DIAGNOSING AND ENHANCING VAE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Although variational autoencoders (VAEs) represent a widely influential deep generative model, many aspects of the underlying energy function remain poorly understood. In particular, it is commonly believed that Gaussian encoder/decoder assumptions reduce the effectiveness of VAEs in generating realistic samples. In this regard, we rigorously analyze the VAE objective, differentiating situations where this belief is and is not actually true. We then leverage the corresponding insights to develop a simple VAE enhancement that requires no additional hyperparameters or sensitive tuning. Quantitatively, this proposal produces crisp samples and stable FID scores that are actually competitive with state-of-the-art GAN models, all while retaining desirable attributes of the original VAE architecture.

1 INTRODUCTION

Our starting point is the desire to learn a probabilistic generative model of observable variables $\boldsymbol{x} \in \boldsymbol{\chi}$, where $\boldsymbol{\chi}$ is a r-dimensional manifold embedded in $\mathbb{R}^d$. Note that if $r = d$, then this assumption places no restriction on the distribution of $\boldsymbol{x} \in \mathbb{R}^d$ whatsoever; however, the added formalism is introduced to handle the frequently encountered case where $\boldsymbol{x}$ possesses low-dimensional structure relative to a high-dimensional ambient space, i.e., $r \ll d$. In fact, the very utility of generative models, and their attendant low-dimensional representations, often hinges on this assumption (Bengio et al., 2013). It therefore behooves us to explicitly account for this situation.

Beyond this, we assume that $\boldsymbol{\chi}$ is a simple Riemannian manifold, which means there exists a diffeomorphism φ between $\boldsymbol{\chi}$ and $\mathbb{R}^r$, or more explicitly, the mapping $\varphi : \boldsymbol{\chi} \mapsto \mathbb{R}^r$ is invertible and differentiable. Denote a ground-truth probability measure on $\boldsymbol{\chi}$ as μ_{gt} such that the probability mass of an infinitesimal $d\boldsymbol{x}$ on the manifold is $\mu_{gt}(d\boldsymbol{x})$ and $\int_{\boldsymbol{\chi}} \mu_{gt}(d\boldsymbol{x}) = 1$.

The variational autoencoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014) attempts to approximate this ground-truth measure using a parameterized density $p_\theta(\boldsymbol{x})$ defined across all of $\mathbb{R}^d$ since any underlying generative manifold is unknown in advance. This density is further assumed to admit the latent decomposition $p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$, where $\boldsymbol{z} \in \mathbb{R}^\kappa$ serves as a low-dimensional representation, with $\kappa \approx r$ and prior $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$.

Ideally we might like to minimize the negative log-likelihood $-\log p_\theta(\boldsymbol{x})$ averaged across the ground-truth measure μ_{gt}, i.e., solve $\min_\theta \int_{\boldsymbol{\chi}} -\log p_\theta(\boldsymbol{x})\mu_{gt}(d\boldsymbol{x})$. Unfortunately though, the required marginalization over $\boldsymbol{z}$ is generally infeasible. Instead the VAE model relies on tractable *encoder* $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and *decoder* $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ distributions, where ϕ represents additional trainable parameters. The canonical VAE cost is a bound on the average negative log-likelihood given by

$$\mathcal{L}(\theta, \phi) \triangleq \int_{\boldsymbol{\chi}} \left\{-\log p_\theta(\boldsymbol{x}) \; + \; \mathbb{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})\right]\right\} \mu_{gt}(d\boldsymbol{x}) \geq \int_{\boldsymbol{\chi}} -\log p_\theta(\boldsymbol{x})\mu_{gt}(d\boldsymbol{x}), \quad (1)$$

where the inequality follows directly from the non-negativity of the KL-divergence. Here ϕ can be viewed as tuning the tightness of bound, while θ dictates the actual estimation of μ_{gt}. Using a few standard manipulations, this bound can also be expressed as

$$\mathcal{L}(\theta, \phi) = \int_{\boldsymbol{\chi}} \left\{-\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta\left(\boldsymbol{x}|\boldsymbol{z}\right)\right] + \mathbb{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right]\right\} \mu_{gt}(d\boldsymbol{x}), \quad (2)$$

which explicitly involves the encoder/decoder distributions and is conveniently amenable to SGD optimization of $\{\theta, \phi\}$ via a reparameterization trick (Kingma & Welling, 2014; Rezende et al., 2014). The first term in (2) can be viewed as a reconstruction cost (or a stochastic analog of a traditional autoencoder), while the second penalizes posterior deviations from the prior $p(\boldsymbol{z})$. Additionally, for any realizable implementation via SGD, the integration over $\boldsymbol{\chi}$ must be approximated via a finite sum across training samples $\{\boldsymbol{x}^{(i)}\}_{i=1}^n$ drawn from μ_{gt}. Nonetheless, examining the true objective $\mathcal{L}(\theta, \phi)$ can lead to important, practically-relevant insights.

At least in principle, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ can be arbitrary distributions, in which case we could simply enforce $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = p_\theta(\boldsymbol{z}|\boldsymbol{x}) \propto p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$ such that the bound from (1) is tight. Unfortunately though, this is essentially always an intractable undertaking. Consequently, largely to facilitate practical implementation, the most commonly adopted distributional assumption is that both $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ are Gaussian. This design choice has previously been cited as a key limitation of VAEs (Burda et al., 2015; Kingma et al., 2016), and existing quantitative tests of generative modeling quality thus far dramatically favor contemporary alternatives such as generative adversarial networks (GAN) (Goodfellow et al., 2014b). Regardless, because the VAE possesses certain desirable properties relative to GAN models (e.g., stable training (Tolstikhin et al., 2018), interpretable encoder/inference network (Brock et al., 2016), outlier-robustness (Dai et al., 2018), etc.), it remains a highly influential paradigm worthy of examination and enhancement.

In Section 2 we closely investigate the implications of VAE Gaussian assumptions leading to a number of interesting diagnostic conclusions. In particular, we differentiate the situation where $r = d$, in which case we prove that recovering the ground-truth distribution is actually possible iff the VAE global optimum is reached, and $r < d$, in which case the VAE global optimum can be reached by solutions that reflect the ground-truth distribution almost everywhere, but not necessarily uniquely so. In other words, there could exist alternative solutions that both reach the global optimum and yet do not assign the same probability measure as μ_{gt}.

Section 3 then further probes this non-uniqueness issue by inspecting necessary conditions of global optima when $r < d$. This analysis reveals that an optimal VAE parameterization will provide an encoder/decoder pair capable of perfectly reconstructing all $\boldsymbol{x} \in \boldsymbol{\chi}$ using *any* $\boldsymbol{z}$ drawn from $q_\phi(\boldsymbol{z}|\boldsymbol{x})$. Moreover, we demonstrate that the VAE accomplishes this using a degenerate latent code whereby only r dimensions are effectively active. Collectively, these results indicate that the VAE global optimum can in fact uniquely learn a mapping to the correct ground-truth manifold when $r < d$, but not necessarily the correct probability measure *within* this manifold, a critical distinction.

Next we leverage these analytical results in Section 4 to motivate an almost trivially-simple, two-stage VAE enhancement for addressing typical regimes when $r < d$. In brief, the first stage just learns the manifold per the allowances from Section 3, and in doing so, provides a mapping to a lower dimensional intermediate representation with no degenerate dimensions that mirrors the $r = d$ regime. The second (much smaller) stage then only needs to learn the correct probability measure on this intermediate representation, which is possible per the analysis from Section 2. Experiments from Section 5 reveal that this procedure can generate high-quality crisp samples, avoiding the blurriness often attributed to VAE models in the past (Dosovitskiy & Brox, 2016; Larsen et al., 2015). And to the best of our knowledge, this is the first demonstration of a VAE pipeline that can produce stable FID scores, an influential recent metric for evaluating generated sample quality (Heusel et al., 2017), that equal or exceed even multiple state-of-the-art GAN models. Moreover, this is accomplished without additional penalty functions, cost function modifications, or sensitive tuning parameters.

2 High-Level Impact of VAE Gaussian Assumptions

Conventional wisdom suggests that VAE Gaussian assumptions will introduce a gap between $\mathcal{L}(\theta, \phi)$ and the ideal negative log-likelihood $\int_{\boldsymbol{\chi}} -\log p_\theta(\boldsymbol{x})\mu_{gt}(d\boldsymbol{x})$, compromising efforts to learn the ground-truth measure. However, we will now argue that this pessimism is in some sense premature. In fact, we will demonstrate that, even with the stated Gaussian distributions, there exist parameters ϕ and θ that can simultaneously: (*i*) Globally optimize the VAE objective and, (*ii*) Recover the ground-truth probability measure in a certain sense described below. This is possible because, at least for some coordinated values of ϕ and θ, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ can indeed become arbitrarily close. Before presenting the details, we first formalize a κ-simple VAE, which is merely a VAE model with explicit Gaussian assumptions and parameterizations:

Definition 1 *A κ-simple VAE is defined as a VAE model with $\dim[\boldsymbol{z}] = \kappa$ latent dimensions, the Gaussian encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$, and the Gaussian decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$. Moreover, the encoder moments are defined as $\boldsymbol{\mu}_z = f_{\mu_z}(\boldsymbol{x};\phi)$ and $\boldsymbol{\Sigma}_z = \boldsymbol{S}_z\boldsymbol{S}_z^\top$ with $\boldsymbol{S}_z = f_{S_z}(\boldsymbol{x};\phi)$. Likewise, the decoder moments are $\boldsymbol{\mu}_x = f_{\mu_x}(\boldsymbol{z};\theta)$ and $\boldsymbol{\Sigma}_x = \gamma\boldsymbol{I}$. Here $\gamma > 0$ is a tunable scalar, while f_{μ_z}, f_{S_z} and f_{μ_x} specify parameterized differentiable functional forms that can be arbitrarily complex, e.g., a deep neural network.*

Equipped with these definitions, we will now demonstrate that a κ-simple VAE, with $\kappa \geq r$, can achieve the optimality criteria (*i*) and (*ii*) from above. In doing so, we first consider the simpler case where $r = d$, followed by the extended scenario with $r < d$. The distinction between these two cases turns out to be significant, with practical implications to be explored in Section 4.

2.1 Manifold Dimension Equal to Ambient Space Dimension ($r = d$)

We first analyze the specialized situation where $r = d$. Assuming $p_{gt}(\boldsymbol{x}) \triangleq \mu_{gt}(d\boldsymbol{x})/d\boldsymbol{x}$ exists everywhere in $\mathbb{R}^d$, then $p_{gt}(\boldsymbol{x})$ represents the ground-truth probability density with respect to the standard Lebesgue measure in Euclidean space. Given these considerations, the minimal possible value of (1) will necessarily occur if

$$\mathbb{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})\right] = 0 \quad \text{and} \quad p_\theta(\boldsymbol{x}) = p_{gt}(\boldsymbol{x}) \text{ almost everywhere.} \tag{3}$$

This follows because by VAE design it must be that $\mathcal{L}(\theta, \phi) \geq -\int p_{gt}(\boldsymbol{x}) \log p_{gt}(\boldsymbol{x}) d\boldsymbol{x}$, and in the present context, this lower bound is achievable iff the conditions from (3) hold. Collectively, this implies that the approximate posterior produced by the encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is in fact perfectly matched to the actual posterior $p_\theta(\boldsymbol{z}|\boldsymbol{x})$, while the corresponding marginalized data distribution $p_\theta(\boldsymbol{x})$ is perfectly matched the ground-truth density $p_{gt}(\boldsymbol{x})$ as desired. Perhaps surprisingly, a κ-simple VAE can actually achieve such a solution:

Theorem 1 *Suppose that $r = d$ and there exists a density $p_{gt}(\boldsymbol{x})$ associated with the ground-truth measure μ_{gt} that is nonzero everywhere on $\mathbb{R}^d$.*[1] *Then for any $\kappa \geq r$, there is a sequence of κ-simple VAE model parameters $\{\theta_t^*, \phi_t^*\}$ such that*

$$\lim_{t\to\infty} \mathbb{KL}\left[q_{\phi_t^*}(\boldsymbol{z}|\boldsymbol{x})||p_{\theta_t^*}(\boldsymbol{x}|\boldsymbol{x})\right] = 0 \quad \textit{and} \quad \lim_{t\to\infty} p_{\theta_t^*}(\boldsymbol{x}) = p_{gt}(\boldsymbol{x}) \textit{ almost everywhere.} \tag{4}$$

All the proofs can be found in the supplementary file. So at least when $r = d$, the VAE Gaussian assumptions need not actually prevent the optimal ground-truth probability measure from being recovered, as long as the latent dimension is sufficiently large (i.e., $\kappa \geq r$). And contrary to popular notions, a richer class of distributions is not required to achieve this. Of course Theorem 1 only applies to a restricted case that excludes $d > r$; however, later we will demonstrate that a key consequence of this result can nonetheless be leveraged to dramatically enhance VAE performance.

2.2 Manifold Dimension Less Than Ambient Space Dimension ($r < d$)

When $r < d$, additional subtleties are introduced that will be unpacked both here and in the sequel. To begin, if both $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ are arbitrary/unconstrained (i.e., not necessarily Gaussian), then $\inf_{\phi,\theta} \mathcal{L}(\theta, \phi) = -\infty$. To achieve this global optimum, we need only choose ϕ such that $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = p_\theta(\boldsymbol{z}|\boldsymbol{x})$ (minimizing the KL term from (1)) while selecting θ such that all probability mass collapses to the correct manifold $\boldsymbol{\chi}$. In this scenario the density $p_\theta(\boldsymbol{x})$ will become unbounded on $\boldsymbol{\chi}$ and zero elsewhere, such that $\int_{\boldsymbol{\chi}} -\log p_\theta(\boldsymbol{x}) \mu_{gt}(d\boldsymbol{x})$ will approach negative infinity.

But of course the stated Gaussian assumptions from the κ-simple VAE model could ostensibly prevent this from occurring by causing the KL term to blow up, counteracting the negative log-likelihood factor. We will now analyze this case to demonstrate that this need not happen. Before proceeding to this result, we first define a manifold density $\tilde{p}_{gt}(\boldsymbol{x})$ as the probability density (assuming it exists) of μ_{gt} with respect to the volume measure of the manifold $\boldsymbol{\chi}$. If $d = r$ then this volume measure reduces to the standard Lebesgue measure in $\mathbb{R}^d$ and $\tilde{p}_{gt}(\boldsymbol{x}) = p_{gt}(\boldsymbol{x})$; however, when $d > r$ a density $p_{gt}(\boldsymbol{x})$ defined in $\mathbb{R}^d$ will not technically exist, while $\tilde{p}_{gt}(\boldsymbol{x})$ is still perfectly well-defined. We then have the following:

Theorem 2 *Assume $r < d$ and that there exists a manifold density $\tilde{p}_{gt}(\boldsymbol{x})$ associated with the ground-truth measure μ_{gt} that is nonzero everywhere on $\boldsymbol{\chi}$. Then for any $\kappa \geq r$, there is a sequence of κ-simple VAE model parameters $\{\theta_t^*, \phi_t^*\}$ such that*

$$i \lim_{t\to\infty} \mathbb{KL}\left[q_{\phi_t^*}(\boldsymbol{z}|\boldsymbol{x})||p_{\theta_t^*}(\boldsymbol{z}|\boldsymbol{x})\right] = 0 \quad \textit{and} \quad \lim_{t\to\infty} \int_{\boldsymbol{\chi}} -\log p_{\theta_t^*}(\boldsymbol{x}) \mu_{gt}(d\boldsymbol{x}) = -\infty, \tag{5}$$

[1] This nonzero assumption can be replaced with a much looser condition. Specifically, if there exists a diffeomorphism between the set $\{\boldsymbol{x}|p_{gt}(\boldsymbol{x}) \neq 0\}$ and $\mathbb{R}^d$, then it can be shown that Theorem 1 still holds even if $p_{gt}(\boldsymbol{x}) = 0$ for some $\boldsymbol{x} \in \mathbb{R}^d$.

$$ii \quad \lim_{t\to\infty} \int_{\boldsymbol{x}\in A} p_{\theta_t^*}(\boldsymbol{x})d\boldsymbol{x} = \mu_{gt}(A \cap \boldsymbol{\chi}) \tag{6}$$

for all measurable sets $A \subseteq \mathbb{R}^d$ with $\mu_{gt}(\partial A \cap \boldsymbol{\chi}) = 0$, where ∂A is the boundary of A.

Technical details notwithstanding, Theorem 2 admits a very intuitive interpretation. First, (5) directly implies that the VAE Gaussian assumptions do not prevent minimization of $\mathcal{L}(\theta, \phi)$ from converging to minus infinity, which can be trivially viewed as a globally optimum solution. Furthermore, based on (6), this solution can be achieved with a limiting density estimate that will assign a probability mass to most all measurable subsets of $\mathbb{R}^d$ that is indistinguishable from the ground-truth measure (which confines all mass to $\boldsymbol{\chi}$). Hence this solution is more-or-less an arbitrarily-good approximation to μ_{gt} for all practical purposes.[2]

Regardless, there is an absolutely crucial distinction between Theorem 2 and the simpler case quantified by Theorem 1. Although both describe conditions whereby the κ-simple VAE can achieve the minimal possible objective, in the $r = d$ case achieving the lower bound (whether the specific parameterization for doing so is unique or not) necessitates that the ground-truth probability measure has been recovered almost everywhere. But the $r < d$ situation is quite different because we have not ruled out the possibility that a different set of parameters $\{\theta, \phi\}$ could push $\mathcal{L}(\theta, \phi)$ to $-\infty$ and yet not achieve (6). In other words, the VAE could reach the lower bound but fail to closely approximate μ_{gt}. And we stress that this uniqueness issue is not a consequence of the VAE Gaussian assumptions per se; even if $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ were unconstrained the same lack of uniqueness can persist.

Rather, the intrinsic difficulty is that, because the VAE model does not have access to the ground-truth low-dimensional manifold, it must implicitly rely on a density $p_\theta(\boldsymbol{x})$ defined across *all* of $\mathbb{R}^d$ as mentioned previously. Moreover, if this density converges towards infinity on the manifold during training without increasing the KL term at the same rate, the VAE cost can be unbounded from below, even in cases where (6) is not satisfied, meaning incorrect assignment of probability mass.

To conclude, the key take-home message from this section is that, at least in principle, VAE Gaussian assumptions need not actually be the root cause of any failure to recover ground-truth distributions. Instead we expose a structural deficiency that lies elsewhere, namely, the non-uniqueness of solutions that can optimize the VAE objective without necessarily learning a close approximation to μ_{gt}. But to probe this issue further and motivate possible workarounds, it is critical to further disambiguate these optimal solutions and their relationship with ground-truth manifolds. This will be the task of Section 3, where we will explicitly differentiate the problem of locating the correct ground-truth manifold, from the task of learning the correct probability measure *within* the manifold.

Note that the only comparable prior work we are aware of related to the results in this section comes from Doersch (2016), where the implications of adopting Gaussian encoder/decoder pairs in the specialized case of $r = d = 1$ are briefly considered. Moreover, the analysis there requires additional much stronger assumptions than ours, namely, that $p_{gt}(\boldsymbol{x})$ should be nonzero and infinitely differentiable everywhere in the requisite 1D ambient space. These requirements of course exclude essentially all practical usage regimes where $d = r > 1$ or $d > r$, or when ground-truth densities are not sufficiently smooth.

3 Optimal Solutions and the Ground Truth Manifold

We will now more closely examine the properties of optimal κ-simple VAE solutions, and in particular, the degree to which we might expect them to at least reflect the true $\boldsymbol{\chi}$, even if perhaps not the correct probability measure μ_{gt} defined within $\boldsymbol{\chi}$. To do so, we must first consider some *necessary* conditions for VAE optima:

Theorem 3 *Let $\{\theta_\gamma^*, \phi_\gamma^*\}$ denote an optimal κ-simple VAE solution (with $\kappa \geq r$) where the decoder variance γ is fixed (i.e., it is the sole unoptimized parameter). Moreover, we assume that μ_{gt} is not a Gaussian distribution when $d = r$.[3] Then for any $\gamma > 0$, there exists a $\gamma' < \gamma$ such that $\mathcal{L}(\theta_{\gamma'}^*, \phi_{\gamma'}^*) < \mathcal{L}(\theta_\gamma^*, \phi_\gamma^*)$.*

[2]Note that (6) is only framed in this technical way to accommodate the difficulty of comparing a measure μ_{gt} restricted to $\boldsymbol{\chi}$ with the VAE density $p_\theta(\boldsymbol{x})$ defined everywhere in $\mathbb{R}^d$. See the supplementary for details.

[3]This requirement is only included to avoid a practically irrelevant form of non-uniqueness that exists with full, non-degenerate Gaussian distributions.

This result implies that we can always reduce the VAE cost by choosing a smaller value of γ, and hence, if γ is not constrained, it must be that $\gamma \to 0$ if we wish to minimize (2). Despite this necessary optimality condition, in existing practical VAE applications, it is standard to fix $\gamma \approx 1$ during training. This is equivalent to simply adopting a non-adaptive squared-error loss for the decoder and, at least in part, likely contributes to unrealistic/blurry VAE-generated samples. Regardless, there are more significant consequences of this intrinsic favoritism for $\gamma \to 0$, in particular as related to reconstructing data drawn from the ground-truth manifold $\boldsymbol{\chi}$:

Theorem 4 *Applying the same conditions and definitions as in Theorem 3, then for all $\boldsymbol{x}$ drawn from μ_{gt}, we also have that*

$$\lim_{\gamma \to 0} f_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x};\phi_\gamma^*) + f_{S_z}(\boldsymbol{x};\phi_\gamma^*)\boldsymbol{\varepsilon};\ \theta_\gamma^*\right] = \lim_{\gamma \to 0} f_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x};\phi_\gamma^*);\ \theta_\gamma^*\right] = \boldsymbol{x}, \quad \forall \boldsymbol{\varepsilon} \in \mathbb{R}^\kappa. \tag{7}$$

By design any random draw $\boldsymbol{z} \sim q_{\phi_\gamma^*}(\boldsymbol{z}|\boldsymbol{x})$ can be expressed as $f_{\mu_z}(\boldsymbol{x};\phi_\gamma^*) + f_{S_z}(\boldsymbol{x};\phi_\gamma^*)\boldsymbol{\varepsilon}$ for some $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\varepsilon}|\boldsymbol{0},\boldsymbol{I})$. From this vantage point then, (7) effectively indicates that any $\boldsymbol{x} \in \boldsymbol{\chi}$ will be perfectly reconstructed by the VAE encoder/decoder pair at globally optimal solutions, achieving this necessary condition despite any possible stochastic corrupting factor $f_{S_z}(\boldsymbol{x};\phi_\gamma^*)\boldsymbol{\varepsilon}$.

But still further insights can be obtained when we more closely inspect the VAE objective function behavior at arbitrarily small but explicitly nonzero values of γ. In particular, when $\kappa = r$ (meaning $\boldsymbol{z}$ has no superfluous capacity), Theorem 4 and attendant analyses in the supplementary ultimately imply that the squared eigenvalues of $f_{S_z}(\boldsymbol{x};\phi_\gamma^*)$ will become arbitrarily small at a rate proportional to γ, meaning $\frac{1}{\sqrt{\gamma}} f_{S_z}(\boldsymbol{x};\phi_\gamma^*) \approx O(1)$ under mild conditions. It then follows that the VAE data term integrand from (2), in the neighborhood around optimal solutions, behaves as $-2\mathbb{E}_{q_{\phi_\gamma^*}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\theta_\gamma^*}(\boldsymbol{x}|\boldsymbol{z})\right] =$

$$2\mathbb{E}_{q_{\phi_\gamma^*}(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{1}{\gamma}\left\|\boldsymbol{x} - f_{\mu_x}\left[\boldsymbol{z};\ \theta_\gamma^*\right]\right\|_2^2\right] + d\log 2\pi\gamma \approx \mathbb{E}_{q_{\phi_\gamma^*}(\boldsymbol{z}|\boldsymbol{x})}\left[O(1)\right] + d\log 2\pi\gamma = d\log\gamma + O(1). \tag{8}$$

This expression can be derived by excluding the higher-order terms of a Taylor series approximation of $f_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x};\phi_\gamma^*) + f_{S_z}(\boldsymbol{x};\phi_\gamma^*)\boldsymbol{\varepsilon};\ \theta_\gamma^*\right]$ around the point $f_{\mu_z}(\boldsymbol{x};\phi_\gamma^*)$, which will be relatively tight under the stated conditions. But because $2\mathbb{E}_{q_{\phi_\gamma^*}(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{1}{\gamma}\left\|\boldsymbol{x} - f_{\mu_x}\left[\boldsymbol{z};\ \theta_\gamma^*\right]\right\|_2^2\right] \geq 0$, a theoretical lower bound on (8) is given by $d\log 2\pi\gamma \equiv d\log\gamma + O(1)$. So in this sense (8) cannot be significantly lowered further.

This observation is significant when we consider the inclusion of addition latent dimensions by allowing $\kappa > r$. Clearly based on the analysis above, adding dimensions to $\boldsymbol{z}$ *cannot* improve the value of the VAE data term in any meaningful way. However, it can have a detrimental impact on the the KL regularization factor in the $\gamma \to 0$ regime, where

$$2\mathbb{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right] \equiv \text{trace}\left[\boldsymbol{\Sigma}_z\right] + \|\boldsymbol{\mu}_z\|_2^2 - \log|\boldsymbol{\Sigma}_z| \approx -\hat{r}\log\gamma + O(1). \tag{9}$$

Here $\hat{r}$ denotes the number of eigenvalues $\{\lambda_j(\gamma)\}_{j=1}^\kappa$ of $f_{S_z}(\boldsymbol{x};\phi_\gamma^*)$ (or equivalently $\boldsymbol{\Sigma}_z$) that satisfy $\lambda_j(\gamma) \to 0$ if $\gamma \to 0$. $\hat{r}$ can be viewed as an estimate of how many low-noise latent dimensions the VAE model is preserving to reconstruct $\boldsymbol{x}$. Based on (9), there is obvious pressure to make $\hat{r}$ as small as possible, at least without disrupting the data fit. The smallest possible value is $\hat{r} = r$, since it is not difficult to show that any value below this will contribute consequential reconstruction errors, causing $2\mathbb{E}_{q_{\phi_\gamma^*}(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{1}{\gamma}\left\|\boldsymbol{x} - f_{\mu_x}\left[\boldsymbol{z};\ \theta_\gamma^*\right]\right\|_2^2\right]$ to grow at a rate of $\Omega\left(\frac{1}{\gamma}\right)$, pushing the entire cost function towards infinity.[4]

Therefore, *in the neighborhood of optimal solutions the VAE will naturally seek to produce perfect reconstructions using the fewest number of clean, low-noise latent dimensions*, meaning dimensions whereby $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ has negligible variance. For superfluous dimensions that are unnecessary for representing $\boldsymbol{x}$, the associated encoder variance in these directions can be pushed to one. This will optimize $\mathbb{KL}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right]$ along these directions, and the decoder can selectively block the residual randomness to avoid influencing the reconstructions per Theorem 4. So in this sense the VAE is capable of learning a minimal representation of the ground-truth manifold $\boldsymbol{\chi}$ when $r < \kappa$.

[4]Note that $\inf_{\gamma>0} \frac{C}{\gamma} + \log\gamma = \infty$ for any $C > 0$.

But we must emphasize that the VAE can learn $\boldsymbol{\chi}$ independently of the actual distribution μ_{gt} within $\boldsymbol{\chi}$. Addressing the latter is a completely separate issue from achieving the perfect reconstruction error defined by Theorem 4. This fact can be understood within the context of a traditional PCA-like model, which is perfectly capable of learning a low-dimensional subspace containing some training data without actually learning the distribution of the data within this subspace. The central issue is that there exists an intrinsic bias associated with the VAE objective such that *fitting the distribution within the manifold will be completely neglected whenever there exists the chance for even an infinitesimally better approximation of the manifold itself*.

Stated differently, if VAE model parameters have learned a near optimal, parsimonious latent mapping onto $\boldsymbol{\chi}$ using $\gamma \approx 0$, then the VAE cost will scale as $(d-r)\log\gamma$ regardless of μ_{gt}. Hence there remains a *huge* incentive to reduce the reconstruction error still further, allowing γ to push even closer to zero and the cost closer to $-\infty$. And if we constrain γ to be sufficiently large so as to prevent this from happening, then we risk degrading/blurring the reconstructions and widening the gap between $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{z}|\boldsymbol{x})$, which can also compromise estimation of μ_{gt}. Fortunately though, as will be discussed next there is a convenient way around this dilemma by exploiting the fact that this dominanting $(d-r)\log\gamma$ factor goes away when $d=r$.

4 From Theory to Practical VAE Enhancements

Sections 2 and 3 have exposed a collection of VAE properties with useful diagnostic value in and of themselves. But the practical utility of these results, beyond the underappreciated benefit of learning γ, warrant further exploration. In this regard, suppose we wish to develop a generative model of high-dimensional data $\boldsymbol{x} \in \boldsymbol{\chi}$ where unknown low-dimensional structure is significant (i.e., the $r < d$ case with r unknown). The results from Section 3 indicate that the VAE can partially handle this situation by learning a parsimonious representation of low-dimensional manifolds, but not necessarily the correct probability measure μ_{gt} within such a manifold. In quantitative terms, this means that a decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ will map all samples from an encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ to the correct manifold such that the reconstruction error is negligible for any $\boldsymbol{x} \in \boldsymbol{\chi}$. But if the measure μ_{gt} on $\boldsymbol{\chi}$ has not been accurately estimated, then

$$q_\phi(\boldsymbol{z}) \triangleq \int_{\boldsymbol{\chi}} q_\phi(\boldsymbol{z}|\boldsymbol{x})\mu_{gt}(d\boldsymbol{x}) \not\approx \int_{\mathbb{R}^d} p_\theta(\boldsymbol{z}|\boldsymbol{x})p_\theta(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathbb{R}^d} p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{x} = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{I}), \quad (10)$$

where $q_\phi(\boldsymbol{z})$ is sometimes referred to as the aggregated posterior (Makhzani et al., 2016). In other words, the distribution of the latent samples drawn from the encoder distribution, when averaged across the training data, will have lingering latent structure that is errantly incongruous with the original isotropic Gaussian prior. This then disrupts the pivotal ancestral sampling capability of the VAE, implying that samples drawn from $\mathcal{N}(\boldsymbol{z}|0,I)$ and then passed through the decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ will *not* closely approximate μ_{gt}. Fortunately, our analysis suggests the following two-stage remedy:

1. Given n observed samples $\{\boldsymbol{x}^{(i)}\}_{i=1}^n$, train a κ-simple VAE, with $\kappa \geq r$, to estimate the unknown r-dimensional ground-truth manifold $\boldsymbol{\chi}$ embedded in $\mathbb{R}^d$ using a minimal number of active latent dimensions. Generate latent samples $\{\boldsymbol{z}^{(i)}\}_{i=1}^n$ via $\boldsymbol{z}^{(i)} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})$. By design, these samples will be distributed as $q_\phi(\boldsymbol{z})$, but likely not $\mathcal{N}(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{I})$.
2. Train a second κ-simple VAE, with independent parameters $\{\theta', \phi'\}$ and latent representation $\boldsymbol{u}$, to learn the unknown distribution $q_\phi(\boldsymbol{z})$, i.e., treat $q_\phi(\boldsymbol{z})$ as a new ground-truth distribution and use samples $\{\boldsymbol{z}^{(i)}\}_{i=1}^n$ to learn it.
3. Samples approximating the *original* ground-truth μ_{gt} can then be formed via the extended ancestral process $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{u}|\boldsymbol{0},\boldsymbol{I})$, $\boldsymbol{z} \sim p_{\theta'}(\boldsymbol{z}|\boldsymbol{u})$, and finally $\boldsymbol{x} \sim p_\theta(\boldsymbol{x}|\boldsymbol{z})$.

The efficacy of the second-stage VAE from above is based on the following. If the first stage was successful, then even though they will not generally resemble $\mathcal{N}(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{I})$, samples from $q_\phi(\boldsymbol{z})$ will nonetheless have nonzero measure across the full ambient space $\mathbb{R}^\kappa$. If $\kappa = r$, this occurs because the entire latent space is needed to represent an r-dimensional manifold, and if $\kappa > r$, then the extra latent dimensions will be naturally filled in via randomness introduced along dimensions associated with nonzero eigenvalues of the decoder covariance $\boldsymbol{\Sigma}_z$ per the analysis in Section 3.

Consequently, as long as we set $\kappa \geq r$, *the operational regime of the second-stage VAE is effectively equivalent to the situation described in Section 2.1 where the manifold dimension is equal to*

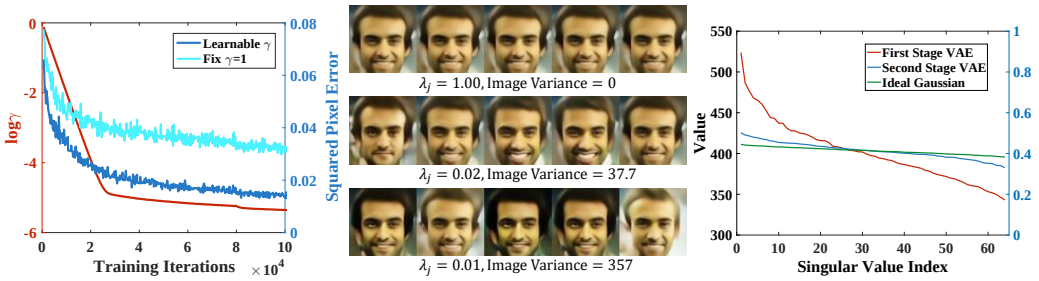


Figure 1: *Demonstrating VAE properties*. (*Left*) Validation of Theorem 3 and the influence on image reconstructions. (*Center*) Validation of Theorem 4. (*Right*) Motivation for two separate VAE stages by comparing the aggregated posteriors $q_\phi(\boldsymbol{z})$ (1st stage) vs. $q_{\phi'}(\boldsymbol{u})$ (enhanced 2nd stage).

the ambient dimension.[5] And as we have already shown there via Theorem 1, the VAE can readily handle this situation, since in the narrow context of the second-stage VAE, $d = r = \kappa$, the troublesome $(d - r)\log\gamma$ factor becomes zero, and any globally minimizing solution is uniquely matched to the new ground-truth distribution $q_\phi(\boldsymbol{z})$. Consequently, the revised aggregated posterior $q_{\phi'}(\boldsymbol{u})$ produced by the second-stage VAE should now closely resemble $\mathcal{N}(\boldsymbol{u}|\mathbf{0}, \boldsymbol{I})$. And finally, because we generally assume that $d \gg \kappa \geq r$, we have found that the second-stage VAE can be quite small.

5 Empirical Evaluation of VAE Two-Stage Enhancement

We initially describe experiments explicitly designed to corroborate some of our previous analytical results using VAE models trained on CelebA (Liu et al., 2015) data; please see the supplementary for training details and more related experiments. First, the leftmost plot of Figure 1 presents support for Theorem 3, where indeed the decoder variance γ does tend towards zero during training. This then allows for tighter image reconstructions with lower average squared error, i.e., a better manifold fit as expected. The center plot bolsters Theorem 4 and the analysis that follows by showcasing the dissimilar impact of noise factors applied to different directions in the latent space before passage through the decoder mean network f_{μ_x}. In a direction where an eigenvalue λ_j of $\boldsymbol{\Sigma}_z$ is large (i.e., a superfluous dimension), a random perturbation is completely muted by the decoder as predicted. In contrast, in directions where such eigenvalues are small (i.e., needed for representing the manifold), varying the input causes large changes in the image space reflecting reasonable movement along the correct manifold. Finally, the rightmost plot of Figure 1 displays the singular value spectrum of latent sample matrices drawn from the first- and second-stage VAE models. As expected, the latter is much closer to the spectrum from an analogous i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ matrix. This indicates a superior latent representation, providing high-level support for our two-stage VAE proposal.

Next we present quantitative evaluation of novel generated samples using the large-scale testing protocol of GAN models from (Lucic et al., 2018). In this regard, GANs are well-known to dramatically outperform existing VAE approaches in terms of the Fréchet Inception Distance (FID) score (Heusel et al., 2017) and related quantitative metrics. For fair comparison, (Lucic et al., 2018) adopted a common neutral architecture for all models, with generator and discriminator networks based on InfoGAN (Chen et al., 2016a); the point here is standardized comparisons, not tuning arbitrarily-large networks to achieve the lowest possible absolute FID values. We applied the same architecture to our first-stage VAE decoder and encoder networks respectively for direct comparison. For the low-dimensional second-stage VAE we used small, 3-layer networks contributing negligible additional parameters beyond the first stage (see the supplementary for further design details).[6]

We compared our proposed two-stage VAE pipeline against three baseline VAE models differing only in the decoder output layer: a Gaussian layer with fixed γ, a Gaussian layer with a learned γ, and a cross-entropy layer as has been adopted in several previous applications involving images

[5] Note that if a regular autoencoder were used to replace the first-stage VAE, then this would no longer be the case, so indeed a VAE is required for both stages.

[6] It should also be emphasized that concatenating the two stages and jointly training does not improve the performance. If trained jointly the few extra second-stage parameters are simply hijacked by the dominant objective from the first stage and forced to work on an incrementally better fit of the manifold. As expected then, on empirical tests (not shown) we have found that this does not improve upon standard VAE baselines.

	MNIST	Fashion	CIFAR-10	CelebA	Mean
MM GAN	9.8 ± 0.9	29.6 ± 1.6	72.7 ± 3.6	65.6 ± 4.2	44.4 ± 2.6
NS GAN	6.8 ± 0.5	26.5 ± 1.6	58.5 ± 1.9	55.0 ± 3.3	36.7 ± 1.8
LSGAN	7.8 ± 0.6	30.7 ± 2.2	87.1 ± 47.5	53.9 ± 2.8	44.9 ± 13.3
WGAN	$\mathbf{6.7 \pm 0.4}$	21.5 ± 1.6	$\mathbf{55.2} \pm 2.3$	41.3 ± 2.0	31.2 ± 1.6
WGAN GP	20.3 ± 5.0	24.5 ± 2.1	55.8 ± 0.9	30.3 ± 1.0	32.7 ± 2.3
DRAGAN	$7.6 \pm \mathbf{0.4}$	27.7 ± 1.2	69.8 ± 2.0	42.3 ± 3.0	36.9 ± 1.7
BEGAN	13.1 ± 1.0	22.9 ± 0.9	71.4 ± 1.6	38.9 ± 0.9	36.6 ± 1.1
VAE (cross-entr.)	23.8 ± 0.6	58.7 ± 1.2	155.7 ± 11.6	85.7 ± 3.8	81 ± 4.3
VAE (fixed γ)	51.2 ± 0.8	104.5 ± 1.3	113.0 ± 0.7	119.8 ± 0.9	97.1 ± 0.9
VAE (learned γ)	47.0 ± 0.9	51.5 ± 1.0	$80.1 \pm \mathbf{0.6}$	67.4 ± 2.1	61.5 ± 1.2
2-Stage VAE	13.4 ± 1.3	22.0 ± 0.6	$71.0 \pm \mathbf{0.6}$	45.9 ± 1.4	38.1 ± 1.0
2-Stage VAE*	11.2 ± 0.5	$\mathbf{21.3} \pm \mathbf{0.4}$	68.0 ± 0.8	$\mathbf{23.8} \pm \mathbf{0.5}$	$\mathbf{31.1} \pm \mathbf{0.6}$

Table 1: *FID score comparisons*. For all GAN-based models, the reported values represent the *best* FID obtained across a large-scale hyperparameter search conducted separately for each dataset; default settings are considerably worse (Lucic et al., 2018). Likewise outlier cases (e.g., severe mode collapse) were omitted, which would have otherwise degraded these GAN scores and increased standard deviations still further. In contrast, for the VAE results we used only default training settings across all models and datasets (no tuning), except for the 2-Stage VAE*. Here we simply tested a couple different values for κ and picked the best result for each dataset. Note that specialized architectures and/or random seed optimization can potentially improve the FID score for all models.

(Chen et al., 2016b). We also present results from (Lucic et al., 2018) involving numerous state-of-the-art GAN models, including MM GAN (Goodfellow et al., 2014a), WGAN (Arjovsky et al., 2017), WGAN-GP (Gulrajani et al., 2017), NS GAN (Fedus et al., 2017), DRAGAN (Kodali et al., 2017), LS GAN (Mao et al., 2017) and BEGAN (Berthelot et al., 2017). Testing is conducted across four significantly different datasets: MNIST (LeCun et al., 1998), Fashion MNIST (Xiao et al., 2017), CIFAR-10 (Krizhevsky & Hinton, 2009) and CelebA (Liu et al., 2015).

For each dataset we executed 10 independent trials and report the mean and standard deviation of the FID scores in Table 1. Despite the fact that all GAN models benefited from a large-scale hyperparameter search executed independently across each dataset to achieve the best results, our proposed two-stage VAE with minimal tuning is capable of equaling or exceeding the performance of all the GAN models and VAE baselines (see Table 1 caption for more details). This is the first demonstration of a VAE pipeline capable of competing with GANs in the arena of generated sample quality. For example, note the poor performance of VAE baselines relative to GANs in Table 1). Representative samples generated using our two-stage VAE model are in the supplementary.

As a final point of reference, although not an exact VAE model per se, an autoencoder-based architecture that substitutes a Wassenstein distance measure for the KL regularizer from (2) has also recently been proposed (Tolstikhin et al., 2018). Two variants of this approach, termed WAE-MMD and WAE-GAN (because different MMD and GAN regularization factors are included), were evaluated using FID scores, with penalty weights and encoder/decoder networks specifically adapted for use with the CelebA dataset (FID values were not provided for other datasets). A baseline VAE using these networks achieved an FID of 63, which is somewhat better than our VAE baselines presumably because of this tuning for CelebA data. In contrast, the corresponding WAE-MMD and WAE-GAN scores were 55 and 42 respectively. Although these values represent an improvement over the VAE baseline, they are considerably worse than the absolute score of 23.8 achieved by our generic two-stage VAE model with neutral architecture borrowed from (Lucic et al., 2018).

6 Conclusion

It is often assumed that there exists an unavoidable trade-off between the stable training, valuable attendant encoder network, and resistance to mode collapse of VAEs, versus the impressive visual quality of images produced by GANs. While we certainly are not claiming that our two-stage VAE model is necessarily superior to the latest and greatest GAN-based model in terms of the realism of generated samples, we do strongly believe that this work at least narrows that gap substantially such that VAEs are worth considering in a broader range of applications.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

David Berthelot, Thomas Schumm, and Luke Metz. BEGAN: Boundary equilibrium generative adversarial networks. *arXiv:1703.10717*, 2017.

Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. *arXiv:1609.07093*, 2016.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv:1509.00519*, 2015.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2172–2180, 2016a.

Xi Chen, Diederik Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv:1611.02731*, 2016b.

Bin Dai, Yu Wang, John Aston, Gang Hua, and David Wipf. Hidden talents of the variational autoencoder. *arXiv:1706.05148*, 2018.

Carl Doersch. Tutorial on variational autoencoders. *arXiv:1606.05908*, 2016.

Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*, pp. 658–666, 2016.

William Fedus, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M Dai, Shakir Mohamed, and Ian Goodfellow. Many paths to equilibrium: GANs do not need to decrease a divergence at every step. *arXiv:1710.08446*, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014a.

I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In *arXiv:1406.2661*, 2014b.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Diederik Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Diederik Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pp. 4743–4751, 2016.

Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of GANs. *arXiv:1705.07215*, 2017.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv:1512.09300*, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? A large-scale study. *arXiv:1711.10337v3*, 2018.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv:1511.05644*, 2016.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision*, pp. 2813–2821, 2017.

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

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein autoencoders. *International Conference on Learning Representations*, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*, 2017.

SUPPLEMENTARY FILE
Diagnosing and Enhancing VAE Models

1. Introduction

This document contains companion technical material regarding our ICLR 2019 submission. Note that herein all equation numbers referencing back to the main submission document will be be prefixed with an 'M-' to avoid confusion, i.e, (M-#) will refer to equation (#) from the main text. Similar notation differentiates sections, tables, and figures, e.g., Section M-#, etc.

2. Contents

The remainder of this document includes the following contents:

- **Section 3** - Comparison of novel samples generated from our model.
- **Section 4** - Example reconstructions of training data.
- **Section 5** - Additional experimental results validating theoretical predictions.
- **Section 6** - Network structure and experimental settings.
- **Section 7** - Proof of Theorem M.1.
- **Section 8** - Proof of Theorem M.2.
- **Section 9** - Proof of Theorem M.3.
- **Section 10** - Proof of Theorem M.4.
- **Section 11** - Further analysis of the VAE cost as γ becomes small.

3. Comparison of Novel Samples Generated from our Model

Generation results for CelebA, MNIST, Fashion-MNIST and CIFAR-10 datasets are shown in Figures 1−4 respectively. When γ is fixed to be one, the generated samples are very blurry. If a learnable γ is used, the samples becomes sharper; however, there are many lingering artifacts as expected. In contrast, the proposed 2-Stage VAE can remove these artifacts and generate more realistic samples. For comparison purposes, we also show the results from WAE-MMD, WAE-GAN (Tolstikhin et al., 2018) and WGAN-GP (Gulrajani et al., 2017) for the CelebA dataset.

(a) WAE-MMD (b) WAE-GAN (c) WGAN-GP

(d) VAE (Fix $\gamma = 1$) (e) VAE (Learnable γ) (f) 2-Stage VAE

Figure 1: Randomly generated samples on the CelebaA dataset (i.e., no cherry-picking).

(a) VAE (Fix $\gamma = 1$) (b) VAE (Learnable γ) (c) 2-Stage VAE

Figure 2: Randomly generated samples on the MNIST dataset (i.e., no cherry-picking).

(a) VAE (Fix $\gamma = 1$) (b) VAE (Learnable γ) (c) 2-Stage VAE

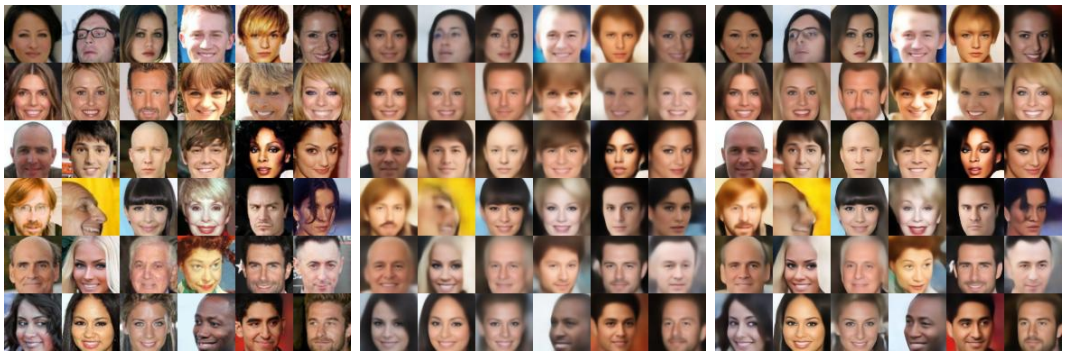
Figure 3: Randomly Generated Samples on Fashion-MNIST Dataset (i.e., no cherry-picking).

(a) VAE (Fix $\gamma = 1$) (b) VAE (Learnable γ) (c) 2-Stage VAE

Figure 4: Randomly Generated Samples on CIFAR-10 Dataset (i.e., no cherry-picking).

4. Example Reconstructions of Training Data

Reconstruction results for MNIST, Fashion-MNIST, CIFAR-10 and CelebA datasets are shown in Figures 5−8 respectively. On relatively simple datasets like MNIST and Fashion-MNIST, the VAE with learnable γ achieves almost exact reconstruction because of a better estimate of the underlying manifold consistent with theory. However, the VAE with fixed $\gamma = 1$ produces blurry reconstructions as expected. Note that the reconstruction of a 2-Stage VAE is the same as that of a VAE with learnable γ because the second-stage VAE has nothing to do with facilitating the reconstruction task.

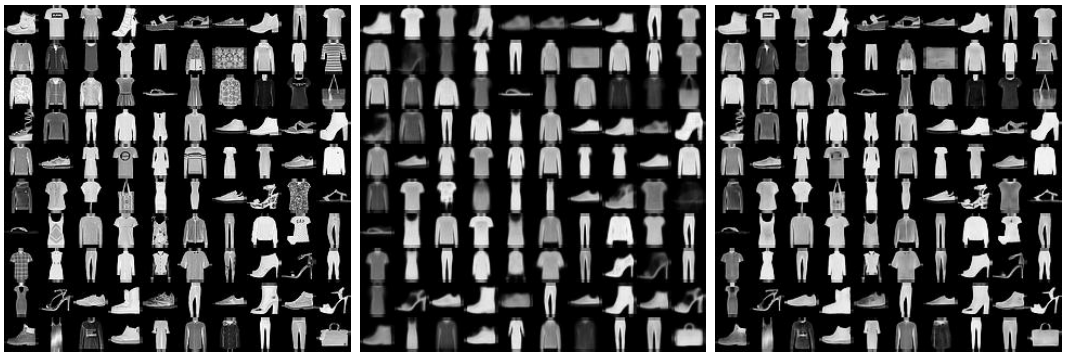


(a) Ground Truth (b) VAE (Fix $\gamma = 1$) (c) VAE (Learnable γ)

Figure 5: Reconstructions on CelebA Dataset.

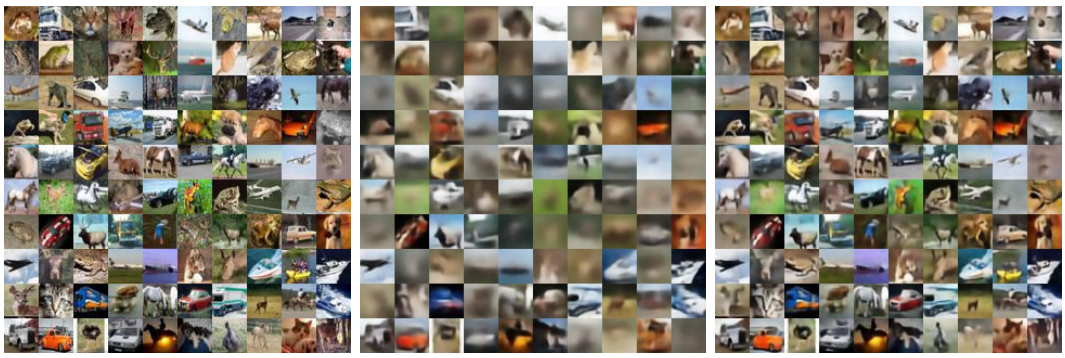
(a) Ground Truth (b) VAE (Fix $\gamma = 1$) (c) VAE (Learnable γ)

Figure 6: Reconstructions on MNIST Dataset.



(a) Ground Truth (b) VAE (Fix $\gamma = 1$) (c) VAE (Learnable γ)

Figure 7: Reconstructions on Fashion-MNIST Dataset.



(a) Ground Truth (b) VAE (Fix $\gamma = 1$) (c) VAE (Learnable γ)

Figure 8: Reconstructions on CIFAR-10 Dataset.

(a) MNIST (b) Fashion-MNIST

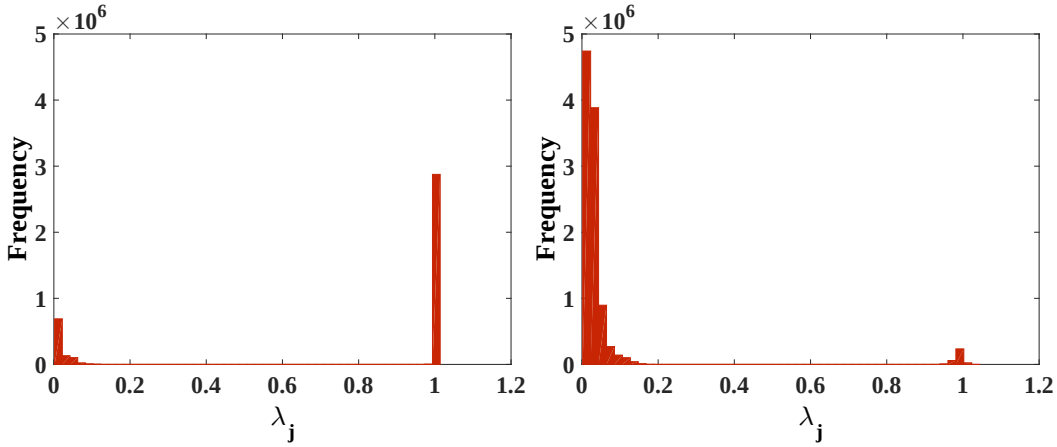
(c) CelebA

Figure 9: More examples similar to Figure M.1(*center*).

5. Additional Experimental Results Validating Theoretical Predictions

We first present more examples similar to Figure M.1(*center*) from the main paper. Random noise is added to $\boldsymbol{\mu}_z$ along different directions and the result is passed through the decoder network. Each row corresponds to a certain direction in the latent space and 15 samples are shown for each direction. These dimensions/rows are ordered by the eigenvalues λ_j of $\boldsymbol{\Sigma}_z$. The larger λ_j is, the less impact a random perturbation along this direction will have as quantified by the reported image variance values. In the first two or three rows, the noise generates some images from different classes/objects/identities, indicating a significant visual difference. For a slightly larger λ_j, the corresponding dimensions encode relatively less significant attributes as predicted. For example, the fifth row of both MNIST and Fashion-MNIST contains images from the same class but with a slightly different style. The images in the fourth row of the CelebA dataset have very subtle differences. When $\lambda_j = 1$, the corresponding dimensions become completely inactive and all the output images are exactly the same, as shown in the last rows for all the three datasets.

Additionally, as discussed in the main text and below in Section 11, there are likely to be r eigenvalues of $\boldsymbol{\Sigma}_z$ converging to zero and $\kappa - r$ eigenvalues converging to one. We plot



(a) Hist of λ_j on MNIST (b) Hist of λ_j on CelebA

Figure 10: Histogram of λ_j values. There are more values around 0 for CelebA because it is more complicated than MNIST and therefore requres more active dimensions to model the underlying manifold.

the histogram of λ_j values for both MNIST and CelebA datasets in Figure 10. For both datasets, λ_j approximately converges to either to zero or one. However, since CelebA is a more complicated dataset than MNIST, the ground-truth manifold dimension of CelebA is likely to be much larger than that of MNIST. So more eigenvalues are expected to be near zero for the CelebA dataset. This is indeed the case, demonstrating that VAE has the ability to detect the manifold dimension and select the proper number of latent dimensions in practical environments.

6. Network Structure and Experimental Settings

We first describe the network and training details used in producing Figure M.1 from the main file, and for generating samples and reconstructions in the supplementary. The first-stage VAE network is shown in Figure 11. Basically we use two Residual Blocks for each resolution scale, and we double the number of channels when downsampling and halve it when upsampling. The specific settings such as the number of channels and the number of scales are specified in the caption. The second VAE is much simpler. Both the encoder and decoder have three 2048-dimensional hidden layers. Finally, the training details are presented below. Note that these settings were not tuned, we simply chose more epochs for more complex data sets and fewer for datasets with larger training samples. For each dataset just a single setting was tested as follows:

- **MNIST and Fashion-MNIST:** The batch size is specified to be 100. We use the ADAM optimizer with the default hyperparameters in TensorFlow. Standard weight decay is set as 5×10^{-4}. The first VAE is trained for 400 epochs. The initial learning rate is 0.0001 and we halve it every 150 epochs. The second VAE is trained for 800 epochs with the same initial learning rate, halved every 300 epochs.

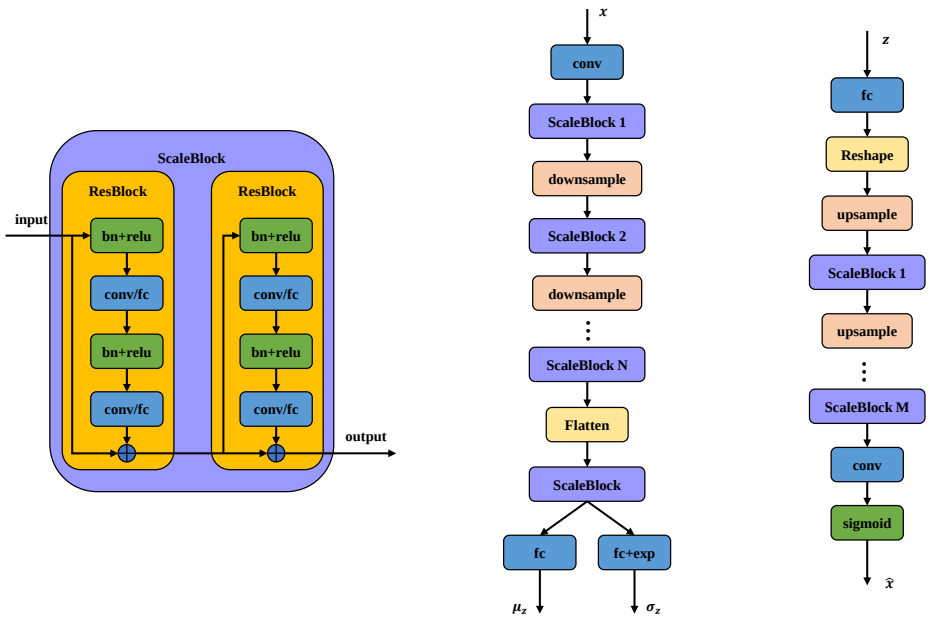


Figure 11: Network structure of the first-stage VAE used in producing Figure M.1, and for generating samples and reconstructions. (*Left*) The basic building block of the network called a Scale Block, which consists of two Residual Blocks. (*Center*) The encoder network. For an input image $\boldsymbol{x}$, we use a convolutional layer to transform it into 32 channels. We then pass it to a Scale Block. After each Scale Block, we downsample using a convolutional layer with stride 2 and double the channels. After N Scale Blocks, the feature map is flattened to a vector. In our experiments, we used $N = 4$ for CelebA dataset and 3 for other datasets. The vector is then passed through another Scale Block, the convolutional layers of which are replaced with fully connected layers of 512 dimensions. The output of this Scale Block is used to produce the κ-dimensional latent code, with $\kappa = 64$. (*Right*) The decoder network. A latent code $\boldsymbol{z}$ is first passed through a fully connected layer. The dimension is 4096 for CelebA dataset and 2048 for other datasets. Then it is reshaped to 2×2 resolution. We upsample the feature map using a deconvolution layer and half the number of channels at the same time. It then goes through some Scale Blocks and upsampling layers until the feature map size becomes the desired value. Then we use a convolutional layer to transform the feature map, which should have 32 channels, to 3 channels for RGB datasets and 1 channel for gray scale datasets.

- **CIFAR-10:** Since CIFAR-10 is more complicated than MNIST and Fashion-MNIST, we use more epochs for training. Specifically, we use 1000 and 2000 epochs for the two VAEs respectively and half the learning rate every 300 and 600 epochs for the two stages. The other settings are the same as that for MNIST.

- **CelebA:** Because CelebA has many more examples, in the first stage we train 120 epochs and half the learning rate every 48 epochs. In the second stage, we train 300 epochs and half the learning rate every 120 epochs. The other settings are the same as that for MNIST, etc.

Finally, to fairly compare against various GAN models and VAE baselines using FID scores on a neutral architecture (i.e., the results from Table M.1), we simply adopt the InfoGAN network structure consistent with the neutral setup from (Lucic et al., 2018) for the first-stage VAE. For the second-stage VAE we just use three 1024-dimensional hidden layers, which contribute less than 5% to the total number of parameters. The only hyperparameter we considered tuning was κ, coarsely testing just a few different values based on dataset complexity. Note that the small number of additional parameters contributing to the second stage do not improve the other VAE baselines when aggregated and trained jointly.

7. Proof of Theorem M.1

We first consider the case where the latent dimension κ equals the manifold dimension r and then extend the proof to allow for $\kappa > r$. The intuition is to build a bijection between $\boldsymbol{\chi}$ and $\mathbb{R}^r$ that transforms the ground-truth distribution $p_{gt}(\boldsymbol{x})$ to a normal Gaussian distribution. The way to build such a bijection is shown in Figure 12. We now fill in the details.

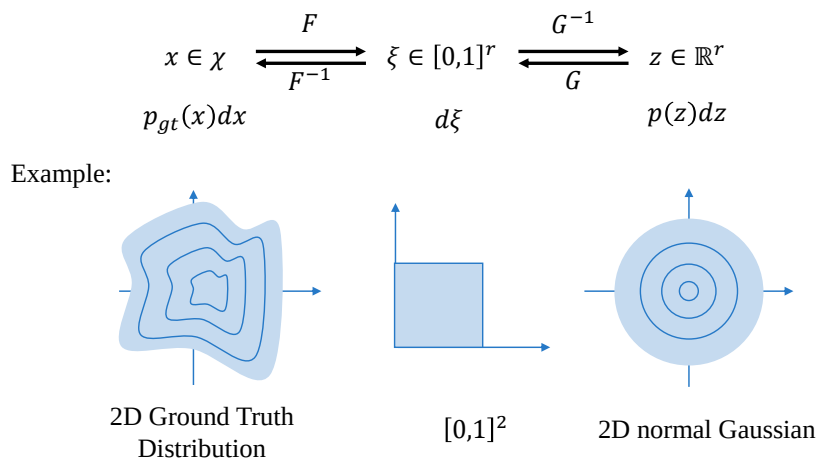


Figure 12: The relationship between different variables.

7.1 Finding a Sequence of Decoders such that $p_{\theta_t^*}(\boldsymbol{x})$ Converges to $p_{gt}(\boldsymbol{x})$

Define the function $F : \mathbb{R}^r \mapsto [0,1]^r$ as

$$F(\boldsymbol{x}) = [F_1(\boldsymbol{x}_1), F_2(\boldsymbol{x}_2; \boldsymbol{x}_1), ..., F_r(\boldsymbol{x}_r; \boldsymbol{x}_{1:r-1})]^\top, \tag{1}$$

$$F_i(\boldsymbol{x}_i; \boldsymbol{x}_{1:i-1}) = \int_{\boldsymbol{x}_i'=-\infty}^{\boldsymbol{x}_i} p_{gt}(\boldsymbol{x}_i' | \boldsymbol{x}_{1:i-1}) d\boldsymbol{x}_i'. \tag{2}$$

Per this definition, we have that

$$dF(\boldsymbol{x}) = p_{gt}(\boldsymbol{x}) d\boldsymbol{x}. \tag{3}$$

Also, since $p_{gt}(\boldsymbol{x})$ is nonzero everywhere, $F(\cdot)$ is invertible. Similarly, we define another differentiable and invertible function $G : \mathbb{R}^r \mapsto [0,1]^r$ as

$$G(\boldsymbol{z}) = [G_1(\boldsymbol{z}_1), G_2(\boldsymbol{z}_2), ..., G_r(\boldsymbol{z}_r)]^\top, \tag{4}$$

$$G_i(\boldsymbol{z}_i) = \int_{\boldsymbol{z}'_i=-\infty}^{\boldsymbol{z}_i} \mathcal{N}(\boldsymbol{z}_i|0,1)d\boldsymbol{z}'_i. \tag{5}$$

Then

$$dG(\boldsymbol{z}) = p(\boldsymbol{z})d\boldsymbol{z} = \mathcal{N}(\boldsymbol{z}|0,\boldsymbol{I})d\boldsymbol{z}. \tag{6}$$

Now let the decoder be

$$f_{\mu_x}(\boldsymbol{z};\theta_t^*) = F^{-1} \circ G(z), \tag{7}$$

$$\gamma_t^* = \frac{1}{t}. \tag{8}$$

Then we have

$$p_{\theta_t^*}(\boldsymbol{x}) = \int_{\mathbb{R}^r} p_{\theta_t^*}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \int_{\mathbb{R}^r} \mathcal{N}\left(\boldsymbol{x}|F^{-1} \circ G(\boldsymbol{z}), \gamma_t^* \boldsymbol{I}\right) dG(\boldsymbol{z}). \tag{9}$$

Additionally, let $\boldsymbol{\xi} = G(\boldsymbol{z})$ such that

$$p_{\theta_t^*}(\boldsymbol{x}) = \int_{[0,1]^r} \mathcal{N}\left(\boldsymbol{x}|F^{-1}(\boldsymbol{\xi}), \gamma_t^* \boldsymbol{I}\right) d\boldsymbol{\xi}, \tag{10}$$

and let $\boldsymbol{x}' = F^{-1}(\boldsymbol{\xi})$ such that $d\boldsymbol{\xi} = dF(\boldsymbol{x}') = p_{gt}(\boldsymbol{x}')d\boldsymbol{x}'$. Plugging this expression into the previous $p_{\theta^*}(\boldsymbol{x})$ we obtain

$$p_{\theta_t^*}(\boldsymbol{x}) = \int_{\mathbb{R}^r} \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{x}', \gamma_t^* \boldsymbol{I}\right) p_{gt}(\boldsymbol{x}')d\boldsymbol{x}'. \tag{11}$$

As $t \to \infty$, γ_t^* becomes infinitely small and $\mathcal{N}\left(\boldsymbol{x}|\boldsymbol{x}', \gamma_t^* \boldsymbol{I}\right)$ becomes a Dirac-delta function, resulting in

$$\lim_{t\to\infty} p_{\theta_t^*}(\boldsymbol{x}) = \int_{\boldsymbol{\chi}} \delta(\boldsymbol{x}' - \boldsymbol{x})p_{gt}(\boldsymbol{x}')d\boldsymbol{x}' = p_{gt}(\boldsymbol{x}). \tag{12}$$

7.2 Finding a Sequence of Encoders such that $\mathbb{KL}\left[q_{\phi_t^*}(\boldsymbol{z}|\boldsymbol{x})||p_{\theta_t^*}(\boldsymbol{z}|\boldsymbol{x})\right]$ Converges to 0

Assume the encoder networks satisfy

$$f_{\mu_z}(\boldsymbol{x};\phi_t^*) = G^{-1} \circ F(\boldsymbol{x}) = f_{\mu_x}^{-1}(\boldsymbol{x};\theta_t^*), \tag{13}$$

$$f_{S_z}(\boldsymbol{x};\phi_t^*) = \sqrt{\gamma_t^* \left(f'_{\mu_x}\left(f_{\mu_z}(\boldsymbol{x};\phi_t^*);\theta_t^*\right)^\top f'_{\mu_x}\left(f_{\mu_z}(\boldsymbol{x};\phi_t^*);\theta_t^*\right)\right)^{-1}}, \tag{14}$$

where $f'_{\mu_x}(\cdot)$ is a $d \times r$ Jacobian matrix. We omit the arguments θ_t^* and ϕ_t^* in $f_{\mu_z}(\cdot)$, $f_{S_z}(\cdot)$ and $f_{\mu_x}(\cdot)$ hereafter to avoid unnecessary clutter. We first explain why $f_{\mu_x}(\cdot)$ is differentiable. Since $f_{\mu_x}(\cdot)$ is a composition of $F^{-1}(\cdot)$ and $G(\cdot)$ according to (7), we only

need to explain that both functions are differentiable. For $F^{-1}(\cdot)$, it is the inverse of a differentiable function $F(\cdot)$. Moreover, the derivative of $F(\boldsymbol{x})$ is $p_{gt}(\boldsymbol{x})$, which is nonzero everywhere. So $F^{-1}(\cdot)$ and therefore $f_{\mu_x}(\cdot)$ are both differentiable.

The true posterior $p_{\theta_t^*}(\boldsymbol{z}|\boldsymbol{x})$ and the approximate posterior are

$$p_{\theta_t^*}(\boldsymbol{z}|\boldsymbol{x}) = \frac{\mathcal{N}(\boldsymbol{z}|0,\boldsymbol{I})\mathcal{N}(\boldsymbol{x}|f_{\mu_x}(\boldsymbol{z}),\gamma_t^*\boldsymbol{I})}{p_{\theta_t^*}(\boldsymbol{x})}, \tag{15}$$

$$q_{\phi_t^*}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}|f_{\mu_z}(\boldsymbol{x}),\gamma_t^*\left(f'_{\mu_x}\left(f_{\mu_z}(\boldsymbol{x})\right)^\top f'_{\mu_x}\left(f_{\mu_z}(\boldsymbol{x})\right)\right)^{-1}\right) \tag{16}$$

respectively. We now prove that $q_{\phi_t^*}(\boldsymbol{z}|\boldsymbol{x})/p_{\theta_t^*}(\boldsymbol{z}|\boldsymbol{x})$ converges to a constant not related to $\boldsymbol{z}$ as t goes to ∞. If this is true, the constant must be 1 since both $q_{\phi_t^*}(\boldsymbol{z}|\boldsymbol{x})$ and $p_{\theta_t^*}(\boldsymbol{z}|\boldsymbol{x})$ are probability distributions. Then the KL divergence between them converges to 0 as $t\to\infty$.

We denote $\left(f'_{\mu_x}\left(f_{\mu_z}(\boldsymbol{x})\right)^\top f'_{\mu_x}\left(f_{\mu_z}(\boldsymbol{x})\right)\right)^{-1}$ as $\tilde{\boldsymbol{\Sigma}}_z(\boldsymbol{x})$ for short. In addition, we define $\boldsymbol{z}^* = f_{\mu_z}(\boldsymbol{x})$. Given these definitions, it follows that

$$\begin{aligned}\frac{q_{\phi_t^*}(\boldsymbol{z}|\boldsymbol{x})}{p_{\theta_t^*}(\boldsymbol{z}|\boldsymbol{x})} &= \frac{\mathcal{N}\left(\boldsymbol{z}|\boldsymbol{z}^*,\gamma_t^*\tilde{\boldsymbol{\Sigma}}_z\right)p_{\theta_t^*}(\boldsymbol{x})}{\mathcal{N}(\boldsymbol{z}|0,I)\mathcal{N}(\boldsymbol{x}|f_{\mu_x}(\boldsymbol{z}),\gamma_t^*\boldsymbol{I})} \\ &= (2\pi)^{d/2}\gamma_t^{*(d-r)/2}\left|\tilde{\boldsymbol{\Sigma}}_z\right|^{-1/2}\exp\left\{-\frac{(\boldsymbol{z}-\boldsymbol{z}^*)^\top\tilde{\boldsymbol{\Sigma}}_z^{-1}(\boldsymbol{z}-\boldsymbol{z}^*)}{2\gamma_t^*}\right. \\ &\qquad \left.+\frac{||\boldsymbol{z}||_2^2}{2}+\frac{||\boldsymbol{x}-f_{\mu_x}(\boldsymbol{z})||_2^2}{2\gamma_t^*}\right\}p_{\theta_t^*}(\boldsymbol{x}).\end{aligned} \tag{17}$$

At this point, let

$$\boldsymbol{z} = \boldsymbol{z}^* + \sqrt{\gamma_t^*}\tilde{\boldsymbol{z}} \tag{18}$$

According to Lagrangian's mean value theorem, there exists a $\boldsymbol{z}'$ between $\boldsymbol{z}$ and $\boldsymbol{z}^*$ such that

$$f_{\mu_x}(\boldsymbol{z}) = f_{\mu_x}(\boldsymbol{z}^*) + f'_{\mu_x}(\boldsymbol{z}')(\boldsymbol{z}-\boldsymbol{z}^*) = \boldsymbol{x} + f'_{\mu_x}(\boldsymbol{z}')\sqrt{\gamma_t^*}\tilde{\boldsymbol{z}}, \tag{19}$$

where $\boldsymbol{z}' = \boldsymbol{z}^* + \eta\sqrt{\gamma_t^*}\tilde{\boldsymbol{z}}$ is between $\boldsymbol{z}$ and $\boldsymbol{z}^*$ and η is a value between 0 and 1 ($\boldsymbol{z}' = \boldsymbol{z}^*$ if $\eta = 0$ and $\boldsymbol{z}' = \boldsymbol{z}$ if $\eta = 1$). Use $C(\boldsymbol{x})$ to represent the terms not related to $\boldsymbol{z}$, i.e., $(2\pi)^{d/2}\gamma_t^{*(d-r)/2}\left|\tilde{\boldsymbol{\Sigma}}_z\right|^{-1/2}p_{\theta_t^*}(\boldsymbol{x})$. Plug (18) and (19) into (17) and consider the limit given by

$$\begin{aligned}\lim_{t\to\infty}\frac{q_{\phi_t^*}(\boldsymbol{z}|\boldsymbol{x})}{p_{\theta_t^*}(\boldsymbol{z}|\boldsymbol{x})} &= \lim_{t\to\infty}C(\boldsymbol{x})\exp\left\{-\frac{\tilde{\boldsymbol{z}}^\top\tilde{\boldsymbol{\Sigma}}_z^{-1}\tilde{\boldsymbol{z}}}{2}+\frac{||\boldsymbol{z}^*+\sqrt{\gamma_t^*}\tilde{\boldsymbol{z}}||_2^2}{2}+\frac{||f'_{\mu_x}\left(\boldsymbol{z}^*+\eta\sqrt{\gamma_t^*}\tilde{\boldsymbol{z}}\right)\tilde{\boldsymbol{z}}||_2^2}{2}\right\} \\ &= C(\boldsymbol{x})\exp\left\{-\frac{\tilde{\boldsymbol{z}}^\top\tilde{\boldsymbol{\Sigma}}_z^{-1}\tilde{\boldsymbol{z}}}{2}+\frac{||\boldsymbol{z}^*||_2^2}{2}+\frac{||f'_{\mu_x}(\boldsymbol{z}^*)\tilde{\boldsymbol{z}}||_2^2}{2}\right\} \\ &= C(\boldsymbol{x})\exp\left\{-\frac{\tilde{\boldsymbol{z}}^\top\tilde{\boldsymbol{\Sigma}}_z^{-1}\tilde{\boldsymbol{z}}}{2}+\frac{||\boldsymbol{z}^*||_2^2}{2}+\frac{\tilde{\boldsymbol{z}}^\top f'_{\mu_x}(\boldsymbol{z}^*)^\top f'_{\mu_x}(\boldsymbol{z}^*)\tilde{\boldsymbol{z}}}{2}\right\} \\ &= C(\boldsymbol{x})\exp\left\{\frac{||\boldsymbol{z}^*||_2^2}{2}\right\}\end{aligned} \tag{20}$$

The fourth equality comes from the fact that $f'_{\mu_x}(\boldsymbol{z}^*)^\top f'_{\mu_x}(\boldsymbol{z}^*) = f'_{\mu_x}(f_{\mu_z}(\boldsymbol{x}))^\top f'_{\mu_x}(f_{\mu_z}(\boldsymbol{x})) = \tilde{\boldsymbol{\Sigma}}_z(\boldsymbol{x})^{-1}$. This expression is not related to $\boldsymbol{z}$. Considering both $q_{\phi_t^*}(\boldsymbol{z}|\boldsymbol{x})$ and $p_{\theta_t^*}(\boldsymbol{z}|\boldsymbol{x})$ are probability distributions, the ratio should be equal to 1. The KL divergence between them thus converges to 0 as $t \to \infty$.

7.3 Generalization to the Case with $\kappa > r$

When $\kappa > r$, we use the first r latent dimensions to build a projection between $\boldsymbol{z}$ and $\boldsymbol{x}$ and leave the remaining $\kappa - r$ latent dimensions unused. Specifically, let $f_{\mu_x}(\boldsymbol{z}) = \tilde{f}_{\mu_x}(\boldsymbol{z}_{1:r})$, where $\tilde{f}_{\mu_x}(\boldsymbol{z}_{1:r})$ is defined as in (7) and $\gamma_t^* = 1/t$. Again consider the case that $t \to \infty$. Then this decoder can also satisfy $\lim_{t\to\infty} p_{\theta_t^*}(\boldsymbol{x}) = p_{gt}(\boldsymbol{x})$ because it produces exactly the same distribution as the decoder defined by (7) and (8). The last $\kappa - r$ dimensions contribute nothing to the generation process.

Now define the encoder as

$$f_{\mu_z}(\boldsymbol{x})_{1:r} = \tilde{f}_{\mu_x}^{-1}(\boldsymbol{x}) \tag{21}$$

$$f_{\mu_z}(\boldsymbol{x})_{r+1:\kappa} = 0 \tag{22}$$

$$f_{S_z}(\boldsymbol{x}) = \begin{bmatrix} \tilde{f}_{S_z}(\boldsymbol{x}) \\ \boldsymbol{n}_{r+1}^\top \\ \cdots \\ \boldsymbol{n}_\kappa^\top \end{bmatrix} \tag{23}$$

where $\tilde{f}_{S_z}(\boldsymbol{x})$ is defined as (14). Denote $\{\boldsymbol{n}_i\}_{i=r+1}^\kappa$ as a set of κ-dimensional column vectors satisfying

$$\tilde{f}_{S_z}(\boldsymbol{x})\boldsymbol{n}_i = 0 \tag{24}$$

$$\boldsymbol{n}_i^\top \boldsymbol{n}_j = \mathbf{1}_{i=j} \tag{25}$$

Such a set always exists because $\tilde{f}_{S_z}(\boldsymbol{x})$ is a $r \times \kappa$ matrix. So the dimension of the null space of $\tilde{f}_{S_z}(\boldsymbol{x})$ is at least $\kappa - r$. Assuming that $\{\boldsymbol{n}_i\}_{i=r+1}^\kappa$ are $\kappa - r$ basis vectors of null$(\tilde{f}_{S_z})$, then the conditions (24) and (25) will be satisfied. The variance of the approximate posterior then becomes

$$\boldsymbol{\Sigma}_z = f_{S_z}(\boldsymbol{x}) f_{S_z}(\boldsymbol{x})^\top = \begin{bmatrix} \tilde{f}_{S_z}(\boldsymbol{x})\tilde{f}_{S_z}(\boldsymbol{x})^\top & 0 \\ 0 & \boldsymbol{I}_{\kappa-r} \end{bmatrix} \tag{26}$$

The first r dimensions can exactly match the true posterior as we have already shown. The remaining $\kappa - r$ dimensions follow a standardized Gaussian distribution. Since these dimensions contribute nothing to generating $\boldsymbol{x}$, the true posterior should be the same as the prior, *i.e.* a standardized Gaussian distribution. Moreover, any of these dimensions is independent of all the other dimensions, so the corresponding off-diagonal elements of the covariance of the true posterior should equal 0. Thus the approximate posterior also matches the true posterior for the last $\kappa - r$ dimensions. As a result, we again have $\lim_{t\to\infty} \mathbb{KL}\left[q_{\phi_t^*}(\boldsymbol{z}|\boldsymbol{x})||p_{\theta_t^*}(\boldsymbol{z}|\boldsymbol{x})\right] = 0$.

8. Proof of Theorem M.2

Similar to Section 7, we also construct a bijection between $\boldsymbol{\chi}$ and $\mathbb{R}^r$ which transforms the ground-truth measure μ_{gt} to a normal Gaussian distribution. But in this construction, we

need one more step that bijects between $\boldsymbol{\chi}$ and $\mathbb{R}^r$ using the diffeomorphism $\varphi(\cdot)$, as shown in Figure 13. We will now go into the details.

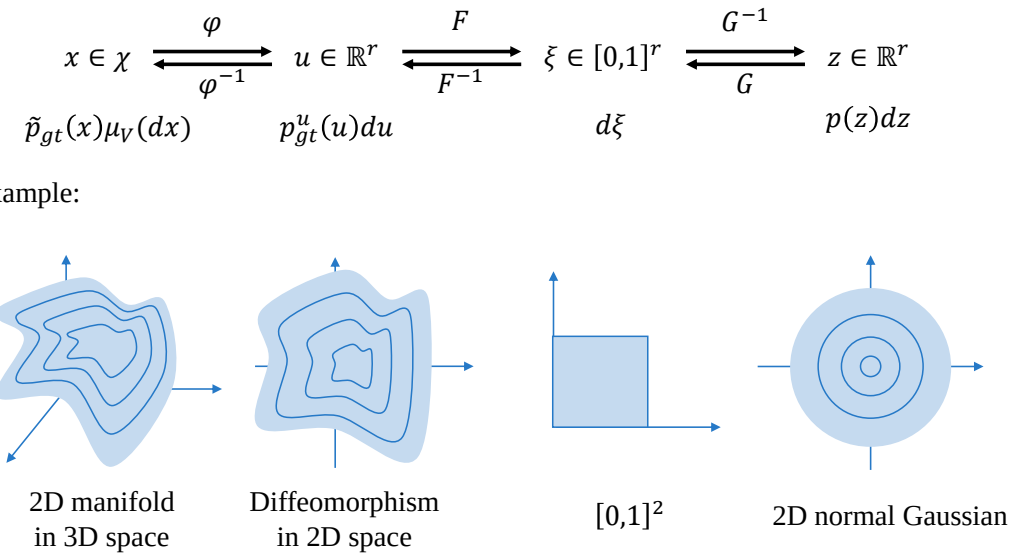


Figure 13: The relationship between different variables.

8.1 Finding a Sequence of Decoders such that $p_{\theta_t^*}(\boldsymbol{x})$ Converges to $-\infty$

$\varphi(\cdot)$ is a diffeomorphism between $\boldsymbol{\chi}$ and $\mathbb{R}^r$. So it transforms the ground-truth probability distribution $\tilde{p}_{gt}(\boldsymbol{x})$ to another distribution $p^u_{gt}(\boldsymbol{u})$, where $\boldsymbol{u} \in \mathbb{R}^r$. The relationship between the two distributions is

$$p^u_{gt}(\boldsymbol{u})d\boldsymbol{u} \;=\; \tilde{p}_{gt}(\boldsymbol{x})\mu_V(d\boldsymbol{x})|_{\boldsymbol{x}=\varphi^{-1}(\boldsymbol{u})} \;=\; \mu_{gt}(d\boldsymbol{x}), \tag{27}$$

where $\mu_V(d\boldsymbol{x})$ is the volume measure with respect to $\mathcal{X}$. Because $\varphi(\cdot)$ is a diffeomorphism, both $\varphi(\cdot)$ and $\varphi^{-1}(\cdot)$ are differentiable. Thus $d\boldsymbol{x}/d\boldsymbol{u}$ is nonzero everywhere on the manifold. Considering $\tilde{p}_{gt}(\boldsymbol{x})$ is also nonzero everywhere, $p^u_{gt}(\boldsymbol{u})$ is nonzero everywhere.

Analogous to the previous proof, define a function $F : \mathbb{R}^r \mapsto [0,1]^r$ as

$$F(\boldsymbol{u}) \;=\; [F_1(\boldsymbol{u}_1), F_2(\boldsymbol{u}_2;\boldsymbol{u}_1), ..., F_r(\boldsymbol{u}_r;\boldsymbol{u}_{1:r-1})]^\top, \tag{28}$$

$$F_i(\boldsymbol{u}_i;\boldsymbol{u}_{1:i-1}) \;=\; \int_{\boldsymbol{u}'_i=-\infty}^{\boldsymbol{u}_i} p^u_{gt}(\boldsymbol{u}'_i|\boldsymbol{u}_{1:i-1})d\boldsymbol{u}'_i. \tag{29}$$

According to this definition, we have

$$dF(\boldsymbol{u}) = p^u_{gt}(\boldsymbol{u})d\boldsymbol{u}. \tag{30}$$

Since $p^u_{gt}(\boldsymbol{u})$ is nonzero everywhere, $F(\cdot)$ is invertible. We also define another differentiable and invertible function $G : \mathbb{R}^r \mapsto [0,1]^r$ as (5).

Now let the decoder mean function be given by

$$f_{\mu_x}(\boldsymbol{z};\theta_t^*) \;=\; \varphi^{-1} \circ F^{-1} \circ G(z), \tag{31}$$

$$\gamma_t^* \;=\; \frac{1}{t}. \tag{32}$$

Then we have

$$\begin{aligned} p_{\theta_t^*}(\boldsymbol{x}) &= \int_{\mathbb{R}^r} p_{\theta_t^*}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \\ &= \int_{\mathbb{R}^r} \mathcal{N}\left(\boldsymbol{x}|\varphi^{-1}(\boldsymbol{u}), \gamma_t^*\boldsymbol{I}\right) p_{gt}^u(\boldsymbol{u})d\boldsymbol{u}. \end{aligned} \tag{33}$$

We next show that $p_{\theta_t^*}(\boldsymbol{x})$ diverges to infinite as $t \to \infty$ for any $\boldsymbol{x}$. For a given $\boldsymbol{x}$, let $\boldsymbol{u}^* = \varphi(\boldsymbol{x})$ and $B(\boldsymbol{u}^*, \sqrt{\gamma_t^*})$ be the closed ball centered at $\boldsymbol{u}^*$ with radius $\sqrt{\gamma_t^*}$. Then

$$\begin{aligned} p_{\theta_t^*}(\boldsymbol{x}) &\geq \int_{B(\boldsymbol{u}^*, \sqrt{\gamma_t^*})} \mathcal{N}\left(\boldsymbol{x}|\varphi^{-1}(\boldsymbol{u}), \gamma_t^*\boldsymbol{I}\right) p_{gt}^u(\boldsymbol{u})d\boldsymbol{u} \\ &= \int_{B(\boldsymbol{u}^*, \gamma_t^*)} (2\pi\sqrt{\gamma_t^*})^{-d/2} \exp\left\{-\frac{\left|\left|\boldsymbol{x} - \varphi^{-1}(\boldsymbol{u})\right|\right|_2^2}{2\gamma_t^*}\right\} p_{gt}^u(\boldsymbol{u})d\boldsymbol{u}. \end{aligned} \tag{34}$$

According to the Lagrangian's mean value theorem, there exists a $\boldsymbol{u}'$ between $\boldsymbol{u}$ and $\boldsymbol{u}^*$ such that

$$\varphi^{-1}(\boldsymbol{u}) \;=\; \varphi^{-1}(\boldsymbol{u}^*) + \frac{d\varphi^{-1}(\boldsymbol{u})}{d\boldsymbol{u}}|_{\boldsymbol{u}=\boldsymbol{u}'}(\boldsymbol{u} - \boldsymbol{u}^*) \;=\; \boldsymbol{x} + \frac{d\varphi^{-1}(\boldsymbol{u})}{d\boldsymbol{u}}|_{\boldsymbol{u}=\boldsymbol{u}'}(\boldsymbol{u} - \boldsymbol{u}^*). \tag{35}$$

If we denote $\Lambda(\boldsymbol{u}') = \left(\frac{d\varphi^{-1}(\boldsymbol{u})}{d\boldsymbol{u}}|_{\boldsymbol{u}=\boldsymbol{u}'}\right)^\top \left(\frac{d\varphi^{-1}(\boldsymbol{u})}{d\boldsymbol{u}}|_{\boldsymbol{u}=\boldsymbol{u}'}\right)$, we then have that

$$\begin{aligned} \left|\left|\boldsymbol{x} - \varphi^{-1}(\boldsymbol{u})\right|\right|_2^2 &= (\boldsymbol{u} - \boldsymbol{u}^*)^\top \Lambda(\boldsymbol{u}')(\boldsymbol{u} - \boldsymbol{u}^*) \;=\; \sum_{i,j} \Lambda(\boldsymbol{u}')_{i,j}(\boldsymbol{u}_i - \boldsymbol{u}_i^*)^\top(\boldsymbol{u}_j - \boldsymbol{u}_j^*) \\ &\leq \sum_i \left(\sum_j \Lambda(\boldsymbol{u}')_{i,j}\right)(\boldsymbol{u}_i - \boldsymbol{u}_i^*)^2 \;\leq\; \left|\left|\Lambda(\boldsymbol{u}')\right|\right|_1 \cdot ||\boldsymbol{u} - \boldsymbol{u}^*||_2^2. \end{aligned} \tag{36}$$

And after defining

$$D(\boldsymbol{u}^*) \;=\; \max_{\boldsymbol{u}\in B(\boldsymbol{u}^*,1)} ||\Lambda(\boldsymbol{u})||_1 \;\geq\; \max_{\boldsymbol{u}\in B(\boldsymbol{u}^*,\sqrt{\gamma_t^*})} ||\Lambda(\boldsymbol{u})||\,, \tag{37}$$

it also follows that

$$\left|\left|\boldsymbol{x} - \varphi^{-1}(\boldsymbol{u})\right|\right|_2^2 \;\leq\; \left|\left|\Lambda(\boldsymbol{u}')\right|\right|_1 \cdot ||\boldsymbol{u} - \boldsymbol{u}^*||_2^2 \;\leq\; D(\boldsymbol{u}^*)\gamma_t^* \quad \forall \boldsymbol{u} \in B(\boldsymbol{u}^*, \sqrt{\gamma_t^*}). \tag{38}$$

Plugging this inequality into (34) gives

$$\begin{aligned} p_{\theta_t^*}(\boldsymbol{x}) &\geq \int_{B(\boldsymbol{u}^*, \sqrt{\gamma_t^*})} (2\pi\gamma_t^*)^{-d/2} \exp\left\{-\frac{D(\boldsymbol{u}^*)\gamma_t^*}{2\gamma_t^*}\right\} p_{gt}^u(\boldsymbol{u})d\boldsymbol{u} \\ &\geq (2\pi\gamma_t^*)^{-d/2} \exp\left\{-\frac{D(\boldsymbol{u}^*)}{2}\right\} \left(\min_{\boldsymbol{u}\in B(\boldsymbol{u}^*,1)} p_{gt}^u(\boldsymbol{u})\right) \int_{B(\boldsymbol{u}^*, \sqrt{\gamma_t^*})} d\boldsymbol{u} \\ &= (2\pi\gamma_t^*)^{-d/2} \exp\left\{-\frac{D(\boldsymbol{u}^*)}{2}\right\} \left(\min_{\boldsymbol{u}\in B(\boldsymbol{u}^*,1)} p_{gt}^u(\boldsymbol{u})\right) V\left(B(\boldsymbol{u}^*, \sqrt{\gamma_t^*})\right), \end{aligned} \tag{39}$$

where $V\left(B(\boldsymbol{u}^*, \sqrt{\gamma_t^*})\right)$ is the volume of the r-dimensional ball $B(\boldsymbol{u}^*, \sqrt{\gamma^*})$. The volume should be $a_r {\gamma^*}^{r/2}$ where a_r is a constant related to the dimension r. So

$$p_{\theta_t^*}(\boldsymbol{x}) \geq (2\pi)^{-d/2} {\gamma^*}^{-(d-r)/2} a_r \exp\left\{-\frac{D(\boldsymbol{u}^*)}{2}\right\} \left(\min_{\boldsymbol{u}\in B(\boldsymbol{u}^*,1)} p_{gt}^u(\boldsymbol{u})\right). \tag{40}$$

Since $\varphi(\cdot)$ defines a diffeomorphism, $D(\boldsymbol{u}^*) < \infty$. Moreover, $\left(\min_{\boldsymbol{u}\in B(\boldsymbol{u}^*,1)} p_{gt}^u(\boldsymbol{u})\right) > 0$ because $p_{gt}^u(\boldsymbol{u})$ is nonzero and continuous everywhere. We may then conclude that

$$\lim_{t\to\infty} -\log p_{\theta^*}(\boldsymbol{x}) = -\infty. \tag{41}$$

for $\boldsymbol{x} \in \mathcal{X}$. This then implies that the stated average across $\mathcal{X}$ with respect to μ_{gt} will also be $-\infty$.

8.2 Finding a Sequence of Encoders such that $\mathbb{KL}\left[q_{\phi_t^*}(\boldsymbol{z}|\boldsymbol{x})||p_{\theta_t^*}(\boldsymbol{z}|\boldsymbol{x})\right]$ Converges to 0

Similar to (13) and (14), let the encoder be

$$f_{\mu_z}(\boldsymbol{x};\phi_t^*) = G^{-1}\circ F\circ\varphi(\boldsymbol{x}) = f_{\mu_x}^{-1}(\boldsymbol{x};\theta_t^*), \tag{42}$$

$$f_{S_z}(\boldsymbol{x};\phi_t^*) = \sqrt{\gamma_t^*\left(f'_{\mu_x}\left(f_{\mu_z}(\boldsymbol{x};\phi_t^*);\theta_t^*\right)^\top f'_{\mu_x}\left(f_{\mu_z}(\boldsymbol{x};\phi_t^*);\theta_t^*\right)\right)^{-1}}. \tag{43}$$

Following the proofs in Section 7.2, we can prove the KL divergence between $q_{\phi_t^*}(\boldsymbol{z}|\boldsymbol{x})$ and $p_{\theta_t^*}(\boldsymbol{z}|\boldsymbol{x})$ converges to 0.

8.3 The Relationship between $\lim_{t\to\infty} p_{\theta_t^*}(\boldsymbol{x})$ and $\mu_{gt}(\boldsymbol{x})$

We then prove our construction from (31) and (32) satisfies (M.6). Unlike the case $d = r$ where we can compare $p_{\theta_t^*}(\boldsymbol{x})$ and $p_{gt}(\boldsymbol{x})$ directly, here $p_{\theta_t^*}(\boldsymbol{x})$ is a density defined everywhere in $\mathbb{R}^d$ while μ_{gt} is a probability measure defined only on the r-dimensional manifold $\boldsymbol{\chi}$. Consequently, to assess $p_{\theta_t^*}(\boldsymbol{x})$ relative to μ_{gt}, we evaluate the respective probability mass assigned to any measurable subset of $\mathbb{R}^d$ denoted as A. For $p_{\theta_t^*}(\boldsymbol{x})$, we integrate the density over A while for μ_{gt} we compute the measure of the intersection of A with $\boldsymbol{\chi}$, i.e., μ_{gt} confines all mass to the manifold.

We begin with the probability distribution given by $p_{\theta_t^*}(\boldsymbol{x})$:

$$\begin{aligned}
p_{\theta_t^*}(\boldsymbol{x}) &= \int_{\mathbb{R}^r} p_{\theta_t^*}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \int_{\mathbb{R}^r} \mathcal{N}\left(\boldsymbol{x}|\varphi^{-1}\circ F^{-1}\circ G(z), \gamma_t^*\boldsymbol{I}\right) dG(\boldsymbol{z}) \\
&= \int_{[0,1]^r} \mathcal{N}\left(\boldsymbol{x}|\varphi^{-1}\circ F^{-1}(\boldsymbol{\xi}), \gamma_t^*\boldsymbol{I}\right) d\boldsymbol{\xi} \\
&= \int_{\mathbb{R}^r} \mathcal{N}\left(\boldsymbol{x}|\varphi^{-1}(\boldsymbol{u}), \gamma_t^*\boldsymbol{I}\right) p_{gt}^u(\boldsymbol{u})d\boldsymbol{u} \\
&= \int_{\boldsymbol{x}'\in\boldsymbol{\chi}} \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{x}', \gamma_t^*\right) \mu_{gt}(d\boldsymbol{x}').
\end{aligned} \tag{44}$$

Consider a measurable set $A \in \mathbb{R}^d$,

$$\begin{aligned}\lim_{t\to\infty}\int_{\boldsymbol{x}\in A} p_{\theta_t^*}(\boldsymbol{x})d\boldsymbol{x} &= \lim_{t\to\infty}\int_{\boldsymbol{x}\in A}\left[\int_{\boldsymbol{x}'\in\boldsymbol{\chi}}\mathcal{N}\left(\boldsymbol{x}|\boldsymbol{x}',\gamma_t^*\right)\mu_{gt}(d\boldsymbol{x}')\right]d\boldsymbol{x}\\ &= \lim_{t\to\infty}\int_{\boldsymbol{x}'\in\boldsymbol{\chi}}\left[\int_{\boldsymbol{x}\in A}\mathcal{N}\left(\boldsymbol{x}|\boldsymbol{x}',\gamma_t^*\right)d\boldsymbol{x}\right]\mu_{gt}(d\boldsymbol{x}')\\ &= \int_{\boldsymbol{x}'\in\boldsymbol{\chi}}\lim_{t\to\infty}\left[\int_{\boldsymbol{x}\in A}\mathcal{N}\left(\boldsymbol{x}|\boldsymbol{x}',\gamma_t^*\right)d\boldsymbol{x}\right]\mu_{gt}(d\boldsymbol{x}'). \qquad (45)\end{aligned}$$

The second equation that interchanges the order of the integrations admitted by Fubini's theorem. The third equation that interchanges the order of the integration and the limit is justified by the bounded convergence theorem. We now note that the term inside the first integration, $\mathcal{N}(\boldsymbol{x}|\boldsymbol{x}',\gamma_t^*\boldsymbol{I})$, converges to a Dirac-delta function as $\gamma_t^*\to 0$. So the integration over A depends on whether $\boldsymbol{x}'$ is inside A or not, i.e.,

$$\lim_{t\to\infty}\left[\int_{\boldsymbol{x}\in\boldsymbol{A}}\mathcal{N}\left(\boldsymbol{x}|\boldsymbol{x}',\gamma_t^*\boldsymbol{I}\right)d\boldsymbol{x}\right] = \begin{cases}1 & \text{if } \boldsymbol{x}'\in\boldsymbol{A}-\partial\boldsymbol{A},\\ 0 & \text{if } \boldsymbol{x}'\in\boldsymbol{A}^c-\partial\boldsymbol{A}.\end{cases} \qquad (46)$$

We separate the manifold $\boldsymbol{\chi}$ into three parts: $\boldsymbol{\chi}\cap(\boldsymbol{A}-\partial\boldsymbol{A})$, $\boldsymbol{\chi}\cap(\boldsymbol{A}^c-\partial\boldsymbol{A})$ and $\boldsymbol{\chi}\cap\partial\boldsymbol{A}$. Then (45) can be separated into three parts accordingly. The first two parts can be derived as

$$\int_{\boldsymbol{\chi}\cap(\boldsymbol{A}-\partial\boldsymbol{A})}\lim_{t\to\infty}\left[\int_{\boldsymbol{A}}\mathcal{N}(\boldsymbol{x}|\boldsymbol{x}',\gamma_t^*\boldsymbol{I})d\boldsymbol{x}\right]\mu_{gt}(d\boldsymbol{x}') = \int_{\boldsymbol{\chi}\cap(\boldsymbol{A}-\partial\boldsymbol{A})}1\,\mu_{gt}(d\boldsymbol{x}') = \mu_{gt}\left(\boldsymbol{\chi}\cap(\boldsymbol{A}-\partial\boldsymbol{A})\right), \qquad (47)$$

$$\int_{\boldsymbol{\chi}\cap(\boldsymbol{A}^c-\partial\boldsymbol{A})}\lim_{t\to\infty}\left[\int_{\boldsymbol{A}}\mathcal{N}(\boldsymbol{x}|\boldsymbol{x}',\gamma_t^*\boldsymbol{I})d\boldsymbol{x}\right]\mu_{gt}(d\boldsymbol{x}') = \int_{\boldsymbol{\chi}\cap(\boldsymbol{A}-\partial\boldsymbol{A})}0\,\mu_{gt}(d\boldsymbol{x}') = 0. \qquad (48)$$

For the third part, given the assumption that $\mu_{gt}(\partial\boldsymbol{A}) = 0$, we have

$$0 \le \int_{\boldsymbol{\chi}\cap\partial\boldsymbol{A}}\lim_{t\to\infty}\left[\int_{\boldsymbol{A}}\mathcal{N}(\boldsymbol{x}|\boldsymbol{x}',\gamma_t^*\boldsymbol{I})d\boldsymbol{x}\right]\mu_{gt}(d\boldsymbol{x}') \le \int_{\boldsymbol{\chi}\cap\partial\boldsymbol{A}}1\,\mu_{gt}(d\boldsymbol{x}') = \mu_{gt}\left(\boldsymbol{\chi}\cap\partial\boldsymbol{A}\right) = 0. \qquad (49)$$

Therefore we have

$$\int_{\boldsymbol{\chi}\cap\partial\boldsymbol{A}}\lim_{t\to\infty}\left[\int_{\boldsymbol{A}}\mathcal{N}(\boldsymbol{x}|\boldsymbol{x}',\gamma_t^*\boldsymbol{I})d\boldsymbol{x}\right]\mu_{gt}(d\boldsymbol{x}') = 0 \qquad (50)$$

and thus

$$\begin{aligned}\lim_{t\to\infty}\int_{\boldsymbol{A}}p_{gt}(\boldsymbol{x};\gamma_t^*\boldsymbol{I})d\boldsymbol{x} &= \int_{\boldsymbol{x}'\in\boldsymbol{\chi}}\lim_{t\to\infty}\left[\int_{\boldsymbol{x}\in A}\mathcal{N}\left(\boldsymbol{x}|\boldsymbol{x}',\gamma_t^*\right)d\boldsymbol{x}\right]\mu_{gt}(d\boldsymbol{x}')\\ &= \mu_{gt}\left(\boldsymbol{\chi}\cap(\boldsymbol{A}-\partial\boldsymbol{A})\right)+0+0\\ &= \mu_{gt}(\boldsymbol{\chi}\cap\boldsymbol{A}), \qquad (51)\end{aligned}$$

leading to (M.6).

Note that this result involves a subtle requirement involving the boundary ∂A. This condition is only included to handle a minor, practically-inconsequential technicality. In brief, as a density $p_{\theta_t^*}(\boldsymbol{x})$ will apply zero mass exactly on any low-dimensional manifold, although it can apply all of its mass to any region in the neighborhood of $\boldsymbol{\chi}$. But suppose we choose some A is a subset of $\boldsymbol{\chi}$, i.e, it is exclusively confined to the ground-truth manifold. Then the probability mass within A assigned by μ_{gt} will be nonzero while that given by $p_{\theta_t^*}(\boldsymbol{x})$ can still be zero. Of course this does not mean that $p_{\theta_t^*}(\boldsymbol{x})$ and μ_{gt} do not match each other in any practical sense. This is because if we expand this specialized A by an arbitrary small d-dimensional volume, then $p_{\theta_t^*}(\boldsymbol{x})$ and μ_{gt} will now supply essentially the same probability mass on this infinitesimally expanded set (which is arbitrary close to A).

9. Proof of Theorem M.3

From the main text, $\{\theta_\gamma^*, \phi_\gamma^*\}$ is the optimal solution with a fixed γ. The true posterior and the approximate posterior are

$$p_{\theta_\gamma^*}(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{z})p_{\theta_\gamma^*}(\boldsymbol{x}|\boldsymbol{z})}{p_{\theta_\gamma^*}(\boldsymbol{x})}, \tag{52}$$

$$q_{\phi_\gamma^*}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}|\boldsymbol{\mu}_z(\boldsymbol{x};\phi_\gamma^*), \boldsymbol{\Sigma}_z\big(\boldsymbol{x};\phi_\gamma^*\big)\right). \tag{53}$$

9.1 Case 1: $r = d$

We first argue that the KL divergence between $p_{\theta_\gamma^*}(\boldsymbol{z}|\boldsymbol{x})$ and $q_{\phi_\gamma^*}(\boldsymbol{z}|\boldsymbol{x})$ is always strictly greater than zero. This can be proved by contradiction. Suppose the KL divergence exactly equals zero. Then $p_{\theta_\gamma^*}(\boldsymbol{z}|\boldsymbol{x})$ must also be a Gaussian distribution, meaning that the logarithm of $p_{\theta_\gamma^*}(\boldsymbol{z}|\boldsymbol{x})$ is a quadratic form in $\boldsymbol{z}$. In particular, we have

$$\begin{aligned}\log p_{\theta_\gamma^*}(\boldsymbol{z}|\boldsymbol{x}) &= \log\mathcal{N}(\boldsymbol{z}|0,\boldsymbol{I}) + \log\mathcal{N}(\boldsymbol{x}|f_{\mu_x}(\boldsymbol{z}),\gamma\boldsymbol{I}) - \log p(\boldsymbol{x}) \\ &= -\frac{1}{2}||\boldsymbol{z}||_2^2 - \frac{1}{2\gamma}||\boldsymbol{x} - f_{\mu_x}(\boldsymbol{z})||_2^2 + \text{constant},\end{aligned} \tag{54}$$

where we have absorbed all the terms not related to $\boldsymbol{z}$ into a constant, and it must be that

$$f_{\mu_x}(\boldsymbol{z}) = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}, \tag{55}$$

for some matrix $\boldsymbol{W}$ and vector $\boldsymbol{b}$. Then we have

$$\begin{aligned}p_{\theta_\gamma^*}(\boldsymbol{x}) &= \int_{\mathbb{R}^\kappa} p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \\ &= \int_{\mathbb{R}^\kappa} \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z}+\boldsymbol{b},\gamma\boldsymbol{I})\mathcal{N}(\boldsymbol{z}|0,\boldsymbol{I})d\boldsymbol{z}.\end{aligned} \tag{56}$$

This is a Gaussian distribution in $\mathbb{R}^d$ which contradicts our assumption that $p_{gt}(\boldsymbol{x})$ is not Gaussian. So the KL divergence between $p_{\theta_\gamma^*}(\boldsymbol{z}|\boldsymbol{x})$ and $q_{\phi_\gamma^*}(\boldsymbol{z}|\boldsymbol{x})$ is always greater than 0. As a result, $\mathcal{L}(\theta_\gamma^*, \phi_\gamma^*)$ cannot reach the theoretical optimal solution, i.e., $\int_{\boldsymbol{\chi}} -p_{gt}(\boldsymbol{x})\log p_{gt}(\boldsymbol{x})d\boldsymbol{x}$. Denote the gap between $\mathcal{L}(\theta_\gamma^*, \phi_\gamma^*)$ and $\int_{\boldsymbol{\chi}} -p_{gt}(\boldsymbol{x})\log p_{gt}(\boldsymbol{x})d\boldsymbol{x}$ as ϵ. According to the proof

in Section 7, there exists a t_0 such that for any $t > t_0$, the gap between the proposed solution in Section 7 and the theoretical optimal solution is smaller than ϵ. Pick some $t > t_0$ such that $1/t < \gamma$ and let $\gamma' = 1/t$. Then

$$\mathcal{L}(\theta^*_{\gamma'}, \phi^*_{\gamma'}) \leq \mathcal{L}(\theta^*_t, \phi^*_t) < \mathcal{L}(\theta^*_\gamma, \phi^*_\gamma). \tag{57}$$

The first inequality comes from the fact that $\{\theta^*_{\gamma'}, \phi^*_{\gamma'}\}$ is the optimal solution when γ is fixed at γ' while $\{\theta^*_t, \phi^*_t\}$ is just one solution with $\gamma = 1/t = \gamma'$. The second inequality holds because we chose $\{\theta^*_t, \phi^*_t\}$ to be a better solution than $\{\theta^*_\gamma, \phi^*_\gamma\}$.

9.2 Case 2: $r < d$

In this case, $\mathbb{KL}\left[q_{\phi^*_\gamma}(\boldsymbol{z}|\boldsymbol{x})||p_{\theta^*_\gamma}(\boldsymbol{z}|\boldsymbol{x})\right]$ does not need to be zero because it is possible that $-\log p_{\theta^*_\gamma}(\boldsymbol{x})$ diverges to negative infinity and absorbs the positive cost caused by the KL divergence. Consider the objective function expression from (M.2). In can be bounded by

$$\begin{aligned}
\mathcal{L}(\theta^*_\gamma, \phi^*_\gamma) &= \int_{\boldsymbol{\mathcal{X}}} \left\{ -\mathbb{E}_{q_{\phi^*_\gamma}(\boldsymbol{z}|\boldsymbol{x})} \left[\log p_{\theta^*_\gamma}(\boldsymbol{x}|\boldsymbol{z})\right] + \mathbb{KL}\left[q_{\phi^*_\gamma}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right] \right\} \mu_{gt}(d\boldsymbol{x}) \\
&\geq \int_{\boldsymbol{\mathcal{X}}} \left\{ -\mathbb{E}_{q_{\phi^*_\gamma}(\boldsymbol{z}|\boldsymbol{x})} \left[\frac{||\boldsymbol{x} - f_{\mu_x}(\boldsymbol{z})||_2^2}{2\gamma} + \frac{d}{2}\log(2\pi\gamma)\right] \right\} \mu_{gt}(d\boldsymbol{x}) \\
&\geq \frac{d}{2}\log\gamma > -\infty.
\end{aligned} \tag{58}$$

The first inequality holds discarding the KL term, which is non-negative. The second inequality holds because a quadratic term is removed. Furthermore, according to the proof in Section 8, there exists a t_0 such that for any $t > t_0$,

$$\mathcal{L}(\theta^*_t, \phi^*_t) < \frac{d}{2}\log\gamma. \tag{59}$$

Again, we select a $t > t_0$ such that $1/t < \gamma$ and let $\gamma' = 1/t$. Then

$$\mathcal{L}(\theta^*_{\gamma'}, \phi^*_{\gamma'}) \leq \mathcal{L}(\theta^*_t, \phi^*_t) < \mathcal{L}(\theta^*_\gamma, \phi^*_\gamma). \tag{60}$$

10. Proof of Theorem M.4

Recall that

$$q_{\phi^*_\gamma}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}|f_{\mu_z}(\boldsymbol{x};\phi^*_\gamma), f_{S_z}(\boldsymbol{x};\phi^*_\gamma) f_{S_z}(\boldsymbol{x};\phi^*_\gamma)^\top\right), \tag{61}$$

$$p_{\theta^*_\gamma}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{x}|f_{\mu_x}(\boldsymbol{z};\theta^*_\gamma), \gamma\boldsymbol{I}\right). \tag{62}$$

Plugging these expressions into (M.2) we obtain

$$\mathcal{L}(\theta^*_\gamma, \phi^*_\gamma) = \mathbb{E}_{\boldsymbol{z}\sim q_{\phi^*_\gamma}(\boldsymbol{z}|\boldsymbol{x})} \left[\frac{1}{2\gamma}||f_{\mu_x}(\boldsymbol{z}) - \boldsymbol{x}||_2^2 + \frac{d}{2}\log(2\pi\gamma)\right] + \mathbb{KL}\left[q_{\phi^*_\gamma}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right] \tag{63}$$

$$\geq \frac{1}{2\gamma}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[||f_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x}) + f_{S_z}(\boldsymbol{x})\epsilon\right] - \boldsymbol{x}||_2^2\right] + \frac{d}{2}\log(2\pi\gamma), \tag{64}$$

where we have omitted explicit inclusion of the parameters ϕ_γ^* and θ_γ^* in the functions $f_{\mu_z}(\cdot), f_{S_z}(\cdot)$ and $f_{\mu_x}(\cdot)$ to avoid undue clutter. Now suppose

$$\lim_{\gamma\to 0} \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[||f_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x}) + f_{S_z}(\boldsymbol{x})\epsilon\right] - \boldsymbol{x}||_2^2\right] = \Delta \neq 0. \tag{65}$$

It then follows that

$$\lim_{\gamma\to 0} \mathcal{L}(\theta_\gamma^*, \phi_\gamma^*) \geq \lim_{\gamma\to 0} \frac{\Delta}{2\gamma} + \frac{d}{2}\log(2\pi\gamma) = +\infty, \tag{66}$$

which contradicts the fact that $\mathcal{L}(\theta_\gamma^*, \phi_\gamma^*)$ converges to $-\infty$. So we must have that

$$\lim_{\gamma\to 0} \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\boldsymbol{I})}\left[||f_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x}) + f_{S_z}(\boldsymbol{x})\epsilon\right] - \boldsymbol{x}||_2^2\right] = 0. \tag{67}$$

Because the term inside the expectation, i.e., $||f_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x}) + f_{S_z}(\boldsymbol{x})\epsilon\right] - \boldsymbol{x}||_2^2$, is always non-negative, we can conclude that

$$\lim_{\gamma\to 0} f_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x}) + f_{S_z}(\boldsymbol{x})\epsilon\right] = \boldsymbol{x}. \tag{68}$$

And if we let $\epsilon = 0$, this equation then becomes

$$\lim_{\gamma\to 0} f_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right] = \boldsymbol{x}. \tag{69}$$

11. Further Analysis of the VAE Cost as γ becomes small

In the main paper, we mentioned that the squared eigenvalues of $f_{S_z}(\boldsymbol{x};\phi_\gamma^*)$ will become arbitrary small at a rate proportional to γ. To justify this, we borrow the simplified notation from the proof of Theorem M.4 and expand $f_{\mu_x}(\boldsymbol{z})$ at $\boldsymbol{z} = f_{\mu_z}(\boldsymbol{x})$ using a Taylor series. Omitting the high order terms (in the present narrow context around the neighborhood of VAE global optima these will be small), this gives

$$f_{\mu_x}(\boldsymbol{z}) \approx f_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right] + f'_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right](\boldsymbol{z} - f_{\mu_z}) \approx \boldsymbol{x} + f'_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right](\boldsymbol{z} - f_{\mu_z}). \tag{70}$$

Plug this expression and (61) into (63), we obtain

$$\begin{aligned}
\mathcal{L}(\theta_\gamma^*, \phi_\gamma^*) \approx\ & \mathbb{E}_{\boldsymbol{z}\sim q_{\phi_\gamma^*}(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{1}{2\gamma}||f'_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right](\boldsymbol{z} - f_{\mu_z}(\boldsymbol{x}))||_2^2 + \frac{d}{2}\log(2\pi\gamma)\right] \\
& + \frac{1}{2}\left\{||f_{\mu_z}(\boldsymbol{x})||_2^2 + \mathrm{tr}\left(f_{S_z}(\boldsymbol{x})f_{S_z}(\boldsymbol{x})^\top\right) - \log|f_{S_z}(\boldsymbol{x})f_{S_z}(\boldsymbol{x})^\top| - \kappa\right\} \\
=\ & \frac{1}{2\gamma}\mathrm{tr}\left(\mathbb{E}_{\boldsymbol{z}\sim q_{\phi_\gamma^*}(\boldsymbol{z}|\boldsymbol{x})}\left[(\boldsymbol{z} - f_{\mu_z}(\boldsymbol{x}))^\top(\boldsymbol{z} - f_{\mu_z}(\boldsymbol{x}))\right] f'_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right]^\top f'_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right]\right) \\
& + \frac{d}{2}\log(2\pi\gamma) + \frac{1}{2}\left\{||f_{\mu_z}(\boldsymbol{x})||_2^2 + tr\left(f_{S_z}(\boldsymbol{x})f_{S_z}(\boldsymbol{x})^\top\right) - \log|f_{S_z}(\boldsymbol{x})f_{S_z}(\boldsymbol{x})^\top| - \kappa\right\} \\
=\ & tr\left(f_{S_z}(\boldsymbol{x})f_{S_z}(\boldsymbol{x})^\top\left[\frac{1}{2}\boldsymbol{I} + \frac{1}{2\gamma}f'_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right]^\top f'_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right]\right]\right) \\
& + \frac{d}{2}\log(2\pi\gamma) + \frac{1}{2}\left\{||f_{\mu_z}(\boldsymbol{x})||_2^2 - \log|f_{S_z}(\boldsymbol{x})f_{S_z}(\boldsymbol{x})^\top| - \kappa\right\}.
\end{aligned} \tag{71}$$

From these manipulations we may conclude that the optimal value of $f_{S_z}(\boldsymbol{x})f_{S_z}(\boldsymbol{x})^\top$ must satisfy

$$\left[\frac{1}{2}\boldsymbol{I} + \frac{1}{2\gamma} f'_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right]^\top f'_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right]\right] - \frac{1}{2}\left(f_{S_z}(\boldsymbol{x})f_{S_z}(\boldsymbol{x})^\top\right)^{-1} = 0. \tag{72}$$

So

$$f_{S_z}(\boldsymbol{x})f_{S_z}(\boldsymbol{x})^\top = \left[\boldsymbol{I} + \frac{1}{\gamma} f'_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right]^\top f'_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right]\right]^{-1}. \tag{73}$$

Note that $f'_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right]$ is the tangent space of the manifold $\boldsymbol{\chi}$ at $f_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right]$, so the rank must be r. $f'_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right]^\top f'_{\mu_x}\left[f_{\mu_z}(\boldsymbol{x})\right]$ can be decomposed as $\boldsymbol{U}^\top\boldsymbol{S}\boldsymbol{U}$, where $\boldsymbol{U}$ is a κ-dimensional orthogonal matrix and $\boldsymbol{S}$ is a κ-dimensional diagonal matrix with r nonzero elements. Denote $\text{diag}[\boldsymbol{S}] = [S_1, S_2, ..., S_r, 0, ..., 0]$. Then

$$f_{S_z}(\boldsymbol{x})f_{S_z}(\boldsymbol{x})^\top = \left[\boldsymbol{U}^\top \text{diag}\left[1 + \frac{S_1}{\gamma}, ..., 1 + \frac{S_r}{\gamma}, 1, ..., 1\right]\boldsymbol{U}\right]^{-1}. \tag{74}$$

Case 1: $r = \kappa$**.** In this case, $\boldsymbol{S}$ has no nonzero diagonal elements, and therefore

$$f_{S_z}(\boldsymbol{x})f_{S_z}(\boldsymbol{x})^\top = \left[\boldsymbol{U}^\top \text{diag}\left[\frac{1}{1+\frac{S_1}{\gamma}}, ..., \frac{1}{1+\frac{S_r}{\gamma}}\right]\boldsymbol{U}\right]. \tag{75}$$

As $\gamma \to 0$, the eigenvalues of $f_{S_z}(\boldsymbol{x})f_{S_z}(\boldsymbol{x})^\top$, which are given by $\frac{1}{1+\frac{S_i}{\gamma}}$, converge to 0 at a rate of $O(\gamma)$.

Case 2: $r < \kappa$**.** In this case, the first r eigenvalues also converge to 0 at a rate of $O(\gamma)$, but the remaining $\kappa - r$ eigenvalues will be 1, meaning the redundant dimensions are simply filled with noise matching the prior $p(\boldsymbol{z}$ as desired.

References

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? A large-scale study. *arXiv:1711.10337v3*, 2018.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *International Conference on Learning Representations*, 2018.