frac{1}{2} (z - \mu_\phi(x))^\top \Sigma_\phi^{-1}(x)(z - \mu_\phi(x)) \right) \leq \exp\left(-\frac{1}{4C_\Sigma} \left(\|z\|^2 - C_\mu^2 \right) \right) \,.$$

Combining these results, we obtain:

$$\frac{1}{(2\pi C_\Sigma)^{d_z/2}} \exp\left(-\frac{1}{c_\Sigma} \left(\|z\|^2 + C_\mu^2 \right) \right) \leq q_\phi(z|x) \leq \frac{1}{(2\pi c_\Sigma)^{d_z/2}} \exp\left(-\frac{1}{4C_\Sigma} \left(\|z\|^2 - C_\mu^2 \right) \right) \,.$$

Since the activation function in the final layer is bounded, there exists a constant $C_G > 0$ such that for all $z \in \mathsf{Z}$, $\|G_\theta(z)\| \leq C_G$. Proceeding similarly for the Gaussian decoder, where the density function is given by $x \mapsto p_\theta(x|z) = \mathcal{N}(x; G_\theta(z), c^2 \mathrm{I}_{d_x})$ yields:

$$\frac{1}{(2\pi c^2)^{d_z/2}} \exp\left(-\frac{1}{c^2}\left(\|x\|^2 + C_G^2\right)\right) \leq p_\theta(x|z) \leq \frac{1}{(2\pi c^2)^{d_z/2}} \exp\left(-\frac{1}{4c^2}\left(\|x\|^2 - C_G^2\right)\right) .$$

This implies that Assumption 1 is verified. □

Lemma D.5 shows that Assumption 2 holds without explicitly specifying the smoothness constant. However, Lemma D.6 provides the smoothness constant L^{S}, and also shows that it is well-defined and finite.

Lemma D.5. *Let $c_0 > 0$, and consider*

$$\mathcal{F}_{dd} = \{(x,z) \mapsto p_\theta(x|z) = \mathcal{N}(x; G_\theta(z), c^2 \mathrm{I}_{d_x}) \, ; \, G_\theta \in \mathcal{F}_G, \theta \in \Theta \subseteq \mathbb{R}^{d_\theta}, c > c_0\} ,$$
$$\mathcal{F}_{ed} = \{(x,z) \mapsto q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \Sigma_\phi(x)) \, ; \, (\mu_\phi(x), \Sigma_\phi(x)) \in \mathcal{F}_{\mu,\Sigma}, \phi \in \Phi \subseteq \mathbb{R}^{d_\phi}\} .$$

Assume that there exists $C_{rec} \in \mathsf{M}(\mathsf{X} \times \mathsf{Z})$ such that $\|x - G_\theta(z)\| \leq C_{rec}(x,z)$ for all $\theta \in \Theta$ and $(x,z) \in \mathsf{X} \times \mathsf{Z}$. and that there exists some constant a such that for any $\theta \in \Theta$ and $\phi \in \Phi$,

$$\|\theta\|_\infty + \|\phi\|_\infty \leq a .$$

Then, there exist $L_1 \in \mathsf{M}(\mathsf{X} \times \mathsf{Z})$ and $L_2 \in \mathsf{M}(\mathsf{X} \times \mathsf{Z})$ such that for all $\theta, \theta' \in \Theta$, $\phi, \phi' \in \Phi$, $x \in \mathsf{X}$ and $z \in \mathsf{Z}$:

$$\|\nabla_\theta \log p_\theta(x|z) - \nabla_\theta \log p_{\theta'}(x|z)\| \leq L_1(x,z) \|\theta - \theta'\| ,$$
$$\|\nabla_\phi \log q_\phi(z|x) - \nabla_\phi \log q_{\phi'}(z|x)\| \leq L_2(x,z) \|\phi - \phi'\| .$$

Proof. **Gaussian Decoder.** First, the gradient of $\log p_\theta(x|z)$ is given by:

$$\nabla_\theta \log p_\theta(x|z) = \frac{1}{c^2} \nabla_\theta G_\theta(z)^\top (x - G_\theta(z)) .$$

We have:

$$\begin{aligned}
\|\nabla_\theta \log p_\theta(x|z) - \nabla_\theta \log p_{\theta'}(x|z)\| &= \frac{1}{c^2} \left\|\nabla_\theta G_\theta(z)^\top (x - G_\theta(z)) - \nabla_\theta G_{\theta'}(z)^\top (x - G_{\theta'}(z))\right\| \\
&\leq \frac{1}{c^2} \left\|\nabla_\theta G_\theta(z)^\top (x - G_\theta(z)) - \nabla_\theta G_\theta(z)^\top (x - G_{\theta'}(z))\right\| \\
&\quad + \frac{1}{c^2} \left\|\nabla_\theta G_\theta(z)^\top (x - G_{\theta'}(z)) - \nabla_\theta G_{\theta'}(z)^\top (x - G_{\theta'}(z))\right\| \\
&\leq \frac{1}{c^2} \left(\|\nabla_\theta G_\theta(z)\| \, \|G_\theta(z) - G_{\theta'}(z)\| + \|x - G_{\theta'}(z)\| \, \|\nabla_\theta G_\theta(z) - \nabla_\theta G_{\theta'}(z)\|\right) ,
\end{aligned}$$

Let M_{f_i} and L_{f_i} represent the Lipschitz constant and the smoothness parameter of the activation function f_i in the i-th layer, respectively. Using Lemma I.5, we get:

$$\|\nabla_\theta G_\theta(z)\| \leq (\|z\| + 1) \, a^{N_{dd}-1} \prod_{j=1}^{N_{dd}} M_{f_j} =: M_{\nabla G}(z) ,$$

$$\|\nabla_\theta G_\theta(z) - \nabla_\theta G_{\theta'}(z)\| \leq L_{\nabla G}(z) \|\theta - \theta'\| ,$$

where $L_{\nabla G}(z) = N_{dd}\left(\|z\|^2 + 1\right) \sum_{k=1}^{N_{dd}} L_{f_k} a^{N_{dd}-2+k} \prod_{i=1}^{k-1} M_{f_i}^2 \prod_{i=k+1}^{N_{dd}} M_{f_i}$. Therefore,

$$\begin{aligned}
\|\nabla_\theta \log p_\theta(x|z) - \nabla_\theta \log p_{\theta'}(x|z)\| &\leq \frac{1}{c^2}\left(\|\nabla_\theta G_\theta(z)\|^2 + L_{\nabla G} \|x - G_{\theta'}(z)\|\right) \|\theta - \theta'\| \\
&\leq L^G(x,z) \|\theta - \theta'\| ,
\end{aligned}$$

where $L^G(x,z) = \left(M_{\nabla G}(z)^2 + C_{rec}(x,z) L_{\nabla G}(z)\right)/c^2$.

Gaussian Encoder. Consider the Gaussian encoder E_ϕ, parameterized by a mean function $\mu_\phi(x)$ and a diagonal covariance matrix $\Sigma_\phi(x) = \mathrm{Diag}(\sigma_1^2(x), \dots, \sigma_{d_z}^2(x))$. We define $l_\phi(x) = (\log \sigma_1^2(x), \dots, \log \sigma_{d_z}^2(x))$ as the logarithm of the variance. The proof for the Gaussian encoder can be treated similarly to that of the decoder. However, it differs in that we also learn the variance of the Gaussian distribution. Since we use the same neural network architecture for both $\mu_\phi(x)$ and $l_\phi(x)$ as is used for $G_\theta(z)$, the mappings $\phi \mapsto \mu_\phi(x)$ and $\phi \mapsto l_\phi(x)$ are smooth functions with bounded gradients. First, the log density of the variational distribution can be expressed as:

$$\begin{aligned}\log q_\phi(z|x) &= -\frac{1}{2}(\mu_\phi(x) - z)^\top \Sigma_\phi^{-1}(x)(\mu_\phi(x) - z) - \frac{1}{2}\log(2\pi \det \Sigma_\phi(x)) \\ &= -\frac{1}{2}(\mu_\phi(x) - z)^\top e^{-l_\phi(x)} \odot (\mu_\phi(x) - z) - \mathrm{tr}(L_\phi(x)) - \frac{1}{2}\log(2\pi) \ ,\end{aligned}$$

where $L_\phi(x) = \mathrm{Diag}(l_\phi(x))$. Then, the gradient is given by:

$$\nabla_\phi \log q_\phi(z|x) = \nabla_\phi \mu_\phi(x)^\top e^{-l_\phi(x)} \odot (z - \mu_\phi(x)) + \frac{1}{2}\nabla_\phi l_\phi(x)^\top e^{-l_\phi(x)} \odot (z - \mu_\phi(x))^2 - \nabla_\phi l_\phi(x)^\top \mathbf{1} \ .$$

Therefore, we have:

$$\|\nabla_\phi \log q_\phi(z|x) - \nabla_\phi \log q_{\phi'}(z|x)\| \le A_1 + A_2 + A_3 \ ,$$

where

$$\begin{aligned}A_1 &= \|\nabla_\phi \mu_\phi(x)^\top e^{-l_\phi(x)} \odot (z - \mu_\phi(x)) - \nabla_\phi \mu_{\phi'}(x)^\top e^{-l_{\phi'}(x)} \odot (z - \mu_{\phi'}(x))\| \ , \\ A_2 &= \frac{1}{2}\|\nabla_\phi l_\phi(x)^\top e^{-l_\phi(x)} \odot (z - \mu_\phi(x))^2 - \nabla_\phi l_{\phi'}(x)^\top e^{-l_{\phi'}(x)} \odot (z - \mu_{\phi'}(x))^2\| \ , \\ A_3 &= \|\nabla_\phi l_\phi(x) - \nabla_\phi l_{\phi'}(x)\| \ .\end{aligned}$$

Term A_1 is upper bounded as follows

$$\begin{aligned}A_1 &= \|\nabla_\phi \mu_\phi(x)^\top \mathrm{e}^{-l_\phi(x)} \odot (z - \mu_\phi(x)) - \nabla_\phi \mu_{\phi'}(x)^\top \mathrm{e}^{-l_{\phi'}(x)} \odot (z - \mu_{\phi'}(x))\| \\ &\le \|\nabla_\phi \mu_\phi(x)^\top \mathrm{e}^{-l_\phi(x)} \odot (z - \mu_\phi(x)) - \nabla_\phi \mu_{\phi'}(x)^\top \mathrm{e}^{-l_\phi(x)} \odot (z - \mu_\phi(x))\| \\ &\quad + \|\nabla_\phi \mu_{\phi'}(x)^\top \mathrm{e}^{-l_\phi(x)} \odot (z - \mu_\phi(x)) - \nabla_\phi \mu_{\phi'}(x)^\top \mathrm{e}^{-l_{\phi'}(x)} \odot (z - \mu_{\phi'}(x))\| \\ &\le \|\mathrm{e}^{-l_\phi(x)}\|\|z - \mu_\phi(x)\|\|\nabla_\phi \mu_\phi(x) - \nabla_\phi \mu_{\phi'}(x)\| \\ &\quad + \|\nabla_\phi \mu_{\phi'}(x)\|\|\mathrm{e}^{-l_\phi(x)} \odot (z - \mu_\phi(x)) - \mathrm{e}^{-l_{\phi'}(x)} \odot (z - \mu_{\phi'}(x))\| \\ &\le \|\mathrm{e}^{-l_\phi(x)}\|\|z - \mu_\phi(x)\|\|\nabla_\phi \mu_\phi(x) - \nabla_\phi \mu_{\phi'}(x)\| + \|\nabla_\phi \mu_{\phi'}(x)\|\|z - \mu_\phi(x)\|\|\mathrm{e}^{-l_\phi(x)} - \mathrm{e}^{-l_{\phi'}(x)}\| \\ &\quad + \|\nabla_\phi \mu_{\phi'}(x)\|\|\mathrm{e}^{-l_{\phi'}(x)}\|\|\mu_\phi(x) - \mu_{\phi'}(x))\| \\ &\le \|\mathrm{e}^{-l_\phi(x)}\|\|z - \mu_\phi(x)\|\|\nabla_\phi \mu_\phi(x) - \nabla_\phi \mu_{\phi'}(x)\| + \|\nabla_\phi \mu_{\phi'}(x)\|\|z - \mu_\phi(x)\|\big(\sup_{\phi \in \Phi} \|\mathrm{e}^{-l_\phi(x)}\|\big)\|l_\phi(x) - l_{\phi'}(x))\| \\ &\quad + \|\nabla_\phi \mu_{\phi'}(x)\|\|\mathrm{e}^{-l_{\phi'}(x)}\|\|\mu_\phi(x) - \mu_{\phi'}(x))\| \ ,\end{aligned}$$

where we used the Mean Value Theorem in the last inequality. Given the conditions that $\|\mathrm{e}^{-l_\phi(x)}\| \le 1/c_\Sigma(x)$, $\|z - \mu_\phi(x)\| \le \|z\| + C_\mu$, and considering the boundedness and smoothness of the functions $\phi \mapsto \mu_\phi(x)$ and $\phi \mapsto l_\phi(x)$, there exists $L_1^{\mu,\Sigma}$ such that

$$A_1 \le L_1^{\mu,\Sigma}(x, z)\|\phi - \phi'\| \ .$$

For the second term A_2, write

$$
\begin{aligned}
2A_2 &= \|\nabla_\phi l_\phi(x)^\top \mathrm{e}^{-l_\phi(x)} \odot (z-\mu_\phi(x))^2 - \nabla_\phi l_{\phi'}(x)^\top \mathrm{e}^{-l_{\phi'}(x)} \odot (z-\mu_{\phi'}(x))^2\| \\
&\leq \|\nabla_\phi l_\phi(x)^\top \mathrm{e}^{-l_\phi(x)} \odot (z-\mu_\phi(x))^2 - \nabla_\phi l_{\phi'}(x)^\top \mathrm{e}^{-l_\phi(x)} \odot (z-\mu_\phi(x))^2\| \\
&\quad + \|\nabla_\phi l_{\phi'}(x)^\top \mathrm{e}^{-l_\phi(x)} \odot (z-\mu_\phi(x))^2 - \nabla_\phi l_{\phi'}(x)^\top \mathrm{e}^{-l_{\phi'}(x)} \odot (z-\mu_{\phi'}(x))^2\| \\
&\leq \|\mathrm{e}^{-l_\phi(x)}\|\|z-\mu_\phi(x)\|^2\|\nabla_\phi l_\phi(x) - \nabla_\phi l_{\phi'}(x)\| \\
&\quad + \|\nabla_\phi l_{\phi'}(x)\|\|\mathrm{e}^{-l_\phi(x)} \odot (z-\mu_\phi(x))^2 - \mathrm{e}^{-l_{\phi'}(x)} \odot (z-\mu_{\phi'}(x))^2\| \\
&\leq \|\mathrm{e}^{-l_\phi(x)}\|\|z-\mu_\phi(x)\|^2\|\nabla_\phi l_\phi(x) - \nabla_\phi l_{\phi'}(x)\| + \|\nabla_\phi l_{\phi'}(x)\|\|z-\mu_\phi(x)\|^2\|\mathrm{e}^{-l_\phi(x)} - \mathrm{e}^{-l_{\phi'}(x)}\| \\
&\quad + \|\nabla_\phi l_{\phi'}(x)\|\|\mathrm{e}^{-l_{\phi'}(x)}\|\|(z-\mu_\phi(x))^2 - (z-\mu_{\phi'}(x))^2\| \\
&\leq \|\mathrm{e}^{-l_\phi(x)}\|\|z-\mu_\phi(x)\|^2\|\nabla_\phi l_\phi(x) - \nabla_\phi l_{\phi'}(x)\| + \|\nabla_\phi l_{\phi'}(x)\|\|z-\mu_\phi(x)\|^2\big(\sup_{\phi\in\Phi}\|\mathrm{e}^{-l_\phi(x)}\|\big)\|l_\phi(x) - l_{\phi'}(x))\| \\
&\quad + \|\nabla_\phi l_{\phi'}(x)\|\|\mathrm{e}^{-l_{\phi'}(x)}\|\|\mu_\phi(x) - \mu_{\phi'}(x)\|\|2z - \mu_\phi(x) - \mu_{\phi'}(x)\| \,,
\end{aligned}
$$

where we used the Mean Value Theorem and the equality $\|(z-\mu_\phi(x))^2 - (z-\mu_{\phi'}(x))^2\| = \|\mu_\phi(x) - \mu_{\phi'}(x)\|\|2z - \mu_\phi(x)) - \mu_{\phi'}(x)\|$ in the last inequality. Given the conditions that $\|\mathrm{e}^{-l_\phi(x)}\| \leq 1/c_\Sigma(x)$, $\|z-\mu_\phi(x)\| \leq \|z\| + C_\mu$, $\|2z - \mu_\phi(x)) - \mu_{\phi'}(x)\| \leq 2(\|z\| + C_\mu)$ and considering the boundedness and smoothness of the functions $\phi \mapsto \mu_\phi(x)$ and $\phi \mapsto l_\phi(x)$, there exists $L_2^{\mu,\Sigma}$ such that

$$
A_2 \leq L_2^{\mu,\Sigma}(x,z)\|\|\phi - \phi'\| \,.
$$

Finally, for the last term A_3, using the smoothness of $\phi \mapsto l_\phi(x)$, there exists $L_3^{\mu,\Sigma}$ such that

$$
A_3 \leq L_3^{\mu,\Sigma}(x,z)\|\phi - \phi'\| \,.
$$

□

Lemma D.6. *Under the assumptions of Lemma D.5, the smoothness constant L^{S} of the expected ELBO is well-defined and finite.*

Proof. Using B.1, the smoothness constant L^{S} is defined by:

$$
L^{\mathsf{S}} = \sup_{\phi\in\Phi} \mathbb{E}_{\pi,\phi}\left[L_1(x,z) + 2\alpha(x,z)L_2(x,z) + 4M(x,z)^2 + 4\alpha(x,z)M(x,z)^2\right] \,.
$$

Let L_{dd}^{S} and L_{ed}^{S} denote the smoothness constants of the decoder and the encoder, respectively. For the smoothness constant of the decoder, applying Lemma I.5, we have:

$$
\begin{aligned}
L_{dd} &= \mathbb{E}_{\pi,\phi}\left[L_1(x,z) + M(x,z)^2\right] \\
&= \frac{1}{c^2}\mathbb{E}_{\pi,\phi}\left[2M_{\nabla G}(z)^2 + C_{rec}(x,z)L_{\nabla G}(z)\right] \\
&= \frac{1}{c^2}\mathbb{E}_{\pi,\phi}\left[2\left(\|z\|+1\right)^2 a^{2(N_{dd}-1)}\prod_{i=1}^{N_{dd}} M_{f_i}^2 + N_{dd}C_{rec}(x,z)\left(\|z\|^2+1\right)\sum_{k=1}^{N_{dd}} L_{f_k} a^{N_{dd}-2+k}\prod_{i=1}^{k-1} M_{f_i}^2 \prod_{i=k+1}^{N_{dd}} M_{f_i}\right] \,.
\end{aligned}
$$

Since

$$
\mathbb{E}_{q_\phi(\cdot|x)}\left[\|z\|^2\right] = \mathrm{Tr}(\Sigma_\phi(x)) + \|\mu_\phi(x)\|^2 \leq d_z\|\Sigma_\phi(x)\| + C_\mu \leq d_z C_\Sigma + C_\mu \,,
$$

it follows that L_{dd}^{S} is well-defined and finite.

For the smoothness constant of the encoder, we have:

$$
L_{ed}^{\mathsf{S}} = \mathbb{E}_{\pi,\phi}\left[2\alpha(x,z)L_2(x,z) + 3M(x,z)^2 + 4\alpha(x,z)M(x,z)^2\right] \,,
$$

where

$$M(x,z) = (\|x\|+1)\, a^{N_{ed}-1} \prod_{i=1}^{N_{ed}} M_{f_i} \ ,$$

$$L_2(x,z) = \frac{N_{ed}}{c_\Sigma}\left(2\|z\|^2 + 2C_\mu^2 + \|z\| + C_\mu + c_\Sigma\right)\|x\|^2 \sum_{k=1}^{N_{ed}} L_{f_k} a^{N_{ed}-2+k} \prod_{i=1}^{k-1} M_{f_i}^2 \prod_{i=k+1}^{N_{ed}} M_{f_i}$$
$$+ \frac{N_{ed}}{c_\Sigma}\left(2\|z\|^2 + 2C_\mu^2 + 3\|z\| + 3C_\mu + 1\right)\|x\|^2 a^{2(N_{ed}-1)} \prod_{i=1}^{N} M_{f_i}^2 \ ,$$

$$\alpha(x,z) = \max\left\{\frac{d_z}{2}\log(2\pi C_\Sigma) + \frac{1}{c_\Sigma}\left(\|z\|^2 + C_\mu^2\right), \frac{d_z}{2}\log(2\pi c^2) + \frac{1}{c^2}\left(\|x\|^2 + C_G^2\right)\right\} \ .$$

Since π has a finite fourth moment, it is evident that L_{ed}^{S} is well-defined and finite.

Moreover, under the conditions $M_{f_i} \leq 1$ and $L_{f_i} \leq 1$ for all $1 \leq i \leq N$, we can establish the following bounds:

$$L_{dd}^{\mathsf{S}} = \mathcal{O}(d_z N_{dd} a^{2(N_{dd}-1)}) \ \text{ and } \ L_{ed}^{\mathsf{S}} = \mathcal{O}(d_z^2 N_{ed} a^{2(N_{ed}-1)}) \ ,$$

where constants in these bounds depend on additional terms, but we focus here only on the dimensions of the latent space and the number of layers in the model architecture. □

Proof of Theorem D.3. Using Lemmas D.4 and D.5, we ensure that Assumptions 1 and 2(i) are satisfied. The proof is then completed by applying Theorem 3.3. □

D.3.2 Analysis with the pathwise gradient

We now analyze the convergence rate using the pathwise gradient in the Gaussian case.

Theorem D.7. *Let $c_0 > 0$, and consider*

$$\mathcal{F}_{dd} = \{(x,z) \mapsto p_\theta(x|z) = \mathcal{N}(x; G_\theta(z), c^2 \mathrm{I}_{d_x}) \,|\, G_\theta(z) \in \mathcal{F}_G, \theta \in \Theta \subseteq \mathbb{R}^{d_\theta}, c > c_0\} \ ,$$
$$\mathcal{F}_{ed} = \{(x,z) \mapsto q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \Sigma_\phi(x)) \,|\, (\mu_\phi(x), \Sigma_\phi(x)) \in \mathcal{F}_{\mu,\Sigma}, \phi \in \Phi \subseteq \mathbb{R}^{d_\phi}\} \ .$$

Assume that there exists $C_{rec} \in \mathsf{M}(\mathsf{X} \times \mathsf{Z})$ such that $\|x - G_\theta(z)\| \leq C_{rec}(x,z)$ for all $\theta \in \Theta$ and $(x,z) \in \mathsf{X} \times \mathsf{Z}$. Assume also that the data distribution π has a finite fourth moment, and that there exists some constant a such that for all $\theta \in \Theta$ and $\phi \in \Phi$,

$$\|\theta\|_\infty + \|\phi\|_\infty \leq a \ .$$

Let $(\theta_n, \phi_n) \in \Theta \times \Phi$ be the n-th iterate of the recursion in Algorithm 1, where $\gamma_n = C_\gamma n^{-1/2}$ with $C_\gamma > 0$. Assume that $\beta_1 < \sqrt{\beta_2} < 1$. For any $n \geq 1$, let $R \in \{0, \ldots, n\}$ be a uniformly distributed random variable. Then,

$$\mathbb{E}\left[\|\nabla \mathcal{L}(\theta_R, \phi_R)\|^2\right] = \mathcal{O}\left(\frac{\mathcal{L}^*}{\sqrt{n}} + d_z \frac{N_{total} a^{2(N_{total}-1)}}{1-\beta_1} \frac{d^* \log n}{\sqrt{n}}\right) \ ,$$

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, $N_{total} = N_{ed} + N_{dd}$ represents the total number of layers in the model architecture.

Lemma D.8. *Let $c_0 > 0$, and consider*

$$\mathcal{F}_{dd} = \{(x,z) \mapsto p_\theta(x|z) = \mathcal{N}(x; G_\theta(z), c^2 \mathrm{I}_{d_x}) \,|\, G_\theta(z) \in \mathcal{F}_G, \theta \in \Theta \subseteq \mathbb{R}^{d_\theta}, c > c_0\} \ ,$$
$$\mathcal{F}_{ed} = \{(x,z) \mapsto q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \Sigma_\phi(x)) \,|\, (\mu_\phi(x), \Sigma_\phi(x)) \in \mathcal{F}_{\mu,\Sigma}, \phi \in \Phi \subseteq \mathbb{R}^{d_\phi}\} \ .$$

Assume that there exists $C_{rec} \in \mathsf{M}(\mathsf{X} \times \mathsf{Z})$ such that $\|x - G_\theta(z)\| \leq C_{rec}(x,z)$ for all $\theta \in \Theta$ and $(x,z) \in \mathsf{X} \times \mathsf{Z}$. Assume also that the data distribution π has a finite fourth moment, and that there exists some constant a such that for all $\theta \in \Theta$ and $\phi \in \Phi$,

$$\|\theta\|_\infty + \|\phi\|_\infty \leq a \ .$$

Then, Assumptions 2(ii) and 3 are satisfied.

Proof. From the analysis of the boundedness and smoothness of the neural network (Lemma I.5), we can establish both the boundedness of the gradient and the smoothness of the functions $\phi \mapsto \mu_\phi(x)$ and $\phi \mapsto \Sigma_\phi(x)$. Notably, by following a similar reasoning to that used in the proof of Lemma D.5, we can handle $\Sigma_\phi(x) = \mathrm{Diag}(\sigma_1^2(x), \ldots, \sigma_{d_z}^2(x))$ by introducing the transformed variable $l_\phi(x) = (\log \sigma_1^2(x), \ldots, \log \sigma_{d_z}^2(x))$, which represents the logarithm of the variances. Although the proof is formulated using $\Sigma_\phi(x)$, it can be easily adapted to the case of $l_\phi(x)$. In the following, each point of Assumptions 2(ii) and 3 is verified one by one.

Boundedness of the gradient of log density. For all $\theta \in \Theta$, $\phi \in \Phi$, $x \in \mathsf{X}$ and $\varepsilon, z \in \mathsf{Z}$ such that $z = g(\varepsilon, \phi)$, we have:

$$\begin{aligned}\|\nabla_z \log p_\theta(x,z)\| &\leq \|\nabla_z \log p_\theta(x|z)\| + \|\nabla_z \log p_\theta(z)\| \\ &\leq \left\|\nabla_z G_\theta(z)^\top (x - G_\theta(z))\right\| + \|z\| \\ &\leq \|\nabla_z G_\theta(z)\| \, \|x - G_\theta(z)\| + \|z\| \ .\end{aligned}$$

Using Lemma I.5, we get:

$$\begin{aligned}\|\nabla_z \log p_\theta(x,z)\| &\leq C_{rec}(x,z) \, \|z\| \, a^{N_{dd}} \prod_{i=1}^{N_{dd}} M_{f_i} + \|z\| \\ &\leq \left(\|\mu_\phi(x)\| + \|\Sigma_\phi(x)\|^{1/2} \, \|\varepsilon\|\right) \left(1 + C_{rec}(x, g(\varepsilon,\phi)) a^{N_{dd}} \prod_{i=1}^{N_{dd}} M_{f_i}\right) \\ &\leq \left(C_\mu + C_\Sigma^{1/2} \, \|\varepsilon\|\right) \left(1 + C_{rec}(x, g(\varepsilon,\phi)) a^{N_{dd}} \prod_{i=1}^{N_{dd}} M_{f_i}\right) \ ,\end{aligned}$$

where we used $z = \mu_\phi(x) + \Sigma_\phi(x)^{1/2}\varepsilon$. For the variational density, the gradient is bounded as follows:

$$\begin{aligned}\|\nabla_z \log q_\phi(z|x)\| &= \left\|\Sigma_\phi(x)^{-1}(z - \mu_\phi(x))\right\| \\ &= \left\|\Sigma_\phi(x)^{-1/2}\varepsilon\right\| \\ &\leq c_\Sigma^{-1/2} \, \|\varepsilon\| \ .\end{aligned}$$

Lipschitz condition on the gradient of log density.

$$\|\nabla_z \log p_\theta(x,z) - \nabla_z \log p_{\theta'}(x,z')\| \leq \|\nabla_z \log p_\theta(x,z) - \nabla_z \log p_{\theta'}(x,z)\| + \|\nabla_z \log p_{\theta'}(x,z) - \nabla_z \log p_{\theta'}(x,z')\| \ . \tag{10}$$

Given that $\nabla_z \log p_\theta(x,z) = \nabla_z G_\theta(z)^\top (x - G_\theta(z))$, we can derive the following bound for the first term in (10):

$$\begin{aligned}\|\nabla_z \log p_\theta(x,z) - \nabla_z \log p_{\theta'}(x,z)\| &\leq \left\|\nabla_z G_\theta(z)^\top (x - G_\theta(z)) - \nabla_z G_{\theta'}(z)^\top (x - G_{\theta'}(z))\right\| \\ &\leq \left\|\nabla_z G_\theta(z)^\top (x - G_\theta(z)) - \nabla_z G_{\theta'}(z)^\top (x - G_\theta(z))\right\| \\ &\quad + \left\|\nabla_z G_{\theta'}(z)^\top (x - G_\theta(z)) - \nabla_z G_{\theta'}(z)^\top (x - G_{\theta'}(z))\right\| \\ &\leq \|x - G_\theta(z)\| \, \|\nabla_z G_\theta(z) - \nabla_z G_{\theta'}(z)\| + \|\nabla_z G_{\theta'}(z)\| \, \|G_\theta(z) - G_{\theta'}(z)\| \\ &\leq C_{rec}(x,z) \, \|\nabla_z G_\theta(z) - \nabla_z G_{\theta'}(z)\| + \|\nabla_z G_{\theta'}(z)\| \, \|G_\theta(z) - G_{\theta'}(z)\| \ .\end{aligned}$$

For the second term in (10), we get:

$$\begin{aligned}\|\nabla_z \log p_\theta(x,z) - \nabla_z \log p_\theta(x,z')\| &\leq \left\|\nabla_z G_\theta(z)^\top (x - G_\theta(z)) - \nabla_z G_\theta(z')^\top (x - G_\theta(z'))\right\| \\ &\leq \left\|\nabla_z G_\theta(z)^\top (x - G_\theta(z)) - \nabla_z G_\theta(z')^\top (x - G_\theta(z))\right\| \\ &\quad + \left\|\nabla_z G_\theta(z')^\top (x - G_\theta(z)) - \nabla_z G_{\theta'}(z')^\top (x - G_\theta(z'))\right\| \\ &\leq \|x - G_\theta(z)\| \, \|\nabla_z G_\theta(z) - \nabla_z G_\theta(z')\| + \|\nabla_z G_\theta(z')\| \, \|G_\theta(z) - G_\theta(z')\| \\ &\leq C_{rec}(x,z) \, \|\nabla_z G_\theta(z) - \nabla_z G_\theta(z')\| + \|\nabla_z G_\theta(z')\| \, \|G_\theta(z) - G_\theta(z')\| \ .\end{aligned}$$

By combining these two terms and applying Lemmas I.2 and I.6, we obtain:

$$\|\nabla_z \log p_\theta(x,z) - \nabla_z \log p_{\theta'}(x,z')\| \leq C_{rec}(x, g(\varepsilon,\phi)) \sum_{k=1}^{N_{dd}} L_{f_k} a^{N_{dd}+k} \prod_{i=1}^{k-1} M_{f_i}^2 \prod_{i=k+1}^{N_{dd}} M_{f_i} \|z - z'\|$$
$$+ a^{2N_{dd}} \prod_{i=1}^{N_{dd}} M_{f_i}^2 \|z - z'\| + C_{rec}(x, g(\varepsilon,\phi)) \left(C_\mu + \|\varepsilon\| C_\Sigma^{1/2}\right) \sum_{k=1}^{N_{dd}} L_{f_k} a^{N_{dd}-1+k} \prod_{i=1}^{N_{dd}-1} M_{f_i}^2 \prod_{i=k+1}^{N_{dd}} M_{f_i} \|\theta - \theta'\|$$
$$+ C_{rec}(x, g(\varepsilon,\phi)) a^{N_{dd}-1} \prod_{i=1}^{N_{dd}} M_{f_i} \|\theta - \theta'\| + \left(C_\mu^2 + \|\varepsilon\|^2 C_\Sigma\right) a^{2N_{dd}-1} \prod_{i=1}^{N_{dd}} M_{f_i}^2 \|\theta - \theta'\| \ .$$

For the variational density, using Lemma I.7, we obtain:

$$\begin{aligned}
\|\nabla_z \log q_\phi(z|x) - \nabla_z \log q_{\phi'}(z'|x)\| &\leq \|\nabla_z \log q_\phi(z|x) - \nabla_z \log q_{\phi'}(z|x)\| + \|\nabla_z \log q_{\phi'}(z|x) - \nabla_z \log q_{\phi'}(z'|x)\| \\
&\leq \|\varepsilon\| \left\|\Sigma_\phi(x)^{-1/2} - \Sigma_{\phi'}(x)^{-1/2}\right\| \\
&\quad + \left\|\Sigma_{\phi'}(x)^{-1}(z - \mu_{\phi'}(x)) - \Sigma_{\phi'}(x)^{-1}(z' - \mu_{\phi'}(x))\right\| \\
&\leq \|\varepsilon\| \frac{1}{2} c_\Sigma^{-3/2} \|\Sigma_\phi(x) - \Sigma_{\phi'}(x)\| + c_\Sigma^{-1} \|z - z'\| \\
&\leq \|\varepsilon\| \frac{1}{2} c_\Sigma^{-3/2} \|x\| \, a^{N_{ed}-1} \prod_{i=1}^{N_{ed}} M_{f_i} \|\phi - \phi'\| + c_\Sigma^{-1} \|z - z'\| \ ,
\end{aligned}$$

where we used the Lipschitz condition of $\phi \mapsto \Sigma_\phi(x)$.

Lipschitz and smoothness condition on g.

$$\begin{aligned}
\|\nabla_\phi g(\varepsilon,\phi)\| &\leq \left\|\nabla_\phi \mu_\phi(x) + \frac{1}{2}\varepsilon \Sigma_\phi(x)^{-1/2} \nabla_\phi \Sigma_\phi(x)\right\| \\
&\leq \|\nabla_\phi \mu_\phi(x)\| + \frac{1}{2} \|\varepsilon\| \left\|\Sigma_\phi(x)^{-1/2}\right\| \|\nabla_\phi \Sigma_\phi(x)\| \\
&\leq \|\nabla_\phi \mu_\phi(x)\| + \frac{1}{2} \|\varepsilon\| \, c_\Sigma^{-1/2} \|\nabla_\phi \Sigma_\phi(x)\| \\
&\leq \left(1 + \frac{1}{2} \|\varepsilon\| \, c_\Sigma^{-1/2}\right) \|x\| \, a^{N_{ed}-1} \prod_{i=1}^{N_{ed}} M_{f_i} \ .
\end{aligned}$$

$$\begin{aligned}
\|\nabla_\phi g(\varepsilon,\phi) - \nabla_\phi g(\varepsilon,\phi')\| &\leq \|\nabla_\phi \mu_\phi(x) - \nabla_\phi \mu_{\phi'}(x)\| + \frac{1}{2} \|\varepsilon\| \left\|\Sigma_\phi(x)^{-1/2} \nabla_\phi \Sigma_\phi(x) - \Sigma_{\phi'}(x)^{-1/2} \nabla_\phi \Sigma_{\phi'}(x)\right\| \\
&\leq \|\nabla_\phi \mu_\phi(x) - \nabla_\phi \mu_{\phi'}(x)\| + \frac{1}{2} \|\varepsilon\| \, c_\Sigma^{-1/2} \|\nabla_\phi \Sigma_\phi(x) - \nabla_\phi \Sigma_{\phi'}(x)\| \\
&\quad + \frac{1}{2} \|\varepsilon\| \|\nabla_\phi \Sigma_\phi(x)\| \left\|\Sigma_\phi(x)^{-1/2} - \Sigma_{\phi'}(x)^{-1/2}\right\| \\
&\leq \|\nabla_\phi \mu_\phi(x) - \nabla_\phi \mu_{\phi'}(x)\| + \frac{1}{2} \|\varepsilon\| \, c_\Sigma^{-1/2} \|\nabla_\phi \Sigma_\phi(x) - \nabla_\phi \Sigma_{\phi'}(x)\| \\
&\quad + \frac{1}{4} \|\varepsilon\| \|\nabla_\phi \Sigma_\phi(x)\| \, c_\Sigma^{-3/2} \|\Sigma_\phi(x) - \Sigma_{\phi'}(x)\| \\
&\leq \left(1 + \frac{1}{2} \|\varepsilon\| \, c_\Sigma^{-1/2}\right) N_{ed} \left(\|x\|^2 + 1\right) \sum_{k=1}^{N_{ed}} L_{f_k} a^{N_{ed}-2+k} \prod_{i=1}^{k-1} M_{f_i}^2 \prod_{i=k+1}^{N_{ed}} M_{f_i} \|\phi - \phi'\| \\
&\quad + \frac{1}{4} \|\varepsilon\| \, c_\Sigma^{-3/2} \|x\|^2 \, a^{2(N_{ed}-1)} \prod_{i=1}^{N_{ed}} M_{f_i}^2 \|\phi - \phi'\| \ ,
\end{aligned}$$

where we used the Lipschitz and the smoothness of $\mu_\phi(x)$ and $\Sigma_\phi(x)$. □

Lemma D.9. *Under the assumptions of Lemma D.5, the smoothness constant L^{P} of the expected ELBO is well-defined and finite.*

Proof. Using B.1, the smoothness constant L^{P} is defined by:

$$L^{\mathsf{P}} = \mathbb{E}_{\pi,p_\varepsilon}\left[L_p(x,\varepsilon) + M_g(x,\varepsilon)^2\left(L_p(x,\varepsilon) + 2L_q(x,\varepsilon)\right) + 3L_g(x,\varepsilon)M(x,\varepsilon) + 2M_g(x,\varepsilon)L_q(x,\varepsilon)\right] + \mathbb{E}_{\pi,p_\varepsilon}\left[L_p(x,\varepsilon)M_g(x,\varepsilon)\right] .$$

Let L^{P}_{dd} and L^{P}_{ed} denote the smoothness constants of the decoder and the encoder, respectively. For the smoothness constant of the decoder, applying Lemma I.5, we have:

$$L^{\mathsf{P}}_{dd} = \mathbb{E}_{\pi,p_\varepsilon}\left[L_p(x,\varepsilon)\right] ,$$

where

$$L_p(x,\varepsilon) = C_{rec}(x,g(\varepsilon,\phi))\sum_{k=1}^{N_{dd}} L_{f_k} a^{N_{dd}+k}\prod_{i=1}^{k-1} M_{f_i}^2\prod_{i=k+1}^{N_{dd}} M_{f_i} + a^{2N_{dd}}\prod_{i=1}^{N_{dd}} M_{f_i}^2 + \left(C_\mu^2 + \|\varepsilon\|^2 C_\Sigma\right) a^{2N_{dd}-1}\prod_{i=1}^{N_{dd}} M_{f_i}^2$$
$$+ C_{rec}(x,g(\varepsilon,\phi))a^{N_{dd}-1}\prod_{i=1}^{N_{dd}} M_{f_i} + C_{rec}(x,g(\varepsilon,\phi))\left(C_\mu + \|\varepsilon\| C_\Sigma^{1/2}\right)\sum_{k=1}^{N_{dd}} L_{f_k} a^{N_{dd}-1+k}\prod_{i=1}^{N_{dd}-1} M_{f_i}^2\prod_{i=k+1}^{N_{dd}} M_{f_i} .$$

Since

$$\mathbb{E}_{p_\varepsilon}\left[\|\varepsilon\|\right] = \sqrt{2}\frac{\Gamma\left(\frac{d_z+1}{2}\right)}{\Gamma\left(\frac{d_z}{2}\right)} ,$$

it follows that L^{P}_{ed} is well-defined and finite.

For the smoothness constant of the encoder, we have:

$$L^{\mathsf{P}}_{ed} = \mathbb{E}_{\pi,p_\varepsilon}\left[M_g(x,\varepsilon)^2\left(L_p(x,\varepsilon) + 2L_q(x,\varepsilon)\right) + 3L_g(x,\varepsilon)M(x,\varepsilon) + 2M_g(x,\varepsilon)L_q(x,\varepsilon)\right] ,$$

where

$$M(x,\varepsilon) = \left(C_\mu + C_\Sigma^{1/2}\|\varepsilon\|\right)\left(1 + C_{rec}(x,g(\varepsilon,\phi))a^{N_{dd}}\prod_{i=1}^{N_{dd}} M_{f_i}\right) + c_\Sigma^{-1/2}\|\varepsilon\| ,$$

$$L_q(x,\varepsilon) = c_\Sigma^{-1} + \|\varepsilon\|\frac{1}{2}c_\Sigma^{-3/2}\|x\| a^{N_{ed}-1}\prod_{i=1}^{N_{ed}} M_{f_i} ,$$

$$M_g(x,\varepsilon) = \left(1 + \frac{1}{2}\|\varepsilon\| c_\Sigma^{-1/2}\right)\|x\| a^{N_{ed}-1}\prod_{i=1}^{N_{ed}} M_{f_i} ,$$

$$L_g(x,\varepsilon) = \left(1 + \frac{1}{2}\|\varepsilon\| c_\Sigma^{-1/2}\right) N_{ed}\left(\|x\|^2+1\right)\sum_{k=1}^{N_{ed}} L_{f_k} a^{N_{ed}-2+k}\prod_{i=1}^{k-1} M_{f_i}^2\prod_{i=k+1}^{N_{ed}} M_{f_i}$$
$$+ \frac{1}{4}\|\varepsilon\| c_\Sigma^{-3/2}\|x\|^2 a^{2(N_{ed}-1)}\prod_{i=1}^{N_{ed}} M_{f_i}^2 .$$

Since all the terms are well-defined, it is evident that L^{P}_{dd} is also well-defined and finite. Moreover, under the conditions $M_{f_i} \le 1$ and $L_{f_i} \le 1$ for all $1 \le i \le N$, we can establish the following bounds:

$$L^{\mathsf{P}}_{dd} = \mathcal{O}(\sqrt{d_z} N_{dd} a^{2N_{dd}}) \text{ and } L^{\mathsf{P}}_{ed} = \mathcal{O}(d_z N_{ed} a^{2(N_{ed}-1)} a^{2N_{dd}}) = \mathcal{O}(d_z N_{ed} a^{2(N_{ed}+N_{dd}-1)}) ,$$

where the constants in these bounds depend on additional terms. However, we focus here solely on the dimensions of the latent space and the number of layers in the model architecture, similar to the case of the score function. □

Proof of Theorem D.7. Using Lemma D.8, we ensure that Assumptions 2(ii) and 3 are satisfied. The proof is then completed by applying Theorem 3.3. □

E IMPORTANCE WEIGHTED AUTOENCODER

E.1 Convergence Rate of IWAE with Respect to Marginal Log Likelihood

Denoting the normalized importance weights as

$$\widetilde{w}_{\theta,\phi}(x, z^{(\ell)}) = \frac{w_{\theta,\phi}(x, z^{(\ell)})}{\sum_{\ell=1}^{K} w_{\theta,\phi}(x, z^{(\ell)})} ,$$

the gradient of the ELBO in IWAE can be expressed as:

$$\nabla_{\theta,\phi}\mathcal{L}_K^{\mathsf{IS}}(\theta,\phi) = \mathbb{E}_{\pi}\left[\mathbb{E}_{q_\phi^{\otimes K}(\cdot|x)}\left[\sum_{\ell=1}^{K} \widetilde{w}_{\theta,\phi}(x, z^{(\ell)})\nabla_{\theta,\phi}\log w_{\theta,\phi}(x, z^{(\ell)})\right]\right] .$$

Using the reparametrization trick, this can be rewritten as:

$$\nabla_{\theta,\phi}\mathcal{L}_K^{\mathsf{IS}}(\theta,\phi) = \mathbb{E}_{\pi}\left[\mathbb{E}_{p_\varepsilon^{\otimes K}}\left[\sum_{\ell=1}^{K} \widetilde{w}_{\theta,\phi}(x, g(\varepsilon^{(\ell)},\phi))\nabla_{\theta,\phi}\log w_{\theta,\phi}(x, g(\varepsilon^{(\ell)},\phi))\right]\right] .$$

Given a batch of observations $\{x_i\}_{i=1}^B$, the estimator for the gradient of this ELBO can be expressed as:

$$\widehat{\nabla}_{\theta,\phi}\mathcal{L}_K^{\mathsf{IS}}(\theta,\phi;\{x_i\}_{i=1}^B) = \frac{1}{B}\sum_{i=1}^{B}\sum_{\ell=1}^{K}\frac{w_{\theta,\phi}(x_i, z_i^{(\ell)})}{\sum_{\ell=1}^{k} w_{\theta,\phi}(x_i, z_i^{(\ell)})}\nabla_{\theta,\phi}\log w_{\theta,\phi}(x_i, z_i^{(\ell)}) .$$

where, for all $1 \le i \le B$ and $1 \le \ell \le K$, $z_i^{(\ell)} = g(\varepsilon_i^{(\ell)},\phi)$ with $\varepsilon_i^{(\ell)}$ being independent samples drawn from p_ε. This estimator can be viewed as a biased gradient estimator for the marginal log-likelihood $\nabla_\theta \log p_\theta(x)$. The non-asymptotic bound on the bias of this gradient estimator, as established in (Surendran et al., 2024, Theorem B.1), indicates that

$$\left\|\mathbb{E}_{q_\phi^{\otimes k}(\cdot|x)}\left[\widehat{\nabla}_\theta\mathcal{L}_k^{\mathsf{IS}}(\theta,\phi;x) - \nabla_\theta\log p_\theta(x)\right]\right\| = \mathcal{O}\left(\frac{1}{k}\right) .$$

This result indicates that as the number of samples K increases, the bias in the gradient estimator diminishes at a rate inversely proportional to K. Furthermore, Surendran et al. (2024) treats IWAE as a biased gradient and establishes a convergence rate of $\mathcal{O}(\log n/\sqrt{n} + b_n)$, where b_n is related to the bias at iteration n. However, they do not explicitly verify all assumptions; thus, our results enable us to theoretically derive the following convergence rate for the IWAE with respect to the marginal log-likelihood:

$$\mathbb{E}\left[\left\|\nabla_\theta \log p_{\theta_R}(x)\right\|^2\right] = \mathcal{O}\left(\frac{\log n}{\sqrt{n}} + b_n\right) ,$$

where $R \in \{0,\dots,n\}$ is a uniformly distributed random variable.

E.2 Proof of Theorem 3.7

Proof. First, for all $K \in \mathbb{N}$, $x \in \mathsf{X}$, $z \in \mathsf{Z}$ and $1 \le \ell \le K$, the Lipschitz condition and smoothness of $w_{\theta,\phi}(x, z^{(\ell)})$ with respect θ and ϕ follow from those of $\theta \mapsto \log p_\theta(x, z)$ and $\phi \mapsto \log q_\phi(z|x)$. Next, we establish that $\widetilde{w}_{\theta,\phi}(x, z^{(\ell)})$ is also Lipschitz continuous and smooth with respect to θ and ϕ for all $K \in \mathbb{N}$, $x \in \mathsf{X}$, $z \in \mathsf{Z}$ and $1 \le \ell \le K$. For all $1 \le \ell \le K$, let $\widetilde{w}_{\theta,\phi}^{(\ell)} := \widetilde{w}_{\theta,\phi}(x, z^{(\ell)})$. For $K = 1$, $|\widetilde{w}_{\theta,\phi}^{(\ell)} - \widetilde{w}_{\theta',\phi'}^{(\ell)}| = 0$. For $K > 1$,

$$\begin{aligned}
\left|\widetilde{w}_{\theta,\phi}^{(\ell)} - \widetilde{w}_{\theta',\phi}^{(\ell)}\right| &= \left|\frac{w_{\theta,\phi}^{(\ell)}}{\sum_{\ell=1}^K w_{\theta,\phi}^{(\ell)}} - \frac{w_{\theta',\phi}^{(\ell)}}{\sum_{\ell=1}^K w_{\theta',\phi}^{(\ell)}}\right| \\
&\le \left|\frac{w_{\theta,\phi}^{(\ell)}}{\sum_{\ell=1}^K w_{\theta,\phi}^{(\ell)}} - \frac{w_{\theta',\phi}^{(\ell)}}{\sum_{\ell=1}^K w_{\theta,\phi}^{(\ell)}}\right| + \left|\frac{w_{\theta',\phi}^{(\ell)}}{\sum_{\ell=1}^K w_{\theta,\phi}^{(\ell)}} - \frac{w_{\theta',\phi}^{(\ell)}}{\sum_{\ell=1}^K w_{\theta',\phi}^{(\ell)}}\right| \\
&\le \left|\frac{1}{\sum_{\ell=1}^K w_{\theta,\phi}^{(\ell)}}\right|\left|w_{\theta,\phi}^{(\ell)} - w_{\theta',\phi}^{(\ell)}\right| + \left|\frac{w_{\theta',\phi}^{(\ell)}}{\sum_{\ell=1}^K w_{\theta,\phi}^{(\ell)}\sum_{\ell=1}^K w_{\theta',\phi}^{(\ell)}}\right|\left|\sum_{\ell=1}^K w_{\theta,\phi}^{(\ell)} - \sum_{\ell=1}^K w_{\theta',\phi}^{(\ell)}\right| \\
&\le \frac{1}{K}\frac{\alpha^+(x,\varepsilon)}{\alpha^-(x,\varepsilon)}M(x,\varepsilon)\left\|\theta-\theta'\right\| + 2\frac{1}{K}\frac{\alpha^+(x,\varepsilon)^3}{\alpha^-(x,\varepsilon)^3}M(x,\varepsilon)\left\|\theta-\theta'\right\| ,
\end{aligned}$$

where we used Assumption 4. This concludes the Lipschitz continuity with respect to θ. The Lipschitz condition of $\widetilde{w}^{(\ell)}_{\theta,\phi}$ with respect to ϕ is treated in the same manner as with θ.

For a given observation $x \in \mathsf{X}$, using the Lipschitz continuity of $(\theta,\phi) \mapsto \widetilde{w}^{(\ell)}_{\theta,\phi}$, we get:

$$
\begin{aligned}
&\left\|\nabla_{\theta,\phi}\mathcal{L}^{\mathsf{IS}}_K(\theta,\phi;x) - \nabla_{\theta,\phi}\mathcal{L}^{\mathsf{IS}}_K(\theta',\phi';x)\right\| \\
&= \left\| \mathbb{E}_{p_\varepsilon^{\otimes K}}\left[\sum_{\ell=1}^K \widetilde{w}_{\theta,\phi}(x, g(\varepsilon^{(\ell)},\phi)) \nabla_{\theta,\phi} \log w_{\theta,\phi}(x, g(\varepsilon^{(\ell)},\phi)) - \sum_{\ell=1}^K \widetilde{w}_{\theta',\phi'}(x, g(\varepsilon^{(\ell)},\phi')) \nabla_{\theta,\phi} \log w_{\theta',\phi'}(x, g(\varepsilon^{(\ell)},\phi')) \right] \right\| \\
&\leq \mathbb{E}_{p_\varepsilon^{\otimes K}}\left[\left\| \sum_{\ell=1}^K \widetilde{w}_{\theta,\phi}(x, g(\varepsilon^{(\ell)},\phi)) \nabla_{\theta,\phi} \log w_{\theta,\phi}(x, g(\varepsilon^{(\ell)},\phi)) - \widetilde{w}_{\theta,\phi}(x, g(\varepsilon^{(\ell)},\phi)) \nabla_{\theta,\phi} \log w_{\theta',\phi'}(x, g(\varepsilon^{(\ell)},\phi')) \right\| \right] \\
&\quad + \mathbb{E}_{p_\varepsilon^{\otimes K}}\left[\left\| \sum_{\ell=1}^K \widetilde{w}_{\theta,\phi}(x, g(\varepsilon^{(\ell)},\phi)) \nabla_{\theta,\phi} \log w_{\theta,\phi'}(x, g(\varepsilon^{(\ell)},\phi')) - \widetilde{w}_{\theta',\phi'}(x, g(\varepsilon^{(\ell)},\phi')) \nabla_{\theta,\phi} \log w_{\theta',\phi'}(x, g(\varepsilon^{(\ell)},\phi')) \right\| \right] \\
&\leq \mathbb{E}_{p_\varepsilon^{\otimes K}}\left[\sum_{\ell=1}^K \widetilde{w}_{\theta,\phi}(x, g(\varepsilon^{(\ell)},\phi)) \left\| \nabla_{\theta,\phi} \log w_{\theta,\phi}(x, g(\varepsilon^{(\ell)},\phi)) - \nabla_{\theta,\phi} \log w_{\theta',\phi'}(x, g(\varepsilon^{(\ell)},\phi')) \right\| \right] \\
&\quad + \mathbb{E}_{p_\varepsilon^{\otimes K}}\left[\sum_{\ell=1}^K \left\| \nabla_{\theta,\phi} \log w_{\theta,\phi'}(x, g(\varepsilon^{(\ell)},\phi')) \right\| \left\| \widetilde{w}_{\theta,\phi}(x, g(\varepsilon^{(\ell)},\phi)) - \widetilde{w}_{\theta',\phi'}(x, g(\varepsilon^{(\ell)},\phi')) \right\| \right] \\
&\leq L^{\mathsf{P}} \left\|(\theta,\phi) - (\theta',\phi')\right\| + \mathbf{1}_{K>1} \frac{1}{K} \mathbb{E}_{p_\varepsilon}\left[\frac{\alpha^+(x,\varepsilon)}{\alpha^-(x,\varepsilon)} M(x,\varepsilon) + 2\frac{\alpha^+(x,\varepsilon)^3}{\alpha^-(x,\varepsilon)^3} M(x,\varepsilon) \right] \left\|(\theta,\phi) - (\theta',\phi')\right\| \ ,
\end{aligned}
$$

where in the last inequality, we applied Lemma B.2 to the first term, and for the second term, we used Assumption 2(ii) along with the Lipschitz continuity of $\widetilde{w}^{(\ell)}_{\theta,\phi}$. Taking the expectation over π then leads to the conclusion that the ELBO for the IWAE is L_K-smooth, where

$$
L_K = L^{\mathsf{P}} + \mathbf{1}_{K>1} \mathbb{E}_{\pi,p_\varepsilon}\left[M(x,\varepsilon)\alpha^+(x,\varepsilon)/\alpha^-(x,\varepsilon) + 2M(x,\varepsilon)\alpha^+(x,\varepsilon)^3/\alpha^-(x,\varepsilon)^3 \right]/K \ . \tag{11}
$$

□

F BLACK-BOX VARIATIONAL INFERENCE

In the following, we refer to the ELBO in BBVI as ELBO-BBVI to distinguish it from the ELBO in VAE and avoid any confusion.

F.1 Previous Work on the Convergence of BBVI

Some existing results on the convergence of BBVI, such as (Regier et al., 2017, Theorem 1) and (Buchholz et al., 2018, Theorem 1), rely on the assumption of the complete smoothness of the ELBO-BBVI to establish their guarantees. In Domke (2020), it is shown that ELBO-BBVI is strongly concave under certain conditions, specifically when the posterior is strongly log-concave and linear parameterizations are used. However, their analysis also demonstrates that while the energy function (the expectation of the joint likelihood) is smooth, the ELBO-BBVI itself is not smooth. This conclusion is reached by decomposing ELBO-BBVI into the sum of two terms, where the entropic regularization term lacks smoothness. In contrast, Kim et al. (2023) shows that ELBO-BBVI can be smooth under specific conditions. The smoothness of ELBO-BBVI in their work (Theorem 1, Corollary 1) is established through the Hessian and the location-scale parameterization. Our approach yields new results for BBVI without assuming the location-scale family and apply to a broader range of reparameterization families.

F.2 Comparison of Our Results with Existing Work

The following corollary provides a more detailed version of Corollary F.1.

Corollary F.1. *Assume that the following conditions hold: There exist M, M_g, L_g, L_p, and $L_q \in \mathsf{M}(\mathsf{X}\times\mathsf{Z})$ such that for all $\phi \in \Phi$, $x \in \mathsf{X}$ and $\varepsilon \in \mathsf{Z}$ with $z = g(\varepsilon,\phi)$,*

(i) $z \mapsto \log p(x, z)$ *is* $L_p(x, \varepsilon)$*-smooth.*

(ii) $z \mapsto \log q_\phi(z|x)$ *is* $M(x, \varepsilon)$*-Lipchitz and* $L_q(x, \varepsilon)$*-smooth.*

(iii) $\phi \mapsto g(\phi, \varepsilon)$ *is* $M_g(x, \varepsilon)$*-Lipchitz and* $L_g(x, \varepsilon)$*-smooth.*

Then, $\phi \mapsto \mathcal{L}^{\mathsf{BBVI}}(\phi)$ *is* L^{BBVI}*-smooth, where the smoothness constant* L^{BBVI} *is given by:*

$$L^{\mathsf{BBVI}} = \mathbb{E}_{\pi, p_\varepsilon}\left[M_g(x,\varepsilon)^2\left(L_p(x,\varepsilon) + 2L_q(x,\varepsilon)\right) + 3L_g(x,\varepsilon)M(x,\varepsilon) + 2M_g(x,\varepsilon)L_q(x,\varepsilon)\right] . \tag{12}$$

The proof follows a similar analysis to that of the second term in the "Lipschitz Condition of $\nabla^P_\phi \mathcal{L}(\theta, \phi)$" in Lemma B.2.

Assumption (i) is consistent with the assumptions used in prior works, such as (Domke et al., 2023, Theorem 1) and (Kim et al., 2023, Corollary 1). Unlike other previous works, our results do not rely on any specific reparameterization trick function. In contrast, prior studies often assume a location-scale parameterization.

Deep Gaussian Case: Verifying Assumptions (ii) and (iii).

To clarify the assumptions (ii) and (iii), we consider the Deep Gaussian case. In this context, we define $\Sigma_\phi(x)$ as the diagonal conditioner and focus on analyzing this component, as the mean $\mu_\phi(x)$ is assumed to follow a linear parameterization in the related works.

To verify assumptions (ii) and (iii), we require that $\Sigma_\phi(x)$ is both Lipschitz and smooth, and that $\Sigma_\phi(x) \geq c$ for some constant $c > 0$. In Kim et al. (2023), the diagonal conditioner is assumed to be 1-Lipschitz and smooth (as shown in Theorem 1 and Corollary 1). However, they do not impose the lower bound $\Sigma_\phi(x) \geq c$. In contrast, (Domke et al., 2023, Theorem 7) shows that this condition is verified with the Gaussian case and the location-scale parameterization.

The assumptions in Domke et al. (2023) ensure that our conditions are satisfied in the Deep Gaussian case. Notably, our assumptions are less restrictive than those in existing works. While our analysis focuses on the diagonal case, the results can be extended to the full-rank case. In general, our results apply to a wider range of reparameterization families than those considered in the existing literature.

G SEQUENTIAL VARIATIONAL AUTOENCODERS

G.1 Introduction

Consider an unobserved state sequence $z_{0:T} = (z_0, \ldots, z_T)$ and an observation sequence $x_{0:T} = (x_0, \ldots, x_T)$. At each time $t \in \mathbb{N}$, the unobserved state z_t and the observation x_t are assumed to take values in some general measurable spaces $(\mathsf{Z}_t, \mathcal{Z}_t)$ and $(\mathsf{X}_t, \mathcal{X}_t)$, respectively. Without any assumption on the sequential latent-variable model, the complete likelihood of the observation sequence $x_{0:T} = (x_0, \ldots, x_T)$ and the latent sequence $z_{0:T} = (z_0, \ldots, z_T)$ is defined as:

$$p(x_{0:T}, z_{0:T}) = p(z_0)p(x_0|z_0)\prod_{t=1}^{T} p(x_t|x_{0:t-1}, z_{0:t})p(z_t|z_{0:t-1}, x_{0:t-1}) .$$

Then, the posterior distribution of this model can be factorized as:

$$p(z_{0:T}|x_{0:T}) = \prod_{t=0}^{T} p(z_t|z_{0:t-1}, x_{0:T}) ,$$

with the convention $p(z_0|z_{0:-1}, x_{0:T}) = p(z_0|x_{0:T})$. The ELBO for a sequential VAE is defined as follows:

$$\mathcal{L}_T = \mathbb{E}_{q(\cdot|x_{0:T})}\left[\log \frac{p(x_{0:T}, z_{0:T})}{q(z_{0:T}|x_{0:T})}\right] = \ell_T - \mathrm{KL}(q(\cdot|x_{0:T}) \parallel p(\cdot|x_{0:T})) ,$$

where $\ell_T = \log p(x_{0:T})$ corresponds to the log evidence. The variational family $q(z_{0:T}|x_{0:T})$ can be factorized using several specific graphical models. The most commonly used variational decompositions are listed in Table 1.

Table 1: An Overview of Variational Approximation for Sequential Inference Networks in the Literature.

Model	Variational Approximation for z_t
q-INDEP	$q(z_t\|x_t)$
q-LR	$q(z_t\|x_{t-1:t+1})$
q-RNN	$q(z_t\|x_{0:t})$
q-BRNN	$q(z_t\|x_{0:T})$
VRNN (Chung et al., 2015)	$q(z_t\|z_{0:t-1}, x_{0:t})$
DVBF (Karl et al., 2017)	$q(z_t\|z_{t-1}, x_t)$
DKF (Krishnan et al., 2015)	$q(z_t\|z_{t-1}, x_{0:T})$
DKS (Krishnan et al., 2017)	$q(z_t\|z_{t-1}, x_{t:T})$
VF (Forward) (Marino et al., 2018)	$q(z_t\|z_{t-1}, x_{0:t})$
VF (Backward) (Campbell et al., 2021)	$q(z_t\|z_{t+1}, x_{0:t})$

In the first four models listed, the latent variable z_t depends solely on the observed data, without any dependence on other latent variables $z_0, \dots, z_T$. These models are referred to as conditionally independent latent variable models, reflecting their simple structure. Specifically, each model uses different parameterizations for $q(z_t)$ based on the context of the observed data:

- q-INDEP where $q(z_t|x_t)$ is parameterized by an MLP,
- q-LR where $q(z_t|x_{t-1:t+1})$ is parameterized by an MLP,
- q-RNN where $q(z_t|x_{0:t})$ is parameterized by an RNN,
- q-BRNN where $q(z_t|x_{0:T})$ is parameterized by a bi-directional RNN.

The remaining models use more complex structured variational approximations. For instance, VRNN (Chung et al., 2015) conditions z_t on past latent states and observations through a recurrent structure, while other models assume a Markovian structure for the latent variables with varying approaches to handling the observed data. The model of particular focus here is the Variational Backward Model, where the latent variable is conditioned on both future latent states and past observations. This backward conditioning enables better smoothing by incorporating information from future observations.

G.2 Convergence Results in a General Setting

First, we compute the gradient of the expected ELBO with respect to θ and ϕ in this sequential setting. The gradient with respect to θ is given by:

$$\nabla_\theta \mathcal{L}_T(\theta, \phi) = \mathbb{E}_\pi \left[\mathbb{E}_{q_\phi(\cdot|x_{0:T})} \left[s_{0:T,\theta} \right] \right], \tag{13}$$

where $s_{0:T,\theta} = \sum_{t=0}^{T} s_{t,\theta}$ with $s_{t,\theta} : \mathsf{X}_{0:t} \times \mathsf{Z}_{0:t} \ni (x_{0:t}, z_{0:t}) \mapsto \nabla_\theta \log\{p_\theta(x_t|x_{0:t-1}, z_{0:t}) p_\theta(z_t|z_{0:t-1}, x_{0:t-1})\}$, with the conventions $p_\theta(x_0|x_{0:-1}, z_0) = p_\theta(x_0|z_0)$ and $p_\theta(z_0|z_{0:-1}, x_{0:-1}) = p_\theta(z_0)$.

Now, for the score function gradient with respect to ϕ, using Proposition A.1, we obtain:

$$\nabla_\phi \mathcal{L}_T(\theta, \phi) = \mathbb{E}_\pi \left[\mathbb{E}_{q_\phi(\cdot|x_{0:T})} \left[\log \frac{p_\theta(x_{0:T}, z_{0:T})}{q_\phi(z_{0:T}|x_{0:T})} \nabla_\phi \log q_\phi(z_{0:T}|x_{0:T}) \right] \right]. \tag{14}$$

To cover all possible scenarios mentioned previously, we consider the variational family $q_\phi(z_{0:T}|x_{0:T})$ which can be factorized as $\prod_t q_\phi(z_t|\bar{z}_t, \bar{x}_t)$. Here, $\bar{z}_t$ can represent $\emptyset$, $z_{0:t-1}$, z_{t-1} or z_{t+1}, and $\bar{x}_t$ can represent $\emptyset$, x_t, $x_{t-1:t+1}$, $x_{0:t}$, $x_{t:T}$ or $x_{0:T}$. In the following, we write $\bar{x}_t \in \bar{\mathsf{X}}_t$ and $\bar{z}_t \in \bar{\mathsf{Z}}_t$.

In this sequential framework, we work with the following assumptions.

Assumption 5. *(Strong Mixing)*
For every $t \in \mathbb{N}$, there exist $0 < \sigma_t^- < \sigma_t^+ < \infty$ such that for all $\theta \in \Theta$ and $\phi \in \Phi$,

(i) $\sigma_t^- \leq p_\theta(x_t|x_{0:t-1}, z_{0:t}) \leq \sigma_t^+$ *for every* $(x_{0:t}, z_{0:t}) \in \mathsf{X}_{0:\mathsf{t}} \times \mathsf{Z}_{0:\mathsf{t}}$,

(ii) $\sigma_t^- \leq p_\theta(z_t|z_{0:t-1}, x_{0:t-1}) \leq \sigma_t^+$ *for every* $(x_{0:t}, z_{0:t}) \in \mathsf{X}_{0:\mathsf{t}} \times \mathsf{Z}_{0:\mathsf{t}}$,

(iii) $\sigma_t^- \leq q_\phi(z_t|\bar{z}_t, \bar{x}_t) \leq \sigma_t^+$ *for every* $z_t, \bar{z}_t, \bar{x}_t \in \mathsf{Z}_\mathsf{t} \times \bar{\mathsf{Z}}_t \times \bar{\mathsf{X}}_t$.

Assumption 5 is quite strong, but it is typically satisfied in models with a compact state space. This assumption is well-established in the Sequential Monte Carlo literature (Douc et al., 2011; Olsson and Westerborn, 2017; Gloaguen et al., 2022; Cardoso et al., 2023), where it is used to obtain quantitative bounds for the errors or variances of estimators. Additionally, it is used to derive variational excess risk bounds for general state space models (Chagneux et al., 2024; Gassiat and Le Corff, 2024). It is worth noting that in the context of approximating filtering distributions in the SMC literature, weaker assumptions, such as pseudo-mixing, are sometimes sufficient to obtain quantitative bounds on estimator errors or variances (see Chigansky and Liptser (2004); Douc et al. (2009)). However, extending these results to the smoothing context remains an open challenge. Consequently, obtaining convergence rates for sequential VAE within a general framework, particularly involving a backward kernel, under this weaker assumption is still far from being fully achieved.

Assumption 6. *(Lipschitz Condition)*

(i) For all $t \in \mathbb{N}$, *there exists* $L_t^s \in \mathsf{M}(\mathsf{X}_{0:\mathsf{t}} \times \mathsf{Z}_{0:\mathsf{t}})$ *such that for all* $(x_{0:t}, z_{0:t}) \in \mathsf{X}_{0:\mathsf{t}} \times \mathsf{Z}_{0:\mathsf{t}}$, *the function* $\theta \mapsto s_{t,\theta}(x_{0:t}, z_{0:t})$ *is* $L_t^s(x_{0:t}, z_{0:t})$*-Lipschitz, and* $\theta \mapsto s_{t,\theta}(x_{0:t}, z_{0:t})$ *is bounded by* $\|s_t(\theta)\|_\infty$. *Furthermore,* $\|L_t^s\|_\infty < \infty$.

(ii) For all $t \in \mathbb{N}$, *there exists* $L_t^q \in \mathsf{M}(\mathsf{Z}_\mathsf{t} \times \bar{\mathsf{Z}}_t \times \bar{\mathsf{X}}_t)$ *such that* $\|L_t^q\|_\infty < \infty$ *and that for all* $(z_t, \bar{z}_t, \bar{x}_t) \in \mathsf{Z}_\mathsf{t} \times \bar{\mathsf{Z}}_t \times \bar{\mathsf{X}}_t$, $\phi \mapsto \log q_\phi(z_t|\bar{z}_t, \bar{x}_t)$ *is* $L_t^q(z_t, \bar{z}_t, \bar{x}_t)$*-Smooth, and* $\phi \mapsto \nabla_\phi \log q_\phi(z_t|\bar{z}_t, \bar{x}_t)$ *is bounded.*

Assumption 6(i) is analogous to the one used in Cardoso et al. (2023) (see Assumption A B.9 (i)), which was employed to establish the convergence rate of their proposed algorithm, the PARIS Particle Gibbs (PPG) sampler. Their method is based on Sequential Monte Carlo (SMC) techniques for the online approximation of posterior distributions in state space models. In contrast, our approach uses variational methods. Consequently, we introduce condition (ii), which is similar to condition (i), but is relevant to the variational distribution.

Theorem G.1. *Let Assumptions 5 and 6 hold. Let* $(\theta_n, \phi_n) \in \Theta \times \Phi$ *be the* n*-th iterate of the recursion in Algorithm 1 where* $\gamma_n = C_\gamma n^{-1/2}$ *with* $C_\gamma > 0$. *For all* $n \geq 1$, *let* $R \in \{0, \ldots, n\}$ *be a uniformly distributed random variable. Then,*

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

where $d^* = d_\theta + d_\phi$ *represents the total dimension of the parameters, and* C_T *is a constant that depends on* T.

In this convergence rate, the factor C_T depends on T and increases as T grows. Since the exact structure of the model is unknown, deriving a bound that depends on T is challenging. Nevertheless, we establish the convergence rate, with dependence on T, for Deep Gaussian Non-Linear State-Space Models within the context of Variational Smoothing below.

G.3 Application to Variational Smoothing in Deep Gaussian Non-Linear State-Space Models

We consider a general form of state-space models (SSMs) where the filtering and smoothing distributions, as well as the log evidence, are typically not available in closed form and thus need to be approximated.

Let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ be a parameter of interest. In the context of SSMs, it is assumed that the sequence $\{z_t\}_{t\in\mathbb{N}}$ forms a Markov chain with an initial distribution ν_θ and transition kernels $\{M_{\theta,t}\}_{t\in\mathbb{N}}$ where each kernel has a transition density m_θ with respect to some reference measure. Given the states $\{z_t\}_{t\in\mathbb{N}}$, the observations $\{x_t\}_{t\in\mathbb{N}}$ are assumed to be independent and such that for all $t \in \mathbb{N}$, the conditional distribution of the observation x_t depends only on the current state z_t. This distribution is assumed to admit a density $g_\theta(z_t, \cdot)$ with respect to some reference measure. In summary, these correspond to the generative model:

$$z_0 \sim \nu_\theta(z_0)\ , \quad z_{t+1}|z_t \sim m_\theta(z_t, z_{t+1})\ , \quad x_t|z_t \sim g_\theta(z_t, x_t)\ .$$

The complete likelihood of the observation sequence $x_{0:T} = (x_0, \dots, x_T)$ and the latent sequence $z_{0:T} = (z_0, \dots, z_T)$ is defined as:

$$p_\theta(x_{0:T}, z_{0:T}) = \nu_\theta(z_0) g_\theta(z_0, x_0) \prod_{t=0}^{T-1} m_\theta(z_t, z_{t+1}) g_\theta(z_{t+1}, x_{t+1}) \, .$$

The posterior distribution of this model can also be written as:

$$p_\theta(z_{0:T}|x_{0:T}) = p_\theta(z_T|x_T) \prod_{t=1}^{T-1} p_\theta(z_t|x_t, z_{t+1}) \, , \quad \text{where} \quad p_\theta(x_t|x_{0:t}, z_{t+1}) = \frac{m_\theta(z_t, z_{t+1}) p_\theta(z_t|x_{0:t})}{p_\theta(z_{t+1}|x_{0:t})} \, .$$

Given x_T, this shows that the sequence $(z_t)_{t=0}^T$ forms a reverse-time Markov chain with the initial distribution $p_\theta(z_T|x_T)$ and backward Markov transition kernels $p_\theta(z_t|x_{0:t}, z_{t+1})$ (Kantas et al., 2015).

Backward Decomposition of the Variational Smoothing Distribution. We consider a variational smoothing distribution of the form

$$q_{\phi_{0:T}}(z_{0:T}|x_{0:T}) = q_{T,\phi_T}(z_T) \prod_{t=0}^{T-1} q_{t|t+1,\phi_t}(z_{t+1}, z_t) \, , \tag{15}$$

where $q_{T,\phi_T}(z_T)$ and $q_{t|t+1,\phi_t}(z_{t+1}, z_t)$ are variational approximations of the filtering density $p_\theta(z_T|x_{0:T})$ and the backward transition density $p_\theta(z_t|x_{0:t}, z_{t+1})$ respectively.

Convergence Results. We establish the convergence rate for this specific Sequential VAE structure, where the transition, emission, and backward kernel densities follow Gaussian distributions. In this framework, both the mean and variance are parameterized by neural networks, utilizing the same architecture for the mean and variance as specified in Theorem 3.6, which is commonly used in practice.

Theorem G.2. *Let $T \in \mathbb{N}$ and for all $0 \leq t \leq T$, consider*

$$\begin{aligned}
\mathcal{F}_m &= \{(z, z') \mapsto m_\theta(z, z') = \mathcal{N}(z'; \mu_\theta^g(z), \tau_m^2 \mathrm{I}_{d_z}) \,|\, \mu_\theta^m(z) \in \mathcal{F}_G, \theta \in \Theta \subseteq \mathbb{R}^{d_\theta}\} \, , \\
\mathcal{F}_g &= \{(z, x) \mapsto g_\theta(z, x) = \mathcal{N}(x; \mu_\theta^g(z), \tau_g^2 \mathrm{I}_{d_x}) \,|\, \mu_\theta^g(z) \in \mathcal{F}_G, \theta \in \Theta \subseteq \mathbb{R}^{d_\theta}\} \, , \\
\mathcal{F}_t &= \{(z, z') \mapsto q_{\phi_t, t|t+1}(z, z') = \mathcal{N}(z'; \mu_{\phi_t}(z), \Sigma_{\phi_t}(z)) \,|\, (\mu_{\phi_t}(z), \Sigma_{\phi_t}(z)) \in \mathcal{F}_{\mu,\Sigma}, \phi_t \in \Phi \subseteq \mathbb{R}^{d_\phi}\} \, .
\end{aligned}$$

For all $t \in \mathbb{N}$, assume that there exists $C_\infty > 0$ such that $\|z\|_\infty \leq C_\infty$ for all $z \in \mathsf{Z}$. Assume also that the data distribution π has a finite fourth moment, and that there exists some constant a such that for all $\theta \in \Theta$ and $\phi \in \Phi$,

$$\|\theta\|_\infty + \|\phi\|_\infty \leq a \, .$$

Let $(\theta_n, \phi_n) \in \Theta \times \Phi$ be the n-th iterate of the recursion in Algorithm 1, where $\gamma_n = C_\gamma n^{-1/2}$ with $C_\gamma > 0$. Assume that $\beta_1 < \sqrt{\beta_2} < 1$. For all $n \geq 1$, let $R \in \{0, \dots, n\}$ be a uniformly distributed random variable. Then,

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

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, $N = \max\{N_m, N_g\} + \max_t\{N_t\}$ the total number of layers in the encoder and decoder.

Theorem G.2 provides the convergence rate of $\mathcal{O}(\log n/\sqrt{n})$ for Deep Gaussian Non-Linear State-Space Models with variational smoothing, similar to 3.6 in the independent case, but with an additional dependence on the time series length T. Notably, the convergence rate scales almost linearly with T. In this context, the gradient is computed over the entire sequence of length T. Alternatively, the time series can be divided into smaller segments, with the gradient computed for each segment to potentially reduce the convergence rate term. However, this introduces bias into the gradient estimator, affecting the overall convergence rate. We leave this exploration for future work.

G.4 Convergence Proofs for the Sequential VAE

G.4.1 Proof of Theorem G.1

Proof. The strong mixing assumption (Assumption 5) ensures that Assumption 1 is satisfied with $\alpha(x, z)$ defined by:

$$\alpha(x, z) = \sum_{t=0}^{T} 2 \max\left(\left|\log \sigma_t^-\right|, \left|\log \sigma_t^+\right|\right).$$

By using the decomposition of the joint likelihood and the variational distribution, along with the boundedness of $s_{t,\theta}(x_{0:t}, z_{0:t})$ and $\log q_\phi(z_t|\bar{z}_t, \bar{x}_t)$, we derive the boundedness of the gradients of $\log p_\theta(x|z)$ and $\log q_\phi(z|x)$. Next, we will establish the smoothness of these quantities, addressing each one separately.

Lipschitz condition of $\nabla_\theta \log p_\theta(x_{0:T}|z_{0:T})$:

$$\begin{aligned}
\|\nabla_\theta \log p_\theta(x_{0:T}|z_{0:T}) - \nabla_\theta \log p_{\theta'}(x_{0:T}|z_{0:T})\| &= \|s_{0:T,\theta}(x_{0:T}, z_{0:T}) - s_{0:T,\theta'}(x_{0:T}, z_{0:T})\| \\
&\leq \sum_{t=0}^{T} \|s_{t,\theta}(x_{0:t}, z_{0:t}) - s_{t,\theta'}(x_{0:t}, z_{0:t})\| \\
&\leq \sum_{t=0}^{T} L_t^s(x_{0:t}, z_{0:t}) \|\theta - \theta'\|.
\end{aligned}$$

Lipschitz condition of $\nabla_\phi \log q_\phi(z_{0:T}|x_{0:T})$:

$$\begin{aligned}
\|\nabla_\phi \log q_\phi(z_{0:T}|x_{0:T}) - \nabla_\phi \log q_{\phi'}(z_{0:T}|x_{0:T})\| &= \left\|\sum_{t=0}^{T} \nabla_\phi \log q_\phi(z_t|\bar{z}_t, \bar{x}_t) - \nabla_\phi \log q_{\phi'}(z_t|\bar{z}_t, \bar{x}_t)\right\| \\
&\leq \sum_{t=0}^{T} \|\nabla_\phi \log q_\phi(z_t|\bar{z}_t, \bar{x}_t) - \nabla_\phi \log q_{\phi'}(z_t|\bar{z}_t, \bar{x}_t)\| \\
&\leq \sum_{t=0}^{T} L_t^q(z_t, \bar{z}_t, \bar{x}_t) \|\phi - \phi'\|,
\end{aligned}$$

which provides the smoothness condition, and thus Assumption 2 is satisfied with $L_1(x, z) = \sum_{t=0}^{T} L_t^s(x_{0:t}, z_{0:t})$ and $L_2(x, z) = \sum_{t=0}^{T} L_t^q(z_t, \bar{z}_t, \bar{x}_t)$ for $x = x_{0:T}$ and $z = z_{0:T}$. We conclude the proof by applying Theorem 3.3. □

G.4.2 Proof of Theorem G.2

Proof. For all $T \in \mathbb{N}$, writing $\nu_\theta(z_0) = m_\theta(z_{-1}, z_0)$ and $q_{T,\phi_T}(z_T) = q_{T|T+1,\phi_T}(z_{T+1}, z_T)$, the gradient of the ELBO with respect to θ is given by:

$$\begin{aligned}
\nabla_\theta \mathcal{L}_T(\theta, \phi; \mathbf{x}_{0:T}) &= \sum_{t=0}^{T} \mathbb{E}_{q_\phi(\cdot|x_{0:T})} \left[\nabla_\theta \left(\log m_\theta(z_{t-1}, z_t) + \log g_\theta(z_t, x_t)\right)\right] \\
&= \sum_{t=0}^{T} \mathbb{E}_{q_\phi(\cdot|x_{0:T})} \left[\nabla_\theta \left(-\frac{1}{2\tau_m^2}\|z_t - \mu_\theta^m(z_{t-1})\|^2 - \frac{1}{2\tau_g^2}\|x_t - \mu_\theta^g(z_t)\|^2\right)\right] \\
&= \sum_{t=0}^{T} \mathbb{E}_{q_\phi(\cdot|x_{0:T})} \left[\frac{1}{\tau_m^2} \nabla_\theta \mu_\theta^m(z_{t-1})^\top (z_t - \mu_\theta^m(z_{t-1})) + \frac{1}{\tau_g^2} \nabla_\theta \mu_\theta^g(z_t)^\top (x_t - \mu_\theta^g(z_t))\right].
\end{aligned}$$

For simplicity, we consider the case where the Lipschitz constant M_{f_i} and the smoothness constant L_{f_i} are both less than 1. This can be easily adapted for any Lipschitz and smoothness constants, similar to the case of independent data.

Lipschitz condition of $\nabla_\theta \mathcal{L}_T(\theta, \phi)$. First, as in the independent case, we have:

$$\|\nabla_\theta \mathcal{L}(\theta,\phi) - \nabla_\theta \mathcal{L}(\theta',\phi')\| \leq \|\nabla_\theta \mathcal{L}(\theta,\phi) - \nabla_\theta \mathcal{L}(\theta',\phi)\| + \|\nabla_\theta \mathcal{L}(\theta',\phi) - \nabla_\theta \mathcal{L}(\theta',\phi')\| \ . \tag{16}$$

Now, we bound each of these terms individually. For all $\theta, \theta' \in \Theta$, $\phi \in \Phi$,

$$\begin{aligned}
&\|\nabla_\theta \mathcal{L}_T(\theta,\phi;\mathbf{x}_{0:T}) - \nabla_\theta \mathcal{L}_T(\theta',\phi;\mathbf{x}_{0:T})\| \\
&\leq \frac{1}{\tau_m^2}\sum_{t=0}^{T} \mathbb{E}_{q_\phi(\cdot|x_{0:T})}\left[\left\|\nabla_\theta \mu_\theta^m(z_{t-1})^\top (z_t - \mu_\theta^m(z_{t-1})) - \nabla_\theta \mu_{\theta'}^m(z_{t-1})^\top (z_t - \mu_{\theta'}^m(z_{t-1}))\right\|\right] \\
&\quad + \frac{1}{\tau_g^2}\sum_{t=0}^{T} \mathbb{E}_{q_\phi(\cdot|x_{0:T})}\left[\left\|\nabla_\theta \mu_\theta^g(z_t)^\top (x_t - \mu_\theta^g(z_t)) - \nabla_\theta \mu_{\theta'}^g(z_t)^\top (x_t - \mu_{\theta'}^g(z_t))\right\|\right] \\
&\leq \frac{1}{\tau_m^2}\sum_{t=0}^{T} \mathbb{E}_{q_\phi(\cdot|x_{0:T})}\left[\|z_t - \mu_\theta^m(z_{t-1})\| \, \|\nabla_\theta \mu_\theta^m(z_{t-1}) - \nabla_\theta \mu_{\theta'}^m(z_{t-1})\| + \|\nabla_\theta \mu_{\theta'}^m(z_{t-1})\| \, \|\mu_\theta^m(z_{t-1}) - \mu_{\theta'}^m(z_{t-1})\|\right] \\
&\quad + \frac{1}{\tau_g^2}\sum_{t=0}^{T} \mathbb{E}_{q_\phi(\cdot|x_{0:T})}\left[\|x_t - \mu_\theta^g(z_t)\| \, \|\nabla_\theta \mu_\theta^g(z_t) - \nabla_\theta \mu_{\theta'}^g(z_t)\| + \|\nabla_\theta \mu_{\theta'}^g(z_t)\| \, \|\mu_\theta^g(z_t) - \mu_{\theta'}^g(z_t)\|\right] \\
&\leq L_T^{dd,1} \, \|\theta - \theta'\| \ ,
\end{aligned}$$

where $L_T^{dd,1} = (T+1)(a^{2(N_m-1)}/\tau_m^2 + a^{2(N_g-1)}/\tau_g^2) + (T+1)N_m a^{2(N_m-1)} C_\infty^2 \left(C_G + C_\infty\right)/\tau_m^2 + (T+1)N_g a^{2(N_g-1)} C_\infty^2 \left(C_G + \|x_{0:T}\|\right)/\tau_g^2$ using the boundedness of the latent state space and the Lipschitz continuity and smoothness of $\theta \mapsto \mu_\theta^m$ and $\theta \mapsto \mu_\theta^g$ (Lemma I.5). This concludes the proof for the first term in (16).

For the second term in (16), we have:

$$\|\nabla_\theta \mathcal{L}_T(\theta,\phi;\mathbf{x}_{0:T}) - \nabla_\theta \mathcal{L}_T(\theta,\phi';\mathbf{x}_{0:T})\| \leq \sum_{t=0}^{T} A_t^1 + \sum_{t=0}^{T} A_t^2 \ ,$$

where,

$$\begin{aligned}
A_t^1 &= \frac{1}{\tau_m^2}\left\| \int \left(\nabla_\theta \mu_\theta^m(z_{t-1})^\top (z_t - \mu_\theta^m(z_{t-1}))\right) \left(q_\phi(z_{0:T}|x_{0:T}) - q_{\phi'}(z_{0:T}|x_{0:T})\right) \mathrm{d}z_{0:T} \right\| \ , \\
A_t^2 &= \frac{1}{\tau_g^2}\left\| \int \left(\nabla_\theta \mu_\theta^g(z_t)^\top (x_t - \mu_\theta^g(z_t))\right) \left(q_\phi(z_{0:T}|x_{0:T}) - q_{\phi'}(z_{0:T}|x_{0:T})\right) \mathrm{d}z_{0:T} \right\| \ .
\end{aligned}$$

Next, we define the constants from Lemma D.4 that satisfy the strong mixing condition (Assumption 5) for the Gaussian density:

$$\sigma^- = \frac{1}{(2\pi C_\Sigma)^{d_z/2}} \exp\left(-\frac{1}{c_\Sigma}\left(C_\infty^2 + C_\mu^2\right)\right) \ , \quad \text{and} \quad \sigma^+ = \frac{1}{(2\pi c_\Sigma)^{d_z/2}} \exp\left(\frac{C_\mu^2}{4C_\Sigma}\right) \ .$$

For all $t \in \mathbb{N}$, let $f_\theta(z_{t-1}, z_t) = \nabla_\theta \mu_\theta^m(z_{t-1})^\top (z_t - \mu_\theta^m(z_{t-1}))/\tau_m^2$ for A_t^1 and $f_\theta(z_{t-1}, z_t) = \nabla_\theta \mu_\theta^g(z_t)^\top (x_t -$

$\mu_\theta^g(z_t))/\tau_g^2$ for A_t^2. In addition, for all $u \leq v$, write $\bar{q}_{u:v+1,\phi} = \prod_{s=u}^{v} q_{s|s+1,\phi_s}$. Then,

$$
\begin{aligned}
&\left\| \int f_\theta(z_{t-1}, z_t) \left(q_\phi(z_{0:T}|x_{0:T}) - q_{\phi'}(z_{0:T}|x_{0:T}) \right) \mathrm{d}z_{0:T} \right\| \\
&= \left\| \int f_\theta(z_{t-1}, z_t) \left(\prod_{s=t-1}^{T} q_{s|s+1,\phi_s}(z_{s+1}, z_s) - \prod_{s=t-1}^{T} q_{s|s+1,\phi'_s}(z_{s+1}, z_s) \right) \mathrm{d}z_{0:T} \right\| \\
&= \left\| \int f_\theta(z_{t-1}, z_t) \sum_{s=t-1}^{T} \left(\bar{q}_{s:T+1,\phi}(z_{s:T}) \bar{q}_{t-1:s,\phi'}(z_{t-1:s}) - \bar{q}_{s+1:T+1,\phi}(z_{s+1:T}) \bar{q}_{t-1:s+1,\phi'}(z_{t-1:s+1}) \right) \mathrm{d}z_{0:T} \right\| \\
&\leq \sum_{s=t-1}^{T} \left\| \int f_\theta(z_{t-1}, z_t) \left(\bar{q}_{s:T+1,\phi}(z_{s:T}) \bar{q}_{t-1:s,\phi'}(z_{t-1:s}) - \bar{q}_{s+1:T+1,\phi}(z_{s+1:T}) \bar{q}_{t-1:s+1,\phi'}(z_{t-1:s+1}) \right) \mathrm{d}z_{0:T} \right\| \\
&\leq \|f_\theta\|_\infty \sum_{s=t-1}^{T} \rho^{s-t+1} \left\| \mu_{s:T,\phi} - \widetilde{\mu}_{s:T,\phi,\phi'} \right\|_{\mathrm{TV}} ,
\end{aligned}
$$

where for all measurable set A, $\mu_{s:T,\phi}(A) = \int \prod_{\ell=s+1}^{T} q_{\ell|\ell+1,\phi_\ell}(z_{\ell+1}, z_\ell) q_{s|s+1,\phi_s}(z_{s+1}, z_s) \mathbf{1}_A(z_s) \mathrm{d}z_{s:T}$, $\widetilde{\mu}_{s:T,\phi,\phi'}(A) = \int \prod_{\ell=s+1}^{T} q_{\ell|\ell+1,\phi_\ell}(z_{\ell+1}, z_\ell) q_{s|s+1,\phi'_s}(z_{s+1}, z_s) \mathbf{1}_A(z_s) \mathrm{d}z_{s:T}$ and where we used (Gloaguen et al., 2022, Theorem 4.10) with $\rho = 1 - \sigma^-/\sigma^+$. Then, similar as in the independent case, using the inequality for all $x \geq 1$, $x - 1 \leq x \log x \leq |x \log x|$, we have:

$$
\begin{aligned}
&\left\| \mu_{s:T,\phi} - \widetilde{\mu}_{s:T,\phi,\phi'} \right\|_{\mathrm{TV}} \\
&= \frac{1}{2} \int \left| \bar{q}_{s+1:T+1,\phi}(z_{s+1:T}) \left(q_{s|s+1,\phi_s}(z_{s+1}, z_s) - q_{s|s+1,\phi'_s}(z_{s+1}, z_s) \right) \right| \mathrm{d}z_{s:T} \\
&\leq \frac{1}{2} \int \bar{q}_{s+1:T+1,\phi}(z_{s+1:T}) \left| q_{s|s+1,\phi_s}(z_{s+1}, z_s) - q_{s|s+1,\phi'_s}(z_{s+1}, z_s) \right| \mathrm{d}z_{s:T} \\
&\leq \frac{1}{2} \int \bar{q}_{s+1:T+1,\phi}(z_{s+1:T}) \left(\frac{q_{s|s+1,\phi_s}(z_{s+1}, z_s)}{q_{s|s+1,\phi'_s}(z_{s+1}, z_s)} - 1 \right) 1_{q_{s|s+1,\phi_s}(z_{s+1}, z_s) \geq q_{s|s+1,\phi'_s}(z_{s+1}, z_s)} q_{s|s+1,\phi'_s}(z_{s+1}, z_s) \mathrm{d}z_{s:T} \\
&\quad + \frac{1}{2} \int \bar{q}_{s+1:T+1,\phi}(z_{s+1:T}) \left(\frac{q_{s|s+1,\phi_s}(z_{s+1}, z_s)}{q_{s|s+1,\phi'_s}(z_{s+1}, z_s)} - 1 \right) 1_{q_{s|s+1,\phi_s}(z_{s+1}, z_s) > q_{s|s+1,\phi'_s}(z_{s+1}, z_s)} q_{s|s+1,\phi'_s}(z_{s+1}, z_s) \mathrm{d}z_{s:T} \\
&\leq \frac{1}{2} \int \bar{q}_{s+1:T+1,\phi}(z_{s+1:T}) \left| \frac{q_{s|s+1,\phi_s}(z_{s+1}, z_s)}{q_{s|s+1,\phi'_s}(z_{s+1}, z_s)} \log \frac{q_{s|s+1,\phi_s}(z_{s+1}, z_s)}{q_{s|s+1,\phi'_s}(z_{s+1}, z_s)} \right| q_{s|s+1,\phi'_s}(z_{s+1}, z_s) \mathrm{d}z_{s:T} \\
&\quad + \frac{1}{2} \int \bar{q}_{s+1:T+1,\phi}(z_{s+1:T}) \left| \frac{q_{s|s+1,\phi'_s}(z_{s+1}, z_s)}{q_{s|s+1,\phi_s}(z_{s+1}, z_s)} \log \frac{q_{s|s+1,\phi'_s}(z_{s+1}, z_s)}{q_{s|s+1,\phi_s}(z_{s+1}, z_s)} \right| q_{s|s+1,\phi_s}(z_{s+1}, z_s) \mathrm{d}z_{s:T} \\
&\leq \frac{1}{2} \int \bar{q}_{s+1:T+1,\phi}(z_{s+1:T}) \left| \log \frac{q_{s|s+1,\phi_s}(z_{s+1}, z_s)}{q_{s|s+1,\phi'_s}(z_{s+1}, z_s)} \right| q_{s|s+1,\phi_s}(z_{s+1}, z_s) \mathrm{d}z_{s:T} \\
&\quad + \frac{1}{2} \int \bar{q}_{s+1:T+1,\phi}(z_{s+1:T}) \left| \log \frac{q_{s|s+1,\phi'_s}(z_{s+1}, z_s)}{q_{s|s+1,\phi_s}(z_{s+1}, z_s)} \right| q_{s|s+1,\phi'_s}(z_{s+1}, z_s) \mathrm{d}z_{s:T} \\
&\leq C_\infty a^{\max\{N_t\}-1} \left\| \phi_s - \phi'_s \right\| ,
\end{aligned}
$$

where we used the Lipschitz condition of $\phi_s \mapsto \log q_{s|s+1,\phi_s}$. We then derive the following bounds:

$$
\begin{aligned}
A_t^1 &\leq \frac{1}{(1-\rho)\tau_m^2} C_\infty^2 \left(C_G + C_\infty \right) a^{N_m - 1} a^{\max\{N_t\}-1} \left\| \phi - \phi' \right\| , \\
A_t^2 &\leq \frac{1}{(1-\rho)\tau_g^2} C_\infty^2 \left(C_G + \|x_{0:T}\| \right) a^{N_g - 1} a^{\max\{N_t\}-1} \left\| \phi - \phi' \right\| .
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\left\| \nabla_\theta \mathcal{L}_T(\theta, \phi; \mathbf{x}_{0:T}) - \nabla_\theta \mathcal{L}_T(\theta, \phi'; \mathbf{x}_{0:T}) \right\| \\
&\qquad \leq (T+1) C_\infty^2 a^{\max\{N_t\}-1} \left(\frac{C_G + C_\infty}{(1-\rho)\tau_m^2} a^{N_m - 1} + \frac{C_G + \|x_{0:T}\|}{(1-\rho)\tau_g^2} a^{N_g - 1} \right) \left\| \phi - \phi' \right\| ,
\end{aligned}
$$

which concludes the proof of the Lipschitz condition for $\nabla_\theta \mathcal{L}_T(\theta, \phi)$ by taking the expectation over $\mathbf{x}_{0:T}$.

Lipschitz condition of $\nabla_{\phi_t}\mathcal{L}_T(\theta,\phi)$. Given an observation sequence $x_{0:T}$, the ELBO for a sequential VAE in this setting, along with the score function gradient of the ELBO with respect to ϕ is defined as:

$$\mathcal{L}_T(\theta,\phi;\mathbf{x}_{0:T}) = \sum_{t=0}^{T} \mathbb{E}_{q_\phi(\cdot|x_{0:T})}\left[\log m_\theta(z_t,z_{t+1}) + \log g_\theta(z_t,x_t) - \log q_{t|t+1,\phi_t}(z_{t+1},z_t)\right] ,$$
$$\nabla_{\phi_t}\mathcal{L}_T(\theta,\phi;\mathbf{x}_{0:T}) = \mathbb{E}_{q_\phi(\cdot|x_{0:T})}\left[\log \frac{m_\theta(z_t,z_{t+1})g_\theta(z_t,x_t)}{q_{t|t+1,\phi_t}(z_{t+1},z_t)} \nabla_{\phi_t}\log q_{t|t+1,\phi_t}(z_{t+1},z_t)\right] .$$

This gradient shares structural similarities with the independent case but differs due to the appearance of both the transition density m and the emission density g. Following a similar procedure to that in Lemma B.1, we obtain:

$$\begin{aligned}
&\|\nabla_{\phi_t}\mathcal{L}_T(\theta,\phi_t;\mathbf{x}_{0:T}) - \nabla_{\phi_t}\mathcal{L}_T(\theta',\phi_t;\mathbf{x}_{0:T})\| \\
&\leq \mathbb{E}_{q_\phi(\cdot|x_{0:T})}\left[\left\|\nabla_{\phi_t}\log q_{t|t+1,\phi_t}(z_{t+1},z_t)\right\| \|\log m_\theta(z_t,z_{t+1}) - \log m_{\theta'}(z_t,z_{t+1})\|\right] \\
&\quad + \mathbb{E}_{q_\phi(\cdot|x_{0:T})}\left[\left\|\nabla_{\phi_t}\log q_{t|t+1,\phi_t}(z_{t+1},z_t)\right\| \|\log g_\theta(z_t,x_t) - \log g_{\theta'}(z_t,x_t)\|\right] \\
&\leq C_\infty^2 a^{N_t-1}\left((C_G + C_\infty)\, a^{N_m-1} + (C_G + \|x_{0:T}\|)\, a^{N_g-1}\right) .
\end{aligned}$$

Now, let the unnormalized weights be denoted as follows:

$$w_{\theta,\phi_t}(x_t,z_t,z_{t+1}) = \frac{m_\theta(z_t,z_{t+1})g_\theta(z_t,x_t)}{q_{t|t+1,\phi_t}(z_{t+1},z_t)} .$$

Following the same approach as in the Lipschitz condition for $\nabla_\phi^{\mathsf{S}}\mathcal{L}(\theta,\phi)$ in the independent case (Lemma B.1), we have:

$$\|\nabla_{\phi_t}\mathcal{L}_T(\theta,\phi_t;\mathbf{x}_{0:T}) - \nabla_{\phi_t}\mathcal{L}_T(\theta,\phi_t';\mathbf{x}_{0:T})\| \leq A_t^1 + A_t^2 ,$$

where

$$\begin{aligned}
A_t^1 &= \left\|\mathbb{E}_\phi\left[\log w_{\theta,\phi_t}(x_t,z_t,z_{t+1})\nabla_{\phi_t}\log q_{t|t+1,\phi_t}(z_{t+1},z_t)\right] - \mathbb{E}_\phi\left[\log w_{\theta,\phi_t'}(x_t,z_t,z_{t+1})\nabla_{\phi_t}\log q_{t|t+1,\phi_t'}(z_{t+1},z_t)\right]\right\| , \\
A_t^2 &= \left\|\mathbb{E}_{\phi_t}\left[\log w_{\theta,\phi_t}(x_t,z_t,z_{t+1})\nabla_{\phi_t}\log q_{t|t+1,\phi_t}(z_{t+1},z_t)\right] - \mathbb{E}_{\phi_t'}\left[\log w_{\theta,\phi_t}(x_t,z_t,z_{t+1})\nabla_{\phi_t}\log q_{t|t+1,\phi_t}(z_{t+1},z_t)\right]\right\| .
\end{aligned}$$

Furthermore, denoting $x = x_{0:T}$ and $z = z_{0:T}$, we also obtain the following bounds:

$$\begin{aligned}
A_t^1 &\leq 2\mathbb{E}_{\pi,\phi}\left[\alpha(x,z)L_2(x,z)\right]\|\phi - \phi'\| + \mathbb{E}_{\pi,\phi}\left[M_q(x,z)^2\right]\|\phi - \phi'\| , \\
A_t^2 &\leq 2\left(\mathbb{E}_{\pi,\phi}\left[\alpha(x,z)M_q(x,z)^2\right] + \mathbb{E}_{\pi,\phi'}\left[\alpha(x,z)M_q(x,z)^2\right]\right)\|\phi - \phi'\| ,
\end{aligned}$$

where

$$\begin{aligned}
\alpha(x,z) &= \max\left\{\frac{d_z}{2}\log(2\pi C_\Sigma) + \frac{1}{c_\Sigma}\left(C_\infty^2 + C_\mu^2\right), \frac{d_z}{2}\log(2\pi c^2) + \frac{1}{c^2}\left(\|x_{0:T}\|^2 + C_G^2\right)\right\} , \\
L_2(x,z) &= \frac{N_t}{c_\Sigma}\|x_{0:T}\|^2 a^{2(N_t-1)}\left(4C_\infty^2 + 4C_\mu^2 + 4C_\infty + 4C_\mu + c_\Sigma + 1\right) , \\
M_q(x,z) &= C_\infty a^{N_t-1} .
\end{aligned}$$

We establish the Lipschitz condition for $\nabla_\phi\mathcal{L}_T(\theta,\phi)$ by combining all the derived inequalities and then taking the expectation over $x_{0:T}$. The proof is then completed by applying Theorem 3.3. □

H ADDITIONAL EXPERIMENTS

H.1 Additional Experiments details on CelebA

In this section, we provide further details regarding the experiments conducted on the CelebA dataset. Specifically, we examine how architectural modifications impact the convergence rate. Figure 5 illustrates the effect of

adding layers to the model on the squared norm of the gradients, Figure 5 illustrates the impact of adding layers to the model on the squared norm of the gradients $\|\nabla\mathcal{L}(\theta_n, \phi_n)\|^2$. The figure compares the baseline model, which has 22,607,435 parameters, with two variants: one that includes an additional fully connected layer (24,706,635 parameters) and another that adds a convolutional layer (30,993,483 parameters). The results indicate that adding an extra layer generally slows down convergence. Notably, the difference in convergence rates between the fully connected layer and the convolutional layer is relatively small, primarily due to the variation in the number of parameters introduced by each layer type.

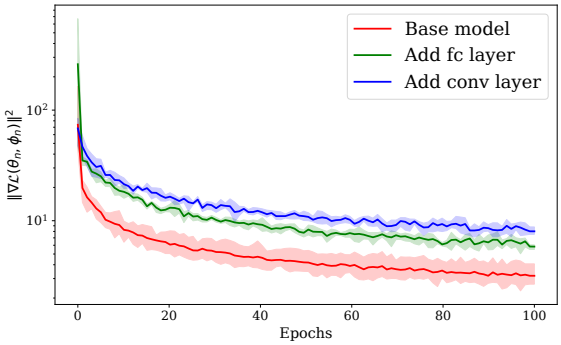

Figure 5: $\|\nabla\mathcal{L}(\theta_n, \phi_n)\|^2$ in VAE trained with Adam for the baseline model, a model with an additional fully connected layer, and a model with an additional convolutional layer. Bold lines represent the mean over 5 independent runs. Figures are plotted on a logarithmic scale for better visualization.

H.2 Experiments on CIFAR-100

Dataset and Model. We conduct our experiments on the CIFAR-100 dataset (Krizhevsky et al., 2009) and use a Convolutional Neural Network (CNN) architecture with ReLU and generalized soft-clipping activation functions for both the encoder and decoder networks. All other model, optimizer, and training parameters are consistent with those used for the CelebA dataset.

In the first experiment, we illustrate the convergence results of the standard VAE using our choice of activation functions, similar to those applied in the CelebA dataset. Figure 6 shows the squared norm of the gradients $\|\nabla\mathcal{L}(\theta_n, \phi_n)\|^2$ and the negative log-likelihood on the test dataset for both ReLU and the generalized soft-clipping activation function with various values of s. We observe a comparable convergence rate for all values of s. Notably, $s = 5$ emerges as a reasonable choice, balancing convergence rate and numerical stability, consistent with observations from the CelebA case. However, $s = 10$ also performs adequately, contrary to the results from the CelebA dataset.

Next, we estimate the squared norm of the gradients and the negative log-likelihood using the β-VAE and IWAE models. Figure 7 shows both the squared gradient norm and the negative log-likelihood for the β-VAE across different values of β. Additionally, Figure 8 displays the same quantities for the IWAE, evaluated with different values of K. As with the CelebA dataset, we observe that smaller values of β for the β-VAE, lead to faster convergence in both cases. Similarly, increasing the value of K for the IWAE results in faster convergence.

The simulations in this paper were conducted using the NVIDIA RTX 6000 GPUs with 48GB of VRAM. The total computing hours required for the results presented in this paper are estimated to be around 100 to 200 hours of GPU usage.

I TECHNICAL LEMMAS

Lemma I.1. *Let $G_\theta : z \mapsto \mathsf{NN}(z; \theta, f, N)$ denote a neural network with N layers, where the parameters are $\theta = \{W_i, b_i\}_{i=1}^N$, and activation functions are $f = \{f_i\}_{i=1}^N$. For all $\theta \in \Theta$ and $z \in \mathsf{Z}$, the gradient of $G_\theta(z)$ with*

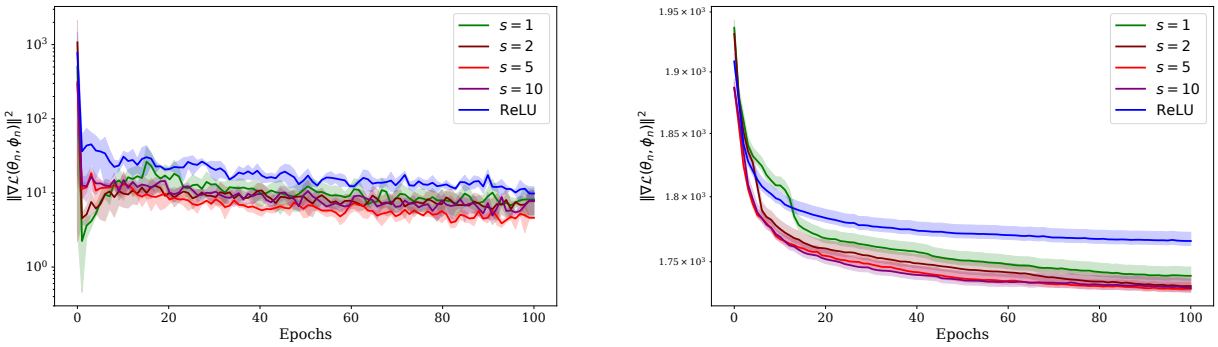

Figure 6: Squared norm of gradients and Negative ELBO on the test set of the CelebA for VAE trained with Adam and generalized soft-clipping activation function. Bold lines represent the mean over 5 independent runs.

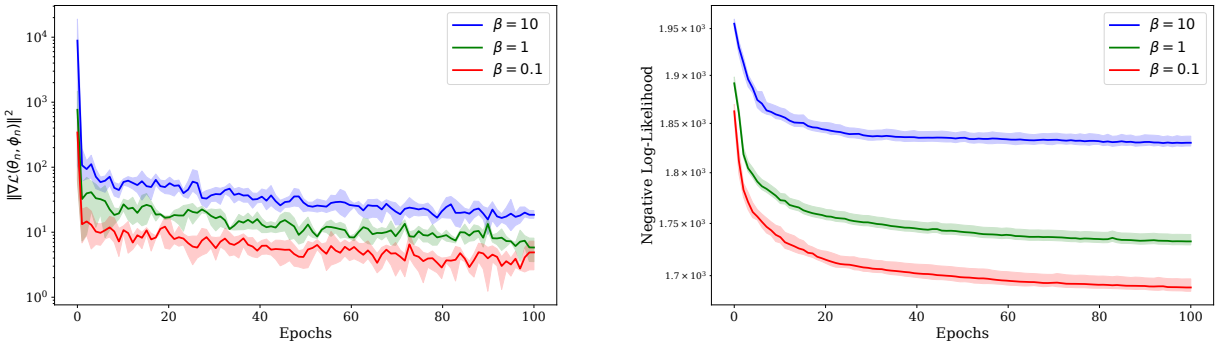

Figure 7: Squared norm of gradients (on the left) and Negative Log-Likelihood (on the right) for β-VAE trained on CIFAR-100 dataset using Adam. Bold lines represent the mean over 5 independent runs.

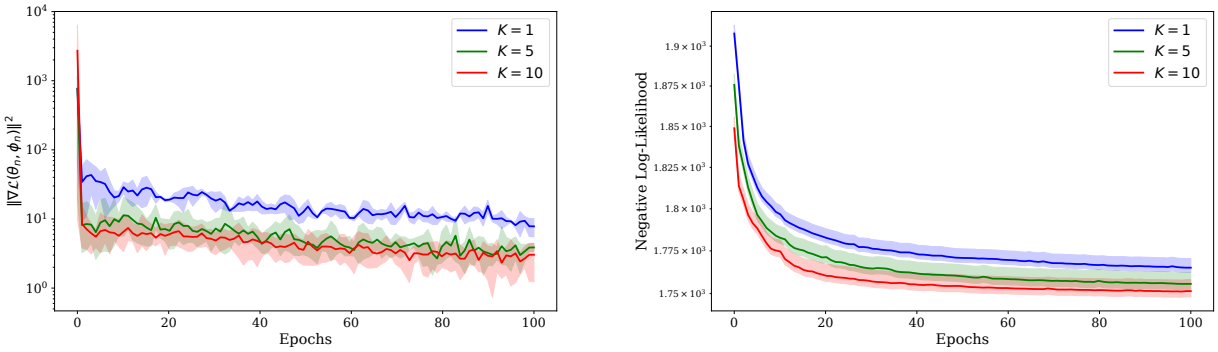

Figure 8: Squared norm of gradients (on the left) and Negative Log-Likelihood (on the right) for IWAE trained on CIFAR-100 dataset using Adam. Bold lines represent the mean over 5 independent runs.

respect to W_i *for all* $1 \leq i \leq N$*, can be expressed as:*

$$\nabla_{W_i} G_\theta(z) = \left(\prod_{j=i+1}^{N} f_j'(u_j) \cdot W_j \right) f_i'(u_i) f_{i-1}(u_{i-1})^\top \text{ and } \nabla_{b_i} G_\theta(z) = \left(\prod_{j=i+1}^{N} f_j'(u_j) \cdot W_j \right) f_i'(u_i) \,,$$

where $f_0(z) = u_0 = z, u_1 = W_1 z + b_1$ *and* $u_j = W_j f_{j-1}(u_{j-1}) + b_j$ *for all* $2 \leq j \leq N$*.*

Proof. To derive the gradient of $G_\theta(z)$ with respect to W_i and b_i, we use the chain rule. First, for the final layer,

i.e., $i = N$, the gradient of $G_\theta(z)$ with respect to W_N and b_N are given by:

$$\nabla_{W_N} G_\theta(z) = \frac{\partial G_\theta(z)}{\partial u_N} \cdot \frac{\partial u_N}{\partial W_N} \quad \text{and} \quad \nabla_{b_N} G_\theta(z) = \frac{\partial G_\theta(z)}{\partial u_N} \cdot \frac{\partial u_N}{\partial b_N} .$$

Note that $\nabla_{W_N} G_\theta(z) \in \mathbb{R}^{d_N \times d_N \times d_{N-1}}$, where d_i denotes the dimension of u_i. For all $p, q, r \in \mathbb{N}$, we observe that $(\nabla_{W_N} G_\theta(z))_{p,q,r} = 0$ if $p \neq q$. Thus, we can conventionally simplify this tensor to a matrix of size $\mathbb{R}^{d_N \times d_{N-1}}$ by ignoring the second dimension. Since $u_N = W_N f_{N-1}(u_{N-1}) + b_N$ and $G_\theta(z) = f_N(u_N)$, we have:

$$\nabla_{W_N} G_\theta(z) = f_N'(u_N) f_{N-1}(u_{N-1})^\top ,$$

$$\nabla_{b_N} G_\theta(z) = \mathrm{Diag}(f_N'(u_N)) .$$

For $i < N$, by recursively applying the chain rule, the gradient of $G_\theta(z)$ with respect to W_i for all $1 \leq i \leq N$, can be expressed as:

$$\nabla_{W_i} G_\theta(z) = \left(\prod_{j=i+1}^{N} f_j'(u_j) \cdot W_j \right) f_i'(u_i) f_{i-1}(u_{i-1})^\top \text{ and } \nabla_{b_i} G_\theta(z) = \left(\prod_{j=i+1}^{N} f_j'(u_j) \cdot W_j \right) f_i'(u_i) ,$$

where $f_0(z) = u_0 = z, u_1 = W_1 z + b_1$ and $u_j = W_j f_{j-1}(u_{j-1}) + b_j$ for all $2 \leq j \leq N$. □

Lemma I.2. *Let $G_\theta : z \mapsto \mathsf{NN}(z; \theta, f, N)$ denote a neural network with N layers, where the parameters are $\theta = \{W_i, b_i\}_{i=1}^N$, and activation functions are $f = \{f_i\}_{i=1}^N$ such that $f_i \in \mathcal{F}_{SL}$ for all $1 \leq i \leq N$. Let M_{f_i} and L_{f_i} represent the Lipschitz constant and the smoothness parameter of the activation function f_i in the i-th layer, respectively. Assume that there exists some constant a such that for any $\theta \in \Theta$, $\|\theta\| \leq a$. Then, for all $z, z' \in \mathsf{Z}$, and $\theta \in \Theta$:*

$$\|\nabla_z G_\theta(z)\| \leq a^N \prod_{j=1}^{N} M_{f_j} ,$$

$$\|\nabla_z G_\theta(z) - \nabla_z G_\theta(z')\| \leq \sum_{k=1}^{N} L_{f_k} a^{N+k} \prod_{j=1}^{k-1} M_{f_j}^2 \prod_{j=k+1}^{N} M_{f_j} \|z - z'\| .$$

Proof. Similar to Lemma I.1, the gradient of $G_\theta(z)$ with respect to z is defined as:

$$\nabla_z G_\theta(z) = \left(\prod_{j=1}^{N} f_j'(u_j) \cdot W_j \right) ,$$

where $u_1 = W_1 z + b_1$ and $u_j = W_j f_{j-1}(u_{j-1}) + b_j$ for all $2 \leq j \leq N$. We will now show by induction on the number of layers N that:

$$\|\nabla_z G_\theta(z) - \nabla_z G_\theta(z')\| \leq \sum_{k=1}^{N} L_{f_k} a^{N+k} \prod_{j=1}^{k-1} M_{f_j}^2 \prod_{j=k+1}^{N} M_{f_j} \|z - z'\| .$$

For the base case, consider the neural network with a single layer, where $G_\theta(z) = f_1(W_1 z + b_1)$. The gradient with respect to z is given by:

$$\nabla_z G_\theta(z) = f_1'(W_1 z + b_1) \cdot W_1 .$$

Note that

$$\begin{aligned} \|\nabla_z G_\theta(z) - \nabla_z G_\theta(z')\| &= \|f_1'(W_1 z + b_1) \cdot W_1 - f_1'(W_1 z' + b_1) \cdot W_1\| , \\ &\leq \|W_1\| \, \|f_1'(W_1 z + b_1) - f_1'(W_1 z' + b_1)\| , \\ &\leq a^2 L_{f_1} \|z - z'\| , \end{aligned}$$

which completes the base case.

Now, for the inductive step, assume that the Lipschitz constant holds for a network with $N-1$ layers. We show that it also holds for a neural network with N layer. Let u and u' be sequences defined as follows: $u_0 = z, u'_0 = z', u_1 = W_1 z' + b_1, u'_1 = W_1 z' + b_1, u_j = W_j f_{j-1}(u_{j-1}) + b_j$ and $u'_j = W_j f_{j-1}(u'_{j-1}) + b_j$ for all $2 \leq j \leq N$. We have:

$$\left\| \prod_{j=1}^{N} f'_j(u_j) \cdot W_j - \prod_{j=1}^{N} f'_j(u'_j) \cdot W_j \right\| \leq A_1 + A_2 \ ,$$

where

$$A_1 = \left\| f'_N(u_N) \cdot W_N \left(\prod_{j=1}^{N-1} f'_j(u_j) \cdot W_j \right) - f'_N(u'_N) \cdot W_N \left(\prod_{j=1}^{N-1} f'_j(u_j) \cdot W_j \right) \right\| \ ,$$
$$A_2 = \left\| f'_N(u'_N) \cdot W_N \left(\prod_{j=1}^{N-1} f'_j(u_j) \cdot W_j \right) - f'_N(u'_N) \cdot W_N \left(\prod_{j=1}^{N-1} f'_j(u'_j) \cdot W_j \right) \right\| \ .$$

First, we have:

$$\begin{aligned}
A_1 &= \left\| f'_N(u_N) \cdot W_N \left(\prod_{j=1}^{N-1} f'_j(u_j) \cdot W_j \right) - f'_N(u'_N) \cdot W_N \left(\prod_{j=1}^{N-1} f'_j(u_j) \cdot W_j \right) \right\| \\
&\leq \left\| \prod_{j=1}^{N-1} f'_j(u_j) \cdot W_j \right\| \|W_N\| \, \|f'_N(u_N) - f'_N(u'_N)\| \\
&\leq a^{N-1} a \prod_{j=1}^{N-1} M_{f_j} \, \|f'_N(u_N) - f'_N(u'_N)\| \\
&\leq a^N \prod_{j=1}^{N-1} M_{f_j} a^N L_{f_N} \prod_{j=1}^{N-1} M_{f_j} \, \|z - z'\| \\
&\leq a^{2N} L_{f_N} \prod_{j=1}^{N-1} M^2_{f_j} \, \|z - z'\| \ ,
\end{aligned}$$

where we used the Lipschitz and smoothness conditions of the activation functions in the second last inequality. For A2, using the induction hypothesis:

$$\begin{aligned}
A_2 &= \left\| f'_N(u'_N) \cdot W_N \left(\prod_{j=1}^{N-1} f'_j(u_j) \cdot W_j \right) - f'_N(u'_N) \cdot W_N \left(\prod_{j=1}^{N-1} f'_j(u'_j) \cdot W_j \right) \right\| \\
&\leq \|f'_N(u'_N)\| \, \|W_N\| \left\| \prod_{j=1}^{N-1} f'_j(u_j) \cdot W_j - \prod_{j=1}^{N-1} f'_j(u'_j) \cdot W_j \right\| \\
&\leq M_{f_N} a \sum_{k=1}^{N-1} L_{f_k} a^{N-1+k} \prod_{j=1}^{k-1} M^2_{f_j} \prod_{j=k+1}^{N-1} M_{f_j} \, \|z - z'\| \\
&\leq \sum_{k=1}^{N-1} L_{f_k} a^{N+k} \prod_{j=1}^{k-1} M^2_{f_j} \prod_{j=k+1}^{N} M_{f_j} \, \|z - z'\| \ .
\end{aligned}$$

By combining these two terms, we obtain:

$$\|\nabla_z G_\theta(z) - \nabla_z G_\theta(z')\| \leq \sum_{k=1}^{N} L_{f_k} a^{N+k} \prod_{j=1}^{k-1} M^2_{f_j} \prod_{j=k+1}^{N} M_{f_j} \, \|z - z'\| \ ,$$

which concludes the proof of the smoothness condition. For the boundedness of the gradient, we have:

$$\|\nabla_z G_\theta(z)\| = \left\| \prod_{j=1}^{N} f_j'(u_j) \cdot W_j \right\| \leq a^N \prod_{j=1}^{N} M_{f_j} \ ,$$

which completes the proof. □

Lemma I.2 establishes the boundedness and smoothness of a neural network with respect to its input. The following lemmas extends these properties to the parameters, proving their boundedness and smoothness. Specifically, the boundedness and smoothness with respect to W_1 are addressed in Lemma I.3, W_i in Lemma I.4 and θ in Lemma I.5.

Lemma I.3. *Let $G_\theta : z \mapsto \mathsf{NN}(z; \theta, f, N)$ denote a neural network with N layers, where the parameters are $\theta = \{W_i, b_i\}_{i=1}^N$, and activation functions are $f = \{f_i\}_{i=1}^N$ such that $f_i \in \mathcal{F}_{SL}$ for all $1 \leq i \leq N$. Let M_{f_i} and L_{f_i} represent the Lipschitz constant and the smoothness parameter of the activation function f_i in the i-th layer, respectively. Assume that there exists some constant a such that for any $\theta \in \Theta$, $\|\theta\| \leq a$. Then, for all $\theta, \theta' \in \Theta$, and $z \in \mathsf{Z}$:*

$$\|\nabla_{W_1} G_\theta(z)\| \leq \|z\| \, a^{N-1} \prod_{j=1}^{N} M_{f_j} \ ,$$

$$\left\|\nabla_{W_1} G_{W_1}(z) - \nabla_{W_1} G_{W_1'}(z)\right\| \leq \sum_{k=1}^{N} \|z\|^2 \, L_{f_k} a^{N-2+k} \prod_{j=1}^{k-1} M_{f_j}^2 \prod_{j=k+1}^{N} M_{f_j} \|W_1 - W_1'\| \ .$$

Proof. Using Lemma I.1, the gradient of $G_\theta(z)$ with respect to W_1 and b_1 are defined as:

$$\nabla_{W_1} G_\theta(z) = \left(\prod_{j=2}^{N} f_j'(u_j) \cdot W_i \right) f_1'(u_1) u_0^\top \quad \text{and} \quad \nabla_{b_1} G_\theta(z) = \left(\prod_{j=2}^{N} f_j'(u_j) \cdot W_j \right) f_1'(u_1) \ ,$$

where $u_0 = z, u_1 = W_1 z + b_1$ and $u_j = W_j f_{j-1}(u_{j-1}) + b_j$ for all $2 \leq j \leq N$. We now show by induction on the number of layers that the Lipschitz constant with respect to W_1 is $\sum_{k=1}^{N} \|z\|^2 L_{f_k} a^{N-2+k} \prod_{j=1}^{k-1} M_{f_j}^2 \prod_{j=k+1}^{N} M_{f_j}$. For the base case, consider the neural network with a single layer, where $G_\theta(z) = f_1(W_1 z + b_1)$. The gradient with respect to W_1 is given by:

$$\nabla_{W_1} G_\theta(z) = f_1'(W_1 z + b_1) z^\top \ .$$

Note that

$$\begin{aligned} \left\|\nabla_{W_1} G_{W_1}(z) - \nabla_{W_1} G_{W_1'}(z)\right\| &= \left\| f_1'(W_1 z + b_1) z^\top - f_1'(W_1' z + b_1) z^\top \right\| , \\ &\leq \|z\| \, \|f_1'(W_1 z + b_1) - f_1'(W_1' z + b_1)\| , \\ &\leq \|z\|^2 \, L_{f_1} \|W_1 - W_1'\| \ , \end{aligned}$$

which completes the base case.

Now, for the inductive step, assume that the Lipschitz constant holds for a network with $N-1$ layer. We show that it also holds for a neural network with N layers. Let u and u' be sequences defined as follows: $u_0 = u_0' = z, u_1 = W_1 z + b_1, u_1' = W_1' z + b_1, u_j = W_j f_{j-1}(u_{j-1}) + b_j$ and $u_j' = W_j f_{j-1}(u_{j-1}') + b_j$ for all $2 \leq j \leq N$. We have:

$$\begin{aligned} \left\|\nabla_{W_1} G_{W_1}(z) - \nabla_{W_1} G_{W_1'}(z)\right\| &= \left\| \left(\prod_{j=2}^{N} f_j'(u_j) \cdot W_j \right) f_1'(u_1) z^\top - \left(\prod_{j=2}^{N} f_j'(u_j') \cdot W_j \right) f_1'(u_1') z^\top \right\| \\ &\leq A_1 + A_2 \ , \end{aligned}$$

where

$$A_1 = \left\| f_N'(u_N) \cdot W_N \left(\prod_{j=2}^{N-1} f_j'(u_j) \cdot W_j \right) f_1'(u_1) z^\top - f_N'(u_N') \cdot W_N \left(\prod_{j=2}^{N-1} f_j'(u_j) \cdot W_j \right) f_1'(u_1) z^\top \right\| ,$$

$$A_2 = \left\| f_N'(u_N') \cdot W_N \left(\prod_{j=2}^{N-1} f_j'(u_j) \cdot W_j \right) f_1'(u_1) z^\top - f_N'(u_N') \cdot W_N \left(\prod_{j=2}^{N-1} f_j'(u_j') \cdot W_j \right) f_1'(u_1') z^\top \right\| .$$

First, we have:

$$\begin{aligned}
A_1 &= \left\| f_N'(u_N) \cdot W_N \left(\prod_{j=2}^{N-1} f_j'(u_j) \cdot W_j \right) f_1'(u_1) z^\top - f_N'(u_N') \cdot W_N \left(\prod_{j=2}^{N-1} f_j'(u_j) \cdot W_j \right) f_1'(u_1) z^\top \right\| \\
&\leq \left\| \left(\prod_{j=2}^{N-1} f_j'(u_j) \cdot W_j \right) f_1'(u_1) z^\top \right\| \|W_N\| \, \|f_N'(u_N) - f_N'(u_N')\| \\
&\leq a^{N-2} a \prod_{j=1}^{N-1} M_{f_j} \|z\| \, \|f_N'(u_N) - f_N'(u_N')\| \\
&\leq a^{N-1} \prod_{j=1}^{N-1} M_{f_j}^2 \|z\|^2 a^{N-1} L_{f_N} \|W_1 - W_1'\| \\
&\leq \|z\|^2 a^{2(N-1)} L_{f_N} \prod_{j=1}^{N-1} M_{f_j}^2 \|W_1 - W_1'\| ,
\end{aligned}$$

where we used the Lipschitz and smoothness conditions of the activation functions in the second last inequality. For A2, using the induction hypothesis,

$$\begin{aligned}
A_2 &= \left\| f_N'(u_N') \cdot W_N \left(\prod_{j=2}^{N-1} f_j'(u_j) \cdot W_j \right) f_1'(u_1) z^\top - f_N'(u_N') \cdot W_N \left(\prod_{j=2}^{N-1} f_j'(u_j') \cdot W_j \right) f_1'(u_1') z^\top \right\| \\
&\leq \|f_N'(u_N')\| \, \|W_N\| \left\| \left(\prod_{j=2}^{N-1} f_j'(u_j) \cdot W_j \right) f_1'(u_1) z^\top - \left(\prod_{j=2}^{N-1} f_j'(u_j') \cdot W_j \right) f_1'(u_1') z^\top \right\| \\
&\leq M_{f_N} a \|z\|^2 \sum_{k=1}^{N-1} L_{f_k} a^{N-3+k} \prod_{j=1}^{k-1} M_{f_j}^2 \prod_{j=k+1}^{N} M_{f_j} \|W_1 - W_1'\| \\
&\leq \|z\|^2 \sum_{k=1}^{N-1} L_{f_k} a^{N-2+k} \prod_{j=1}^{k-1} M_{f_j}^2 \prod_{j=k+1}^{N} M_{f_j} \|W_1 - W_1'\| .
\end{aligned}$$

By combining these two terms, we obtain:

$$\left\| \nabla_{W_1} G_{W_1}(z) - \nabla_{W_1} G_{W_1'}(z) \right\| \leq \|z\|^2 \sum_{k=1}^{N} L_{f_k} a^{N-2+k} \prod_{j=1}^{k-1} M_{f_j}^2 \prod_{j=k+1}^{N} M_{f_j} \|W_1 - W_1'\| ,$$

which concludes the proof of the smoothness condition. For the boundedness of the gradient, we have:

$$\|\nabla_{W_1} G_\theta(z)\| = \left\| \left(\prod_{j=2}^{N} f_j'(u_j) \cdot W_j \right) f_1'(u_1) z^\top \right\| \leq \|z\| \, a^{N-1} \prod_{j=1}^{N} M_{f_j} ,$$

which completes the proof. □

Lemma I.4. *Let $G_\theta : z \mapsto \mathsf{NN}(z;\theta,f,N)$ denote a neural network with N layers, where the parameters are $\theta = \{W_i, b_i\}_{i=1}^N$, and activation functions are $f = \{f_i\}_{i=1}^N$ such that $f_i \in \mathcal{F}_{SL}$ for all $1 \le i \le N$. Let M_{f_i} and L_{f_i} represent the Lipschitz constant and the smoothness parameter of the activation function f_i in the i-th layer, respectively. Assume that there exists some constant a such that for any $\theta \in \Theta$, $\|\theta\| \le a$. Then, for all $\theta, \theta' \in \Theta$, and $z \in \mathsf{Z}$:*

$$\|\nabla_{W_i} G_\theta(z)\| \le \|z\|\, a^{N-1} \prod_{j=1}^N M_{f_j} \ ,$$

$$\left\|\nabla_{W_i} G_{W_i}(z) - \nabla_{W_i} G_{W_i'}(z)\right\| \le \sum_{k=i}^N \|z\|^2 L_{f_k} a^{N-2i+k} \prod_{j=i}^{k-1} M_{f_j}^2 \prod_{j=k+1}^N M_{f_j} \|W_i - W_i'\| \ .$$

Proof. For the boundedness of the gradient of $G_\theta(z)$ with respect to W_i, we have:

$$\begin{aligned}\|\nabla_{W_i} G_\theta(z)\| \le \left\| \left(\prod_{j=i+1}^N f_j'(u_j) \cdot W_j \right) f_i'(u_i) f_{i-1}(u_{i-1})^\top \right\| &\le \|f_{i-1}(u_{i-1})\|\, a^{N-i} \prod_{j=i}^N M_{f_j} \\ &\le \|z\|\, a^{N-1} \prod_{j=1}^N M_{f_j} \ ,\end{aligned}$$

where we used the fact that $\|f_{i-1}(u_{i-1})\| \le \|z\|\, a^{i-1} \prod_{j=1}^{i-1} M_{f_j}$. For the smoothness condition, using a similar argument as in Lemma I.3, we define the sequences u and u' as follows: $u_0 = u_0' = z, u_1 = W_1 z + b_1, u_1' = W_1' z + b_1, u_j = W_j f_{j-1}(u_{j-1}) + b_j$ and $u_j' = W_j' f_{j-1}(u_{j-1}') + b_j$ for all $2 \le j \le N$. Thus, we obtain:

$$\begin{aligned}\left\|\nabla_{W_i} G_{W_i}(z) - \nabla_{W_i} G_{W_i'}(z)\right\| &\le \left\| \left(\prod_{j=i+1}^N f_j'(u_j) \cdot W_j \right) f_i'(u_i) f_{i-1}(u_{i-1})^\top - \left(\prod_{j=i+1}^N f_j'(u_j') \cdot W_j \right) f_i'(u_i') f_{i-1}(u_{i-1})^\top \right\| \\ &\le \|f_{i-1}(u_{i-1})\|^2 \sum_{k=i}^N L_{f_k} a^{N-2i+k} \prod_{j=i}^{k-1} M_{f_j}^2 \prod_{j=k+1}^N M_{f_j} \|W_i - W_i'\| \\ &\le \sum_{k=i}^N L_{f_k} a^{N-2+k} \prod_{j=1}^{k-1} M_{f_j}^2 \prod_{j=k+1}^N M_{f_j} \|W_i - W_i'\| \ .\end{aligned}$$

which concludes the proof. □

Lemma I.5. *Let $G_\theta : z \mapsto \mathsf{NN}(z;\theta,f,N)$ denote a neural network with N layers, where the parameters are $\theta = \{W_i, b_i\}_{i=1}^N$, and activation functions are $f = \{f_i\}_{i=1}^N$ such that $f_i \in \mathcal{F}_{SL}$ for all $1 \le i \le N$. Let M_{f_i} and L_{f_i} represent the Lipschitz constant and the smoothness parameter of the activation function f_i in the i-th layer, respectively. Assume that there exists some constant a such that for any $\theta \in \Theta$, $\|\theta\| \le a$. Then, for all $\theta, \theta' \in \Theta$, and $z \in \mathsf{Z}$:*

$$\|\nabla_\theta G_\theta(z)\| \le (\|z\| + 1)\, a^{N-1} \prod_{j=1}^N M_{f_j} \ ,$$

$$\|\nabla_\theta G_\theta(z) - \nabla_\theta G_{\theta'}(z)\| \le N \left(\|z\|^2 + 1\right) \sum_{k=1}^N L_{f_k} a^{N-2+k} \prod_{j=1}^{k-1} M_{f_j}^2 \prod_{j=k+1}^N M_{f_j} \|\theta - \theta'\| \ .$$

Proof. For the gradient of $G_\theta(z)$ with respect to θ, we have:

$$\begin{aligned}\|\nabla_\theta G_\theta(z)\| &= \max_{1 \le i \le N} \left\{ \|\nabla_{W_i} G_\theta(z)\|, \|\nabla_{b_i} G_\theta(z)\| \right\} \\ &\le \max_{1 \le i \le N} \|\nabla_{W_i} G_\theta(z)\| + \max_{1 \le i \le N} \|\nabla_{b_i} G_\theta(z)\| \\ &\le \|z\| a^{N-1} \prod_{j=1}^N M_{f_j} + a^{N-1} \prod_{j=1}^N M_{f_j} \ ,\end{aligned}$$

where we used Lemma I.4. Using Lemma I.4, we obtain the following smoothness condition:

$$\|\nabla_\theta G_\theta(z) - \nabla_\theta G_{\theta'}(z)\| = \max_{1\le i\le N} \|\nabla_{W_i} G_\theta(z) - \nabla_{W_i} G_{\theta'}(z)\| + \max_{1\le i\le N} \|\nabla_{b_i} G_\theta(z) - \nabla_{b_i} G_{\theta'}(z)\| \ .$$

For each weight W_i, we have:

$$\begin{aligned}
\|\nabla_{W_i} G_\theta(z) - \nabla_{W_i} G_{\theta'}(z)\| &\le \sum_{j=1}^{N} \left\|\nabla_{W_i} G_{W_j}(z) - \nabla_{W_i} G_{W_j'}(z)\right\| \\
&\le \sum_{j=1}^{N} L_{W_j} \left\|W_j - W_j'\right\| \\
&\le L_{\max} \sum_{j=1}^{N} \left\|W_j - W_j'\right\| \\
&\le L_{\max} N \|\theta - \theta'\| \ ,
\end{aligned}$$

where $L_{\max} = \|z\|^2 \max_{1\le i\le N} \left\{\sum_{k=i}^{N} L_{f_k} a^{N-2+k} \prod_{j=1}^{k-1} M_{f_j}^2 \prod_{j=k+1}^{N} M_{f_j}\right\}$. Similarly, for the bias terms b_i, we can use the same reasoning, obtaining an analogous bound. Thus, combining both the weight and bias terms, and noting that all terms in the sum for $L_{\max}$ are positive, we conclude:

$$\|\nabla_\theta G_\theta(z) - \nabla_\theta G_{\theta'}(z)\| \le N\left(\|z\|^2 + 1\right) \sum_{k=1}^{N} \|z\|^2 L_{f_k} a^{N-2+k} \prod_{j=1}^{k-1} M_{f_j}^2 \prod_{j=k+1}^{N} M_{f_j} \|\theta - \theta'\| \ .$$

□

Lemma I.6. *Let $G_\theta : z \mapsto \mathsf{NN}(z;\theta,f,N)$ denote a neural network with N layers, where the parameters are $\theta = \{W_i, b_i\}_{i=1}^N$, and activation functions are $f = \{f_i\}_{i=1}^N$ such that $f_i \in \mathcal{F}_{SL}$ for all $1 \le i \le N$. Let M_{f_i} and L_{f_i} represent the Lipschitz constant and the smoothness parameter of the activation function f_i in the i-th layer, respectively. Assume that there exists some constant a such that for any $\theta \in \Theta$, $\|\theta\| \le a$. Then, for all $1 \le i \le N$, $W_i, W_i' \in \Theta$, and $z \in \mathsf{Z}$,*

$$\left\|\nabla_z G_{W_i}(z) - \nabla_z G_{W_i'}(z)\right\| \le \prod_{j=1}^{i-1} M_{f_j} \left(a^{N-1} \prod_{j=i}^{N} M_{f_j} + \sum_{k=i}^{N} \|z\| L_{f_k} a^{N-i+k} \prod_{j=i}^{k-1} M_{f_j}^2 \prod_{j=k+1}^{N} M_{f_j}\right) \|W_i - W_i'\| \ .$$

Proof. The gradient of $G_\theta(z)$ with respect to z is defined as:

$$\nabla_z G_\theta(z) = \left(\prod_{j=1}^{N} f_j'(u_j) \cdot W_j\right),$$

where $u_1 = W_1 z + b_1$ and $u_j = W_j f_{j-1}(u_{j-1}) + b_j$ for all $2 \le j \le N$. Let u and u' be sequences defined as follows: $u_0 = u_0' = z, u_i = W_i z + b_i, u_i' = W_i' z + b_i$ and $u_j = u_j' = W_j f_{j-1}(u_{j-1}) + b_j$ for all $j \ne i$. We have: We have:

$$\begin{aligned}
\left\|\nabla_z G_{W_i}(z) - \nabla_z G_{W_i'}(z)\right\| &= \left\|\prod_{j=1}^{N} f_j'(u_j) \cdot W_j - \prod_{j=1}^{N} f_j'(u_j') \cdot W_j'\right\| \\
&\le \left\|\left(\prod_{j=1}^{i-1} f_j'(u_j) \cdot W_i\right)\left(\prod_{j=i}^{N} f_j'(u_j) \cdot W_j\right) - \left(\prod_{j=1}^{i-1} f_j'(u_j) \cdot W_j\right)\left(\prod_{j=i}^{N} f_j'(u_j') \cdot W_j'\right)\right\| \\
&\le \left\|\prod_{j=1}^{i-1} f_j'(u_j) \cdot W_j\right\| \left\|\prod_{j=i}^{N} f_j'(u_j) \cdot W_j - \prod_{j=i}^{N} f_j'(u_j') \cdot W_j'\right\| \\
&\le a^{i-1} \prod_{j=1}^{i-1} M_{f_j} \left\|\prod_{j=i}^{N} f_j'(u_j) W_j - \prod_{j=i}^{N} f_j'(u_j') W_j'\right\| \ .
\end{aligned}$$

We now show by induction on the number of layers that:

$$\left\|\prod_{j=i}^{N} f_j'(u_j)\cdot W_j - \prod_{j=i}^{N} f_j'(u_j')\cdot W_j'\right\| \leq \left(a^{N-i}\prod_{j=i}^{N} M_{f_j} + \sum_{k=i}^{N}\|z\|\, L_{f_k} a^{N-2i+k+1}\prod_{j=i}^{k-1} M_{f_j}^2 \prod_{j=k+1}^{N} M_{f_j}\right)\|W_i - W_i'\| \ .$$

For the base case, consider the neural network with a single layer, where $G_\theta(z) = f_1(W_1 z + b_1)$. The gradient with respect to z is given by:

$$\nabla_z G_\theta(z) = f_1'(W_1 z + b_1)\cdot W_1 \ .$$

We have:

$$\begin{aligned}\left\|\nabla_z G_{W_1}(z) - \nabla_z G_{W_1'}(z)\right\| &= \|f_1'(W_1 z + b_1)\cdot W_1 - f_1'(W_1' z + b_1)\cdot W_1'\| \\ &\leq \|f_1'(W_1 z + b_1)\cdot W_1 - f_1'(W_1 z + b_1)\cdot W_1'\| + \|f_1'(W_1 z + b_1)\cdot W_1' - f_1'(W_1' z + b_1)\cdot W_1'\| \\ &\leq M_{f_1}\|W_1 - W_1'\| + \|z\|\, a L_{f_1}\|W_1 - W_1'\| \ ,\end{aligned}$$

which completes the base case.

Now, for the inductive step, assume that the Lipschitz constant holds for a network with $N-1$ layers. We show that it also holds for a neural network with N layers. Let u and u' be sequences defined as follows: $u_0 = u_0' = z, u_i = W_i z + b_i, u_i' = W_i' z + b_i$ and $u_j = u_j' = W_j f_{j-1}(u_{j-1}) + b_j$ for all $j \neq i$. We have:

$$\left\|\prod_{j=i}^{N} f_j'(u_j)\cdot W_j - \prod_{j=i}^{N} f_j'(u_j')\cdot W_j'\right\| \leq A_1 + A_2 \ ,$$

where

$$\begin{aligned}A_1 &= \left\|f_N'(u_N)\cdot W_N\left(\prod_{j=i}^{N-1} f_j'(u_j)\cdot W_j\right) - f_N'(u_N')\cdot W_N\left(\prod_{j=i}^{N-1} f_j'(u_j)\cdot W_j\right)\right\| \ , \\ A_2 &= \left\|f_N'(u_N')\cdot W_N\left(\prod_{j=i}^{N-1} f_j'(u_j)\cdot W_j\right) - f_N'(u_N')\cdot W_N\left(\prod_{j=i}^{N-1} f_j'(u_j')\cdot W_j'\right)\right\| \ .\end{aligned}$$

First, we have:

$$\begin{aligned}A_1 &= \left\|f_N'(u_N)\cdot W_N\left(\prod_{j=i}^{N-1} f_j'(u_j)\cdot W_j\right) - f_N'(u_N')\cdot W_N\left(\prod_{j=i}^{N-1} f_j'(u_j)\cdot W_j\right)\right\| \\ &\leq \left\|\prod_{j=i}^{N-1} f_j'(u_j)\cdot W_j\right\|\|W_N\|\,\|f_N'(u_N) - f_N'(u_N')\| \\ &\leq a^{N-j} a \prod_{j=i}^{N-1} M_{f_j}\|f_N'(u_N) - f_N'(u_N')\| \\ &\leq a^{N-i+1}\prod_{j=i}^{N-1} M_{f_j}\|z\|\, a^{N-i} L_{f_N}\prod_{j=i}^{N-1} M_{f_j}\|W_i - W_i'\| \\ &\leq a^{2(N-i)+1}\|z\|\, L_{f_N}\prod_{j=i}^{N-1} M_{f_j}^2\|W_i - W_i'\| \ ,\end{aligned}$$

where we used the Lipschitz and smoothness conditions of the activation functions in the second last inequality.

For A2, using the induction hypothesis:

$$\begin{aligned}
A_2 &= \left\| f'_N(u'_N) \cdot W_N \left(\prod_{j=i}^{N-1} f'_j(u_j) \cdot W_j \right) - f'_N(u'_N) \cdot W_N \left(\prod_{j=i}^{N-1} f'_j(u'_j) \cdot W'_j \right) \right\| \\
&\leq \|f'_N(u'_N)\| \, \|W_N\| \left\| \prod_{j=i}^{N-1} f'_j(u_j) \cdot W_j - \prod_{j=i}^{N-1} f'_j(u'_j) \cdot W'_j \right\| \\
&\leq M_{f_N} a \left(a^{N-1-i} \prod_{j=i}^{N-1} M_{f_j} + \sum_{k=i}^{N-1} \|z\| \, L_{f_k} a^{N-2i+k} \prod_{j=i}^{k-1} M_{f_j}^2 \prod_{j=k+1}^{N-1} M_{f_j} \right) \|W_i - W'_i\| \\
&\leq \left(a^{N-i} \prod_{j=i}^{N} M_{f_j} + \sum_{k=i}^{N-1} \|z\| \, L_{f_k} a^{N-2i+k+1} \prod_{j=i}^{k-1} M_{f_j}^2 \prod_{j=k+1}^{N} M_{f_j} \right) \|W_i - W'_i\| \ .
\end{aligned}$$

By combining these two terms, we obtain:

$$\left\| \prod_{j=i}^{N} f'_j(u_j) \cdot W_j - \prod_{j=i}^{N} f'_j(u'_j) \cdot W'_j \right\| \leq \left(a^{N-i} \prod_{j=i}^{N} M_{f_j} + \sum_{k=i}^{N} \|z\| \, L_{f_k} a^{N-2i+k+1} \prod_{j=i}^{k-1} M_{f_j}^2 \prod_{j=k+1}^{N} M_{f_j} \right) \|W_i - W'_i\| \ ,$$

which concludes the proof. □

Lemma I.7. *Assume that there exists a constant $c_\Sigma > 0$ such that for all $\phi \in \Phi$ and $x \in \mathsf{X}$, $\lambda_{\min}(\Sigma_\phi(x)) \geq c_\Sigma$. Then, for all $\phi, \phi' \in \Phi$, and $x \in \mathsf{X}$:*

$$\left\| \Sigma_\phi(x)^{-1/2} - \Sigma_{\phi'}(x)^{-1/2} \right\| \leq c_\Sigma^{-3/2} \left\| \Sigma_\phi(x) - \Sigma_{\phi'}(x) \right\| \ .$$

Proof. Using the mean value theorem for matrix functions, we get:

$$\Sigma_\phi(x)^{-1/2} - \Sigma_{\phi'}(x)^{-1/2} = \int_0^1 \left((1-t)\Sigma_\phi(x) + t\Sigma_{\phi'}(x) \right)^{-3/2} \left(\Sigma_{\phi'}(x) - \Sigma_\phi(x) \right) \mathrm{d}t \ .$$

Therefore,

$$\left\| \Sigma_\phi(x)^{-1/2} - \Sigma_{\phi'}(x)^{-1/2} \right\| \leq \left\| \Sigma_\phi(x) - \Sigma_{\phi'}(x) \right\| \int_0^1 \left\| \left((1-t)\Sigma_\phi(x) + t\Sigma_{\phi'}(x) \right)^{-3/2} \right\| \mathrm{d}t \ .$$

Since $\lambda_{\min}(\Sigma_\phi(x)) \geq c_\Sigma$, for all $t \in [0,1]$, it follows that:

$$\lambda_{\min}\left((1-t)\Sigma_\phi(x) + t\Sigma_{\phi'}(x) \right) \geq (1-t)\lambda_{\min}(\Sigma_\phi(x)) + t\lambda_{\min}(\Sigma_{\phi'}(x)) \geq c_\Sigma \ .$$

Then,

$$\left\| \left((1-t)\Sigma_\phi(x) + t\Sigma_{\phi'}(x) \right)^{-3/2} \right\| \leq c_\Sigma^{-3/2} \ ,$$

and

$$\left\| \Sigma_\phi(x)^{-1/2} - \Sigma_{\phi'}(x)^{-1/2} \right\| \leq c_\Sigma^{-3/2} \left\| \Sigma_\phi(x) - \Sigma_{\phi'}(x) \right\| \ .$$

□