Controlling Posterior Collapse by an Inverse Lipschitz Constraint on the Decoder Network

Yuri Kinoshita [1] **Kenta Oono** [2] **Kenji Fukumizu** [3] **Yuichi Yoshida** [4] **Shin-ichi Maeda** [2]

Abstract

Variational autoencoders (VAEs) are one of the deep generative models that have experienced enormous success over the past decades. However, in practice, they suffer from a problem called posterior collapse, which occurs when the encoder coincides, or collapses, with the prior taking no information from the latent structure of the input data into consideration. In this work, we introduce an inverse Lipschitz neural network into the decoder and, based on this architecture, provide a new method that can control in a simple and clear manner the degree of posterior collapse for a wide range of VAE models equipped with a concrete theoretical guarantee. We also illustrate the effectiveness of our method through several numerical experiments.

1. Introduction

1.1. Background and Organization

Over the past decades, generative models that aim to capture the distribution of a given data have intensively contributed to the creation of many performant algorithms in the field of machine learning and artificial intelligence. The recent surge of interest in incorporating expressive neural networks into statistical and probabilistic methods has even more enhanced their ability to handle high-dimensional data such as image, text and speech. Especially, variational autoencoders (VAEs) are one of these deep generative models that have experienced enormous success (Kingma & Welling, 2014; Rezende et al., 2014). They can draw low-dimensional latent random variables from a predefined prior distribution and transform them into meaningful data using deep neural networks trained on a tractable objective function called the evidence lower bound (ELBO).

While VAEs are nowadays omnipresent in the field of machine learning, it is also widely recognized that there remain in practice some major challenges that still require effective solutions. Notably, they suffer from the problem of *posterior collapse*, which occurs when the distribution corresponding to the encoder coincides, or collapses, with the prior taking no information from the latent structure of the input data into consideration. Also known as *KL vanishing* or *over-pruning*, this phenomenon makes VAEs incapable to produce pertinent representations and has been reportedly observed in many fields (e.g., Bowman et al. (2016); Fu et al. (2019); Wang & Ziyin (2022); Yeung et al. (2017)). There exists now a large body of literature that examines its underlying causes and presents various techniques to prevent it (e.g., Bowman et al. (2016); He et al. (2019); Razavi et al. (2019)). Please refer to Subsection 1.3 for further details.

Despite this abundant number of studies conducted so far, the mechanism of posterior collapse is not completely understood, and many different approaches, such as β-VAE and δ-VAE, have been suggested over time based on varied hypotheses and theories. Some blame the variational inference (Burda et al., 2015; Bowman et al., 2016; Chen et al., 2017; Fu et al., 2019; Huang et al., 2018; Havrylov & Titov, 2020; Sønderby et al., 2016; Zhao et al., 2018), some focus on the optimization procedure (He et al., 2019; Kim et al., 2018; Li et al., 2019), and others hold the formulation of the model responsible (Dai et al., 2020; Gulrajani et al., 2017; Yang et al., 2017; van den Oord et al., 2017; Zhao et al., 2020; Dieng et al., 2019; Yeung et al., 2017; Razavi et al., 2019). Nevertheless, most proposed methods are based on heuristics and crucially lack convincing theoretical guarantees. That is why a line of work also tries to rigorously analyse the mechanism of posterior collapse (Dai et al., 2020; Wang & Ziyin, 2022; Lucas et al., 2019). For example, the recent work of Wang et al. (2021) showed that posterior collapse and latent variable non-identifiability are equivalent. These theoretical works have indeed helped us recognize the primary cause of this phenomenon. However, they require the explicit formulation of the VAE and objective function or too rigid a definition of posterior col-

[1]The University of Tokyo, Tokyo, Japan. Work done at Preferred Networks. [2]Preferred Networks, Inc., Tokyo, Japan [3]The Institute of Statistical Mathematics, Tokyo, Japan [4]National Institute of Informatics (NII), Tokyo, Japan. Correspondence to: Yuri Kinoshita <yuri-kinoshita111@g.ecc.u-tokyo.ac.jp>, Shin-ichi Maeda <ichi@preferred.jp>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

lapse. This means guarantees and suggested techniques are either only applicable to simple problems or lack practical usefulness.

Therefore, a technique that has both a theoretical guarantee and a broad applicable spectrum is first and foremost required. In this paper, we investigate a method that can control in a simple and clear manner the degree of posterior collapse for a wide range of VAE models equipped with a concrete analysis that assures this control.

1.2. Contributions

The major contributions of this paper can be summarized as follows:

- We introduce the concept of *inverse Lipschitzness* into the underlying decoder and prove under minor assumptions that the degree of posterior collapse can be controlled by this property.
- Based on this theoretical guarantee, we provide the first method that can not only directly adjust the degree of posterior collapse but is also simple and applicable to a broad type of models.
- We explain and illustrate with experiments that our method is effective and can outperform prior works.

1.3. Related Works

The first held responsible for posterior collapse was the Kullback-Leibler (KL) divergence between the encoder and prior, present in the formulation of the ELBO (Bowman et al., 2016). It would force the model to prioritize its minimization, leading to posterior collapse. As a result, many previous works have tried to attenuate its influence during the training with an annealing scheme (Huang et al., 2018; Sønderby et al., 2016; Fu et al., 2019). These works are often summarized as β-VAE originally created for other goals (Higgins et al., 2017). More loosely, the problem arises in a sense because optimization fails and is driven into undesired minima. This has encouraged some researchers to find tighter bounds than the ELBO (Burda et al., 2015) or other objective functions (Zhao et al., 2018; Chen et al., 2017; Havrylov & Titov, 2020). There have even been attempts to find a more suited optimization procedure (He et al., 2019; Kim et al., 2018; Li et al., 2019). Others have pointed out that this phenomenon mainly happens when VAEs involve high flexibility due to neural networks (Dai et al., 2020). This point of view has incited researchers to restrict the flexibility of VAEs (Gulrajani et al., 2017; Yang et al., 2017) or to modify their architecture (van den Oord et al., 2017; Zhao et al., 2020; Dieng et al., 2019; Yeung et al., 2017; Li et al., 2022). Although posterior collapse can be regarded as the optimization falling into local minima, the recent theoretical investigation of Lucas et al. (2019) underlined that these spurious stationary points are not created by the ELBO but are actually inherent in the exact maximization of the marginal log-likelihood.

δ-VAE (Razavi et al., 2019) is a model that constrains the variational family of the posterior to assure a minimum distance from the prior in terms of KL-divergence. That way, they avoid, by definition, posterior collapse. However, finding parameters that satisfy this structural constraint involves an additional tedious optimization in general. In this paper, we will provide a method that can similarly control the discrepancy between the posterior and prior but without any calculations needed at all.

As the most relevant work to this paper, Wang et al. (2021) showed that posterior collapse occurs if and only if latent variables are non-identifiable in the generative model. They proposed an architecture called latent identifiable VAE (LID-VAE) that solves this problem of non-identifiability without losing the flexibility of VAEs. Their definition of posterior collapse requires the posterior and prior to be exactly equal, which means their model cannot avoid the usual case in practice where the posterior and prior are *nearly* equal. In this paper, we will work in the same framework as Wang et al. (2021) but provide a simple model that can handle broader situations closer to reality.

In the context of generative adversarial networks, Yamaguchi & Koyama (2019) introduced into the transport map a concept equivalent to the inverse Lipschitzness in order to promote the entropy of the generator distribution and to avoid a phenomenon called mode collapse. In contrast to inverse Lipschitzness, the role of Lipschitz continuity in VAEs and other machine learning methods has been the subject of much research (Yang et al., 2020; Barrett et al., 2022). For example, Barrett et al. (2022) constrained many components of the VAE including the encoder and the decoder to satisfy Lipschitz continuity and showed this to be useful to increase VAE robustness against adversarial attacks.

Organization In Section 2, we will first explain the basic formulation of VAEs and clarify necessary mathematical backgrounds. Section 3 will be devoted to the description of our theoretical analysis, and Section 4 to the implementation of our proposed model. Finally, Section 5 will illustrate its effectiveness with synthetic and real-world data.

Notation The Euclidean norm is denoted by $\|\cdot\|$ for vectors. The exponential family $h(x)\exp\left\{T(x)^\top\xi - A(\xi)\right\}$, where $T(x) = (T_1(x), \ldots, T_t(x))^\top$ is called the sufficient statistic, is represented as $\mathrm{EF}_T(x \mid \xi)$, abbreviating the dependence on h since it does not appear in our analysis.

2. Preliminaries

In this section, we briefly explain the mathematical background for VAEs and posterior collapse.

2.1. Variational Autoencoders

Let $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^{m \times n}$ be n i.i.d. data points in $\mathbb{R}^m$, and suppose they were generated from an underlying structure. In other words, VAEs assume there exists a latent variable $z \in \mathbb{R}^l$ sampled from a pre-defined prior $p(z)$ that creates the observed data through a conditional distribution $p_\theta(x \mid z)$ parameterized over θ, also called a generative model. In short,

$$z_i \sim p(z), \quad x_i \sim p_\theta(x \mid z_i) \quad \forall i = 1, \ldots, n. \tag{1}$$

Under this problem setting, the ultimate goal of VAE is to maximize the following marginal log-likelihood:

$$\log p_\theta(\boldsymbol{x}) = \sum_{i=1}^{n} \log \int p_\theta(x_i \mid z) p(z) \mathrm{d}z.$$

While this optimization can in theory determine values of θ, it requires some intractable computations involving the marginalization over z. Therefore, a tractable lower bound called evidence lower bound (ELBO) has been proposed to avoid this problem. The idea is to first introduce a recognition model $q_\phi(z \mid x)$, which will approximate the true posterior $p_\theta(z \mid x)$. Then, the marginal log-likelihood can be formulated as follows:

$$\log p_\theta(\boldsymbol{x}) = \sum_{i=1}^{n} \mathcal{L}_{\theta,\phi}(x_i) + D(q_\phi(z \mid x_i) \mid\mid p_\theta(z \mid x_i)),$$

where

$$\mathcal{L}_{\theta,\phi}(x) := \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - D(q_\phi(z \mid x) \mid\mid p(z)), \tag{2}$$

and $D(\cdot \mid\mid \cdot)$ refers to the KL-divergence. Consequently, we obtain the ELBO defined as the right-hand side of the following inequality:

$$\log p_\theta(\boldsymbol{x}) \geq \sum_{i=1}^{n} \mathcal{L}_{\theta,\phi}(x_i).$$

Here, we used the fact that KL-divergence is non-negative, and the equality holds only if $D(q_\phi(z \mid x_i) \mid\mid p_\theta(z \mid x_i)) = 0$ for all x_i. Via a reparameterization trick, the optimization over the ELBO in terms of θ and ϕ becomes more tractable than that of the log-likelihood (Kingma & Welling, 2014). In this context, the recognition model $q_\phi(z \mid x)$ is often called an encoder, and the generative model $p_\theta(x \mid z)$ a decoder.

2.2. Posterior Collapse

Posterior collapse refers to the situation when the encoder coincides, or collapses, with the prior taking no information from the latent structure of the input data into consideration. Mathematically, this can be translated into the following general definition.

Definition 2.1 (ϵ-posterior collapse)**.** For a given parameter $\hat{\phi}$, a data set $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^{m \times n}$ and a closeness criterion $d(\cdot, \cdot)$, an *ϵ-posterior collapse* is defined for a given $\epsilon \geq 0$ by the condition

$$d(q_{\hat{\phi}}(z \mid x_i), p(z)) \leq \epsilon \quad \forall i = 1, \ldots, n. \tag{3}$$

For example, Razavi et al. (2019) concentrated on the case where the closeness criterion is KL-divergence. The problem setting of Wang et al. (2021) corresponds to a particular situation with ϵ set to 0 (see Definition 2.2). In fact, theoretical analyses that can treat such a level of abstract definition as ϵ-posterior collapse are quite rare since most use the explicit formulation of the model, ELBO and marginal log-likelihood in order to draw conclusions. An even broader definition was introduced by Lucas et al. (2019) for measurement purpose named (ϵ, δ)-posterior collapse which substitutes equation (3) with the stochastic formula $\Pr_x \left(d(p_{\hat{\theta}}(z \mid x_i), p(z)) \leq \epsilon\right) > 1 - \delta$. This formulation is outside the scope of this work, while an extension to this case is an interesting future direction.

Many elements of a VAE are held responsible for this phenomenon, namely the KL term in equation (2), the variational approximation, the optimization scheme, and the model itself. Particularly, Wang et al. (2021) showed that the 0-posterior collapse of Definition 2.2 and latent variable non-identifiability (Definition 2.3) are equivalent. This equivalence implies that it is sufficient to make the likelihood function $p_\theta(x \mid z)$ injective in terms of z for all parameters θ in order to avoid posterior collapse of Definition 2.2. This led to the model called LIDVAE proposed by Wang et al. (2021).

Definition 2.2 (posterior collapse, Wang et al. (2021))**.** For a given parameter $\hat{\theta}$ and data set $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^{m \times n}$, *posterior collapse* occurs if $p_{\hat{\theta}}(z \mid \boldsymbol{x}) = p(z)$.

Definition 2.3 (latent variable non-identifiability, Wang et al. (2021))**.** For a given parameter $\hat{\theta}$ and data set $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^{m \times n}$, the latent variable is *non-identifiable* if $p_{\hat{\theta}}(\boldsymbol{x} \mid z) = p_{\hat{\theta}}(\boldsymbol{x} \mid z')$ for all z, z', i.e., the likelihood of the data set $\boldsymbol{x}$ does not depend on the latent variable.

Note they assumed that the variational approximation is exact. That is, the encoder can represent the posterior $p_\theta(z \mid x)$. This assumption is sensible; if the exact inference already presents symptoms of posterior collapse, this is the first problem to tackle before concentrating on the approximated case, which can only aggravate the situation.

In the remainder of this paper, we will treat this exact case as well. In other words, we will investigate methods to mitigate the ϵ-posterior collapse inherent in the formulation of the model and not that caused by any sort of approximation.

However, their definition of posterior collapse is clearly too restrictive because it requires the posterior and prior to be exactly equal, which means their model cannot avoid the usual case in practice where these are *nearly* equal. This motivates us to extend Definition 2.2 to the general one (Definition 2.1) and discuss how to prevent it.

3. Theoretical Analysis

In this section, we will show that the general posterior collapse of Definition 2.1 in terms of the relative Fisher information divergence can be controlled by a simple inverse Lipschitz constraint on the decoder network.

3.1. Assumptions and Problem Setting

Let us first clarify our problem setting.

3.1.1. Generative Model

We introduce the concept of inverse Lipschitzness.

Definition 3.1 (inverse Lipschitzness)**.** Let $L \geq 0$. $f : \mathbb{R}^l \to \mathbb{R}^t$ is L-inverse Lipschitz if $\|f(x) - f(y)\| \geq L\|x - y\|$ holds for all $x, y \in \mathbb{R}^l$.

If f is L-inverse Lipschitz with $L > 0$, then it is injective. Therefore, the restriction of f on its image possesses an inverse which is $1/L$-Lipschitz. Inverse Lipschitz functions can be regarded as a stronger condition than injectivity, which is the property of the decoder network suggested by Wang et al. (2021) in order to avoid latent variable non-identifiability, and consequently posterior collapse. The motivation to introduce this stronger concept is to capture more nuances in the latent variable identifiability with the inverse Lipschitz constant L, which is not possible with simple injective functions.

Construction of inverse Lipschitz functions Inverse Lipschitz neural networks can be generated by Brenier maps (Ball, 2004; Wang et al., 2021). A *Brenier map* is a function that is the gradient of a real-valued convex function. The gradient of a real-valued L-strongly convex function F (i.e., $F(x) - L\|x\|^2/2$ is convex) becomes L-inverse Lipschitz. Therefore, theoretical results will be proved for this type of inverse Lipschitz functions, derivatives of strongly convex functions. Notably, this means we can only handle functions with the same input dimension and output dimension. Please refer to Section 4 for further details.

We can now state our main assumption.

Assumption 3.2. The generative model $p_\theta(x \mid z)$ is an exponential family so that $p_\theta(x \mid z) = \mathrm{EF}_T(x \mid f_\theta(z))$, where $f_\theta : \mathbb{R}^l \to \mathbb{R}^t$ is constructed as follows. If $l = t$, f_θ is an L-inverse Lipschitz function generated from a Brenier map, and we denote Θ_L as the set of parameters θ that achieve this property. If $l < t$, $f_\theta = f_\theta^{(2)}(B^\top f_\theta^{(1)}(z))$, where $f_\theta^{(1)} : \mathbb{R}^l \to \mathbb{R}^l$ and $f_\theta^{(2)} : \mathbb{R}^t \to \mathbb{R}^t$ are respectively inverse Lipschitz with constant L_1 and L_2 generated from Brenier maps. B is a $t \times l$ diagonal matrix with all diagonal elements with value 1. Likewise, we define the set Θ_{L_1,L_2}.

This will be the sole condition that we will impose on the model. We restrict neither the type of prior nor the optimization scheme. The assumption that the generative model or likelihood function is an exponential family is keeping large liberty to the model. It is even less restrictive than some prior works on posterior collapse, which often necessitate all components, including the generative model, to be Gaussian (e.g., Dai et al. (2020); Lucas et al. (2019)). The sole limitation is our requirement of inverse Lipschitzness, which may restrict the expression of the likelihood function. However, the effect of this restriction is precisely one of the interests of this paper. Moreover, it is shown that when f_θ is only injective, $\mathrm{EF}_T(x \mid f_\theta(z))$ can model any distributions of the form $\mathrm{EF}_T(x \mid f(z))$ where f is an arbitrary function (Wang et al., 2021). Therefore, by adjusting this inverse Lipschitz constant, we can cover the full spectrum, i.e., from the case with no restriction ($L \to 0$) to the extremely restrictive ($L \to \infty$), which implies our problem setting still leaves considerable freedom to the model.

3.1.2. Criterion

As a closeness criterion, we will select the relative Fisher information divergence.

Definition 3.3. We define the *relative Fisher information divergence* $F(\cdot \mid\mid \cdot)$ of $p(x)$ with respect to $q(x)$ as

$$F(p(x) \mid\mid q(x)) := \int \|\nabla \log p(x) - \nabla \log q(x)\|^2 p(x) \mathrm{d}x.$$

This divergence is used for a wide range of statistical analysis and machine learning applications (Yang et al., 2019; Otto & Villani, 2000; Elkhalil et al., 2021; Holmes & Walker, 2017; Walker, 2016; Huggins et al., 2018). It is also intrinsically related to the Hyvärinen score (Hyvärinen, 2005). Furthermore, many types of distributions, such as Gaussian and Gaussian mixture, satisfy the log-Sobolev inequality (LSI), which provides an upper bound of the KL-divergence in terms of the relative Fisher information divergence (see Appendix C for the exact formulation). Many distributions can satisfy this inequality since LSI is robust to bounded perturbation and Lipschitz mapping (Gross, 1975; Holley & Stroock, 1987; Ledoux, 1999). As a consequence, if a posterior that satisfies LSI collapses on the prior in terms of relative Fisher divergence, so will it in terms of

KL-divergence. On the other hand, de Bruijn's identity relates KL divergence to relative Fisher divergence. Under some additional conditions, we can show that the control of the latter results in that of the former (see Proposition 3.9 for further details). Hence, our choice of discrepancy is sensible since it is necessary and sometimes even sufficient to avoid the collapse in terms of relative Fisher divergence in order to prevent that in terms of KL-divergence.

3.2. Theoretical Guarantee

We are now ready to state our main theorem, which shows that under Assumption 3.2, posterior collapse can be efficiently controlled by the inverse Lipschitz constant. We will first discuss in detail the case where $l = t$ for clarity, and then show that similar statements hold for the general case as well.

Theorem 3.4. *Under model (1), Assumption 3.2 and $l = t$, the following holds for all i and $\theta \in \Theta_L$:*

$$F(p_\theta(z \mid x_i) \mid\mid p(z)) \geq L^2 \int \|T(x_i) - \mathbb{E}_{p_\theta(x|z)}[T(x)]\|^2 p(z)\mathrm{d}z.$$

See Appendix A.1 for the proof.

Remark 3.5. The crux of this theorem is the relation

$$\|\nabla_z \log p_\theta(z \mid x) - \nabla_z \log p(z)\| = \|\nabla_z \log p_\theta(x \mid z)\|.$$

This equation relates posterior collapse (left-hand side) to the latent variable non-identifiability (right-hand side). Indeed, if the non-identifiability in Definition 2.3 holds, then $p_\theta(x \mid z)$ would be constant with respect to z, leading to a zero derivative and posterior collapse in the strict sense of the term. This use of derivatives enables us to additionally capture nuances of the latent variable identifiability and essentially contributes to our main result.

Corollary 3.6. *Under model (1), Assumption 3.2 and $l = t$,*

$$F(p_\theta(z \mid x_i) \mid\mid p(z)) \geq L^2 \inf_{\theta \in \Theta_L} \left\{ \int \|T(x_i) - \mathbb{E}_{p_\theta(x|z)}[T(x)]\|^2 p(z)\mathrm{d}z \right\}.$$

Therefore, the lower bound is non-decreasing in terms of L. Moreover, if the infinimum of this lower bound is attained by a parameter $\theta \in \Theta_L$ and $p(z)$ has a positive variance, then the lower bound is monotonically increasing in terms of L.

In other words, we are guaranteed to avoid an ϵ-posterior collapse as long as we take a sufficiently large L. Increasing this value has the effect of moving away the posterior from the prior without any limits. In this sense, we can *control* the degree of posterior collapse only with the inverse Lipschitz constant.

This theorem adjusted for an empirical version of the relative Fisher information divergence provides additional insights concerning another formulation of the lower bound.

Theorem 3.7. *Under model (1), Assumption 3.2 and $l = t$ the following holds for all $\theta \in \Theta_L$:*

$$\bar{F}_\theta(\boldsymbol{x}) := \int \left\| \frac{1}{n} \sum_{i=1}^n \nabla_z \log p_\theta(z \mid x_i) - \nabla_z \log p(z) \right\|^2 p(z)\mathrm{d}z \geq L^2 \int \left\| \frac{1}{n} \sum_{i=1}^n T(x_i) - \mathbb{E}_{p_\theta(x|z)}[T(x)] \right\|^2 p(z)\mathrm{d}z. \tag{4}$$

See Appendix A.2 for the proof.

Bias-Variance Decomposition Now, if we have enough samples (i.e., $n \to \infty$), we can approximate $\frac{1}{n}\sum_{i=0}^n T(x_i)$ as the expectation under the true distribution $p^*(x)$, $\mathbb{E}_{p^*(x)}[T(x)]$. Moreover, consider the generative model contains the true model θ^*. This means there exists θ^* such that $p^*(x) = \int p_{\theta^*}(x \mid z)p(z)\mathrm{d}z$. Let us define $S_\theta := E_{p_\theta(x|z)}[T(x)]$. Then, the integral of the right-hand side of Equation (4) can be reformulated as

$$\begin{aligned} \int \|\mathbb{E}_{p_\theta(x|z)}[T(x)] - \mathbb{E}_{p^*(x)}[T(x)]\|^2 p(z)\mathrm{d}z &= \int \|S_\theta - \mathbb{E}[S_{\theta^*}]\|^2 p(z)\mathrm{d}z \\ &= \mathbb{V}[S_\theta] + \|\mathbb{E}[S_{\theta^*}] - \mathbb{E}[S_\theta]\|^2, \end{aligned}$$

where $\mathbb{V}[\cdot]$ is the variance of S_θ in terms of $z \sim p(z)$. As a result,

$$\bar{F}_\theta(\boldsymbol{x}) \geq L^2 \left(\mathbb{V}[S_\theta] + \|\mathbb{E}[S_{\theta^*}] - \mathbb{E}[S_\theta]\|^2 \right).$$

Interestingly, the lower bound can be written as the sum of the variance of S_θ and its bias with the true parameter.

General Case Now, let us state a similar theorem that holds for the general case $l < t$ under an additional condition.

Theorem 3.8. *Under Assumptions 3.2, $l < t$ and that $\xi \mapsto \nabla_\xi A(\xi) = \mathbb{E}_{\mathrm{EF}_T(x|\xi)}[T(x)]$ is a diffeomorphism, $F(p_\theta(z|x)||p(z))$ is lower-bounded by a term that is increasing in terms of L_1.*

See Appendix A.3 for the precise formulation and for the proof.

Expansion to KL divergence Finally, although it may require stronger assumptions, we can derive a lower bound of KL divergence between the posterior and prior from the bound of Fisher divergence as follows.

Proposition 3.9. *Suppose that the lower bound of Fisher divergence $F(p(x)||q(x)) \geq \varepsilon$ holds for any small perturbations of p and q to some extent. More precisely, let p_t (or q_t) denote the convolution between p (q, resp.) and $N(0,t)$. Assume that there is $\delta > 0$ such that $F(p_t||q_t) \geq \epsilon$ for any $t \in [0,\delta]$. Then, the bound $D(p||q) \geq \frac{1}{2}\delta\epsilon$ holds.*

See Appendix A.4 for the proof. This suggests that a lower bound of the Fisher divergence can also control the lower bound of KL divergence, and thus our method avoids posterior collapse in terms of KL divergence as well.

3.3. Discussion

The theoretical analysis led in the previous subsection considerably contributes to the research on posterior collapse in several aspects. First of all, we proved this phenomenon, inherent in the formulation of the model, can be controlled by the inverse Lipschitz constant of the underlying decoder. Indeed, increasing this constant forces the discrepancy between the prior and posterior to become larger. This kind of analysis against the general ϵ-posterior collapse was provided neither in the work of Wang et al. (2021) nor in most previous works at all. In fact, theoretical guarantee is often missing in previously proposed heuristic methods, such as β-VAE. δ-VAE similarly requires finding parameters of prior and decoder that satisfy a discrepancy constraint in terms of KL-divergence, but this needs some intractable and heavy optimization in general. On the contrary, our method is far easier and simpler since we know in advance what kind of parameter is required: the inverse Lipschitz constant. Finally, while previous analyses only treated marginal or simple cases such as the Gaussian VAE (Lucas et al., 2019), it is also important to note that our model can be used for the general exponential family. In short, we provided the first solution based on an inverse Lipschitz constraint that can *control* the degree of posterior collapse equipped with a concrete theoretical guarantee, is simple and is applicable to a broad type of models.

On the other hand, our method also presents some drawbacks. The main limitation of our theory is that we only assure to avoid posterior collapse in terms of the relative Fisher information divergence, a weaker concept than that in terms of KL-divergence. This relaxation was necessary to proceed into rigorous analysis and derive theoretical guarantees. Nonetheless, we showed that posterior collapse in terms of KL divergence can also be controlled under some further assumptions and will show in the experiments that our method can often alleviate this stronger posterior collapse as well even though it is outside the scope of our guarantee. Furthermore, while increasing L can avoid posterior collapse, this has, at the same time, the effect of contracting the set Θ_L at the risk of limiting the flexibility of the model. It is thus essential to find a good balance. However, this trade-off is clear thanks to our analysis, and the tuning is quite simple.

4. Implementation

In this section, we will describe the implementation of our model. It turns out that it is rather simple since it only extends the model LIDVAE proposed by Wang et al. (2021). We just have to focus on the realization of a neural network f_θ that is inverse Lipschitz with respect to its output. The idea is to modify the Input Convex Neural Network (ICNN) of Amos et al. (2017) and compute its Brenier map so that it becomes an L-inverse Lipschitz function.

4.1. Brenier Maps and ICNN

As previously mentioned, a Brenier map is the gradient of a real-valued convex function (Ball, 2004; Wang et al., 2021). It is injective by definition. In addition, that of an L-strongly convex function will become L-inverse Lipschitz as desired.

ICNN is a neural architecture that creates convex functions (Amos et al., 2017). As long as this property is satisfied other algorithms can be chosen, but for clarity, we will use this as an example. Amos et al. (2017) defined the fully connected ICNN $G_\theta : \mathbb{R}^l \to \mathbb{R}$ with k layers and input $z \in \mathbb{R}^l$ as follows:

$$\begin{aligned} y_{i+1} &= g_i(W_i^{(y)} y_i + W_i^{(z)} z + b_i) \ (i = 0, \dots, k-1), \\ G_\theta(z) &= y_k, \end{aligned}$$

where $\{W_i^{(y)}\}_i$ are non-negative, and all functions g_i are convex and non-decreasing. This kind of neural network is not only convex but is also known to be a universal approximator of convex function on a compact domain endowed with the sup norm (Chen et al., 2019). Given an ICNN G_θ, it is thus not difficult to extend it to an L-strongly convex function since it suffices to add a regularization term $L\|z\|^2/2$ to the output as follows:

$$\begin{aligned} y_{i+1} &= g_i(W_i^{(y)} y_i + W_i^{(z)} z + b_i) \ (i = 0, \dots, k-1), \\ G_\theta(z) &= y_k + \frac{L}{2}\|z\|^2, \end{aligned}$$

where $\{W_i^{(y)}\}_i$ are non-negative and all functions g_i are convex and non-decreasing.

Provided an L-strongly convex ICNN G_θ, we can compute its gradient as $f_\theta = \mathrm{d}G_\theta/\mathrm{d}z$ and obtain our desired L-inverse Lipschitz neural network.

4.2. IL-LIDVAE and its Variants

The construction in Subsection 4.1 leads to an extension of LIDVAE, simple but with strong theoretical guarantees, as shown in Section 3. We will call it Inverse Lipschitz LID-

VAE (IL-LIDVAE) in order to distinguish it from LIDVAE and other prior works.

Definition 4.1 (IL-LIDVAE)**.** We define IL-LIDVAE with inverse Lipschitz constants (L_1, L_2) as the following generative model:

$$z \sim p(z), \quad x \sim \mathrm{EF}(x \mid f_\theta^{(2)}(B^\top f_\theta^{(1)}(z))),$$

where $f_\theta^{(1)} : \mathbb{R}^l \to \mathbb{R}^l$ is L_1-inverse Lipschitz and $f_\theta^{(2)} : \mathbb{R}^t \to \mathbb{R}^t$ is L_2-inverse Lipschitz, both generated by Brenier maps. B is an $l \times t$ diagonal matrix with all diagonal elements with value 1.

When $l = t$, we replace $f_\theta^{(2)}(B^\top f_\theta^{(1)}(z))$ with a single L-inverse Lipschitz function generated by a Brenier map.

We can also establish variants of IL-LIDVAE for mixture models (IL-LIDMVAE) and sequential models (IL-LIDSVAE) as Wang et al. (2021) did.

Definition 4.2 (IL-LIDMVAE)**.** We define IL-LIDMVAE with inverse Lipschitz constant (L_1, L_2) as the following generative model:

$$y \sim \mathrm{Categorical}(1/c), z \sim \mathrm{EF}(z \mid B_1^\top y),$$
$$x \sim \mathrm{EF}(x \mid f_\theta^{(2)}(B_2^\top f_\theta^{(1)}(z))),$$

where $f_\theta^{(1)} : \mathbb{R}^l \to \mathbb{R}^l$ is L_1-inverse Lipschitz and $f_\theta^{(2)} : \mathbb{R}^t \to \mathbb{R}^t$ is L_2-inverse Lipschitz, both generated by Brenier maps. y is a one-hot vector that indicates the class. B_1 and B_2 are respectively a $c \times l$ and $l \times t$ diagonal matrix with all diagonal elements with value 1.

Definition 4.3 (IL-LIDSVAE)**.** We define IL-LIDSVAE with inverse Lipschitz constant (L_1, L_2) as the following generative model

$$z_i \sim p(z),\ x_i \sim \mathrm{EF}(x \mid f_\theta^{(2)}(B^\top f_\theta^{(1)}(z_i, h_\theta(x_{1:i-1})))),$$

where $f_\theta^{(1)} : \mathbb{R}^l \to \mathbb{R}^l$ is L_1-inverse Lipschitz and $f_\theta^{(2)} : \mathbb{R}^t \to \mathbb{R}^t$ is L_2-inverse Lipschitz, both generated by Brenier maps. B is an $l \times m$ diagonal matrix with all diagonal elements with value 1.

4.3. Discussion

In practice, the main limitation of this method resides in its computational cost and scalability due to the use of Brenier maps. Since our model is only an extension of LIDVAE Wang et al. (2021) and does not add additional calculation, as mentioned in their paper, fitting LIDVAE or IL-LIDVAE requires a computational complexity of $O(kp^2)$ (that of the classical VAE is $O(kp)$), where k is the number of iterations, and p the number of parameters. The training time is thus longer than the vanilla VAE. While this is the price to pay to assure the identifiability of latent variables and thus to avoid posterior collapse, this algorithm may consequently become difficult to scale for more challenging tasks. Nevertheless, the essential aspect of our model is the inverse Lipschitzness, and the suggested implementation is only the most exact realization of this property. As an approximation, we could consider to replace the Brenier maps and ICNN with a general deep neural network subjected to regularization or constraints, such as $\mathbb{E}[\|\nabla \log p(x|z)\|] > L$ or $\mathbb{E}[\|\log p(x|z) - \log p(x|z')\|/\|z - z'\|] > L$, that encourages this crucial inverse Lipschitzness. This is outside the scope of this work but we believe it is an interesting avenue for future work.

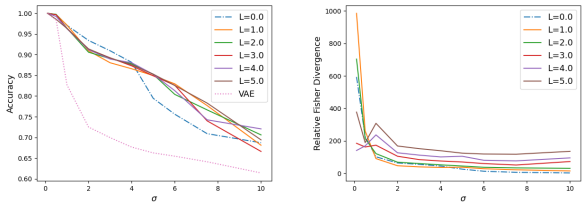


Figure 1: Accuracy of the learned posterior (left) and Relative Fisher divergence between the posterior and prior (right) for different standard deviations σ and inverse Lipschitz constants L. $L = 0$ is also the LIDVAE (Wang et al., 2021). Posterior collapse is happening for LIDVAE but can be controlled with IL-LIDVAE. "VAE" in the legend refers to the GMVAE.

5. Experiments

In this section, we will illustrate our theoretical result with several numerical examples. Our goal is to empirically verify (i) whether IL-LIDVAE can indeed control the discrepancy between the posterior and prior with the inverse Lipschitz constant and (ii) if our model can achieve better performance than the vanilla VAE. In order to answer these questions, we will first use toy data and then switch to high-dimensional text and image data.

5.1. Toy Data

We generated 10,000 samples from $\mathcal{N}((0,0)^\top, \sigma^2 I_2)$ and from $\mathcal{N}((10,10)^\top, \sigma^2 I_2)$ with different values of σ, where I_2 is the 2×2 identity matrix. With these 20,000 data points, we trained our model IL-LIDMVAE (Definition 4.2) defined for Gaussian mixtures with $c = 2$. The dimension of the latent variables l was set to 2, and the decoder was also Gaussian. For this case, we only need to control one inverse Lipschitz constant as $l = t$. $L = 0$ corresponds to the LIDMVAE of Wang et al. (2021). Intuitively, for large σ, the two classes will overlap each other and become similar to a single Gaussian distribution. This is expected

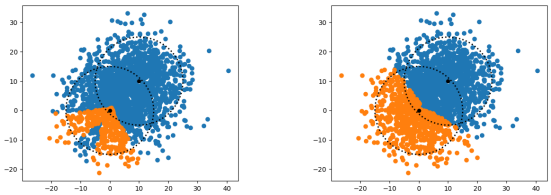


Figure 2: Posterior of GMVAE (left) and IL-LIDMVAE with $L_1 = L_2 = 5.0$ (right) for the toy data with $\sigma = 7.5$. Black points are the means of $N((0,0)^\top, \sigma^2 I_2)$ and $N((10,10)^\top, \sigma^2 I_2)$, and dashed circles delimit the 2σ area of each distributions. IL-LIDMVAE performs better. See Figure 4 for more data.

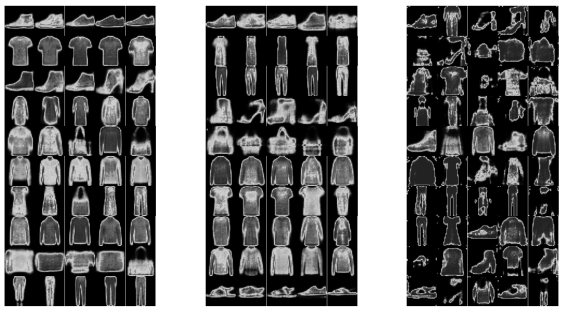
to drive the model to represent the whole data with only one Gaussian distribution even though we set to learn two classes of data. IL-LIDMVAE should be able to avoid this with adequate values of L.

Results are shown in Figure 1. We used the Gaussian mixture version for VAE, called GMVAE (Dilokthanakul et al., 2016). For these complicated settings of high variance, the posterior of LIDVAE ($L = 0.0$) is collapsing on the prior with a relative Fisher divergence of almost 0, and the accuracy becomes lower than other graphs for $\sigma \geq 5$. This supports our claim that LIDVAE can only avoid the exact posterior collapse, and that we need a method that can flexibly avoid any degree of ϵ-posterior collapse. Note, the normal VAE does not perform well at all. On the contrary, IL-LIDVAE can adapt to any degree of posterior collapse as higher values of L achieve higher divergence. This leads in general to better accuracy as it is illustrated in Figure 2.

5.2. Training on Images and Text

We used two image data sets, namely Fashion-MNIST (Xiao et al., 2017) and Omniglot (Lake et al., 2015), and one text data set, namely the synthetic text data provided by Wang et al. (2021). IL-LIDMVAE was applied to the image data, and IL-LIDSVAE to the text data. All distributions were set to Gaussian. We compared the performance in terms of the negative log-likelihood and the KL-divergence between the posterior and the prior. We used RealNVPs (Lake et al., 2015) for images and LSTMs (Hochreiter & Schmidhuber, 1997) for text in order to keep the decoder as flexible as possible. Results are shown in Tables 1 and 2.[1] Further details of the experiments, including additional results, can be found in Appendix B and in the supplementary material.

[1]We did not cite the result of Wang et al. (2021) in Table 2 for Omniglot because they provided values of the negative log-likelihood (around 600) that were too far from ours (around 100) which are of the same order as those reported by other papers such as Tomczak & Welling (2018).



(a) $L_1 = L_2 = 0.0$ (b) $L_1 = L_2 = 1.5$ (c) $L_1 = L_2 = 5.0$

Figure 3: Samples of Fashion-MNIST data generated with different inverse Lipschitz parameters of IL-LIDMVAE with $c = 10$, and all distributions were Gaussian. Each row corresponds to a different category. With $L_1 = L_2 = 1.5$, we obtain the ten true classes with varied images.

As expected, the KL divergence increases in all tables as we augment the value of L and can effectively avoid posterior collapse when it happens. In most cases, the negative log-likelihood is also improved compared to other algorithms. As for Fashion-MNIST, we presented in Figure 3 some samples generated from the trained IL-LIDMVAE. Fashion-MNIST contains 10 classes. We can notice that LIDVAE (a) can outputs high-quality data but cannot learn the category of bags, which collapses on several other classes. On the contrary, IL-LIDVAE that achieves the best loss with $L_1 = L_2 = 1.5$ can reproduce these 10 distinct classes with varied data in each one. With too high inverse Lipschitz constants, IL-LIDVAE cannot output high-quality data since the constraint is too strong and the model cannot find good parameters at all.

For Fashion-MNIST, the classification accuracy of the model based on the posterior of the mixture components was 0.56 for $L = 0$, 0.58 for $L = 0.5$, 0.64 for $L = 1.5$ and 0.12 for $L = 5.0$. Our model is an unsupervised learning method, and the clustering method for Fashion-MNIST using baseline autoencoders achieves an accuracy of 0.54 (Agarap & Azcarraga, 2020), and that using the classical VAE 0.12 in our experiment. As we can observe, the appropriate choice of the inverse Lipschitz constant ($L = 1.5$) improves the accuracy over these methods too.

In short, we can not only control the degree of posterior collapse but also find adequate parameters of inverse Lipschitzness that improve the performance as well.

5.3. Annealing Method

An interesting feature of our method is that we can easily

Table 1: Results for synthetic text data. β was set to 0.2 for β-VAE. The column entitled L refers to the inverse Lipschitz constant of $f_\theta^{(1)}$ and $f_\theta^{(2)}$. NLL stands for negative log-likelihood. See Table 5 for more data.

MODEL	L	NLL	KL
VAE	-	42.56	0.01
β-VAE ($\beta = 0.2$)	-	42.34	0.08
LAGGING VAE	-	45.44	2.13
LIDSVAE	0	56.67	0.24
IL-LIDSVAE	0.5	**40.48**	0.60
IL-LIDSVAE	1.5	44.50	3.80
IL-LIDSVAE	5.0	52.34	8.13
+ANNEALING	-	39.6	0.38

Table 2: Results for Fashion-MNIST (Fashion) and Omniglot. The column entitled L refers to the inverse Lipschitz constant of $f_\theta^{(1)}$ and $f_\theta^{(2)}$. NLL stands for negative log-likelihood. * means that the result is cited from Wang et al. (2021). See Table 4 for more data.

		FASHION		OMNIGLOT	
MODEL	L	NLL	KL	NLL	KL
VAE*	-	258.8	0.2	-	-
SA-VAE*	-	252.2	0.3	-	-
LAGGING VAE*	-	248.5	0.6	-	-
β-VAE* ($\beta = 0.2$)	-	245.3	1.2	-	-
LIDMVAE	0	237.3	9.5	135.0	15.8
IL-LIDMVAE	0.5	240.3	10.0	129.9	20.7
IL-LIDMVAE	1.5	**234.4**	11.0	**126.5**	25.2
IL-LIDMVAE	5.0	243.7	14.3	128.4	26.2
+ANNEALING	-	235.6	8.1	117.7	26.0

consider an annealing scheme that relaxes the inverse Lipschitz constraint when the optimization comes at its boundary (i.e., when the inverse Lipschitz constant of the Brenier map of ICNN becomes close to the set value). That way, we can avoid tuning the inverse Lipschitz constant by hand. Results of this approach are shown in Table 1 and 2, where the negative log-likelihoods are mainly better than results with constant inverse Lipschitz parameters. We only calculated the inverse Lipschitz constant of the first layer $f_\theta^{(1)}$ and estimated it by recycling inputs and outputs acquired during the training, which does not considerably increase the computational complexity. The decrease rate of the inverse Lipschitz constant was set to 0.85. Although we did not evaluate quantitatively, we observed that the final value of the parameter remained around that we found with the best negative log-likelihood without annealing.

6. Conclusion

In conclusion, starting from the recent observation that posterior collapse and latent variable non-identifiability are related to each other, we investigated a method that can guarantee to mitigate any degree of posterior collapse engendered from the formulation of the model itself. This was achieved by introducing an inverse Lipschitz neural network into the decoder that can freely control the degree of latent variable identifiability, and thus that of posterior collapse. Indeed, we theoretically proved that, as this constant increases, the posterior is moved away from the prior in terms of the relative Fisher information divergence in a non-decreasing manner. Based on this theoretical guarantee, we expanded our method to the algorithm IL-LIDVAE and its variants, which are applicable to a broad range of problem settings and can be easily tuned to avoid posterior collapse. We applied them to synthetic and real-world data and showed that they could clearly control the discrepancy between the posterior and the prior in agreement with the theoretical analysis, which was never explicitly realized in any prior work. In most cases, this also had the effect of finding better local minima with lower loss than that of the vanilla VAE.

Acknowledgements

We thank anonymous reviewers for their valuable feedback and advice. KS has been supported in part by JST CREST JPMJCR2015 and JSPS Grant-in-Aid for Transformative Research Areas (A) 22H05106.

References

Agarap, A. F. and Azcarraga, A. P. Improving k-means clustering performance with disentangled internal representations. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2020. doi: 10.1109/IJCNN48605.2020.9207192.

Amos, B., Xu, L., and Kolter, J. Z. Input convex neural networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 146–155. PMLR, 06–11 Aug 2017.

Bakry, D. and Émery, M. Diffusions hypercontractives. In Azéma, J. and Yor, M. (eds.), *Séminaire de Probabilités XIX 1983/84*, pp. 177–206, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg. ISBN 978-3-540-39397-9.

Ball, K. *An Elementary Introduction to Monotone Transportation*, pp. 41–52. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-44489-3. doi: 10.1007/978-3-540-44489-3_5.

Barrett, B., Camuto, A., Willetts, M., and Rainforth, T. Certifiably robust variational autoencoders. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The*

25th International Conference on Artificial Intelligence and Statistics, volume 151 of *Proceedings of Machine Learning Research*, pp. 3663–3683. PMLR, 28–30 Mar 2022.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., and Bengio, S. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1002.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

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

Chen, Y., Shi, Y., and Zhang, B. Optimal control via neural networks: A convex approach. In *International Conference on Learning Representations*, 2019.

Dai, B., Wang, Z., and Wipf, D. The usual suspects? Reassessing blame for VAE posterior collapse. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2313–2322. PMLR, 13–18 Jul 2020.

Dieng, A. B., Kim, Y., Rush, A. M., and Blei, D. M. Avoiding latent variable collapse with generative skip models. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2397–2405. PMLR, 16–18 Apr 2019.

Dilokthanakul, N., Mediano, P. A., Garnelo, M., Lee, M. C., Salimbeni, H., Arulkumaran, K., and Shanahan, M. Deep unsupervised clustering with gaussian mixture variational autoencoders. *International Conference on Learning Representations*, 2016.

Elkhalil, K., Hasan, A., Ding, J., Farsiu, S., and Tarokh, V. Fisher auto-encoders. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 352–360. PMLR, 13–15 Apr 2021.

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., and Carin, L. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 240–250, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1021.

Gross, L. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975. ISSN 00029327, 10806377.

Gulrajani, I., Kumar, K., Ahmed, F., Taiga, A. A., Visin, F., Vazquez, D., and Courville, A. PixelVAE: A latent variable model for natural images. In *International Conference on Learning Representations*, 2017.

Havrylov, S. and Titov, I. Preventing posterior collapse with Levenshtein variational autoencoder. *arXiv preprint arXiv:2004.14758*, 2020.

He, J., Spokoyny, D., Neubig, G., and Berg-Kirkpatrick, T. Lagging inference networks and posterior collapse in variational autoencoders. In *International Conference on Learning Representations*, 2019.

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

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Hoffman, M. D. and Johnson, M. J. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference*, volume 1, 2016.

Holley, R. and Stroock, D. Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46(5):1159–1194, Mar 1987. ISSN 1572-9613. doi: 10.1007/BF01011161.

Holmes, C. C. and Walker, S. G. Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503, 03 2017. ISSN 0006-3444. doi: 10.1093/biomet/asx010.

Huang, C.-W., Tan, S., Lacoste, A., and Courville, A. C. Improving explorability in variational inference with annealed variational objectives. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Huggins, J. H., Campbell, T., Kasprzak, M., and Broderick, T. Practical bounds on the error of Bayesian posterior approximations: A nonasymptotic approach. *arXiv preprint arXiv:1809.09505*, 2018.

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Kim, Y., Wiseman, S., Miller, A., Sontag, D., and Rush, A. Semi-amortized variational autoencoders. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2678–2687. PMLR, 10–15 Jul 2018.

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

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Ledoux, M. Concentration of measure and logarithmic Sobolev inequalities. In Azéma, J., Émery, M., Ledoux, M., and Yor, M. (eds.), *Séminaire de Probabilités XXXIII*, pp. 120–216, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg. ISBN 978-3-540-48407-3.

Li, B., He, J., Neubig, G., Berg-Kirkpatrick, T., and Yang, Y. A surprisingly effective fix for deep latent variable modeling of text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3603–3614, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1370.

Li, Y., Wang, C., Duan, Z., Wang, D., Chen, B., An, B., and Zhou, M. Alleviating "posterior collapse" in deep topic models via policy gradient. In *Advances in Neural Information Processing Systems*, 2022.

Lucas, J., Tucker, G., Grosse, R. B., and Norouzi, M. Don't blame the ELBO! A linear VAE perspective on posterior collapse. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Lyu, S. Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pp. 359–366, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

Otto, F. and Villani, C. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000. ISSN 0022-1236. doi: https://doi.org/10.1006/jfan.1999.3557.

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

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32(2) of *Proceedings of Machine Learning Research*, pp. 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. Ladder variational autoencoders. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Tomczak, J. and Welling, M. Vae with a VampPrior. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1214–1223. PMLR, 09–11 Apr 2018.

van den Oord, A., Vinyals, O., and kavukcuoglu, k. Neural discrete representation learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Vempala, S. and Wibisono, A. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *Advances in neural information processing systems*, volume 32, 2019.

Walker, S. G. Bayesian information in an experiment and the Fisher information distance. *Statistics & Probability Letters*, 112:5–9, 2016. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2016.01.014.

Wang, Y., Blei, D., and Cunningham, J. P. Posterior collapse and latent variable non-identifiability. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 5443–5455. Curran Associates, Inc., 2021.

Wang, Z. and Ziyin, L. Posterior collapse of a linear latent variable model. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yamaguchi, S. and Koyama, M. Distributional concavity regularization for GANs. In *International Conference on Learning Representations*, 2019.

Yang, Y., Martin, R., and Bondell, H. Variational approximations using Fisher divergence. *arXiv preprint arXiv:1905.05284*, 2019.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R. R., and Chaudhuri, K. A closer look at accuracy vs. robustness. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8588–8601. Curran Associates, Inc., 2020.

Yang, Z., Hu, Z., Salakhutdinov, R., and Berg-Kirkpatrick, T. Improved variational autoencoders for text modeling using dilated convolutions. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3881–3890. PMLR, 06–11 Aug 2017.

Yeung, S., Kannan, A., Dauphin, Y., and Fei-Fei, L. Tackling over-pruning in variational autoencoders. *arXiv preprint arXiv:1706.03643*, 2017.

Zhao, T., Lee, K., and Eskenazi, M. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1098–1107, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1101.

Zhao, Y., Yu, P., Mahapatra, S., Su, Q., and Chen, C. Improve variational autoencoder for text generation with discrete latent bottleneck. *arXiv preprint arXiv:2004.10603*, 2020.

A. Proof of Theorems and Propositions

In this appendix, we will prove the theorems stated in the main paper.

A.1. Proof of Theorem 3.4

Let us first state some simple lemmas to prove Theorem 3.4. While the proofs may seem simple, we will still show most of them in order to keep our paper self-contained.

Lemma A.1. *The following holds between the prior $p(z)$, the posterior $p_\theta(z \mid x)$, the likelihood $p_\theta(x \mid z)$ and the marginal likelihood $p_\theta(x)$:*

$$\|\nabla_z \log p_\theta(z \mid x) - \nabla_z \log p(z)\| = \|\nabla_z \log p_\theta(x \mid z)\|.$$

Proof. Bayes' theorem implies

$$p_\theta(x \mid z) = \frac{p_\theta(x, z)}{p(z)} = \frac{p_\theta(z \mid x) p_\theta(x)}{p(z)}.$$

Taking logarithms on both sides,

$$\log p_\theta(x \mid z) = \log p_\theta(z \mid x) + \log p_\theta(x) - \log p(z),$$

which leads to

$$\log p_\theta(x \mid z) - \log p_\theta(x) = \log p_\theta(z \mid x) - \log p(z).$$

Now, differentiating with respect to z and taking the norm gives the desired equality since $p_\theta(x)$ is independent of z. $Q.E.D$

Remark A.2. Note this theorem does not need any specification on the class of distributions.

The following lemma is a fundamental property of the exponential family.

Lemma A.3. *The following holds for the exponential family between the sufficient statistic and log-partition function:*

$$\mathbb{E}_x[T(x)] = \nabla_z A(z).$$

Furthermore, the generation of inverse-Lipschitz functions by Brenier maps implies the following property.

Lemma A.4. *If $f : \mathbb{R}^l \to \mathbb{R}^l$ is an L-inverse Lipschitz function generated by an L-strongly convex real-valued function B, then the following holds:*

$$\nabla f(x) \succeq L I_l,$$

where I_l is the $l \times l$ identity matrix, and $A \succeq B$ means that $A - B$ is positive semi-definite.

Proof. Since $f = \nabla B$ and $\nabla f = \nabla^2 B$, by definition of L-strong convexity, we immediately get $\nabla f(x) \succeq L I_l$. $Q.E.D$

Remark A.5. This statement can be proved thanks to the use of Brenier maps in order to create inverse Lipschitz functions.

Finally, we can prove Theorem 3.4.

Theorem 3.4. *Under model (1), Assumption 3.2 and $l = t$, the following holds for all i and $\theta \in \Theta_L$:*

$$F(p_\theta(z \mid x_i) \mid\mid p(z)) \geq L^2 \int \|T(x_i) - \mathbb{E}_{p_\theta(x|z)}[T(x)]\|^2 p(z) \mathrm{d}z. \tag{5}$$

Proof. Since

$$F(p_\theta(z \mid x) \mid\mid p(z)) = \int \|\nabla_z \log p_\theta(z \mid x) - \nabla_z \log p(z)\|^2 p(z) \mathrm{d}z,$$

Previous lemmas imply

$$\begin{aligned}
F(p_\theta(z \mid x) \mid\mid p(z)) &= \int \|\nabla_z \log p_\theta(x \mid z)\|^2 p(z)\mathrm{d}z \\
&= \int \|\nabla_z \left(\log h(x) + f_\theta(z)^\top T(x) - A(f_\theta(z))\right)\|^2 p(z)\mathrm{d}z \\
&= \int \|\nabla_z f_\theta(z)^\top T(x) - \nabla_z A(f_\theta(z))\|^2 p(z)\mathrm{d}z \\
&= \int \|\nabla_z f_\theta(z)^\top T(x) - \nabla_z f_\theta(z)^\top \nabla A(f_\theta(z))\|^2 p(z)\mathrm{d}z \\
&= \int \|\nabla_z f_\theta(z)^\top \left(T(x) - \nabla A(f_\theta(z))\right)\|^2 p(z)\mathrm{d}z \\
&= \int \left(T(x) - \nabla A(f_\theta(z))\right)^\top \nabla_z f_\theta(z) \nabla_z f_\theta(z)^\top \left(T(x) - \nabla A(f_\theta(z))\right) p(z)\mathrm{d}z \\
&\geq L^2 \int \|T(x) - \nabla A(f_\theta(z))\|^2 p(z)\mathrm{d}z \\
&= L^2 \int \|T(x) - \mathbb{E}_{p_\theta(x|z)}[T(x)]\|^2 p(z)\mathrm{d}z.
\end{aligned}$$

We used Lemma A.1 for the first equality, Lemma A.4 for the inequality, and Lemma A.3 for the last equality. $Q.E.D$

Corollary 3.6. *Under model (1), Assumption 3.2 and $l = t$,*

$$F(p_\theta(z \mid x_i) \mid\mid p(z)) \geq L^2 \inf_{\theta \in \Theta_L} \left\{ \int \|T(x_i) - \mathbb{E}_{p_\theta(x|z)}[T(x)]\|^2 p(z)\mathrm{d}z \right\}.$$

Therefore, the lower bound is non-decreasing in terms of L. Moreover, if the infinimum of this lower bound is attained by a parameter $\theta \in \Theta_L$ and $p(z)$ has a positive variance, then the lower bound is monotonically increasing in terms of L.

Proof. The infimum immediately follows from equation (5) which holds for all $\theta \in \Theta_L$. Now, since if $L \geq L'$, then $\Theta_L \supset \Theta'_L$ by definition of inverse Lipschitzness, the infimum is non-decreasing in terms of L.

Concerning the second half of the statement, let us suppose that the infinimum can be attained by a parameter $\theta \in \Theta_L$ and that $p(z)$ has a positive variance. It is enough to show that the infinimum is not 0, or in other words, that the lower bound is not vacuous. However if

$$\int \|T(x_i) - \mathbb{E}_{p_\theta(x|z)}[T(x)]\|^2 p(z)\mathrm{d}z = 0,$$

this means the random variable $\mathbb{E}_{p_\theta(x|z)}[T(x)]$ satisfies $\mathbb{E}_{p_\theta(x|z)}[T(x)] = T(x_i)$ for all z, which implies $\mathbb{E}_{p_\theta(x|z)}[T(x)]$ is constant. This is only possible if $\mathbb{E}_{p_\theta(x|z)}[T(x)]$ does not depend on z, which means $p_\theta(x \mid z)$ is independent of z. This contradicts our definition of f_θ which is injective. Therefore, the infimum is positive, and the lower bound is, in consequence, increasing in terms of L. $Q.E.D$

A.2. Proof of Theorem 3.7

Theorem 3.7. *Under model (1), Assumption 3.2 and $l = t$, the following holds for all $\theta \in \Theta_L$:*

$$\bar{F}_\theta(\boldsymbol{x}) := \int \left\| \frac{1}{n} \sum_{i=1}^n \nabla_z \log p_\theta(z \mid x_i) - \nabla_z \log p(z) \right\|^2 p(z)\mathrm{d}z \geq L^2 \int \left\| \frac{1}{n} \sum_{i=1}^n T(x_i) - \mathbb{E}_{p_\theta(x|z)}[T(x)] \right\|^2 p(z)\mathrm{d}z.$$

Proof. It suffices to note that

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \nabla_z \log p_\theta(z \mid x_i) &= \frac{1}{n} \sum_{i=1}^n \nabla_z \left(\log h(x_i) + f_\theta(z)^\top T(x_i) - A(f_\theta(z))\right) \\
&= \nabla_z \left(f_\theta(z)^\top \frac{1}{n} \sum_{i=1}^n T(x_i) - A(f_\theta(z)) \right).
\end{aligned}$$

$Q.E.D$

A.3. Proof of Theorem 3.8

This subsection considers IL-LIDVAE in the general case where $t > l$, i.e., the latent dimension l is smaller than that of the sufficient statistic of x (Definition 4.1), and discusses the lower bound of the relative Fisher information divergence between the prior and posterior.

Let us first remind Assumption 3.2 and the problem setting when $t > l$. $\mathrm{EF}(x \mid \xi)$ is an exponential family defined by

$$\mathrm{EF}_T(x \mid \xi) = \exp\{T(x)^\top \xi - A(\xi)\}h(x),$$

where $T(x) = (T_1(x), \ldots, T_t(x))$ is a set of sufficient statistics, ξ a natural parameter, and $h(x)$ a base probability density. We assume $h(x) > 0$ and that the natural parameter ξ is defined on an open set in $\mathbb{R}^l$. The IL-LIDVAE model is defined by $p_\theta(x \mid z) = \mathrm{EF}_T(x \mid f_\theta(z))$, where

$$f_\theta(z) = f_\theta^{(2)}(B^\top f_\theta^{(1)}(z))$$

which is given by inverse Lipschitz functions $f_\theta^{(1)}$ and $f_\theta^{(2)}$ with constant L_1 and L_2, respectively. The model can be regarded as a curved exponential family parameterized with $f_\theta(z)$.

We make a natural assumption on the exponential family.

Assumption A.6. The mapping from the natural parameter to the expectation parameter

$$\xi \mapsto \nabla_\xi A(\xi) = \mathbb{E}_{\mathrm{EF}_T(x|\xi)}[T(x)]$$

is a diffeomorphism.

In light of the relation

$$\nabla_\xi \log \mathrm{EF}_T(x \mid \xi) = T(x) - \nabla_\xi A(\xi) = T(x) - \mathbb{E}_{\mathrm{EF}_T(x|\xi)}[T(x)],$$

the above assumption means that the natural parameter ξ effectively changes the density function for any direction of the parameter space. This holds for many popular exponential families, such as Gaussian distributions.

Under this problem setting, we can prove Theorem 3.8, which is restated more explicitly.

Theorem A.7 (Theorem 3.8 restated)**.** *Under Assumptions 3.2 and A.6, and $l = t$,*

$$F(p_\theta(z|x)||p(z)) \geq L_1^2 \inf_{\theta \in \Theta_{L_1,L_2}} \left\{ \int \left\| \left(T(x) - \mathbb{E}_{p_\theta(x|z)}[T(x)]\right)^\top \nabla f_\theta^{(2)}\left(B^\top f_\theta^{(1)}(z)\right) B^\top \right\|^2 p(z)dz \right\},$$

where the right-hand side of the inequality is increasing in terms of L_1.

Proof. As in the proof of Theorem 3.4, the relative Fisher information divergence $F(p_\theta(z|x)||p(z))$ is lower bounded by

$$\begin{aligned}
F(p_\theta(z|x)||p(z)) &= \int \|\nabla_z \log p_\theta(x \mid z)\|^2 p(z)dz \\
&= \int \|\nabla_\xi \log \mathrm{EF}_T(x \mid \xi)|_{\xi = f_\theta(z)} \nabla_z f_\theta(z)\|^2 p(z)dz \\
&= \int \left\| \left(T(x) - \mathbb{E}_{p_\theta(x|z)}[T(x)]\right)^\top \nabla f_\theta^{(2)}\left(B^\top f_\theta^{(1)}(z)\right) B^\top \nabla f_\theta^{(1)}(z)) \right\|^2 p(z)dz \\
&\geq L_1^2 \int \left\| \left(T(x) - E_{p_\theta(x|z)}[T(x)]\right)^\top \nabla f_\theta^{(2)}\left(B^\top f_\theta^{(1)}(z)\right) B^\top \right\|^2 p(z)dz.
\end{aligned}$$

To guarantee the lower boundedness of the divergence with the control by the inverse-Lipschitz constant L_1, the integral in the last line should be positive. We consider this positiveness under the assumption that the density of the prior $p(z)$ is everywhere positive and continuous. In this case, the integral is positive if and only if the latent space has an open set on which

$$\left(T(x) - \mathbb{E}_{p_\theta(x|z)}[T(x)]\right)^\top \nabla f_\theta^{(2)}\left(B^\top f_\theta^{(1)}(z)\right) B^\top \neq 0.$$

Because $\nabla f_\theta^{(1)}$ is invertible, this is equivalent to

$$\nabla_z \log p_\theta(x \mid z) \neq 0$$

on that open set. This holds under the assumption because the parameter $\xi = f_\theta(z)$ moves t-dimensional directions as z changes, and it in turns changes $\log p_\theta(x \mid z)$ by Assumption A.6. This implies the desired result. $Q.E.D$

Table 3: Details of Experiments

DATA	MODEL	DECODER	LATENT DIMENSION	SIZE OF HIDDEN LAYERS
TOY	IL-LIDMVAE ($c = 2$)	GAUSSIAN	2	10
FASHION-MNIST	IL-LIDMVAE ($c = 10$)	REALNVP (2 LAYERS)	64	512
OMNIGLOT	IL-LIDMVAE ($c = 50$)	REALNVP (2 LAYERS)	32	200
SYNTHETIC	IL-LIDSVAE	LSTM (2 LAYERS)	1024	1024

A.4. Proof of Proposition 3.9

Proposition 3.9. *Suppose that the lower bound of Fisher divergence $F(p(x)||q(x)) \geq \varepsilon$ holds for any small perturbations of p and q to some extent. More precisely, let p_t (or q_t) denote the convolution between p (q, resp.) and $N(0,t)$. Assume that there is $\delta > 0$ such that $F(p_t||q_t) \geq \epsilon$ for any $t \in [0, \delta]$. Then, the bound $D(p||q) \geq \frac{1}{2}\delta\epsilon$ holds.*

Proof. This is a straightforward consequence of the well-known de Bruijn's identity (Lyu, 2009) about the relation between KL and Fisher divergences: $\frac{\mathrm{d}}{\mathrm{d}t}D(p_t||q_t) = -\frac{1}{2}F(p_t||q_t)$. By integrating both sides, we obtain $D(p_\delta||q_\delta) - D(p||q) = -\frac{1}{2}\int_0^\delta F(p_t||q_t)\mathrm{d}t$, and this implies $D(p||q) \geq \frac{1}{2}\int_0^\delta F(p_t||q_t)dt \geq \frac{1}{2}\delta\epsilon$, which concludes the assertion. $Q.E.D$

B. Details of Experiments

In this appendix, we provide further details on experiments conducted in this paper.

B.1. Details of Data and Experiments

Image For image data sets, we used the IL-LIDMVAE (Definition 4.2) with the exponential family defined as Gaussian distribution. For all other methods, we used their Gaussian mixture variant as GMVAE for VAE. The number of categories c was set to the true number of classes of the data. Fashion-MNIST contains 10 classes (T-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag and ankle boot), and Omniglot 50. We parameterized the prior distribution as well. The two Lipschitz constants L_1 and L_2 were equal throughout the experiments. Further details can be found in Table 3.

Text For the text data, we used IL-LIDSVAE (Definition 4.3). The synthetic data set was generated from a two-layer sequential VAE with five-dimensional latent variables by Wang et al. (2021). The two Lipschitz constants L_1 and L_2 were equal throughout the experiments. Further details can be found in Table 3.

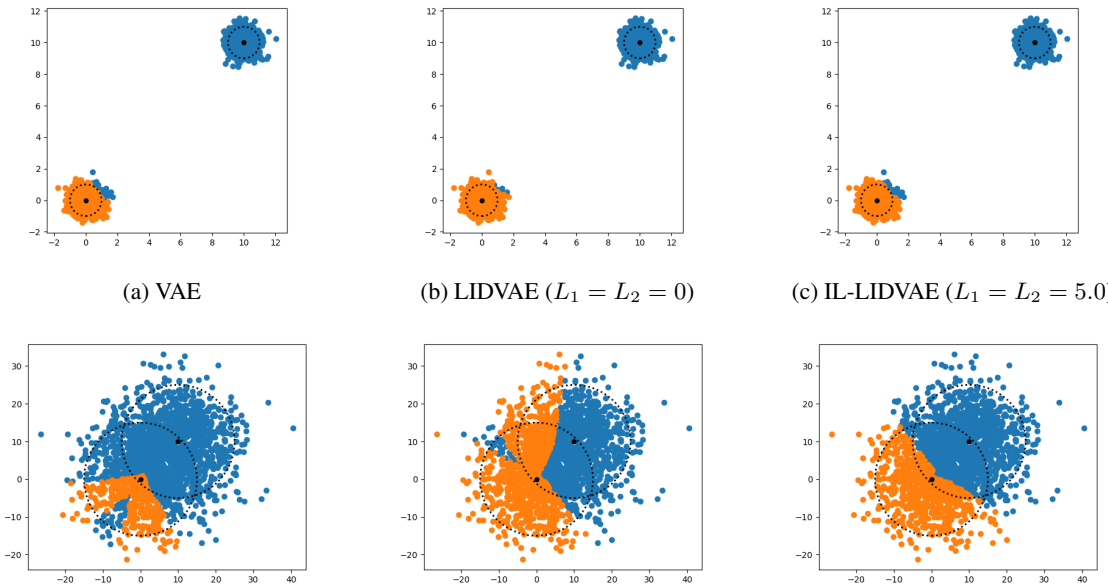
B.2. Further Results of Experiments

B.2.1. TOY DATA

We show in Figure 4 the posterior of some trained models. All models perform well for moderate values of σ, but only IL-LIDVAE can adapt to the more extreme cases. Data were generated from $N((0,0)^\top, \sigma^2 I_2)$ and $N((10,10)^\top, \sigma^2 I_2)$. Therefore, data should be separated along the mediator of the segment connecting $(0,0)^\top$ and $(10,10)^\top$. IL-LIDVAE with $L_1 = L_2 = 5.0$ is the closest to this situation.

B.2.2. IMAGES

We show in Table 4 additional data with different Lipschitz constants. The mutual information (MI) between the data and the latent variables (Hoffman & Johnson, 2016) and the percentage of active units (AU) (Burda et al., 2015), two alternatives measure of posterior collapse, are also presented in the table. For the percentage of active units, the threshold was set to 0.01. We also show some reconstruction and sampling conducted with the trained models of IL-LIDVAE in Figures 5 and 6, respectively. While the reconstruction performance does not change a lot, we can observe that the sampling quality declines with too high values of L_1 and L_2. $L_1 = L_2 = 1.5$ achieves the best loss and is the only one that learned the ten true distinct classes.



(a) VAE (b) LIDVAE ($L_1 = L_2 = 0$) (c) IL-LIDVAE ($L_1 = L_2 = 5.0$)

(d) VAE (e) LIDVAE ($L_1 = L_2 = 0$) (f) IL-LIDVAE ($L_1 = L_2 = 5.0$)

Figure 4: Posterior of VAE (left), LIDVAE (middle) and IL-LIDVAE (right) for the toy data with different standard deviations. $\sigma = 0.5$ (top) and $\sigma = 7.5$ (bottom). The black points are the means of $N((0,0)^\top, \sigma^2 I_2)$ and $N((10,10)^\top, \sigma^2 I_2)$, and the dashed circles delimit the 2σ area of each distributions.

B.2.3. TEXT

We show in Table 5 additional data with different Lipschitz constants.

C. Log-Sobolev Inequality

In this appendix, we briefly describe the Log-Sobolev inequality, mentioned in the main paper, and some well-known properties as well. We only treat distributions absolutely continuous with respect to the Lebesgue measure for simplicity.

Definition C.1. Distribution ν satisfies the *Log-Sobolev inequality (LSI) with a constant α if for all probability density functions ρ absolutely continuous with respect to ν, the following holds:*

$$D(\rho \,||\, \nu) \leq \frac{1}{2\alpha} F(\rho \,||\, \nu),$$

where $D(\rho \,||\, \nu) = \mathbb{E}_\rho \left[\log \frac{\rho}{\nu}\right]$ is the KL-divergence of ρ with respect to ν, and $F(\rho \,||\, \nu) = \mathbb{E}_\rho \left[\left\|\nabla \log \frac{\rho}{\nu}\right\|^2\right]$ is the relative Fisher information divergence of ρ with respect to ν.

The Gaussian distribution satisfies LSI as implied by the following proposition.

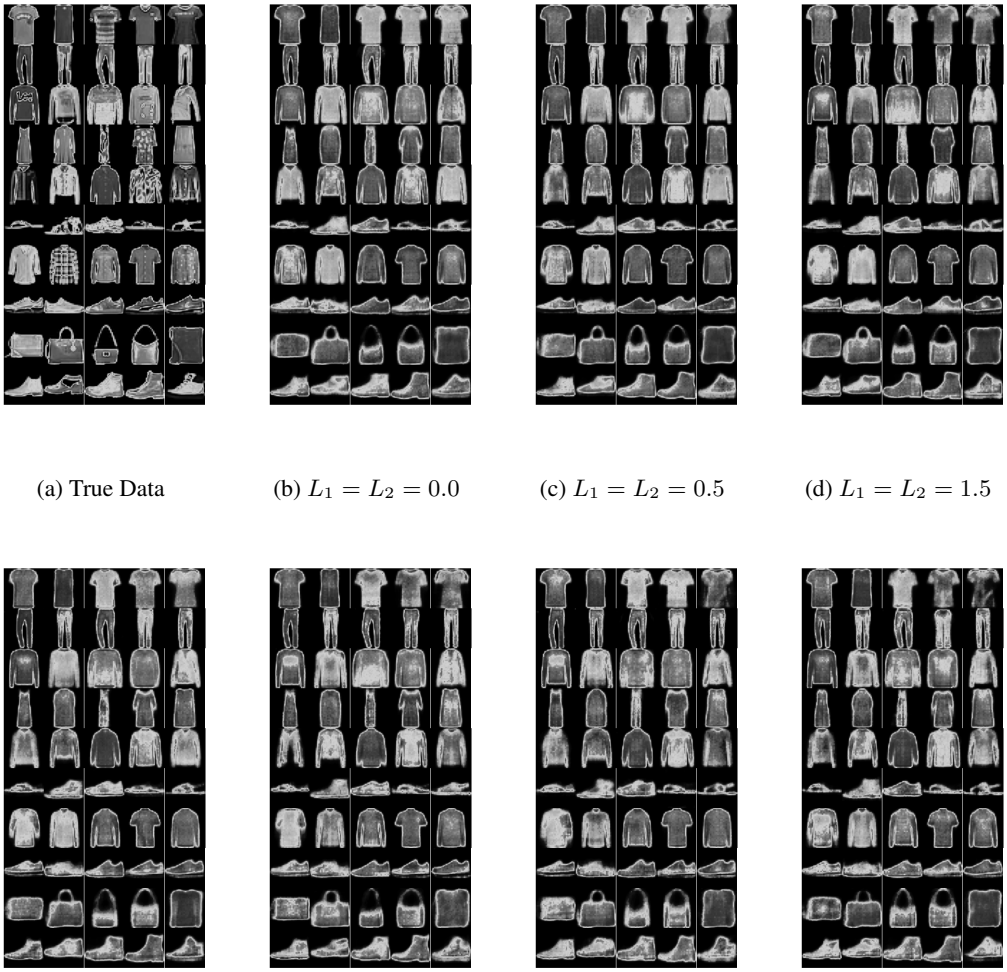
Proposition C.2 (Bakry & Émery (1985))**.** *Suppose $q \propto \mathrm{e}^{-f}$ is a probability density, where $f : \mathbb{R}^d \to \mathbb{R}$ is a smooth function. If there exists a constant $c > 0$ such that $\nabla^2 f \succeq c I_d$, then $q(z)dz$ satisfies LSI with constant c.*

Therefore, if we are only using Gaussian distributions in the VAE, then posterior collapse in terms of the relative Fisher information divergence results in posterior collapse in terms of KL-divergence.

The following statements show that LSI is robust under bounded perturbations and Lipschitz mappings.

Proposition C.3 (Holley & Stroock (1987))**.** *Suppose q is a probability density that satisfies LSI with constant α. For any bounded function $B : \mathbb{R}^d \to \mathbb{R}$, $q_B \propto \mathrm{e}^B q$ satisfies LSI with constant $\alpha \mathrm{e}^{-4\|B\|_\infty}$.*

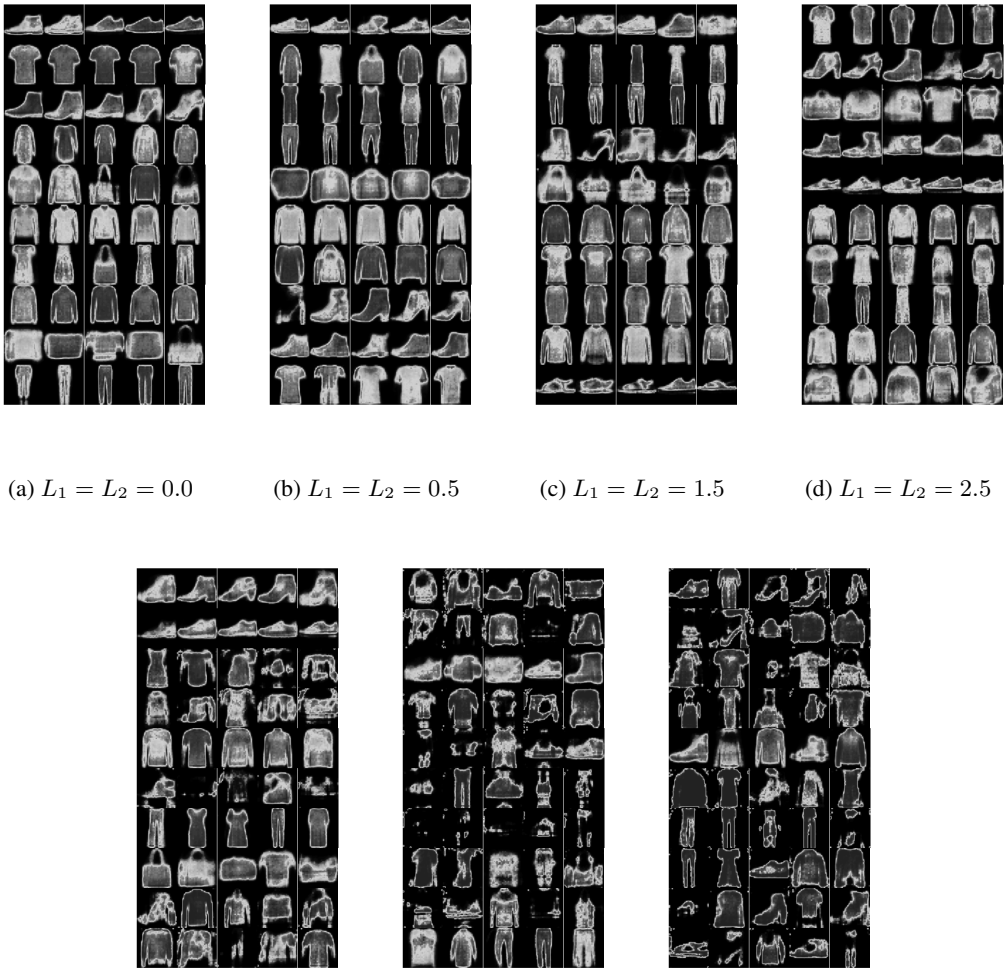
Proposition C.4 (Vempala & Wibisono (2019))**.** *Suppose q is a probability density that satisfies LSI with constant α. If*



(a) True Data (b) $L_1 = L_2 = 0.0$ (c) $L_1 = L_2 = 0.5$ (d) $L_1 = L_2 = 1.5$

(e) $L_1 = L_2 = 2.5$ (f) $L_1 = L_2 = 3.5$ (g) $L_1 = L_2 = 4.5$ (h) $L_1 = L_2 = 5.0$

Figure 5: Reconstruction of randomly chosen data of Fashion-MNIST for different inverse Lipschitz parameters of IL-LIDMVAE. The number of classes was set to 10, and all distributions were Gaussian. Each row corresponds to a different category.



(a) $L_1 = L_2 = 0.0$ (b) $L_1 = L_2 = 0.5$ (c) $L_1 = L_2 = 1.5$ (d) $L_1 = L_2 = 2.5$

(e) $L_1 = L_2 = 3.5$ (f) $L_1 = L_2 = 4.5$ (g) $L_1 = L_2 = 5.0$

Figure 6: Samples of Fashion-MNIST data generated with different inverse Lipschitz parameters of IL-LIDMVAE. The number of classes was set to 10, and all distributions were Gaussian. Each row corresponds to a different category. With $L_1 = L_2 = 1.5$, we obtain the ten true distinct classes with varied images.

Table 4: Results for Fashion-MNIST (Fashion) and Omniglot. The column entitled L refers to the inverse Lipschitz constant of $f_\theta^{(1)}$ and $f_\theta^{(2)}$. NLL stands for negative log-likelihood, MI for mutual information and AU for active units. * means that the result is cited from Wang et al. (2021).

		FASHION				OMNIGLOT			
MODEL	L	NLL	KL	MI	AU	NLL	KL	MI	AU
VAE*	-	258.8	0.2	0.9	0.1	-	-	-	-
SA-VAE*	-	252.2	0.3	1.3	0.2	-	-	-	-
LAGGING VAE*	-	248.5	0.6	1.6	0.4	-	-	-	-
β-VAE* ($\beta = 0.2$)	-	245.3	1.2	2.4	0.6	-	-	-	-
LIDVAE	0	237.3	9.5	8,7	1.0	135.0	15.8	15.2	1.0
IL-LIDVAE	0.5	240.3	10.0	9.4	1.0	129.9	20.7	20.2	1.0
IL-LIDVAE	1.5	**234.4**	11.0	10.7	1.0	126.5	25.2	24.8	1.0
IL-LIDVAE	2.5	236.0	12.6	12.5	1.0	**126.2**	25.9	25.3	1.0
IL-LIDVAE	3.5	239.7	13.0	12.9	0.5	126.5	26.3	26.2	1.0
IL-LIDVAE	4.5	241.7	13.2	13.2	0.4	127.7	26.3	26.2	1.0
IL-LIDVAE	5.0	243.7	14.3	14.2	0.3	128.4	26.2	26.1	1.0
+ANNEALING	-	235.6	8.1	7.8	1.0	117.7	26.0	25.8	1.0

Table 5: Results for synthetic text data. β was set to 0.2 for β-VAE. The column entitled L refers to the inverse Lipschitz constant of $f_\theta^{(1)}$ and $f_\theta^{(2)}$. NLL stands for negative log-likelihood, MI for mutual information and AU for active units.

MODEL	L	NLL	KL	MI	AU
VAE	-	42.56	0.01	0.0	0.0
β-VAE ($\beta = 0.2$)	-	42.34	0.08	0.0	0.0
LAGGING VAE	-	45.44	2.13	1.0	1.0
LIDVAE	0	56.67	0.24	0.2	0.8
IL-LIDVAE	0.2	**40.39**	0.32	0.3	1.0
IL-LIDVAE	0.4	40.48	0.39	0.3	1.0
IL-LIDVAE	0.5	40.48	0.60	0.5	1.0
IL-LIDVAE	0.6	41.65	0.85	0.6	1.0
IL-LIDVAE	1.0	44.06	2.45	0.8	1.0
IL-LIDVAE	1.5	44.50	3.80	0.9	1.0
IL-LIDVAE	5.0	52.34	8.13	1.3	1.0
+ANNEALING	-	39.6	0.38	0.1	1.0

$H : \mathbb{R}^d \to \mathbb{R}^d$ is a differentiable L-Lipschitz mapping, then the distribution of $H(z)$ with $z \sim q(z)$ satisfies LSI with constant α/L^2.