Statistical Guarantees for Variational Autoencoders using PAC-Bayesian Theory

Sokhna Diarra Mbacke, Florence Clerc, Pascal Germain

Statistical Guarantees for Variational Autoencoders using PAC-Bayesian Theory

Sokhna Diarra Mbacke
Université Laval
sokhna-diarra.mbacke.1@ulaval.ca

Florence Clerc
McGill University
florence.clerc@mail.mcgill.ca

Pascal Germain
Université Laval
pascal.germain@ift.ulaval.ca

Abstract

Since their inception, Variational Autoencoders (VAEs) have become central in machine learning. Despite their widespread use, numerous questions regarding their theoretical properties remain open. Using PAC-Bayesian theory, this work develops statistical guarantees for VAEs. First, we derive the first PAC-Bayesian bound for posterior distributions conditioned on individual samples from the data-generating distribution. Then, we utilize this result to develop generalization guarantees for the VAE's reconstruction loss, as well as upper bounds on the distance between the input and the regenerated distributions. More importantly, we provide upper bounds on the Wasserstein distance between the input distribution and the distribution defined by the VAE's generative model.

1 Introduction

In recent years, deep generative models have exhibited tremendous empirical success. Two of the most important families of generative models are Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Variational Autoencoders (Kingma and Welling, 2014; Rezende et al., 2014). GANs take an adversarial approach, whereas VAEs are based on maximum likelihood estimation and variational inference. VAEs comprise two main components: an encoder which parameterizes an approximation of the posterior distribution over the latent variables, and a decoder which parameterizes the likelihood. In addition to generative modelling tasks such as image generation (Vahdat and Kautz, 2020) and text generation (Bowman et al., 2016), VAEs have been successfully applied to other topics such as semi-supervised learning (Kingma et al., 2014), anomaly detection (An and Cho, 2015), and dimensionality reduction (Kaur et al., 2021). However, despite their empirical success, the question of statistical guarantees for the performance of VAEs remains largely open. Namely, how can one certify that VAEs generalize well, both in terms of reconstruction and generation?

PAC-Bayesian theory (McAllester, 1999; Catoni, 2003) is an influential tool of statistical learning theory dedicated to providing generalization bounds for machine learning models. PAC-Bayes has been applied to a wide variety of problems such as classification (Germain et al., 2009; Parrado-Hernández et al., 2012), meta-learning (Amit and Meir, 2018), co-clustering (Seldin and Tishby, 2010), domain adaptation (Germain et al., 2020), and online learning (Haddouche and Guedj, 2022). In recent years, PAC-Bayes has been used to derive non-vacuous generalization bounds for supervised learning algorithms based on neural networks (Dziugaite and Roy, 2018; Pérez-Ortiz et al., 2021). See Guedj (2019) and Alquier (2021) for excellent surveys.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

The objective of this work is to utilize PAC-Bayesian theory to derive statistical guarantees for VAEs. Our generalization bounds investigate the reconstruction, regeneration, as well as the generation properties of VAEs.

1.1 Related Works

In order to explain the empirical success of deep generative models, a lot of attention has been put into deriving theoretical guarantees for these models. Most of the results, however, have been dedicated to GANs and their variants (Arora et al., 2017; Zhang et al., 2018; Liang, 2021; Singh et al., 2018; Schreuder et al., 2021; Biau et al., 2021; Mbacke et al., 2023). A possible explanation for this plethora of theoretical results is the adversarial loss function, which directly offers an estimation of the discrepancy between the input distribution and the generator's distribution. Despite being central tools in modern machine learning, VAEs have not benefited from such a thorough theoretical analysis (Chakrabarty and Das, 2021).

The work of Chakrabarty and Das (2021) studies the regeneration properties of Wasserstein autoencoders (WAEs) (Tolstikhin et al., 2018), which come from the same family as VAEs. Using VC theory, Chakrabarty and Das (2021) derive rates of convergence for the Wasserstein distance between the input distribution and the distribution regenerated by the WAE, as well as the total variation distance between the empirical latent distribution and the latent prior. Taking a more empirical approach, Chérief-Abdellatif et al. (2022) use PAC-Bayes to study the generalization properties of stochastic reconstruction models. They define a $[0, 1]$-bounded reconstruction loss function, then utilize McAllester's bound (McAllester, 2003) to formulate a generalization bound for models with probabilistic neural networks (Langford and Caruana, 2001). Then, they re-scale their loss and compare the empirical results to the reconstruction of standard VAEs on benchmark datasets.

We also mention the work of Mbacke et al. (2023), who developed PAC-Bayesian bounds for the analysis of adversarial generative models. Using McDiarmid's inequality, they proved upper bounds on the distance between the input distribution and the generator's distribution, for WGANs (Arjovsky et al., 2017) and EBGANs (Zhao et al., 2017).

1.2 Our Contributions

In this work, we derive theoretical guarantees for variational autoencoders using PAC-Bayesian theory. We provide three types of guarantees: reconstruction guarantees showing that VAEs can successfully reconstruct unseen samples from the input distribution; regeneration guarantees proving upper bounds on the Wasserstein distance between the input distribution and the distribution regenerated by the VAE, given the training set as input; and finally, generation guarantees showing upper bounds on the Wasserstein distance between the data-generating distribution and the VAE's generated distribution defined by the latent prior and the decoder. To the best of our knowledge, these are the first generalization bounds for the standard VAE's reconstruction and regeneration properties, as well as the first statistical guarantees for the VAE's generative model.

In our analysis, the PAC-Bayesian posterior coincides with the variational posterior, which requires the PAC-Bayesian posterior to be conditional. Since, to the best of our knowledge, such PAC-Bayes bounds do not exist in the literature, we start by developing the first PAC-Bayesian bound for conditional posterior distributions. Then, we provide upper bounds for the VAE's performance under two main assumptions: we start by assuming the instance space is bounded, then we take advantage of the manifold hypothesis. Our bounds are functions of the optimization objective of the VAE, namely, the empirical reconstruction loss, and the empirical KL-loss.

The remainder of this paper is organized as follows. In Section 2, we define some preliminary concepts, then briefly introduce VAEs and PAC-Bayesian theory. Section 3 presents our general PAC-Bayesian theorem for conditional posteriors. Then, in Sections 4 and 5, we present our generalization bounds for the reconstruction loss, and the regeneration and generation guarantees.

2 Preliminaries

2.1 Definitions and Notations

Given metric spaces $(\mathcal{X}, d)$ and $(\mathcal{Y}, d')$, and a real number $K > 0$, a function $f : \mathcal{X} \to \mathcal{Y}$ is K-Lipschitz continuous if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we have

$$d'(f(\mathbf{x}), f(\mathbf{y})) \leq K d(\mathbf{x}, \mathbf{y}).$$

The smallest K such that this condition is satisfied is called the *Lipschitz norm* or *Lipschitz constant* of f and is denoted $\|f\|_{\text{Lip}}$. Moreover, the set of K-Lipschitz continuous functions $f : \mathcal{X} \to \mathcal{Y}$ is denoted $\text{Lip}_K(\mathcal{X}, \mathcal{Y})$ (the underlying metrics will be clear from the context).

Throughout the paper, we use lower case letters p, q to denote both probability distributions and their densities w.r.t. the Lebesgue measure. We may add variables between parentheses to improve readability (e.g. $p(\mathbf{z})$ to emphasize that p is a distribution on the space of variables $\mathbf{z}$, and $q(\mathbf{z}|\mathbf{x})$ to indicate that q is a conditional distribution). The set of probability measures on a space $\mathcal{X}$ is denoted $\mathcal{M}_+^1(\mathcal{X})$. The Kullback–Leibler (KL) divergence between $p, q \in \mathcal{M}_+^1(\mathcal{X})$ is denoted $\text{KL}(p \,||\, q)$. We omit the absolute continuity condition $p \ll q$ in the statements of the results below, since if it is not satisfied, then one may assume the KL divergence is infinite and the bounds hold trivially.

Integral Probability Metrics (IPM, see Müller (1997)) are a class of pseudo-metrics defined on the space of probability measures. Given a family $\mathcal{F}$ of real-valued functions defined on $\mathcal{X}$, the IPM defined by $\mathcal{F}$ is denoted $d_{\mathcal{F}}$ and defined as

$$d_{\mathcal{F}}(p, q) = \sup_{f \in \mathcal{F}} \left| \int f \, dp - \int f \, dq \right|, \quad \forall p, q \in \mathcal{M}_+^1(\mathcal{X}). \tag{1}$$

Stemming from the theory of optimal transportation (Villani, 2009), the Wasserstein distances (see Definition A.2) are a class of metrics between probability measures. The Wasserstein distance of order 1, also referred to simply as *the Wasserstein distance*, is the IPM defined by the set $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R} \text{ s.t. } \|f\|_{\text{Lip}} \leq 1\}$.

Finally, we recall the definition of a *pushforward measure*. Let p be a probability distribution on a space $\mathcal{Z}$ and $g : \mathcal{Z} \to \mathcal{X}$ be a measurable function. The pushforward measure defined by g and p and denoted $g\sharp p$ is a probability distribution on $\mathcal{X}$ defined as $g\sharp p(A) = p(g^{-1}(A))$, for any measurable set $A \subseteq \mathcal{X}$. In other words, sampling $\mathbf{x} \sim g\sharp p$ means sampling $\mathbf{z} \sim p$ first, then setting $\mathbf{x} = g(\mathbf{z})$.

2.2 Variational Autoencoders

We consider a Euclidean observation space $\mathcal{X}$, a data-generating distribution $\mu \in \mathcal{M}_+^1(\mathcal{X})$, and a latent space $\mathcal{Z} = \mathbb{R}^{d_{\mathcal{Z}}}$. VAEs comprise two main components: the encoder network whose parameters are denoted ϕ, and the decoder network whose parameters are denoted θ. For simplicity, we may refer to ϕ and θ as the encoder and decoder respectively. The encoder parameterizes a distribution $q_\phi(\mathbf{z}|\mathbf{x})$ over the latent space $\mathcal{Z}$, which is a variational approximation of the Bayesian posterior $p_\theta(\mathbf{z}|\mathbf{x})$. The likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ is parameterized by the decoder network. In this work, we consider the standard VAE, with a standard Gaussian prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ on $\mathcal{Z}$ and Gaussian latent distributions $q_\phi(\mathbf{z}|\mathbf{x})$. More precisely, for any $\mathbf{x} \in \mathcal{X}$, the distribution $q_\phi(\mathbf{z}|\mathbf{x})$ is a Gaussian distribution with a diagonal covariance matrix $\mathcal{N}(\mu_\phi(\mathbf{x}), \text{diag}(\sigma_\phi^2(\mathbf{x})))$, where

$$\mu_\phi : \mathcal{X} \to \mathcal{Z} = \mathbb{R}^{d_{\mathcal{Z}}} \quad \text{and} \quad \sigma_\phi : \mathcal{X} \to \mathbb{R}^{d_{\mathcal{Z}}}_{\geq 0}.$$

Note that $\text{diag}(\sigma)$ denotes the diagonal matrix whose main diagonal is the vector σ. In order to simplify some of the expressions below, it may be useful to express the encoder network as a function

$$Q_\phi : \mathcal{X} \to \mathbb{R}^{2d_{\mathcal{Z}}}, \quad \text{where } Q_\phi(\mathbf{x}) = \begin{bmatrix} \mu_\phi(\mathbf{x}) \\ \sigma_\phi(\mathbf{x}) \end{bmatrix}. \tag{2}$$

We express the decoder as a parametric function $g_\theta : \mathcal{Z} \to \mathcal{X}$. For any $\mathbf{x} \in \mathcal{X}$, upon receiving $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$, the decoder's output $g_\theta(\mathbf{z})$ is a reconstruction of $\mathbf{x}$. Given a training set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the encoder and decoder networks are jointly trained by minimizing the following objective:

$$\mathcal{L}_{\text{VAE}}(\phi, \theta) = \frac{1}{n} \sum_{i=1}^{n} \left[\mathop{\mathbb{E}}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}_i)} [-\log p_\theta(\mathbf{x}_i|\mathbf{z})] + \beta \text{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i) \,||\, p(\mathbf{z})) \right], \tag{3}$$

where the first part of (3) is the *reconstruction loss* and the second part is the KL-divergence between the latent distributions (associated to the training samples) and the prior over the latent space, weighted by a hyperparameter $\beta > 0$ (Higgins et al., 2017). The reconstruction loss measures the similarity between $\mathbf{x}$ and its reconstruction $g_\theta(\mathbf{z})$, and can be defined in many ways. With a Gaussian likelihood, the reconstruction loss is the squared L_2 norm $\|\mathbf{x} - g_\theta(\mathbf{z})\|^2$.

After training, the VAE defines a generative model using the prior $p(\mathbf{z})$ and the decoder g_θ (Kingma and Welling, 2014). The distribution $g_\theta \sharp p(\mathbf{z}) \in \mathcal{M}_+^1(\mathcal{X})$ allows one to generate new samples by first sampling a latent vector from the prior, then passing it through the decoder. We refer to $g_\theta \sharp p(\mathbf{z})$ as the VAE's generated distribution.

2.3 A Brief Introduction to PAC-Bayesian Theory

Dating back to McAllester (1999), PAC-Bayesian theory develops high-probability generalization bounds for machine learning algorithms. In essence, PAC-Bayes frames the output of such algorithm as a posterior distribution over a class of hypotheses, and provides an upper bound on the discrepancy between a model's empirical risk and its population risk.

PAC-Bayes considers the following concepts: a hypothesis class $\mathcal{H}$, a training set $S = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ iid sampled from an unknown distribution μ over an instance space $\mathcal{X}$[1], and a real-valued loss function $\ell : \mathcal{H} \times \mathcal{X} \to [0, \infty)$. Moreover, the primary goal of PAC-Bayes is to provide generalization bounds uniformly valid for any posterior $q \in \mathcal{M}_+^1(\mathcal{H})$. These bounds are dependent on the empirical performance of q and its closeness to a chosen *prior distribution* $p \in \mathcal{M}_+^1(\mathcal{H})$, as measured by the KL-divergence. The empirical and true risks of a posterior distribution $q \in \mathcal{M}_+^1(\mathcal{H})$ are defined as

$$\hat{\mathcal{R}}_S(q) = \mathop{\mathbb{E}}_{h\sim q(h)}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(h, \mathbf{x}_i)\right] \quad \text{and} \quad \mathcal{R}(q) = \mathop{\mathbb{E}}_{h\sim q(h)}\left[\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\ell(h, \mathbf{x})\right].$$

As an illustration, consider the following PAC-Bayesian bound for bounded loss functions developed by Catoni (2003).

Theorem 2.1. *Given a probability measure μ on $\mathcal{X}$, a hypothesis class $\mathcal{H}$, a prior distribution p on $\mathcal{H}$, a loss function $\ell : \mathcal{H} \times \mathcal{X} \to [0, 1]$, real numbers $\delta \in (0, 1)$ and $\lambda > 0$, with probability at least $1 - \delta$ over the random draw of $S \sim \mu^{\otimes n}$, the following holds for any posterior $q \in \mathcal{M}_+^1(\mathcal{H})$:*

$$\mathcal{R}(q) \leq \hat{\mathcal{R}}_S(q) + \frac{\lambda}{8n} + \frac{\mathrm{KL}(q\,||\,p) + \log\frac{1}{\delta}}{\lambda}.$$

The connection between PAC-Bayesian theory and Bayesian inference was highlighted by Grünwald (2012) and Germain et al. (2016), who showed that with a proper choice of λ and the negative log-likelihood as the loss function ℓ, the optimal posterior minimizing the right-hand side of Catoni's bound is the Bayesian posterior. Note that although the Bayesian posterior is unique (for a given prior and likelihood), a "PAC-Bayesian posterior" could be, in principle, any distribution over $\mathcal{H}$.

In our PAC-Bayesian analysis of VAEs, we will use the latent space $\mathcal{Z}$ as our hypothesis class, so that the VAE's prior will coincide with the PAC-Bayesian prior and the variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ will stand for our PAC-Bayesian posterior. An immediate concern with this approach is that the encoder's distributions are conditioned on individual samples $\mathbf{x} \sim \mu$, whereas the usual PAC-Bayesian bounds hold for unconditional posteriors $q(h)$. We address this issue in the next section, by developing a novel PAC-Bayesian bound for posterior distributions $q(\cdot|\mathbf{x})$. This general result will be later utilized to analyze VAEs.

3 A General PAC-Bayesian Bound with a Conditional Posterior

In this section, we present our general PAC-Bayesian bound with a conditional posterior distribution. Note that the novelty of this result is not the conditioning on observations, since this can be achieved by exploiting the existing PAC-Bayesian bounds. Indeed, Haddouche and Guedj (2022) utilized the general theorem of Rivasplata et al. (2020) to derive bounds for the online learning framework.

[1] In supervised learning, the instance space has the form $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is a set of features, and $\mathcal{Y}$ a set of labels. We use a more general formulation to encompass the unsupervised learning setting.

Instead, the contribution of Theorem 3.1 is to predict the behavior of $q(h|\mathbf{x})$, for any (previously unseen) $\mathbf{x} \sim \mu$, when the posterior q was only learned using the training samples $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. To the best of our knowledge, this is the first PAC-Bayesian bound where the posterior distribution is a conditional distribution conditioned on individual elements from the instance space. This bound will require the posterior q and the loss function ℓ to satisfy the following technical assumption.

Assumption 1. We say that a distribution $q(\cdot|\mathbf{x})$ and a loss function ℓ satisfy Assumption 1 with a constant $K > 0$ if there exists a family $\mathcal{E}$ of functions $\mathcal{H} \to \mathbb{R}$ such that the following properties hold.

1. The function $\mathbf{x} \mapsto q(\cdot|\mathbf{x})$ is continuous in the following sense: for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$,
$$d_{\mathcal{E}}\left(q(h|\mathbf{x}_1), q(h|\mathbf{x}_2)\right) \leq K d(\mathbf{x}_1, \mathbf{x}_2).$$
2. For any $\mathbf{x} \in \mathcal{X}$, the function $\ell(\cdot, \mathbf{x}) : \mathcal{H} \to \mathbb{R}$ is in $\mathcal{E}$:
$$\ell(\cdot, \mathbf{x}) \in \mathcal{E}, \quad \text{for any } \mathbf{x} \in \mathcal{X}.$$

Before stating the general result, let us pause and discuss this assumption. Intuitively, the goal of a generalization bound is to predict the behavior of the posterior distribution $q(h|\mathbf{x})$ on previously unseen examples $\mathbf{x} \sim \mu$. Since the posterior $q(h|\mathbf{x})$ is learned by minimizing the loss function ℓ on the training samples $S = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, one may need two things to be true.

First, the mapping $\mathbf{x} \mapsto q(h|\mathbf{x})$ has to be somewhat continuous. This is ensured by the first part of Assumption 1, which states that the posterior q is Lipschitz-continuous[2] with respect to the IPM $d_{\mathcal{E}}$ and the underlying metric d on $\mathcal{X}$. Indeed, this tells us that if $\mathbf{x}_1$ and $\mathbf{x}_2$ are close w.r.t. the underlying metric on $\mathcal{X}$, then $q(h|\mathbf{x}_1)$ and $q(h|\mathbf{x}_2)$ are close, w.r.t. the IPM $d_{\mathcal{E}}$.

Second, that continuity has to be "understood" by the loss function ℓ, which corresponds to the second part of the assumption. It states that the loss function's discriminative power is weaker than the one defined by the IPM $d_{\mathcal{E}}$. In other words, the discrepancy measure used to measure the similarity between the distributions $q(h|\mathbf{x}_1)$ and $q(h|\mathbf{x}_2)$ needs to be just strong enough to fool the loss function into thinking that the distributions are close to each other. An alternate formulation of Assumption 1 is provided in the supplementary material (Remark F.1).

Finally, we emphasize that Assumption 1 is not as restrictive as it may seem at first. For instance, it is satisfied by a VAE's variational posterior, when the encoder and decoder networks have finite Lipschitz norms and the reconstruction loss is defined with the L_2 norm (see Proposition 4.1). We are ready to state our first result.

Theorem 3.1. *Let $(\mathcal{X}, d)$ be a metric space. Consider a probability measure μ on $\mathcal{X}$, a hypothesis class $\mathcal{H}$, a prior distribution $p(h)$ on $\mathcal{H}$, a loss function $\ell : \mathcal{H} \times \mathcal{X} \to \mathbb{R}$, real numbers $\delta \in (0, 1)$ and $\lambda > 0$. With probability at least $1 - \delta$ over the random draw of $S \sim \mu^{\otimes n}$, the following holds for any conditional posterior $q(h|\mathbf{x})$ such that Assumption 1 is satisfied by $q(h|\mathbf{x})$ and ℓ with constant $K > 0$:*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x})} \ell(h, \mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x}_i)} \ell(h, \mathbf{x}_i) \leq \frac{1}{\lambda}\left[\sum_{i=1}^{n} \mathrm{KL}(q(h|\mathbf{x}_i) \,||\, p(h)) + \frac{\lambda K}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} d(\mathbf{x}, \mathbf{x}_i) + \log\frac{1}{\delta} + \log \mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \mathop{\mathbb{E}}_{h\sim p(h)} e^{\lambda\left(\mathbb{E}_{\mathbf{x}\sim\mu}[\ell(h,\mathbf{x})] - \frac{1}{n}\sum_{i=1}^{n}\ell(h,\mathbf{x}_i)\right)}\right].$$

In order to prove Theorem 3.1, we start by deriving a bound where the expected loss for samples $\mathbf{x} \sim \mu$ is computed w.r.t. distributions $q(h|\mathbf{x}_i)$ associated to the training samples (see Lemma B.1). This result uses standard PAC-Bayesian techniques, with a key difference: we start with n iid hypotheses from the prior $p(h)$, then we perform the change of measure with n posteriors $q_\phi(\mathbf{z}|\mathbf{x}_1), \dots, q_\phi(\mathbf{z}|\mathbf{x}_n)$, and show that the resulting exponential moment is equal to the one in Theorem 3.1. Moreover, one of the original aspects of this work comes from Assumption 1, which enables us to obtain a bound where the expected loss for $\mathbf{x} \sim \mu$ is computed w.r.t. the posterior $q(h|\mathbf{x})$, associated to $\mathbf{x}$ itself instead of all the training samples. However, the price to pay for having a posterior $q(h|\mathbf{x})$ for each $\mathbf{x} \in \mathcal{X}$ is that the bound depends on $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathbf{x}\sim\mu} d(\mathbf{x}, \mathbf{x}_i)$, which we refer to as the *average distance*.

[2] $d_{\mathcal{E}}$ is a pseudo-metric in the general case, so we abuse the definition by calling this Lipschitz continuity, since the latter concept is only defined for metric spaces.

Applied to supervised learning, Theorem 3.1 bounds the expected risk of a Gibbs posterior q which, upon receiving a previously unseen datapoint $\mathbf{x} \sim \mu$, samples a predictor h *dependent* on $\mathbf{x}$, and uses it to make a prediction. Note that the family $\mathcal{E}$ from Assumption 1 does not appear in the bound, which has nice consequences in practice. Indeed one may pick a loss function ℓ that fits the problem, and then find a family $\mathcal{E}$ for which the continuity assumption is satisfied with constant K that is as small as possible.

Note also that, in the tradition of PAC-Bayesian bounds, Theorem 3.1 does not make any assumptions on the nature of the elements of $\mathcal{H}$ (e.g. $\mathcal{H}$ could be a class of functions, a set of neural network's parameters, etc). Therefore, the theorem is very general and could be applied to different domains and models. In the following sections, we will use a specific kind of hypothesis class $\mathcal{H} = \mathcal{Z}$, in order to capture the VAE's latent space.

4 Generalization bounds for the Reconstruction Loss

For the remainder of this work, $\|\cdot\|$ denotes the L_2 norm, and we assume the instance space $\mathcal{X}$ is Euclidean, and the latent space $\mathcal{Z} = \mathbb{R}^{d_\mathcal{Z}}$, where $d_\mathcal{Z} > 0$. Both $\mathcal{X}$ and $\mathcal{Z}$ are equipped with the Euclidean distance as the underlying metric. Therefore, if $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, $d(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|$.

The following assumption states that the encoder and decoder networks have finite Lipschitz norms.

Assumption 2. The encoder and decoder are Lipschitz-continuous w.r.t. their inputs, meaning there exist real numbers $K_\phi, K_\theta > 0$ such that for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}$,

$$\|Q_\phi(\mathbf{x}_1) - Q_\phi(\mathbf{x}_2)\| \leq K_\phi \|\mathbf{x}_1 - \mathbf{x}_2\| \tag{4}$$

and

$$\|g_\theta(\mathbf{z}_1) - g_\theta(\mathbf{z}_2)\| \leq K_\theta \|\mathbf{z}_1 - \mathbf{z}_2\| . \tag{5}$$

Recall the definition of Q_ϕ from Equation (2). Note that in practice, one can estimate the Lipschitz constant of trained networks (Fazlyab et al., 2019; Latorre et al., 2020) or train the VAE with preset Lipschitz constants (Barrett et al., 2022).

Moreover, we define the reconstruction loss ℓ^θ_{rec} with the L_2 norm, instead of the squared L_2 norm, which enables us to exploit the properties of a metric. We discuss this choice in Section 6. In order to be consistent with the PAC-Bayesian framework, we define the loss function as follows: $\ell^\theta_{\text{rec}} : \mathcal{Z} \times \mathcal{X} \to [0, \infty)$,

$$\ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x}) = \|\mathbf{x} - g_\theta(\mathbf{z})\| . \tag{6}$$

Our goal is to apply the general bound of Theorem 3.1 to the VAE model. But first, since Theorem 3.1 requires Assumption 1 to be satisfied, we start by showing that if the encoder and decoder networks have finite Lipschitz norms, then Assumption 1 holds.

Proposition 4.1. *Consider a VAE with parameters ϕ and θ and let $K_\phi, K_\theta \in \mathbb{R}$ be the Lipschitz norms of the encoder and decoder respectively. Then the variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$ satisfies Assumption 1, with $\mathcal{E} = \{f : \mathcal{Z} \to \mathbb{R} \ s.t. \ \|f\|_{Lip} \leq K_\theta\}$, $\ell = \ell^\theta_{rec}$, and $K = K_\phi K_\theta$.*

Proof idea. The proof of Proposition 4.1 is in Appendix C, we provide a brief summary here. To prove the first part of Assumption 1, we first notice that if $\mathcal{E}$ is the set of real-valued K_θ-Lipschitz continuous functions, then $d_\mathcal{E}$ is a scaling of the Wasserstein distance. In addition, since $W_1 \leq W_2$, using the closed form of the Wasserstein-2 distance between Gaussian distributions, one can show that $d_\mathcal{E}(q_\phi(\cdot|\mathbf{x}_1), q_\phi(\cdot|\mathbf{x}_2)) \leq K_\phi K_\theta \|\mathbf{x}_1 - \mathbf{x}_2\|$. Finally, the second part of the assumption is a consequence of the definition of the loss function and the Lipschitz continuity of the decoder. □

Proposition 4.1 tells us that Assumption 1 holds for VAEs. Consequently, we can utilize our general bound of Theorem 3.1 to obtain generalization guarantees. This leads to the following general PAC-Bayesian bound for the VAE's reconstruction loss.

Theorem 4.2. *Let $\mathcal{X}$ be the instance space, $\mu \in \mathcal{M}^1_+(\mathcal{X})$ the data-generating distribution, $\mathcal{Z}$ the latent space, $p(\mathbf{z}) \in \mathcal{M}^1_+(\mathcal{Z})$ the prior distribution on the latent space, θ the decoder's parameters,*

$\delta \in (0,1), \lambda > 0$ be real numbers. With probability at least $1-\delta$ over the random draw of $S \sim \mu^{\otimes n}$, the following holds for any posterior $q_\phi(\mathbf{z}|\mathbf{x})$:

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x})} \ell^\theta_{rec}(\mathbf{z},\mathbf{x}) \leq \frac{1}{n}\sum_{i=1}^n \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \ell^\theta_{rec}(\mathbf{z},\mathbf{x}_i) + \frac{1}{\lambda}\left[\sum_{i=1}^n \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i)\,||\,p(\mathbf{z})) + \right.$$
$$\left. \frac{\lambda K_\phi K_\theta}{n}\sum_{i=1}^n \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} d(\mathbf{x},\mathbf{x}_i) + \log\tfrac{1}{\delta} + \log \mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} e^{\lambda\left(\mathbb{E}_{\mathbf{x}\sim\mu}\left[\ell^\theta_{rec}(\mathbf{z},\mathbf{x})\right] - \frac{1}{n}\sum_{i=1}^n \ell^\theta_{rec}(\mathbf{z},\mathbf{x}_i)\right)}\right],$$

where K_ϕ and K_θ are the Lipschitz norms of the encoder and the decoder (see (4) and (5)) and $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}$ is a shorthand for $\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}$.

Note that the choice of the hyperparameter β in the VAE's optimization objective (3) correlates with the choice of the hyperparameter λ in Theorem 4.2 (e.g. $\lambda = n$ corresponds to $\beta = 1$). Note also that the encoder and decoder are not treated the same way in Theorem 4.2. Indeed, the inequality holds for a given decoder, but uniformly for any encoder. We discuss this subtle difference and its practical consequences in Section 6.

Theorem 4.2 can be seen as a general framework. In order to obtain a useful upper bound, one needs to bound the average distance and the exponential moment on the right-hand side. In the sections below, we provide upper bounds for these terms under various assumptions on the instance space.

4.1 Reconstruction Guarantees for Bounded Instance Spaces

In the following theorem, we provide a special case of Theorem 4.2 when the instance space's diameter $\Delta \stackrel{\text{def}}{=} \sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} d(\mathbf{x},\mathbf{x}')$ is finite (see Section C.2 for the proof).

Theorem 4.3. *Let $\mathcal{X}$ be the instance space, $\Delta < \infty$ its diameter, $\mu \in \mathcal{M}^1_+(\mathcal{X})$ the data-generating distribution, $\mathcal{Z}$ the latent space, $p(\mathbf{z}) \in \mathcal{M}^1_+(\mathcal{Z})$ the prior on the latent space, θ the decoder's parameters, $\delta \in (0,1), \lambda > 0$ be real numbers. With probability at least $1-\delta$ over the random draw of $S \sim \mu^{\otimes n}$, the following holds for any posterior $q_\phi(\mathbf{z}|\mathbf{x})$:*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x})} \ell^\theta_{rec}(\mathbf{z},\mathbf{x}) \leq \frac{1}{n}\sum_{i=1}^n \left\{ \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \ell^\theta_{rec}(\mathbf{z},\mathbf{x}_i)\right\} + \frac{1}{\lambda}\left(\sum_{i=1}^n \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i)\,||\,p(\mathbf{z})) + \right.$$
$$\left. \lambda K_\phi K_\theta \Delta + \log\frac{1}{\delta} + \frac{\lambda^2\Delta^2}{8n}\right).$$

The left-hand side of this inequality is the expected reconstruction loss for samples $\mathbf{x} \sim \mu$, while the right-hand side is the empirical reconstruction and KL losses, plus an additional term depending on the Lipschitz constants of the VAE and the model's diameter.

Note that for real-life datasets, the diameter of the instance space might be very large and non-representative of the structure and complexity of the data. Indeed, it is common to scale image datasets in order to utilize a specific architecture (Radford et al., 2016). In the following section, we provide a special case of Theorem 4.2 under the manifold hypothesis on the data-generating process.

4.2 Reconstruction Guarantees under the Manifold Assumption

The manifold assumption (Fodor, 2002; Narayanan and Mitter, 2010; Fefferman et al., 2016) states that most high-dimensional datasets encountered in practice lie close to low-dimensional manifolds. This assumption is exploited by latent variable generative models such as GANs and VAEs, which approximate high-dimensional datasets using transformations of distributions on a low-dimensional space. The works of Schreuder et al. (2021) and Mbacke et al. (2023) provide generalization bounds for GANs, by assuming that the data-generating distribution is a smooth transformation of the uniform distribution on $[0,1]^{d^*}$, where d^* is the intrinsic dimension. However, since the standard VAE calls for a standard Gaussian prior, in the following theorem, we assume μ is a smooth transformation of the standard Gaussian distribution p^* on $\mathbb{R}^{d^*}$. We consider the case when p^* is the uniform distribution on $[0,1]^{d^*}$ in the supplementary material.

Theorem 4.4. *Let $\mathcal{X}$ be the instance space, $\mu \in \mathcal{M}_+^1(\mathcal{X})$ the data-generating distribution, $\mathcal{Z}$ the latent space, $p(\mathbf{z}) \in \mathcal{M}_+^1(\mathcal{Z})$ the prior distribution on the latent space, θ the decoder's parameters, $\delta \in (0,1), \lambda > 0, a > 0$ real numbers. Assume the data-generating distribution $\mu = g^* \sharp p^*$, where p^* is the standard Gaussian distribution on $\mathbb{R}^{d^*}$ and $g^* \in \mathrm{Lip}_{K_*}(\mathbb{R}^{d^*}, \mathcal{X})$. With probability at least $1 - \delta - \frac{nd^*}{2} e^{-a^2/2}$ over the random draw of $S \sim \mu^{\otimes n}$, the following holds for any posterior $q_\phi(\mathbf{z}|\mathbf{x})$:*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x})} \ell_{rec}^\theta(\mathbf{z},\mathbf{x}) \leq \frac{1}{n}\sum_{i=1}^n \left\{ \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \ell_{rec}^\theta(\mathbf{z},\mathbf{x}_i) \right\} + \frac{1}{\lambda}\left(\sum_{i=1}^n \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i) \,||\, p(\mathbf{z})) + \lambda K_\phi K_\theta K_* \sqrt{(1+a^2)d^*} + \log\frac{1}{\delta} + \frac{\lambda^2 K_*^2}{2n} \right).$$

Let us clarify the role of the new parameter $a > 0$. Each training sample $\mathbf{x}_i \in S$ can be expressed as $\mathbf{x}_i = g^*(\mathbf{w}_i)$, where $\mathbf{w}_i \sim p^*$. Since p^* is the standard Gaussian distribution on $\mathbb{R}^{d^*}$, all samples $\mathbf{w}_i$ will be inside a hypercube $[-a,a]^{d^*}$, with high probability. This uncertainty is reflected in the lowered confidence (from $1-\delta$ in Theorem 4.2 to $1 - \delta - \frac{nd^*}{2} e^{-a^2/2}$ in Theorem 4.4), and can be controlled by choosing a large enough value of a. The proof of Theorem 4.4 is in the supplementary material (Section C.3), we provide a short summary below.

Proof idea. The proof starts with Theorem 4.2, and uses the assumptions of Theorem 4.4 to obtain upper bounds on the exponential moment and the average distance. To derive the upper bound on the exponential moment, we observe that the function $\mathbf{z} \mapsto \ell_{\text{rec}}^\theta(\mathbf{z},\mathbf{x})$ is K_*-Lipschitz continuous, then we use a dimension-free upper bound on the MGF of Lipschitz-continuous functions of Gaussian random variables. Furthermore, we obtain the upper bound on the average distance $\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbf{x}\sim\mu} \|\mathbf{x} - \mathbf{x}_i\|$, by using Holder's inequality and the expectation of a non-central χ^2 distribution. Then, we upper-bound the probability that $\mathbf{w}_i \in [-a,a]^{d^*}$ for all $1 \leq i \leq n$ using the error function and Bernoulli's inequality. Finally, we use the union bound to update the overall confidence. □

5 Generalization Bounds for Regeneration and Generation

Let $\hat{\mu}_{\phi,\theta}$ be the *empirical regenerated distribution*, meaning

$$\hat{\mu}_{\phi,\theta} = \frac{1}{n}\sum_{i=1}^n g_\theta \sharp q_\phi(\mathbf{z}|\mathbf{x}_i). \tag{7}$$

In other words, sampling $\mathbf{x} \sim \hat{\mu}_{\phi,\theta}$ is done by sampling $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}_i)$ where i is uniformly sampled from $\{1,\ldots,n\}$, then passing $\mathbf{z}$ through the decoder: $\mathbf{x} = g_\theta(\mathbf{z})$. It is therefore the distribution regenerated by the VAE, given the training set $S = \{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ as input.

In this section, we provide statistical guarantees on the regenerative and generative properties of VAEs. More precisely, we derive upper bounds for the quantities $W_1(\mu, \hat{\mu}_{\phi,\theta})$ and $W_1(\mu, g_\theta \sharp p(\mathbf{z}))$. Note that the average distance term does not appear in the bounds of this section. This is because instead of relying on Theorem 3.1, the results of this section depend upon a preliminary lemma (Lemma B.1), which does not necessitate Assumption 1.

5.1 Regeneration and Generation Guarantees for Bounded Instance Spaces

The following theorem presents our first upper bound on the distance between the input distribution and the empirical regenerated distribution.

Theorem 5.1. *Under the definitions and assumptions of Theorem 4.3, we have that with probability at least $1-\delta$ over the random draw of $S \sim \mu^{\otimes n}$, the following holds for any posterior $q_\phi(\mathbf{z}|\mathbf{x})$:*

$$W_1(\mu, \hat{\mu}_{\phi,\theta}) \leq \frac{1}{n}\sum_{i=1}^n \left\{ \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \ell_{rec}^\theta(\mathbf{z},\mathbf{x}_i) \right\} + \frac{1}{\lambda}\left(\sum_{i=1}^n \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i) \,||\, p(\mathbf{z})) + \log\frac{1}{\delta} + \frac{\lambda^2 \Delta^2}{8n} \right).$$

As we can see, the right-hand side of Theorem 5.1 depends on the empirical reconstruction loss and KL-divergence. This guarantees that as the VAE's empirical risk decreases, the regenerated distribution gets closer to the data-generating distribution. The proof of Theorem 5.1 exploits the fact

that the underlying metric on $\mathcal{X}$ is the Euclidean distance $d(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|$, which is also used to define the reconstruction loss $\ell^{\theta}_{\text{rec}}$ (see Equation 6). The full proof can be found in Appendix D.

The following theorem provides an upper bound of the distance between the input distribution and the VAE's generated distribution.

Theorem 5.2. *Under the definitions and assumptions of Theorem 4.3, we have that with probability at least* $1 - \delta$ *over the random draw of* $S \sim \mu^{\otimes n}$*, the following holds for any posterior* $q_\phi(\mathbf{z}|\mathbf{x})$*:*

$$W_1(\mu, g_\theta \sharp p(\mathbf{z})) \leq \frac{1}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \ell^{\theta}_{rec}(\mathbf{z}, \mathbf{x}_i)\right\} + \frac{1}{\lambda}\left(\sum_{i=1}^{n}\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i)\,||\,p(\mathbf{z})) + \log\frac{1}{\delta} + \frac{\lambda^2\Delta^2}{8n}\right) + \frac{K_\theta}{n}\sum_{i=1}^{n}\sqrt{\|\mu_\phi(\mathbf{x}_i)\|^2 + \left\|\sigma_\phi(\mathbf{x}_i) - \vec{1}\right\|^2},$$

where $\vec{1} \in \mathbb{R}^{d_Z}$ *denotes the vector whose entries are all* 1*.*

The right-hand side of Theorem 5.2 is equal to the right-hand side of Theorem 5.1, plus an additional term depending on the Wasserstein-2 distance $W_2(q_\phi(\mathbf{z}|\mathbf{x}_i), p(\mathbf{z}))$, which is used in the proof because of its closed form for Gaussian distributions. Hence, the right-hand side of Theorem 5.2 augments the VAE's optimization objective with $W_2(q_\phi(\mathbf{z}|\mathbf{x}_i), p(\mathbf{z}))$, suggesting that a good generative performance may require the latent codes to be even closer to the prior. This is consistent with the findings of Zhao et al. (2019), who showed that in order to improve generative performance, the latent codes need to be much closer to the prior, which may disrupt the balance between reconstruction loss and KL-loss.

5.2 Regeneration and Generation Guarantees under the Manifold Assumption

Similar to what we did in Section 4.2, we assume that the data-generating distribution is a smooth transformation of the standard Gaussian distribution on $\mathbb{R}^{d^*}$, where d^* is the intrinsic dimension of the dataset. This yields the following results.

Theorem 5.3. *Under the definitions and assumptions of Theorem 4.4, with probability at least* $1 - \delta$ *over the random draw of* $S \sim \mu^{\otimes n}$*, the following holds for any posterior* $q_\phi(\mathbf{z}|\mathbf{x})$*:*

$$W_1(\mu, \hat{\mu}_{\phi,\theta}) \leq \frac{1}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \ell^{\theta}_{rec}(\mathbf{z}, \mathbf{x}_i)\right\} + \frac{1}{\lambda}\left(\sum_{i=1}^{n}\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i)\,||\,p(\mathbf{z})) + \log\frac{1}{\delta} + \frac{\lambda^2 K_*^2}{2n}\right).$$

Note that the intrinsic and extrinsic dimensions do not explicitly appear in this inequality, although they may affect the reconstruction and KL loss.

We now present our last result, an upper bound on the Wasserstein distance between the input distribution and the VAE's generated distribution, under the manifold assumption.

Theorem 5.4. *Under the definitions and assumptions of Theorem 4.4, with probability at least* $1 - \delta$ *over the random draw of* $S \sim \mu^{\otimes n}$*, the following holds for any posterior* $q_\phi(\mathbf{z}|\mathbf{x})$ *:*

$$W_1(\mu, g_\theta \sharp p(\mathbf{z})) \leq \frac{1}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \ell^{\theta}_{rec}(\mathbf{z}, \mathbf{x}_i)\right\} + \frac{1}{\lambda}\left(\sum_{i=1}^{n}\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i)\,||\,p(\mathbf{z})) + \log\frac{1}{\delta} + \frac{\lambda^2 K_*^2}{2n}\right) + \frac{K_\theta}{n}\sum_{i=1}^{n}\sqrt{\|\mu_\phi(\mathbf{x}_i)\|^2 + \left\|\sigma_\phi(\mathbf{x}_i) - \vec{1}\right\|^2},$$

where $\vec{1} \in \mathbb{R}^{d_Z}$ *denotes the vector whose entries are all* 1*.*

Theorem 5.2 and Theorem 5.4 show that by minimizing the VAE's objective, one is also minimizing the Wasserstein distance between the input distribution and the VAE's generated distribution.

From the upper bounds given by Theorems 5.1 and 5.3, one can deduce rates of convergence of $O(n^{-1/2})$ (when $\lambda \approx \sqrt{n}$) for the empirical regenerated distribution. Note that $\lambda \approx n$ leads to the much faster rate of n^{-1}, but then the bounds do not converge to the empirical risk, but to a larger positive number, dependent on the input distribution. Similarly, Theorems 5.2 and 5.4 provide rates of convergence of $O(n^{-1/2})$ for the VAE's generated distribution.

6 Discussion and Conclusion

The different treatments of θ and ϕ. The bounds we've presented in this work hold for a given decoder θ, but uniformly for all encoders. In practice, this means that the risk certificate has to be computed using samples different from the ones used to train the VAE. This is different from the usual PAC-Bayesian trick (Germain et al., 2009; Parrado-Hernández et al., 2012; Pérez-Ortiz et al., 2021, see also Remark F.2) of splitting the training set to learn the prior, then training the model on the whole training set, because the decoder and encoder are jointly optimized. Instead, one has to make sure that the model is only trained on samples distinct from the ones used to compute the bound. The same method would be necessary when computing the risk certificates given by the recent PAC-Bayesian bounds of Rivasplata et al. (2020) and Haddouche and Guedj (2022), since those bounds are not uniformly valid for any posterior.

The reconstruction loss. In our bounds, the reconstruction loss is the L_2 norm (RMSE), instead of the squared L_2 norm (MSE). In practice, one can still optimize a VAE with the MSE (or any other reconstruction loss, e.g. the cross entropy loss), and then compute the bounds using the RMSE. However, if the reconstruction loss is not the RMSE, then the optima of the chosen optimization objective might differ from the ones minimizing the right-hand side of the bounds. Therefore, if the goal is to minimize the bounds, one should utilize the RMSE as the reconstruction loss.

Conclusion. It is common, when applying PAC-Bayesian theory to new problems, to add additional stochasticity in order to account for the PAC-Bayesian distributions on the hypothesis class. For instance, Mbacke et al. (2023) added distributions on the parameters of a WGAN's generator, in order to perform a PAC-Bayesian analysis. However, because of the seamless integration of the PAC-Bayesian and VAE frameworks, such modification to the original problem has been avoided in this work. We matched the prior and posterior distributions on the VAE's latent space to the PAC-Bayesian prior and posterior, which allowed us to recover the VAE's optimization objective. We provide preliminary experiments on synthetic datasets in the supplementary material.

This work is a humble contribution to the theoretical understanding of VAEs. We developed novel PAC-Bayesian bounds suited to the analysis of VAEs and provided generalizations bounds for the VAE's reconstruction loss. In addition, we also derived upper bounds on the Wasserstein distance between the input distribution and the VAE's generative model's distribution. These bounds depend on the VAE's empirical optimization objective and the data-generating process. By integrating the VAE and PAC-Bayesian frameworks, we hope to establish PAC-Bayesian theory as a prime tool for the theoretical analysis of VAEs.

Acknowledgements

This research is supported by the Canada CIFAR AI Chair Program, and the NSERC Discovery grant RGPIN-2020- 07223. F. Clerc is funded by IVADO through the DEEL Project CRDPJ 537462 18 and by a grant from NSERC.

References

Alquier, P. (2021). User-friendly introduction to PAC-Bayes bounds. *arXiv preprint arXiv:2110.11216*.

Amit, R. and Meir, R. (2018). Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *International Conference on Machine Learning*, pages 205–214. PMLR.

An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18.

Anil, C., Lucas, J., and Grosse, R. (2019). Sorting out Lipschitz function approximation. In *International Conference on Machine Learning*, pages 291–301. PMLR.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. (2017). Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning*, pages 224–232. PMLR.

Barrett, B., Camuto, A., Willetts, M., and Rainforth, T. (2022). Certifiably robust variational autoencoders. In *International Conference on Artificial Intelligence and Statistics*, pages 3663–3683. PMLR.

Biau, G., Sangnier, M., and Tanielian, U. (2021). Some theoretical insights into Wasserstein GANs. *Journal of Machine Learning Research*.

Björck, Å. and Bowie, C. (1971). An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Józefowicz, R., and Bengio, S. (2016). Generating sentences from a continuous space. In *Conference on Computational Natural Language Learning, CoNLL*, pages 10–21. ACL.

Catoni, O. (2003). A PAC-Bayesian approach to adaptive classification. *preprint LPMA*, 840.

Chakrabarty, A. and Das, S. (2021). Statistical regeneration guarantees of the Wasserstein autoencoder with latent space consistency. In *Advances in Neural Information Processing Systems*.

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

Chu, J. T. (1955). On bounds for the normal integral. *Biometrika*, 42(1/2):263–265.

Donsker, M. D. and Varadhan, S. S. (1976). Asymptotic evaluation of certain markov process expectations for large time—iii. *Communications on pure and applied Mathematics*, 29(4):389–461.

Dziugaite, G. K. and Roy, D. M. (2018). Data-dependent PAC-Bayes priors via differential privacy. In *Advances in Neural Information Processing Systems*, volume 31.

Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G. (2019). Efficient and accurate estimation of Lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32.

Fefferman, C., Mitter, S., and Narayanan, H. (2016). Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049.

Fodor, I. K. (2002). A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Lab., CA (US).

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. (2016). PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems*, volume 29.

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. (2020). PAC-Bayes and domain adaptation. *Neurocomputing*, 379:379–397.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. (2009). PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learning*, page 353–360.

Givens, C. R. and Shortt, R. M. (1984). A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240.

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

Grünwald, P. (2012). The safe Bayesian - learning the learning rate via the mixability gap. In *Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Computer Science*, pages 169–183. Springer.

Guedj, B. (2019). A primer on PAC-Bayesian learning. In *Proceedings of the French Mathematical Society*, volume 33, pages 391–414. Société Mathématique de France.

Haddouche, M. and Guedj, B. (2022). Online PAC-Bayes learning. In *Advances in Neural Information Processing Systems*.

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

Kaur, D., Islam, S. N., and Mahmud, M. A. (2021). 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.

Kingma, D. P., Mohamed, S., Jimenez Rezende, D., and Welling, M. (2014). Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, volume 27.

Kingma, D. P. and Welling, M. (2014). Auto-encoding Variational Bayes. In *International Conference on Learning Representations*.

Langford, J. and Caruana, R. (2001). (not) bounding the true error. In *Advances in Neural Information Processing Systems*, volume 14.

Latorre, F., Rolland, P., and Cevher, V. (2020). Lipschitz constant estimation of neural networks via sparse polynomial optimization. In *International Conference on Learning Representations*.

Liang, T. (2021). How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(228):1–41.

Mbacke, S. D., Clerc, F., and Germain, P. (2023). PAC-Bayesian generalization bounds for adversarial generative models. In *International Conference on Machine Learning*, volume 202, pages 24271–24290. PMLR.

McAllester, D. A. (1999). Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363.

McAllester, D. A. (2003). PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21.

Mitrinovic, D. S. and Vasic, P. M. (1970). *Analytic inequalities*, volume 1. Springer.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443.

Narayanan, H. and Mitter, S. (2010). Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems*, volume 23.

Parrado-Hernández, E., Ambroladze, A., Shawe-Taylor, J., and Sun, S. (2012). PAC-Bayes bounds with data dependent priors. *Journal of Machine Learning Research*, 13(1):3507–3531.

Pérez-Ortiz, M., Rivasplata, O., Shawe-Taylor, J., and Szepesvári, C. (2021). Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22.

Petersen, K. B. and Pedersen, M. S. (2008). The matrix cookbook. Version 20081110.

Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*.

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

Rivasplata, O., Kuzborskij, I., Szepesvári, C., and Shawe-Taylor, J. (2020). PAC-Bayes analysis beyond the usual bounds. In *Advances in Neural Information Processing Systems*, volume 33, pages 16833–16845.

Schreuder, N., Brunel, V.-E., and Dalalyan, A. (2021). Statistical guarantees for generative models without domination. In *Algorithmic Learning Theory*, pages 1051–1071. PMLR.

Seeger, M. (2002). PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3(Oct):233–269.

Seldin, Y. and Tishby, N. (2010). PAC-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11(12).

Singh, S., Uppal, A., Li, B., Li, C.-L., Zaheer, M., and Póczos, B. (2018). Nonparametric density estimation under adversarial losses. In *Advances in Neural Information Processing Systems*, volume 31.

Thiemann, N., Igel, C., Wintenberger, O., and Seldin, Y. (2017). A strongly quasiconvex PAC-Bayesian bound. In *International Conference on Algorithmic Learning Theory*, volume 76, pages 466–492. PMLR.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. (2018). Wasserstein auto-encoders. In *International Conference on Learning Representations*.

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

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Zhang, P., Liu, Q., Zhou, D., Xu, T., and He, X. (2018). On the discrimination-generalization tradeoff in GANs. In *International Conference on Learning Representations*.

Zhao, J., Mathieu, M., and LeCun, Y. (2017). Energy-based generative adversarial networks. In *International Conference on Learning Representations*.

Zhao, S., Song, J., and Ermon, S. (2019). InfoVAE: Balancing learning and inference in variational autoencoders. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 5885–5892.

Statistical Guarantees for Variational Autoencoders using PAC-Bayesian Theory: Supplementary Material

Sokhna Diarra Mbacke
Université Laval
sokhna-diarra.mbacke.1@ulaval.ca

Florence Clerc
McGill University
florence.clerc@mail.mcgill.ca

Pascal Germain
Université Laval
pascal.germain@ift.ulaval.ca

A Preliminaries

Definition A.1 (Coupling)**.** Let $p, q \in \mathcal{M}_+^1(\mathcal{X})$. A distribution γ on $\mathcal{X} \times \mathcal{X}$ is a coupling (Villani, 2009) of p and q if for every measurable set $B \subset \mathcal{X}$, $\gamma(B \times \mathcal{X}) = p(B)$ and $\gamma(\mathcal{X} \times B) = q(B)$. In other words, a coupling of p and q is a distribution on $\mathcal{X} \times \mathcal{X}$ whose marginals are p and q respectively.

For example, the product measure $p \otimes q$ is a coupling of p and q.

Definition A.2 (Wasserstein distances)**.** Let $(\mathcal{X}, d)$ be a Polish metric space and $p, q \in \mathcal{M}_+^1(\mathcal{X})$ Given a real number $k \geq 1$, the Wasserstein-k distance W_k is defined as

$$W_k(p, q) = \left(\inf_{\pi \in \Gamma(p,q)} \int d(\mathbf{x}, \mathbf{y})^k \, d\pi(\mathbf{x}, \mathbf{y}) \right)^{1/k},$$

where $\Gamma(p, q)$ denotes the set of couplings of p and q (see Definition A.1 above). As stated in the main paper, W_1 is referred to as the Wasserstein distance.

Given two Gaussian distributions $p = \mathcal{N}(\mu_1, \Sigma_1)$ and $q = \mathcal{N}(\mu_2, \Sigma_2)$ on $\mathbb{R}^{d^*}$, the Wasserstein-2 distance has the following closed form (Givens and Shortt, 1984):

$$W_2(p, q)^2 = \|\mu_1 - \mu_2\|^2 + \operatorname{Tr}\left(\Sigma_1 + \Sigma_2 - 2 \left(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2} \right)^{1/2} \right). \tag{A.1}$$

This expression can be greatly simplified when the distributions have diagonal covariance matrices. Indeed, if $\Sigma_1 = \operatorname{diag}(\sigma_1^2)$ and $\Sigma_2 = \operatorname{diag}(\sigma_2^2)$ where $\sigma_1, \sigma_2 \in \mathbb{R}^{d^*}$, then the product of the covariance matrices commutes $\Sigma_1 \Sigma_2 = \Sigma_2 \Sigma_1$ and we get

$$\left(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2} \right)^{1/2} = \Sigma_1^{1/2} \Sigma_2^{1/2},$$

which, combined with the symmetry of covariance matrices and the definition of the Frobenius norm $\|\cdot\|_{\text{Fr}}$ (Petersen and Pedersen, 2008), implies

$$\operatorname{Tr}\left(\Sigma_1 + \Sigma_2 - 2 \left(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2} \right)^{1/2} \right) = \left\| \Sigma_1^{1/2} - \Sigma_2^{1/2} \right\|_{\text{Fr}}^2 = \|\sigma_1 - \sigma_2\|^2.$$

Hence, if $p = \mathcal{N}(\mu_1, \operatorname{diag}(\sigma_1^2))$ and $q = \mathcal{N}(\mu_2, \operatorname{diag}(\sigma_2^2))$, then the Wasserstein-2 distance between p and q is

$$W_2(p, q) = \|\mu_1 - \mu_2\|^2 + \|\sigma_1 - \sigma_2\|^2. \tag{A.2}$$

We will use this equality to prove some of the results of Section 5.

The following change of measure theorem dates back to Donsker and Varadhan (1976) and has been used in the proof of many PAC-Bayesian theorems. A proof can be found in Boucheron et al. (2013, Corollary 4.15).

Proposition A.1 (Donsker-Varadhan change of measure)**.** *Let p, q be probability measures on a space $\mathcal{H}$ such that $q \ll p$, and let $g : \mathcal{H} \to \mathbb{R}$ be a function such that $\mathbb{E}_{h\sim p}\, e^{g(h)} < \infty$. Then,*

$$\mathop{\mathbb{E}}_{h\sim p} e^{g(h)} \geq e^{\mathbb{E}_{h\sim q}[g(h)] - \mathrm{KL}(q\,||\,p)}.$$

There are many different formulations of this proposition, we chose a formulation that facilitates readability of the proof of the following lemma.

B Proofs of the results in Section 3

We state and prove our first result. Note that the following lemma does not use Assumption 1. Moreover, the main difference between the inequality of this lemma and the one of Theorem 3.1 is the left-hand side. In Lemma B.1, the expected loss for samples $\mathbf{x} \sim \mu$ is computed w.r.t. distributions $q(h|\mathbf{x}_i)$ associated to the training samples. In contrast, in Theorem 3.1, the expected loss for each $\mathbf{x} \sim \mu$ is computed w.r.t. the distribution $q(h|\mathbf{x})$ associated to $\mathbf{x}$ itself.

Lemma B.1. *Let $\mathcal{X}$ be the instance space, $\mu \in \mathcal{M}_+^1(\mathcal{X})$ the data-generating distribution, $\mathcal{H}$ the hypothesis class, $\ell : \mathcal{H} \times \mathcal{X} \to \mathbb{R}$ the loss function, $p(h) \in \mathcal{M}_+^1(\mathcal{H})$ the prior distribution and $\delta \in (0,1), \lambda > 0$ real numbers. Then with probability at least $1-\delta$ over the random draw of the training set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim \mu^{\otimes n}$, the following holds for any conditional posterior $q(h|\mathbf{x}) \in \mathcal{M}_+^1(\mathcal{H})$:*

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x}_i)}\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\ell(h,\mathbf{x})\right\} \leq \frac{1}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x}_i)}\ell(h,\mathbf{x}_i)\right\} + \frac{1}{\lambda}\left[\sum_{i=1}^{n}\mathrm{KL}(q(h|\mathbf{x}_i)\,||\,p(h)) + \log\frac{1}{\delta} + \log\mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}}\mathop{\mathbb{E}}_{h\sim p(h)} e^{\lambda\left(\mathbb{E}_{\mathbf{x}\sim\mu}[\ell(h,\mathbf{x})] - \frac{1}{n}\sum_{i=1}^{n}\ell(h,\mathbf{x}_i)\right)}\right]. \tag{B.1}$$

Proof. First, we consider a set $H = \{h_1, \ldots, h_n\} \sim p(h)^{\otimes n}$ iid sampled from $p(h)$. By applying Markov's inequality to the positive random variable Y, defined as

$$Y \overset{\text{def}}{=} \mathop{\mathbb{E}}_{H\sim p(h)^{\otimes n}} \exp\left[\frac{\lambda}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}[\ell(h_i,\mathbf{x})] - \ell(h_i,\mathbf{x}_i)\right\}\right],$$

we obtain that with probability at least $1-\delta$ over the draw of $S \sim \mu^{\otimes n}$, $Y \leq \frac{1}{\delta}\,\mathbb{E}\,[Y]$, meaning

$$\mathop{\mathbb{E}}_{H\sim p(h)^{\otimes n}} \exp\left[\frac{\lambda}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}[\ell(h_i,\mathbf{x})] - \ell(h_i,\mathbf{x}_i)\right\}\right] \leq \frac{1}{\delta}\mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}}\mathop{\mathbb{E}}_{H\sim p(h)^{\otimes n}} \exp\left[\frac{\lambda}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}[\ell(h_i,\mathbf{x})] - \ell(h_i,\mathbf{x}_i)\right\}\right]. \tag{B.2}$$

Let us focus on the left-hand side of (B.2). We have

$$\mathop{\mathbb{E}}_{H\sim p(h)^{\otimes n}} \exp\left[\frac{\lambda}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h_i,\mathbf{x})\right]-\ell(h_i,\mathbf{x}_i)\right\}\right]$$

$$=\mathop{\mathbb{E}}_{H\sim p(h)^{\otimes n}}\prod_{i=1}^{n}\exp\left[\frac{\lambda}{n}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h_i,\mathbf{x})\right]-\ell(h_i,\mathbf{x}_i)\right)\right]$$

$$=\prod_{i=1}^{n}\mathop{\mathbb{E}}_{h_i\sim p(h)}\exp\left[\frac{\lambda}{n}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h_i,\mathbf{x})\right]-\ell(h_i,\mathbf{x}_i)\right)\right]$$

$$=\prod_{i=1}^{n}\mathop{\mathbb{E}}_{h\sim p(h)}\exp\left[\frac{\lambda}{n}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h,\mathbf{x})\right]-\ell(h,\mathbf{x}_i)\right)\right]$$

$$\geq\prod_{i=1}^{n}\exp\left[\mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x}_i)}\left[\frac{\lambda}{n}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h,\mathbf{x})\right]-\ell(h,\mathbf{x}_i)\right)\right]-\mathrm{KL}(q(h|\mathbf{x}_i)\,||\,p(h))\right],$$

where the inequality uses the Donsker-Varadhan change of measure theorem (Proposition A.1). Applying the logarithm, we obtain

$$\log\mathop{\mathbb{E}}_{H\sim p(h)^{\otimes n}}\exp\left[\frac{\lambda}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h_i,\mathbf{x})\right]-\ell(h_i,\mathbf{x}_i)\right\}\right]$$

$$\geq\log\prod_{i=1}^{n}\exp\left[\mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x}_i)}\left[\frac{\lambda}{n}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h,\mathbf{x})\right]-\ell(h,\mathbf{x}_i)\right)\right]-\mathrm{KL}(q(h|\mathbf{x}_i)\,||\,p(h))\right]$$

$$=\sum_{i=1}^{n}\left(\mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x}_i)}\left[\frac{\lambda}{n}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h,\mathbf{x})\right]-\ell(h,\mathbf{x}_i)\right)\right]-\mathrm{KL}(q(h|\mathbf{x}_i)\,||\,p(h))\right)$$

$$=\frac{\lambda}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x}_i)}\left[\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h,\mathbf{x})\right]-\ell(h,\mathbf{x}_i)\right]-\sum_{i=1}^{n}\mathrm{KL}(q(h|\mathbf{x}_i)\,||\,p(h)).$$

This, combined with (B.2) yields

$$\frac{\lambda}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x}_i)}\left[\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h,\mathbf{x})\right]-\ell(h,\mathbf{x}_i)\right]-\sum_{i=1}^{n}\mathrm{KL}(q(h|\mathbf{x}_i)\,||\,p(h))\leq \frac{1}{\delta}\mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}}\mathop{\mathbb{E}}_{H\sim p(h)^{\otimes n}}\exp\left[\frac{\lambda}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h_i,\mathbf{x})\right]-\ell(h_i,\mathbf{x}_i)\right\}\right]. \tag{B.3}$$

It remains to show that the exponential moment on the right-hand side of Equation (B.3) can be modified by replacing the expectation w.r.t. $p(h)^{\otimes n}$ with an expectation w.r.t. $p(h)$. Similar to what we did in the first part of the first derivation, we can use Fubini's theorem to obtain

$$\mathop{\mathbb{E}}_{H\sim p(h)^{\otimes n}}\exp\left[\frac{\lambda}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h_i,\mathbf{x})\right]-\ell(h_i,\mathbf{x}_i)\right\}\right]$$

$$=\prod_{i=1}^{n}\mathop{\mathbb{E}}_{h\sim p(h)}\exp\left[\frac{\lambda}{n}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h,\mathbf{x})\right]-\ell(h,\mathbf{x}_i)\right)\right]$$

$$=\mathop{\mathbb{E}}_{h\sim p(h)}\prod_{i=1}^{n}\exp\left[\frac{\lambda}{n}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h,\mathbf{x})\right]-\ell(h,\mathbf{x}_i)\right)\right]$$

$$=\mathop{\mathbb{E}}_{h\sim p(h)}\exp\left[\lambda\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell(h,\mathbf{x})\right]-\frac{1}{n}\sum_{i=1}^{n}\ell(h,\mathbf{x}_i)\right)\right].$$

Combining this equation with Equation (B.3) yields the theorem. □

The reader familiar with PAC-Bayes bounds may notice that the proof of Lemma B.1 is similar to the usual derivation of PAC-Bayesian bounds, with a key difference. We start with an iid set of n hypotheses sampled from the prior, which allows us to apply the change of measure theorem to n posteriors $q(h|\mathbf{x}_1), \ldots, q(h|\mathbf{x}_n)$. Then, we show that the exponential moment obtained with n hypotheses instead of one is equal to the exponential moment obtained with one hypothesis.

B.1 Proof of Theorem 3.1

The first summand on the left-hand side of Lemma B.1 is the risk on samples $\mathbf{x} \sim \mu$, when the hypotheses are uniformly sampled from $q(h|\mathbf{x}_i), 1 \leq i \leq n$. In order to replace $q(h|\mathbf{x}_i)$ by $q(h|\mathbf{x})$ in that term and derive Theorem 3.1, we utilize Assumption 1.

First, recall that Theorem 3.1 states that under the assumptions of Lemma B.1, if Assumption 1 holds with a constant $K > 0$, then the following inequality holds with probability at least $1 - \delta$:

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x})} \ell(h, \mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x}_i)} \ell(h, \mathbf{x}_i) \leq \frac{1}{\lambda}\left[\sum_{i=1}^{n} \mathrm{KL}(q(h|\mathbf{x}_i)\,||\,p(h)) + \frac{\lambda K}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}[d(\mathbf{x}, \mathbf{x}_i)] + \log\frac{1}{\delta} + \log \mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \mathop{\mathbb{E}}_{h\sim p(h)} e^{\lambda\left(\mathbb{E}_{\mathbf{x}\sim\mu}[\ell(h,\mathbf{x})] - \frac{1}{n}\sum_{i=1}^{n}\ell(h,\mathbf{x}_i)\right)}\right]. \tag{B.4}$$

Proof of Theorem 3.1. Using the definition of an IPM and Assumption 1, for any $\mathbf{x}_i \in S, \mathbf{x} \in \mathcal{X}$, we have

$$\mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x})} \ell(h, \mathbf{x}) - \mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x}_i)} \ell(h, \mathbf{x}) \leq d_{\mathcal{E}}(q(h|\mathbf{x}), q(h|\mathbf{x}_i)) \leq K d(\mathbf{x}, \mathbf{x}_i).$$

Combined with Fubini's theorem, we obtain

$$\sum_{i=1}^{n} \mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x}_i)} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \ell(h, \mathbf{x}) = \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x}_i)} \ell(h, \mathbf{x})\right] \geq \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\mathop{\mathbb{E}}_{h\sim q(h|\mathbf{x})} \ell(h, \mathbf{x}) - K d(\mathbf{x}, \mathbf{x}_i)\right].$$

Combining this with Lemma B.1, yields Theorem 3.1. □

C Proofs of the results in Section 4

C.1 Proof of Proposition 4.1

First, we recall the statement of Proposition 4.1.

Proposition C.1 (Restatement of Proposition 4.1)**.** *If there exists positive real numbers K_ϕ and K_θ such that the encoder and decoder are respectively K_ϕ-Lipschitz and K_θ-Lipschitz continuous, then*

$$d_{\mathcal{E}}\left(q_\phi(\mathbf{z}|\mathbf{x}_1), q_\phi(\mathbf{z}|\mathbf{x}_2)\right) \leq K_\phi K_\theta \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|, \tag{C.1}$$

and

$$\ell(\cdot, \mathbf{x}) \in \mathcal{E}, \quad \textit{for any } \mathbf{x} \in \mathcal{X}. \tag{C.2}$$

where $\mathcal{E} = \mathrm{Lip}_{K_\theta}(\mathcal{Z}, \mathbb{R})$ is the set of real-valued K_θ-Lipschitz continuous functions defined on $\mathcal{Z}$.

Proof.

1. Let us prove (C.1). First, since $q_\phi(\mathbf{z}|\mathbf{x}_i) = \mathcal{N}(\mu_\phi(\mathbf{x}_i), \mathrm{diag}(\sigma_\phi^2(\mathbf{x}_i)))$, by (A.2), the Wasserstein-2 distance $W_2(q_\phi(\mathbf{z}|\mathbf{x}_1), q_\phi(\mathbf{z}|\mathbf{x}_2))$ has the following closed form:

$$W_2(q_\phi(\mathbf{z}|\mathbf{x}_1), q_\phi(\mathbf{z}|\mathbf{x}_2))^2 = \left\|\mu_\phi(\mathbf{x}_1) - \mu_\phi(\mathbf{x}_2)\right\|^2 + \left\|\sigma_\phi(\mathbf{x}_1) - \sigma_\phi(\mathbf{x}_2)\right\|^2,$$

which, combined with the definition $Q_\phi(\mathbf{x}) = \begin{bmatrix}\mu_\phi(\mathbf{x}) \\ \sigma_\phi(\mathbf{x})\end{bmatrix}$, yields

$$\left\|Q_\phi(\mathbf{x}_1) - Q_\phi(\mathbf{x}_2)\right\|^2 = W_2(q_\phi(\mathbf{z}|\mathbf{x}_1), q_\phi(\mathbf{z}|\mathbf{x}_2))^2.$$

Since Q_ϕ is K_ϕ-Lipschitz continuous, we have $\|Q_\phi(\mathbf{x}_1) - Q_\phi(\mathbf{x}_2)\| \leq K_\phi \|\mathbf{x}_1 - \mathbf{x}_2\|$, and

$$W_2(q_\phi(\mathbf{z}|\mathbf{x}_1), q_\phi(\mathbf{z}|\mathbf{x}_2)) \leq K_\phi \|\mathbf{x}_1 - \mathbf{x}_2\|. \tag{C.3}$$

On the other hand, the definition $\mathcal{E} = \mathrm{Lip}_{K_\theta}(\mathcal{Z}, \mathbb{R})$ and the Kantorovich duality imply

$$d_{\mathcal{E}}(q_\phi(\mathbf{z}|\mathbf{x}_1), q_\phi(\mathbf{z}|\mathbf{x}_2)) = K_\theta W_1(q_\phi(\mathbf{z}|\mathbf{x}_1), q_\phi(\mathbf{z}|\mathbf{x}_2)).$$

Since $W_1 \leq W_2$, this equation, combined with (C.3) yields

$$d_{\mathcal{E}}(q_\phi(\mathbf{z}|\mathbf{x}_1), q_\phi(\mathbf{z}|\mathbf{x}_2)) \leq K_\theta K_\phi \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

2. Now, we shall prove (C.2), meaning, we show that $\ell(\cdot, \mathbf{x}) \in \mathrm{Lip}_{K_\theta}(\mathcal{Z}, \mathbb{R})$. Let $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}$. We have

$$\begin{aligned}
\ell(\mathbf{z}_1, \mathbf{x}) - \ell(\mathbf{z}_2, \mathbf{x}) &= \|\mathbf{x} - g_\theta(\mathbf{z}_1)\| - \|\mathbf{x} - g_\theta(\mathbf{z}_2)\| \\
&= \|\mathbf{x} - g_\theta(\mathbf{z}_1) + g_\theta(\mathbf{z}_2) - g_\theta(\mathbf{z}_2)\| - \|\mathbf{x} - g_\theta(\mathbf{z}_2)\| \\
&\leq \|\mathbf{x} - g_\theta(\mathbf{z}_2)\| + \|g_\theta(\mathbf{z}_2) - g_\theta(\mathbf{z}_1)\| - \|\mathbf{x} - g_\theta(\mathbf{z}_2)\| \\
&= \|g_\theta(\mathbf{z}_2) - g_\theta(\mathbf{z}_1)\| \\
&\leq K_\theta \|\mathbf{z}_1 - \mathbf{z}_2\|,
\end{aligned}$$

where the first inequality uses the triangle inequality and the second uses the Lipschitz assumption on g_θ.

□

C.2 Proof of Theorem 4.3

Proof of Theorem 4.3. In order to prove Theorem 4.3, we need to upper bound the average distance and the exponential moment of Theorem 4.2, under the finite diameter assumption:

$$\sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} d(\mathbf{x}, \mathbf{x}') = \Delta < \infty. \tag{C.4}$$

More precisely, we need to prove the two following inequalities.

$$\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mu} d(\mathbf{x}, \mathbf{x}_i) \leq n\Delta \tag{C.5}$$

and

$$\mathop{\mathbb{E}}_{\mathbf{z} \sim p(\mathbf{z})} \mathop{\mathbb{E}}_{S \sim \mu^{\otimes n}} \exp\left[\lambda \left(\mathop{\mathbb{E}}_{\mathbf{x} \sim \mu} \left[\ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x})\right] - \frac{1}{n} \sum_{i=1}^{n} \ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x}_i) \right)\right] \leq \exp\left[\frac{\lambda^2 \Delta^2}{8n}\right]. \tag{C.6}$$

First, (C.5) is a direct consequence of the definition of the diameter Δ.

Now, let us prove (C.6). Let $\mathbf{z} \in \mathcal{Z}$. Since $\ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x}) = \|\mathbf{x} - g_\theta(\mathbf{z})\| = d(\mathbf{x}, g_\theta(\mathbf{z}))$ is the distance between $\mathbf{x}$ and $g_\theta(\mathbf{z})$, the definition of Δ implies $\ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x}) \in [0, \Delta]$, for any $\mathbf{x} \in \mathcal{X}$. Hence, applying Hoeffding's lemma on the random variables $\ell_i = \ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x}_i) \in [0, \Delta]$, we obtain

$$\mathop{\mathbb{E}}_{\mathbf{x}_i \sim \mu} \exp\left[\frac{\lambda}{n} \left(\mathop{\mathbb{E}}_{\mathbf{x} \sim \mu} \left[\ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x})\right] - \ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x}_i) \right)\right] \leq \exp\left[\frac{\lambda^2 \Delta^2}{8n^2}\right].$$

Using Fubini's theorem, we have that:

$$\mathop{\mathbb{E}}_{S \sim \mu^{\otimes n}} \exp\left[\lambda \left(\mathop{\mathbb{E}}_{\mathbf{x} \sim \mu} \left[\ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x})\right] - \frac{1}{n} \sum_{i=1}^{n} \ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x}_i) \right)\right] = \prod_{i=1}^{n} \mathop{\mathbb{E}}_{x_i \sim \mu} \exp\left[\frac{\lambda}{n} \left(\mathop{\mathbb{E}}_{\mathbf{x} \sim \mu} \left[\ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x})\right] - \ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x}_i) \right)\right]$$

which leads to

$$\mathop{\mathbb{E}}_{S \sim \mu^{\otimes n}} \exp\left[\lambda \left(\mathop{\mathbb{E}}_{\mathbf{x} \sim \mu} \left[\ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x})\right] - \frac{1}{n} \sum_{i=1}^{n} \ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x}_i) \right)\right] \leq \left(\exp\left[\frac{\lambda^2 \Delta^2}{8n^2}\right] \right)^n = \exp\left[\frac{\lambda^2 \Delta^2}{8n}\right]$$

□

C.3 Proof of Theorem 4.4

We need to bound the average distance and the exponential moment of Theorem 4.2, under the assumption $\mu = g^* \sharp p^*$, with $p^* = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the standard Gaussian distribution on $\mathbb{R}^{d^*}$, and $g^* \in \text{Lip}_{K_*}(\mathbb{R}^{d^*}, \mathcal{X})$.

Lemma C.2. *Under the hypotheses of Theorem 4.4, the following inequality holds:*

$$\log \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \exp\left[\lambda\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell^\theta_{rec}(\mathbf{z},\mathbf{x})\right] - \frac{1}{n}\sum_{i=1}^n \ell^\theta_{rec}(\mathbf{z},\mathbf{x}_i)\right)\right] \leq \frac{\lambda^2 K_*^2}{2n}. \tag{C.7}$$

Proof. We have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \exp\left[\lambda\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x})\right] - \frac{1}{n}\sum_{i=1}^n \ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x}_i)\right)\right] \\
&= \mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \exp\left[\lambda\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\|\mathbf{x} - g_\theta(\mathbf{z})\|\right] - \frac{1}{n}\sum_{i=1}^n \|\mathbf{x}_i - g_\theta(\mathbf{z})\|\right)\right] \\
&= \mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \exp\left[\sum_{i=1}^n \frac{\lambda}{n}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\|\mathbf{x} - g_\theta(\mathbf{z})\|\right] - \|\mathbf{x}_i - g_\theta(\mathbf{z})\|\right)\right] \\
&= \mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \prod_{i=1}^n \exp\left[\frac{\lambda}{n}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\|\mathbf{x} - g_\theta(\mathbf{z})\|\right] - \|\mathbf{x}_i - g_\theta(\mathbf{z})\|\right)\right]
\end{aligned}$$

Since all the $\mathbf{x}_i$ are samples iid, we can use Fubini's theorem to obtain:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} e^{\lambda\left(\mathbb{E}_{\mathbf{x}\sim\mu}\left[\ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x})\right] - \frac{1}{n}\sum_{i=1}^n \ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x}_i)\right)} \\
&= \prod_{i=1}^n \mathop{\mathbb{E}}_{\mathbf{x}_i\sim\mu} \exp\left[\frac{\lambda}{n}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\|\mathbf{x} - g_\theta(\mathbf{z})\|\right] - \|\mathbf{x}_i - g_\theta(\mathbf{z})\|\right)\right]
\end{aligned}$$

Recall that $\mu = g^* \sharp p^*$, where p^* is the standard Gaussian distribution on $\mathbb{R}^{d^*}$ and g^* is K_*-Lipschitz continuous. This means we can rewrite an expectation wrt $\mathbf{x}_i \sim \mu$ as an expectation wrt $\mathbf{w}_i \sim p^*$ as follows:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \exp\left[\lambda\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x})\right] - \frac{1}{n}\sum_{i=1}^n \ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x}_i)\right)\right] \\
&= \prod_{i=1}^n \mathop{\mathbb{E}}_{\mathbf{x}_i\sim\mu}\left[\exp\left[\frac{\lambda}{n}\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\|\mathbf{x} - g_\theta(\mathbf{z})\|\right] - \|\mathbf{x}_i - g_\theta(\mathbf{z})\|\right)\right]\right] \\
&= \prod_{i=1}^n \mathop{\mathbb{E}}_{\mathbf{w}_i\sim p^*}\left[\exp\left[\frac{\lambda}{n}\left(\mathop{\mathbb{E}}_{\mathbf{w}'\sim p^*}\left[\|g^*(\mathbf{w}') - g_\theta(\mathbf{z})\|\right] - \|g^*(\mathbf{w}_i) - g_\theta(\mathbf{z})\|\right)\right]\right] \\
&\overset{(*)}{\leq} \prod_{i=1}^n \exp\left[\frac{\lambda^2 K_*^2}{2n^2}\right] \\
&= \exp\left[\frac{\lambda^2 K_*^2}{2n}\right]
\end{aligned}$$

We still need to justify $\overset{(*)}{\leq}$. Define for any arbitrary $\alpha \in \mathcal{X}$ the function $f : \mathbb{R}^{d^*} \to \mathbb{R}$ as:

$$f(\mathbf{w}) = \|g^*(\mathbf{w}) - \alpha\|.$$

Since $g^* \in \mathrm{Lip}_{K_*}(\mathbb{R}^{d^*}, \mathcal{X})$, the function f is K_*-Lipschitz. Indeed, for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^{d^*}$,

$$\begin{aligned} f(\mathbf{w}_1) - f(\mathbf{w}_2) &= \|g^*(\mathbf{w}_1) - \alpha\| - \|g^*(\mathbf{w}_2) - \alpha\| \\ &= \|g^*(\mathbf{w}_1) - \alpha + g^*(\mathbf{w}_2) - g^*(\mathbf{w}_2)\| - \|g^*(\mathbf{w}_2) - \alpha\| \\ &\leq \|g^*(\mathbf{w}_1) - g^*(\mathbf{w}_2)\| + \|g^*(\mathbf{w}_2) - \alpha\| - \|g^*(\mathbf{w}_2) - \alpha\| \\ &= \|g^*(\mathbf{w}_1) - g^*(\mathbf{w}_2)\| \\ &\leq K_* \|\mathbf{w}_1 - \mathbf{w}_2\| \end{aligned}$$

Moreover, it is known (see Theorem 5.5 of Boucheron et al. (2013)) that if f is a K_*-Lipschitz function of a standard normal random variable $\mathbf{z}$, then

$$\mathbb{E}\, e^{\lambda(\mathbb{E}[f(\mathbf{z})] - f(\mathbf{z}))} \leq e^{\frac{\lambda^2 K_*^2}{2}}.$$

Hence,

$$\mathop{\mathbb{E}}_{\mathbf{w}_i \sim p^*} \left[\exp\left[\frac{\lambda}{n} \left(\mathop{\mathbb{E}}_{\mathbf{w}' \sim p^*} [\|g^*(\mathbf{w}') - g_\theta(\mathbf{z})\|] - \|g^*(\mathbf{w}_i) - g_\theta(\mathbf{z})\| \right) \right] \right] \leq \exp\left[\frac{\lambda^2 K_*^2}{2n^2} \right],$$

which proves $\overset{(*)}{\leq}$ and concludes this proof. □

Lemma C.3. *Under the hypotheses of Theorem 4.4, with probability at least* $1 - \frac{nd^*}{2} e^{\frac{-a^2}{2}}$ *over the random draw of* S*,*

$$\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mu} d(\mathbf{x}, \mathbf{x}_i) \leq n K_* \sqrt{(1 + a^2) d^*} \tag{C.8}$$

Proof. First, since the training set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \overset{\text{iid}}{\sim} \mu$, for each $1 \leq i \leq n$, there exists $\mathbf{w}_i \sim p^*$ such that $\mathbf{x}_i = g^*(\mathbf{w}_i)$. Let $a > 0$ be a positive real number. By definition of p^*, we have

$$\mathbb{P}\left[\forall i, \mathbf{w}_i \in [-a, a]^{d^*} \right] = \left(\mathrm{erf}\left(\frac{a}{\sqrt{2}} \right) \right)^{nd^*},$$

where $\mathrm{erf}(\cdot)$ denotes the error function. Since the error function verifies (see Chu (1955))

$$\mathrm{erf}\left(\frac{a}{\sqrt{2}} \right) \geq \sqrt{1 - e^{\frac{-a^2}{2}}},$$

we can use Bernoulli's inequality (see Section 2.4 of Mitrinovic and Vasic (1970)) to obtain

$$\mathbb{P}\left[\forall i, \mathbf{w}_i \in [-a, a]^{d^*} \right] \geq \left(1 - e^{\frac{-a^2}{2}} \right)^{nd^*/2} \geq 1 - \frac{nd^*}{2} e^{\frac{-a^2}{2}}. \tag{C.9}$$

Now we assume $\mathbf{w}_i \in [-a, a]^{d^*}$ for all $1 \leq i \leq n$ and we shall prove the desired inequality:

$$\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mu} d(\mathbf{x}, \mathbf{x}_i) \leq n K_* \sqrt{(1 + a^2) d^*} \tag{C.10}$$

Let us prove (C.10). We have

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mu} d(\mathbf{x}, \mathbf{x}_i) = \mathop{\mathbb{E}}_{\mathbf{x} \sim \mu} \|\mathbf{x} - \mathbf{x}_i\| = \mathop{\mathbb{E}}_{\mathbf{w} \sim p^*} \|g^*(\mathbf{w}) - g^*(\mathbf{w}_i)\| \leq K_* \mathop{\mathbb{E}}_{\mathbf{w} \sim p^*} \|\mathbf{w} - \mathbf{w}_i\|, \tag{C.11}$$

where the inequality follows from the assumption $g^* \in \mathrm{Lip}_{K_*}(\mathbb{R}^{d^*}, \mathcal{X})$. Using Holder's inequality, the fact that $\|\mathbf{w} - \mathbf{w}_i\|^2$ is a non-central χ^2 random variable with d^* degrees of freedom and non-centrality coefficient $\|\mathbf{w}_i\|^2$, and the assumption $\mathbf{w}_i \in [-a, a]^{d^*}$, we obtain

$$\mathop{\mathbb{E}}_{\mathbf{w} \sim p^*} \|\mathbf{w} - \mathbf{w}_i\| \leq \left(\mathop{\mathbb{E}}_{\mathbf{w} \sim p^*} \|\mathbf{w} - \mathbf{w}_i\|^2 \right)^{1/2} = \left(d^* + \|\mathbf{w}_i\|^2 \right)^{1/2} \leq \left(d^* + a^2 d^* \right)^{1/2}.$$

Hence,

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mu} \|\mathbf{x} - \mathbf{x}_i\| \leq K_* \sqrt{(1 + a^2) d^*}$$

which proves (C.10). □

Proof of Theorem 4.4. Lemmas C.2 and C.3 applied to the result from Theorem 4.2 provide us with the inequality of Theorem 4.4. Finally, the confidence of $1-\delta-\frac{nd^*}{2}e^{\frac{-a^2}{2}}$ is obtained by using the union bound: the inequality in Theorem 4.2 holds with probability at least $1-\delta$, whereas the inequality appearing in Lemma C.3 holds with probability at least $1-\frac{nd^*}{2}e^{\frac{-a^2}{2}}$. □

In the following proposition, we provide an alternate version of Theorem 4.4, where the distribution p^* is the uniform distribution[3] on $[0,1]^{d^*}$, instead of the standard Gaussian distribution on $\mathbb{R}^{d^*}$.

Proposition C.4. *Let $\mathcal{X}$ be the instance space, $\mathcal{Z}$ the latent space, $p(\mathbf{z}) \in \mathcal{M}_+^1(\mathcal{Z})$ the prior distribution, θ the parameters of the decoder, $\delta \in (0,1), \lambda > 0, a > 0$ be real numbers. Assume the data-generating distribution $\mu = g^* \sharp p^*$, where $p^* = \mathcal{U}([0,1]^{d^*})$ is the uniform distribution on $[0,1]^{d^*}$ and $g^* \in \mathrm{Lip}_{K_*}(\mathbb{R}^{d^*}, \mathcal{X})$ is K_*-Lipschitz continuous. With probability at least $1-\delta$ over the random draw of S, the following holds for any posterior $q_\phi(\mathbf{z}|\mathbf{x})$:*

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x})} \ell_{rec}^\theta(\mathbf{z},\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n \left\{ \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \ell_{rec}^\theta(\mathbf{z},\mathbf{x}_i) \right\} \leq \frac{1}{\lambda}\left(\sum_{i=1}^n \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i)\,||\,p(\mathbf{z})) + \lambda K_\phi K_\theta K_* \sqrt{d^*} + \log\frac{1}{\delta} + \frac{\lambda^2 K_*^2}{2n} \right).$$

Proof. Let $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\} \subseteq [0,1]^{d^*}$ be such that for all $1 \leq i \leq n$, $\mathbf{x}_i = g^*(\mathbf{w}_i)$. Since the diameter of $[0,1]^{d^*}$ is $\sqrt{d^*}$, using the assumptions on μ and g^*, we obtain

$$\sum_{i=1}^n \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} d(\mathbf{x},\mathbf{x}_i) = \sum_{i=1}^n \mathop{\mathbb{E}}_{\mathbf{w}\sim p^*} d(g^*(\mathbf{w}), g^*(\mathbf{w}_i)) \leq K_* \sum_{i=1}^n \mathop{\mathbb{E}}_{\mathbf{w}\sim p^*} \|\mathbf{w}-\mathbf{w}_i\| \leq nK_*\sqrt{d^*}.$$

Applying the inequality above to Theorem 4.2 yields the desired result. □

Note that unlike Theorem 4.4, the confidence $1-\delta$ of Theorem 4.2 is not lowered in Proposition C.4.

D Proofs of the results in Section 5

To simplify the proofs of the theorems of Section 5, we start by proving Lemmas D.1 and D.2 below. First, recall the definition of $\hat{\mu}_{\phi,\theta}$:

$$\hat{\mu}_{\phi,\theta} = \frac{1}{n}\sum_{i=1}^n g_\theta \sharp q_\phi(\mathbf{z}|\mathbf{x}_i).$$

The triangle inequality implies

$$W_1(\mu, g_\theta \sharp p(\mathbf{z})) \leq W_1(\mu, \hat{\mu}_{\phi,\theta}) + W_1(\hat{\mu}_{\phi,\theta}, g_\theta \sharp p(\mathbf{z})). \tag{D.1}$$

Let us state and prove the first lemma of this section.

Lemma D.1. *The following inequality holds with probability at least $1-\delta$ over the random draw of $S \sim \mu^{\otimes n}$:*

$$\lambda W_1(\mu, \hat{\mu}_{\phi,\theta}) \leq \frac{\lambda}{n}\sum_{i=1}^n \left(\mathop{\mathbb{E}}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x}_i)} \ell_{rec}^\theta(\mathbf{z},\mathbf{x}_i) \right) + \sum_{i=1}^n \mathrm{KL}(q(\mathbf{z}|\mathbf{x}_i)\,||\,p(\mathbf{z})) + \log\frac{1}{\delta} + \log \mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} e^{\lambda\left(\mathbb{E}_{\mathbf{x}\sim\mu}\left[\ell_{rec}^\theta(\mathbf{z},\mathbf{x})\right] - \frac{1}{n}\sum_{i=1}^n \ell_{rec}^\theta(\mathbf{z},\mathbf{x}_i)\right)}.$$

Proof. Recall the expression for the Wasserstein distance based on couplings:

$$W_1(\mu, \hat{\mu}_{\phi,\theta}) = \inf_{\pi\in\Gamma(\mu,\hat{\mu}_{\phi,\theta})} \int_{\mathcal{X}\times\mathcal{X}} \|\mathbf{x}-\mathbf{y}\|\, d\pi(\mathbf{x},\mathbf{y})$$

[3] Note that the result holds for any distribution on $[0,1]^{d^*}$, not just the uniform distribution.

In particular, $W_1(\mu, \hat{\mu}_{\phi,\theta})$ is less than the right-hand side obtained by the product coupling which can be rewritten, using Fubini's theorem, as:

$$\begin{aligned} W_1(\mu, \hat{\mu}_{\phi,\theta}) &\leq \int_{\mathcal{X}\times\mathcal{X}} \|\mathbf{x}-\mathbf{y}\| \; d\mu(\mathbf{x}) d\hat{\mu}_{\phi,\theta}(\mathbf{y}) \\ &= \mathop{\mathbb{E}}_{\mathbf{y}\sim\hat{\mu}_{\phi,\theta}} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \|\mathbf{x}-\mathbf{y}\| \,. \end{aligned}$$

Using the derivation above and the definition of $\hat{\mu}_{\phi,\theta}$, we obtain

$$\begin{aligned} W_1(\mu, \hat{\mu}_{\phi,\theta}) \leq \mathop{\mathbb{E}}_{\mathbf{y}\sim\hat{\mu}_{\phi,\theta}} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \|\mathbf{x}-\mathbf{y}\| &= \frac{1}{n}\sum_{i=1}^{n} \left(\mathop{\mathbb{E}}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i)} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \|\mathbf{x} - g_\theta(\mathbf{z})\| \right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(\mathop{\mathbb{E}}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i)} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x}) \right). \end{aligned}$$

We can upper bound this expression using Lemma B.1 with $\mathcal{H} = \mathcal{Z}$ and $\ell = \ell^\theta_{\text{rec}}$. We get that with probability at least $1-\delta$ over the random draw of $S \sim \mu^{\otimes n}$:

$$\begin{aligned} \frac{\lambda}{n}\sum_{i=1}^{n} \left(\mathop{\mathbb{E}}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x}_i)} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x}) \right) \leq \frac{\lambda}{n}\sum_{i=1}^{n} \left(\mathop{\mathbb{E}}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x}_i)} \ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x}_i) \right) + \sum_{i=1}^{n} \mathrm{KL}(q(\mathbf{z}|\mathbf{x}_i) \,||\, p(\mathbf{z})) + \\ \log\frac{1}{\delta} + \log \mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} e^{\lambda\left(\mathbb{E}_{\mathbf{x}\sim\mu}\left[\ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x})\right] - \frac{1}{n}\sum_{i=1}^{n} \ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x}_i)\right)}. \end{aligned}$$

□

Therefore, using the upper bounds on the exponential moment from Section 4, we can prove Theorems 5.1 and 5.3 in the following sections.

Next, we prove the following lemma.

Lemma D.2. *The following inequality holds.*

$$W_1(\hat{\mu}_{\phi,\theta}, g_\theta \sharp p(\mathbf{z})) \leq \frac{K_\theta}{n}\sum_{i=1}^{n} \sqrt{\|\mu_\phi(\mathbf{x}_i)\|^2 + \left\|\sigma_\phi(\mathbf{x}_i) - \vec{1}\right\|^2},$$

where $\vec{1} \in \mathbb{R}^{d_\mathcal{Z}}$ *denotes the vector whose entries are all* 1.

Proof. Defining the mixture of measures

$$\hat{q}_\phi(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} q_\phi(\mathbf{z}|\mathbf{x}_i),$$

the definition of $\hat{\mu}_{\phi,\theta}$ and the definition of a pushforward measures yield

$$\hat{\mu}_{\phi,\theta} = \frac{1}{n}\sum_{i=1}^{n} g_\theta \sharp q_\phi(\mathbf{z}|\mathbf{x}_i) = g_\theta \sharp \hat{q}_\phi(\mathbf{z}).$$

Using the dual formulation of the Wasserstein distance, we have

$$\begin{aligned} W_1(\hat{\mu}_{\phi,\theta}, g_\theta \sharp p(\mathbf{z})) &= W_1\left(g_\theta \sharp \hat{q}_\phi(\mathbf{z}), g_\theta \sharp p(\mathbf{z})\right) \\ &= \sup_{f\in \mathrm{Lip}_1(\mathcal{X},\mathbb{R})} \left[\int_\mathcal{Z} f\circ g_\theta(\mathbf{z})\, d\hat{q}_\phi(\mathbf{z}) - \int_\mathcal{Z} f\circ g_\theta(\mathbf{z})\, dp(\mathbf{z}) \right] \\ &= \sup_{g\in\mathcal{G}_\theta} \left[\int_\mathcal{Z} g(\mathbf{z})\, d\hat{q}_\phi(\mathbf{z}) - \int_\mathcal{Z} g(\mathbf{z})\, dp(\mathbf{z}) \right] \\ &\leq \sup_{g\in \mathrm{Lip}_{K_\theta}(\mathcal{Z},\mathbb{R})} \left[\int_\mathcal{Z} g(\mathbf{z})\, d\hat{q}_\phi(\mathbf{z}) - \int_\mathcal{Z} g(\mathbf{z})\, dp(\mathbf{z}) \right] \\ &= K_\theta W_1(\hat{q}_\phi(\mathbf{z}), p(\mathbf{z})), \end{aligned}$$

where $\mathcal{G}_\theta = \{g : \mathcal{Z} \to \mathbb{R} \text{ s.t. } g = f \circ g_\theta \text{ and } f \in \text{Lip}_1(\mathcal{X}, \mathbb{R})\}$ and the inequality holds because $\mathcal{G}_\theta \subseteq \text{Lip}_{K_\theta}(\mathcal{Z}, \mathbb{R})$, since $g_\theta : \mathcal{Z} \to \mathcal{X}$ is K_θ-Lipschitz. Now, since $(p, q) \mapsto W_1(p, q)$ is convex, the definition of $\hat{q}_\phi(\mathbf{z})$ implies

$$W_1(\hat{q}_\phi(\mathbf{z}), p(\mathbf{z})) \leq \frac{1}{n}\sum_{i=1}^{n} W_1(q_\phi(\mathbf{z}|\mathbf{x}_i), p(\mathbf{z})) \leq \frac{1}{n}\sum_{i=1}^{n} W_2(q_\phi(\mathbf{z}|\mathbf{x}_i), p(\mathbf{z})). \tag{D.2}$$

Since, by Equation (A.2),

$$W_2(q_\phi(\mathbf{z}|\mathbf{x}_i), p(\mathbf{z}))^2 = \|\mu_\phi(\mathbf{x}_i)\|^2 + \left\|\sigma_\phi(\mathbf{x}_i) - \vec{1}\right\|^2,$$

we obtain

$$W_1(\hat{\mu}_{\phi,\theta}, g_{\theta\sharp}p(\mathbf{z})) \leq \frac{K_\theta}{n}\sum_{i=1}^{n}\sqrt{\|\mu_\phi(\mathbf{x}_i)\|^2 + \left\|\sigma_\phi(\mathbf{x}_i) - \vec{1}\right\|^2}.$$

□

D.1 Proof of Theorem 5.1

Proof of Theorem 5.1. Recall from Lemma D.1 that with probability at least $1-\delta$ over the random draw of $S \sim \mu^{\otimes n}$,

$$\begin{aligned}\lambda W_1(\mu, \hat{\mu}_{\phi,\theta}) \leq \frac{\lambda}{n}\sum_{i=1}^{n}\left(\mathop{\mathbb{E}}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x}_i)} \ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x}_i)\right) + \sum_{i=1}^{n}\text{KL}(q(\mathbf{z}|\mathbf{x}_i)\,||\,p(\mathbf{z})) + \\ \log\frac{1}{\delta} + \log \mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} e^{\lambda\left(\mathbb{E}_{\mathbf{x}\sim\mu}\left[\ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x})\right] - \frac{1}{n}\sum_{i=1}^{n}\ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x}_i)\right)}.\end{aligned} \tag{D.3}$$

In order to prove Theorem 4.3 in section C.2, we proved that

$$\mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \exp\left[\lambda\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x})\right] - \frac{1}{n}\sum_{i=1}^{n}\ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x}_i)\right)\right] \leq \exp\left[\frac{\lambda^2\Delta^2}{8n}\right].$$

Now, we can reuse this inequality to upper-bound the last term on the right-hand side of Equation (D.3). We obtain the desired theorem: under the assumptions of Theorem 4.3, with probability at least $1-\delta$ over the random draw of $S \sim \mu^{\otimes n}$, the following holds for any posterior $q_\phi(\mathbf{z}|\mathbf{x})$:

$$W_1(\mu, \hat{\mu}_{\phi,\theta}) \leq \frac{1}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x}_i)\right\} + \frac{1}{\lambda}\left(\sum_{i=1}^{n}\text{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i)\,||\,p(\mathbf{z})) + \log\frac{1}{\delta} + \frac{\lambda^2\Delta^2}{8n}\right).$$

□

D.2 Proof of Theorem 5.2

Proof of Theorem 5.2. Theorem 5.2 is a direct consequence of Theorem 5.1 and Lemma D.2 applied to Equation (D.1). □

D.3 Proof of Theorem 5.3

Proof of Theorem 5.3. Recall from Lemma D.1 that with probability at least $1-\delta$ over the random draw of $S \sim \mu^{\otimes n}$,

$$\begin{aligned}\lambda W_1(\mu, \hat{\mu}_{\phi,\theta}) \leq \frac{\lambda}{n}\sum_{i=1}^{n}\left(\mathop{\mathbb{E}}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x}_i)} \ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x}_i)\right) + \sum_{i=1}^{n}\text{KL}(q(\mathbf{z}|\mathbf{x}_i)\,||\,p(\mathbf{z})) + \\ \log\frac{1}{\delta} + \log \mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} e^{\lambda\left(\mathbb{E}_{\mathbf{x}\sim\mu}\left[\ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x})\right] - \frac{1}{n}\sum_{i=1}^{n}\ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x}_i)\right)}.\end{aligned}$$

We can then use Lemma C.2 which stated that

$$\log \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \mathop{\mathbb{E}}_{S\sim\mu^{\otimes n}} \exp\left[\lambda\left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x})\right] - \frac{1}{n}\sum_{i=1}^{n}\ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x}_i)\right)\right] \leq \frac{\lambda^2 K_*^2}{2n}. \tag{D.4}$$

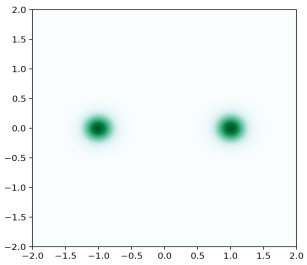
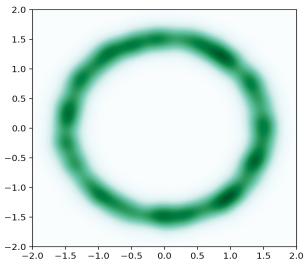

Figure 1: Samples from the real datasets

The expectations over $\mathbf{z}$ and S can be swapped using Fubini's Theorem. Hence, combining Lemma C.2 and Lemma D.1, we obtain Theorem 5.3: with probability at least $1-\delta$ over the random draw of $S \sim \mu^{\otimes n}$, the following holds for any posterior $q_\phi(\mathbf{z}|\mathbf{x})$.

$$W_1(\mu, \hat{\mu}_{\phi,\theta}) \leq \frac{1}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x}_i)} \ell^\theta_{\text{rec}}(\mathbf{z}, \mathbf{x}_i)\right\} + \frac{1}{\lambda}\left(\sum_{i=1}^{n}\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i)\,||\,p(\mathbf{z})) + \log\frac{1}{\delta} + \frac{\lambda^2 K_*^2}{2n}\right).$$

□

D.4 Proof of Theorem 5.4

Proof of Theorem 5.4. Theorem 5.4 is a direct consequence of Theorem 5.3 and Lemma D.2 applied to Equation (D.1). □

E Numerical Experiments

We computed the numerical value of the bound of Theorem 4.3. We performed the experiments on two 2-dimensional synthetic datasets. The first one is a mixture of two isotropic Gaussian distributions on $\mathbb{R}^2$ centered at $(-1, 0)$ and $(1, 0)$ respectively, and with standard deviation $\sigma = 0.1$ and null covariances. The second dataset consists of noisy samples arranged in a circle centered at the origin, with radius 1.5 and standard deviation $\sigma = 0.1$. Both datasets are truncated so that no sample is over 4 standard deviations away from its corresponding mean. This is to formally ensure that the diameter of the instance spaces is finite, as required by Theorem 4.3. The sizes of the training, validation and test sets are respectively 50,000, 20,000 and 20,000. Samples from the two datasets are shown in Figure 1.

We used the same architecture and hyperparameters for both datasets. The encoder and decoder are fully connected networks with 3 hidden layers and 100 hidden units per layer. We also set the Lipschitz constants of the encoder and decoder networks to $K_\phi = K_\theta = 2$. In order to enforce Lipschitz continuity, we used Björk orthonormalization (Björck and Bowie, 1971) with GroupSort activations (Anil et al., 2019), and we utilized the implementation of Lipschitz layers by Anil et al. (2019). Note that Barrett et al. (2022) performed experiments with VAEs with fixed Lipschitz constants, but we did not directly use their implementation because of a difference in the definition of the Lipschitz norm of the encoder, which affects the implementation. Note also that unlike the usual computations of PAC-Bayesian bounds (Pérez-Ortiz et al., 2021), our implementation does not use probabilistic neural networks. It uses deterministic networks, as it is usual for VAEs, because our analysis did not include additional stochasticity. We used the MSE as the reconstruction loss during training, and computed the bounds on validation datasets. The samples from the different models are displayed in Figure 2.

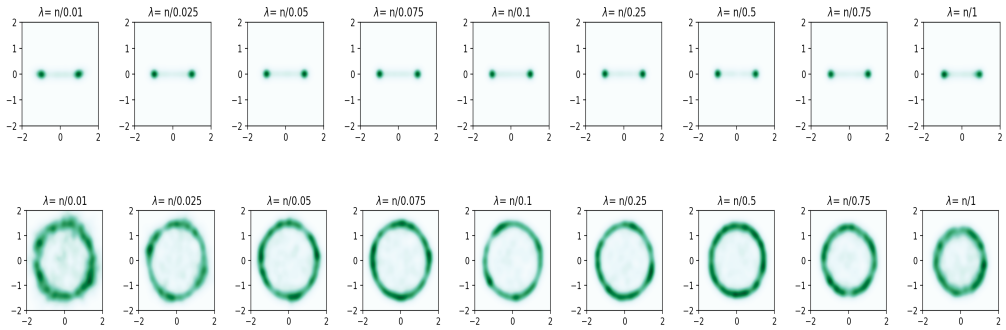

Figure 2: Samples from the models trained on the 2-Gaussian dataset (top) and the Circle dataset (bottom).

λ	**Test Rec. loss**	Emp. Rec. loss	Emp. KL loss	Exp. moment	**Bound**
$n/0.01$	0.1107	0.1110	0.0192	89.00	99.80
$n/0.025$	0.1228	0.1237	0.0505	35.60	46.45
$n/0.05$	0.1299	0.1299	0.1010	17.80	28.70
$n/0.075$	0.1388	0.1403	0.1511	11.867	22.83
$n/0.1$	0.1425	0.1436	0.2003	8.900	19.92
$n/0.25$	0.1707	0.1732	0.4883	3.560	14.89
$n/0.5$	0.2120	0.2162	0.9602	1.780	13.63
$n/0.75$	0.2718	0.2725	1.4122	1.1868	13.54
$n/1$	0.3586	0.3596	1.8593	0.8901	13.78

Table 1: Table showing the values of the different quantities of Equation E.1 for the "2-Gaussian" dataset. The upper bound on the average distance term is 10.67.

Recall the inequality of Theorem 4.3:

$$\underbrace{\mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x})}\ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x})}_{\text{Test Rec. Loss}} \leq \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left\{\mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x}_i)}\ell^\theta_{\text{rec}}(\mathbf{z},\mathbf{x}_i)\right\}}_{\text{Emp. Rec. Loss}} + \underbrace{\frac{1}{\lambda}\sum_{i=1}^{n}\text{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i)\,||\,p(\mathbf{z}))}_{\text{Emp. KL loss}} + \underbrace{K_\phi K_\theta \Delta}_{\text{Avg distance}} + \underbrace{\frac{\lambda\Delta^2}{8n}}_{\text{Exp. moment}} + \frac{1}{\lambda}\log\frac{1}{\delta}. \tag{E.1}$$

Tables 1 and 2 show the numerical values of the bound of Theorem 4.3 for different values of λ. The first column is approximated using the test set, and the last one refers to all the right-hand side of (E.1). The empirical reconstruction and KL losses are computed using the validation set, since, as mentioned in the main paper, the bounds need to be computed using a set independent from the training set.

From Tables 1 and 2, once can see that the bounds are dominated by two terms: the average distance and the exponential moment. Although as λ approaches n, the exponential moment gets smaller and the main influence comes from the upper bound on the average distance. Hence, in order to tighten the bound, one may need to derive tighter upper bounds on the average distance, or derive versions of Theorem 4.3 where this term is replaced by a numerically smaller one.

λ	**Test Rec. loss**	Emp. Rec. loss	Emp. KL loss	Exp. moment	**Bound**
$n/0.01$	0.095	0.0959	0.0197	180.50	195.81
$n/0.025$	0.1354	0.1362	0.0525	72.20	87.59
$n/0.05$	0.1785	0.1783	0.1058	36.10	51.58
$n/0.075$	0.2005	0.2020	0.1587	24.07	39.63
$n/0.1$	0.2245	0.2247	0.2117	18.05	33.69
$n/0.25$	0.3498	0.3486	0.5160	7.220	23.28
$n/0.5$	0.5026	0.4940	0.9997	3.610	20.30
$n/0.75$	0.6171	0.6154	1.4691	2.406	19.691
$n/1$	0.7513	0.7499	1.9314	1.805	19.686

Table 2: Table showing the values of the different quantities of Equation E.1 for the "Circle" dataset. The upper bound on the average distance term is 15.2.

F Additional Results and Remarks

This section contains additional remarks and discussions. We start with possible extensions of our results.

F.1 The variance of the likelihood

Our definition of the decoder network's output (the function $g_\theta : \mathcal{Z} \to \mathcal{X}$) only considers the deterministic part of the decoder. In other words, our results only apply to VAEs whose likelihood has constant variance. However, they can be extended to cases when the variance of the likelihood is optimized, but at a cost. We discuss separately the two cases where the variance depends on individual datapoints or not.

Instance-independent variance. If the standard deviation σ of the decoder is fixed, then we have $\sigma \propto \frac{n}{\lambda}$, (recall the hyperparameter λ from Theorem 3.1 and subsequent theorems). Hence, optimizing σ corresponds to optimizing λ, which is non-trivial in PAC-Bayes. Indeed, most PAC-Bayes bounds (including ours) do not directly allow one to optimize λ (see Section 2.1.4 of Alquier (2021)). Although there are some ways around this restriction, we are not aware of any results that allow one to optimize in the general case (meaning continuous values of λ and unbounded loss). For $[0, 1]$-bounded loss functions, Thiemann et al. (2017) developed a PAC-Bayes bound uniformly valid for a trade-off parameter λ', and show that one can optimize w.r.t. both the posterior and λ', under certain assumptions. For unbounded losses, if one assumes $\lambda \in \Lambda$, where $|\Lambda|$ is finite, a union bound argument allows one to make the bound uniform with respect to λ, at the cost of $\log |\Lambda|$ (see Alquier (2021)). One can still optimize with respect to a continuous set Λ, by considering a grid. For instance, if one considers $\Lambda \cap \{1, \dots, n\}$, then the penalty is $\log n$ and if one considers $\Lambda \cap \{e^k : 1 \leq k \leq n\}$, the penalty is $\log \log n$.

Instance-dependent variance. Now, assume the standard deviation is dependent on individual instances. Say we define the reconstruction loss as $\ell_\theta(\mathbf{z}, \mathbf{x}) = \frac{1}{\sigma_\theta(\mathbf{z})} \|\mathbf{x} - g_\theta(\mathbf{z})\|$, where $\sigma_\theta : \mathcal{Z} \to \mathbb{R}_{>0}$. Because of the division by $\sigma_\theta(\mathbf{z})$, let us assume that there is a fixed upper bound $\sigma_1 > 0$ such that $\sigma_\theta(\mathbf{z}) > \sigma_1$, for any $\mathbf{z} \in \mathcal{Z}$. There are two main tasks: making sure Assumption 1 is satisfied, and bounding the exponential moment of Theorem 4.2, with this new loss function.

Verifying Assumption 1 is equivalent to showing that Proposition 4.1 is verified for this new loss function ℓ_θ. The second part of the proof of Proposition 4.1 tells us that we need to show that ℓ_θ is Lipschitz-continuous. Note that in general, the product of real-valued Lipschitz functions is not Lipschitz. Hence, we assume, in addition, that $\|\mathbf{x} - g_\theta(\mathbf{z})\| \leq M < \infty$. The following proposition shows that Assumption 1 is satisfied with the constant $K = K_\phi \left(\frac{K_\sigma M}{\sigma_1^2} + \frac{K_\theta}{\sigma_1} \right)$.

Proposition F.1. *Consider a VAE with parameters ϕ and θ and let $K_\phi, K_\theta \in \mathbb{R}$ be the Lipschitz norms of the encoder and decoder respectively. Also, consider the loss function $l^\theta_{rec} : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}$ defined as*

$$l^\theta_{rec}(\mathbf{z}, \mathbf{x}) = \frac{1}{\sigma_\theta(\mathbf{z})} \|\mathbf{x} - g_\theta(\mathbf{z})\|$$

where $\sigma_\theta : \mathcal{Z} \to \mathbb{R}_{>0}$ is K_σ-Lipschitz. Assume and for all $\mathbf{z} \in \mathcal{Z}$, $\sigma_\theta(\mathbf{z}) > \sigma_1$ and $\|\mathbf{x} - g_\theta(\mathbf{z})\| \leq M$ for some fixed $0 < \sigma_1 < 1$ and $M > 0$. Then the variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$ satisfies Assumption 1 with $\mathcal{E} = \{f : \mathcal{Z} \to \mathbb{R} : \|f\|_{Lip} \leq \frac{K_\sigma M}{\sigma_1^2} + \frac{K_\theta}{\sigma_1}\}$, $K = K_\phi \left(\frac{K_\sigma M}{\sigma_1^2} + \frac{K_\theta}{\sigma_1}\right)$, and $\ell = l^\theta_{rec}$.

Proof. The first part of Assumption 1 is satisfied, since $\frac{K_\sigma M}{\sigma_1^2} + \frac{K_\theta}{\sigma_1} > K_\theta$. Now, for the second part of Assumption 1, we need to show that l^θ_{rec} is $\frac{K_\sigma M}{\sigma_1^2} + \frac{K_\theta}{\sigma_1}$-Lipschitz continuous. First,

$$\left|\frac{1}{\sigma_\theta(\mathbf{z}_1)} - \frac{1}{\sigma_\theta(\mathbf{z}_2)}\right| = \left|\frac{\sigma_\theta(\mathbf{z}_2) - \sigma_\theta(\mathbf{z}_1)}{\sigma_\theta(\mathbf{z}_1)\sigma_\theta(\mathbf{z}_2)}\right| \leq \frac{K_\sigma \|\mathbf{z}_1 - \mathbf{z}_2\|}{\sigma_1^2}.$$

We have

$$\begin{aligned}
\left|l^\theta_{\text{rec}}(\mathbf{z}_1, \mathbf{x}) - l^\theta_{\text{rec}}(\mathbf{z}_2, \mathbf{x}\right| &= \left|\frac{1}{\sigma_\theta(\mathbf{z}_1)} \|\mathbf{x} - g_\theta(\mathbf{z}_1)\| - \frac{1}{\sigma_\theta(\mathbf{z}_2)} \|\mathbf{x} - g_\theta(\mathbf{z}_2)\|\right| \\
&= \left|\frac{1}{\sigma_\theta(\mathbf{z}_1)} - \frac{1}{\sigma_\theta(\mathbf{z}_2)}\right| \|\mathbf{x} - g_\theta(\mathbf{z}_1)\| + \frac{1}{\sigma_\theta(\mathbf{z}_2)} \left|\|\mathbf{x} - g_\theta(\mathbf{z}_1)\| - \|\mathbf{x} - g_\theta(\mathbf{z}_2)\|\right| \\
&\leq \frac{K_\sigma M}{\sigma_1^2} \|\mathbf{z}_1 - \mathbf{z}_2\| + \frac{K_\theta}{\sigma_1} \|\mathbf{z}_1 - \mathbf{z}_2\| \\
&= \left(\frac{K_\sigma M}{\sigma_1^2} + \frac{K_\theta}{\sigma_1}\right) \|\mathbf{z}_1 - \mathbf{z}_2\|
\end{aligned}$$

□

Now, let us focus on bounding the exponential moment. In this case, when the instance space is bounded, the upper bound on the exponential moment (in the proof of Theorem 4.3) is:

$$\frac{\lambda^2 \Delta^2}{8n\sigma_1^2}, \quad \text{instead of} \quad \frac{\lambda^2 \Delta^2}{8n}.$$

And under the manifold assumption, we get the following upper bound (in the proof of Theorem 4.4):

$$\frac{\lambda^2 K_*^2}{2n\sigma_1^2}, \quad \text{instead of} \quad \frac{\lambda^2 K_*^2}{2n}$$

Note that although the upper bounds on the average distance remain unchanged, the coefficient $K_\phi K_\theta$ is replaced by $K_\phi \left(\frac{K_\sigma M}{\sigma_1^2} + \frac{K_\theta}{\sigma_1}\right)$, which is larger, specially if σ_1 is very small.

F.2 Uniformity with respect to θ

As mentioned in the main paper, although our bounds hold uniformly for any encoder ϕ, they only hold for a given decoder θ. the consequence of this limitation is that the numerical computations of the bounds need to be done on a sample set disjoint from the training set (e.g. a validation or test set). Let Θ denote a set of decoder parameters over which the optimization is performed.

From a theoretical perspective, the union bound can be used to circumvent this issue, when we consider a finite set of parameters Θ. In that case, the $\log \frac{1}{\delta}$ in Theorem 3.1 becomes $\log \frac{|\Theta|}{\delta}$, which loosens the bound. Moreover, since Θ denotes a set of neural network parameters, this assumption may not be appropriate unless one chooses a very large set Θ, which can significantly loosen the bound.

Another option would be to make assumptions on the complexity of the set of loss functions $\{\ell^\theta_{\text{rec}} : \theta \in \Theta\}$ parameterized by decoder parameters $\theta \in \Theta$ (e.g. the Rademacher complexity), in order to obtain uniform bounds in a more general case. We leave such explorations to future works.

F.3 Additional Remarks

Remark F.1 (Alternate formulation of Assumption 1)**.** We can provide an equivalent formulation of Assumption 1. A posterior $q(h|\mathbf{x})$ and a loss function ℓ satisfy Assumption 1 with a constant $K > 0$

if and only if for any $\mathbf{x} \in \mathcal{X}$,

$$\left| \mathop{\mathbb{E}}_{h \sim q(h|\mathbf{x}_1)} \ell(h, \mathbf{x}) - \mathop{\mathbb{E}}_{h \sim q(h|\mathbf{x}_2)} \ell(h, \mathbf{x}) \right| \leq K d(\mathbf{x}_1, \mathbf{x}_2).$$

The formulation given in the paper is more intuitive, but this expression shows that the specific choice of $\mathcal{E}$ does not matter. The equivalence of the two formulations is a consequence of the definition of an IPM.

Remark F.2 (Prior Learning in PAC-Bayes)**.** The majority of PAC-Bayesian bounds (McAllester, 1999; Seeger, 2002; Germain et al., 2009; Mbacke et al., 2023) require the prior distribution p on the hypothesis class to be independent of the training set[4]. In practice, this means one has to use data-free priors when minimizing PAC-Bayes bounds. Since, in that case, the learned posterior is likely very far from the prior, the KL-divergence tends to be orders of magnitude larger than the empirical risk. In practice, this means the optimization is monopolized by the KL-divergence, leading to a poor performance of the learning algorithm. In order to avoid this issue and still obtain a valid certificate, the following "prior learning trick" is used. Split the training set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in two disjoint subsets S_1, S_2, where $|S_1| = n_0, |S_2| = n - n_0$ with $n_0 < n$. Then, learn the prior p on S_1, learn the posterior q on S (the whole training set), and compute the certificate on S_2.

The reason why this trick cannot be directly applied to circumvent the fact that our bounds are valid for a given decoder, is that the encoder and the decoder are jointly optimized in VAEs. Hence, one has to make sure the samples used to learn the encoder (hence, train the model) are not used in the computation of the risk certificate. We emphasize that in our case, the issue does not lie in the learning of the prior (the standard VAE considers a standard Gaussian prior), but of the loss function $\ell^{\theta}_{\text{rec}}$, which is dependent on the decoder's parameters θ.

[4] PAC-Bayesian bounds with data-dependent priors were developed by Dziugaite and Roy (2018); Rivasplata et al. (2020).