Embrace the Gap: VAEs Perform Independent Mechanism Analysis

Patrik Reizinger[*1] **Luigi Gresele**[*2] **Jack Brady**[*1] **Julius von Kügelgen**[2,3] **Dominik Zietlow**[2,4]
Bernhard Schölkopf[2] **Georg Martius**[2] **Wieland Brendel**[1] **Michel Besserve**[† 2]

Abstract

Despite the widespread use of variational autoencoders (VAEs), the consequences of optimizing the evidence lower bound (ELBO) opposed to the exact log likelihood remain poorly understood. We shed light on this matter by studying nonlinear VAEs in the limit of near-deterministic decoders. We first prove that, in this regime, the optimal encoder approximately inverts the decoder—a commonly used but unproven conjecture—which we call *self-consistency*. Leveraging self-consistency, we show that the ELBO converges to a regularized log-likelihood rather than to the exact one. The regularization term allows VAEs to perform what has been termed independent mechanism analysis (IMA): it adds an inductive bias towards decoders with column-orthogonal Jacobians. This connection to IMA allows us to precisely characterize the gap w.r.t. the log-likelihood in near-deterministic VAEs. Furthermore, it elucidates an unanticipated benefit of ELBO optimization for nonlinear representation learning as, unlike the unregularized log-likelihood, the IMA-regularized objective promotes identification of the ground-truth latent factors.

1 INTRODUCTION

Variational Autoencoders (VAEs) [19, 31] are a popular framework for generative modelling and nonlinear representation learning. Instead of optimizing the exact but intractable model evidence, VAEs employ a variational distribution parameterized by a neural network (*encoder*), to optimize the tractable evidence lower bound (ELBO). While empirically successful, the consequences of optimizing the ELBO opposed to the exact log-likelihood in VAEs remain poorly understood [36, 22].

In this work, we analyze the effects of optimizing the ELBO for VAEs in a *near-deterministic* limit for the conditional distribution parameterized by the nonlinear decoder. Our first result concerns the encoder's optimality in this regime. Previous works relied on the intuitive assumption that the optimal encoder inverts the decoder [28, 22, 38]. We formalize this *self-consistency* assumption and prove its validity for the optimal variational posterior in the near-deterministic nonlinear regime.

Using self-consistency, we show that the ELBO tends to a regularized log-likelihood rather than to the exact one as conjectured in [28]. The regularization term allows VAEs to perform what has been termed Independent Mechanism Analysis (IMA) [11]: it adds an inductive bias towards decoders with column-orthogonal Jacobians. This generalizes previous findings based on linearizations or approximations of the ELBO [32, 25, 22], and characterizes the gap w.r.t. the log-likelihood in the near-deterministic limit. Furthermore, our results elucidate the gap between the ELBO and exact log-likelihood as a mechanism through which the ELBO implements a useful inductive bias. Unlike the unregularized log-likelihood, the IMA-regularized objective promotes identification of the ground-truth latent factors under suitable assumptions [11]. Empirically, we verify our theoretical results as well as show that VAEs recover the ground-truth latent factors when the IMA assumptions are met.

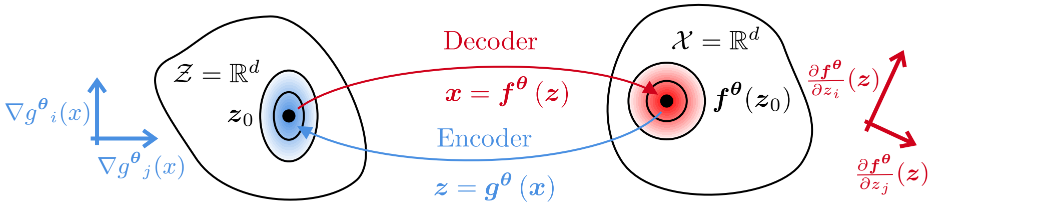
2 BACKGROUND

Variational Autoencoders. Maximizing the data likelihood in deep Latent Variable Models (LVMs) over decoder parameters $\boldsymbol{\theta}$ ($p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p_0(\boldsymbol{z})d\boldsymbol{z}$) is intractable in general, so approximate objectives are required. Variational approximations [34] replace the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ by the approximate *variational posterior* $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$, which is a stochastic mapping $\boldsymbol{x} \mapsto \boldsymbol{z}$ with parameters $\boldsymbol{\phi}$, and yields

[*] Equal contribution. Correspondence to patrik.reizinger@uni-tuebingen.de. [1] University of Tübingen; [2] MPI for Intelligent Systems, Tübingen; [3] University of Cambridge, Cambridge; [4] Amazon Web Services, Tübingen. Code: github.com/rpatrik96/ima-vae

[†] Senior author

Accepted for the 5th Workshop on Tractable Probabilistic Modeling at UAI (TPM 2022).


Decoder
$x = f^{\theta}(z)$
Encoder
$z = g^{\theta}(x)$
$\mathcal{Z} = \mathbb{R}^d$
z_0
$\mathcal{X} = \mathbb{R}^d$
$f^{\theta}(z_0)$
$\nabla g^{\theta}{}_i(x)$
$\nabla g^{\theta}{}_j(x)$
$\frac{\partial f^{\theta}}{\partial z_i}(z)$
$\frac{\partial f^{\theta}}{\partial z_j}(z)$

$$q_{\phi}(z|x) \xrightarrow{\text{Cov}} \text{diag}\left(\sigma_1^{\phi}(x)^2, \dots, \sigma_d^{\phi}(x)^2\right) = \mathbf{J}_{g^{\theta}}(x) \frac{1}{\gamma^2} \mathbf{J}_{g^{\theta}}(x)^T \xleftarrow{\text{Cov}} g^{\theta}{}_*[p_{\theta}(x|z)] \xleftarrow{g^{\theta}} p_{\theta}(x|z)$$

Figure 1: **Modeling choices in VAEs promote *Independent Mechanism Analysis* (IMA) [11].** We assume a Gaussian VAE (3), and prove that in the near-deterministic regime the mean encoder approximately inverts the mean decoder, $g^{\theta} \approx f^{\theta -1}$ (*self-consistency*, Prop. 1). **Bottom:** Closing the gap requires matching the covariances of the variational (LHS, $q_{\phi}(z|x)$) and the true posterior (RHS, approximated by $g^{\theta}{}_*[p_{\theta}(x|z)]$, cf. § 3.2 for details). Under self-consistency, an encoder with diagonal covariance enforces a row-orthogonal encoder Jacobian $\mathbf{J}_{g^{\theta}}(x)$—or equivalently, a column-orthogonal decoder Jacobian $\mathbf{J}_{f^{\theta}}(z)$. This regularization was termed Independent Mechanism Analysis (IMA) [11] and shown to be beneficial for learning the true latent factors. The connection elucidates unintended benefits of using the ELBO for representation learning.

a tractable evidence lower bound (ELBO) [19, 31] of the model's log-likelihood, defined as

$$\begin{aligned}\text{ELBO}(x, \theta, \phi) = \mathbb{E}_{q_{\phi}(z|x)}\left[\log p_{\theta}(x|z)\right] \\ - \text{KL}\left[q_{\phi}(z|x) || p_0(z)\right]. \quad (1)\end{aligned}$$

The variational approximation trades off computational efficiency with a difference w.r.t. the exact log-likelihood, which is expressed alternatively as (see [7, 20] and § 6)

$$\text{ELBO}(x, \theta, \phi) = \log p_{\theta}(x) – \text{KL}\left[q_{\phi}(z|x) || p_{\theta}(z|x)\right], \quad (2)$$

where the Kullback-Leibler Divergence (KL) between variational and true posteriors characterizes the *gap*. VAEs [19] rely on the variational approximation in (1) to train deep LVMs where neural networks parametrize the *encoder* $q_{\phi}(z|x)$ and the *decoder* $p_{\theta}(x|z)$. Common modeling choices include choosing $q_{\phi}(z|x)$ to be a factorized Gaussian with posterior means $\mu_k^{\phi}(x)$ and variances $\sigma_k^{\phi}(x)^2$ for each $z_k|x$, and choosing $p_{\theta}(x|z)$ to be a factorized Gaussian with mean $f^{\theta}(z)$ and an isotropic covariance in d dimensions,

$$z_k | x \sim \mathcal{N}(\mu_k^{\phi}(x), \sigma_k^{\phi}(x)^2);\, x|z \sim \mathcal{N}(f^{\theta}(z), \gamma^{-2}\mathbf{I}_d). \quad (3)$$

Independent Mechanism Analysis. A common assumption is that observations x can be modeled as the mixing f of a latent vector z s.t. $x = f(z)$. A core goal of representation learning is then to learn an unmixing g^{θ} such that the recovered components $y = g^{\theta}(x)$ identify the true ones up to tolerable ambiguities [2, 17]. When f is nonlinear, however, it is well known that identifying z is impossible without further constraints [15, 24]. Recently, Gresele et al. [11] proposed constraining the function class in an approach termed Independent Mechanism Analysis (IMA). IMA postulates that the latent components influence the observations "independently", where influences correspond to the partial derivatives $\partial f^{\theta}/\partial z_k$, and their non-statistical independence amounts to an orthogonality condition. While full identifiability has not been characterized for this model class, it was proven to rule out the most common counterexamples to identifiability, and empirically shown to help recover the ground-truth latent factors. To estimate a model in the IMA class, [11] proposed the IMA-regularized log-likelihood: $\mathcal{L}_{\text{IMA}}(f^{\theta}, z) := \log p_{\theta}(x) - \lambda \cdot c_{\text{IMA}}(f^{\theta}, z)$ where $c_{\text{IMA}}(f^{\theta}, z)$ encourages column-orthogonality of $\mathbf{J}_{f^{\theta}}(z)$ and is defined as

$$c_{\text{IMA}}(f^{\theta}, z) = \sum_k \log \left\| \frac{\partial f^{\theta}}{\partial z_k}(z) \right\| - \log \left| \mathbf{J}_{f^{\theta}}(z) \right| \quad (4)$$

3 THEORY

Our theoretical analysis assumes factorized densities (for $p_0(z)$, $q_{\phi}(z|x)$, and $p_{\theta}(x|z)$) and a Gaussian decoder, matching common modeling practice in VAEs.

Assumption 1 (Factorized VAE class with isotropic Gaussian decoder and log-concave prior)**.** *We are given a fixed latent prior and three parameterized classes of $\mathbb{R}^d \to \mathbb{R}^d$ mappings: the mean decoder class $\theta \mapsto f^{\theta}$, and the mean and standard deviation encoder classes, $\phi \mapsto \mu^{\phi}$ and $\phi \mapsto \sigma^{\phi}$ s.t.*

(i) *$p_0(z) \sim \prod_k m(z_k)$, with a smooth m fully supported on $\mathbb{R}$, having bounded non-positive second-order, and bounded third-order logarithmic derivatives;*

(ii) *the encoder and decoder are of the form in* (3)*, with isotropic decoder covariance $1/\gamma^2 \mathbf{I}_d$;*

(iii) *the variational mean and variance encoder classes are universal approximators;*

(iv) *for all θ, $f^{\theta}: \mathbb{R}^d \to \mathbb{R}^d$ is a bijection with inverse g^{θ}, and both are C^2 with bounded first and second order derivatives.*

Crucially, *both the mean encoder and the mean decoder can be nonlinear*. Moreover, the family of log-concave priors contains the commonly-used Gaussian distribution as a special case. We study the *near-deterministic decoder* regime of such models, where $\gamma \to +\infty$. This regime is expected to model data generating processes with vanishing observation noise well—in line with the typical ICA setting—and is commonly considered in theoretical analyses of VAEs, e.g. in [28] and in [25, 22]. Unlike [28], we consider a large but finite γ, not *at* the limit $\gamma = \infty$, where the decoder is fully deterministic. In fact, for any large but finite γ, the objective is well-behaved and amenable to theoretical analysis, while the KL-divergence is undefined in the deterministic setting. The requirement in assumption *(iv)* deviates from common practice in VAEs—where observations are typically higher-dimensional—but it allows to connect VAEs and exact likelihood methods such as normalizing flows [28].

3.1 SELF-CONSISTENCY

First, we prove *self-consistency* in the near-deterministic regime by characterizing optimal variational posteriors (i.e., those minimizing the ELBO gap w.r.t. the likelihood) for a *particular point* $\boldsymbol{x}$ and *fixed decoder parameters* $\boldsymbol{\theta}$. Based on (2), any associated optimal choice of encoder parameters satisfies

$$\begin{aligned}\widehat{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{\theta}) &\in \arg\max{}_{\boldsymbol{\phi}} \, \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi}) \\ &= \arg\min{}_{\boldsymbol{\phi}} \, \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) || p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \,. \qquad (5)\end{aligned}$$

We call *self-consistent* ELBO the resulting achieved value:

$$\mathrm{ELBO}^*(\boldsymbol{x}; \boldsymbol{\theta}) = \mathrm{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \widehat{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{\theta}))\,. \qquad (6)$$

The expression in (5) corresponds to a problem of *information projection* [4, 27] of $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ onto the set of factorized Gaussian distributions. This means that given a variational family, we search for the optimal $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ to minimize the KL to $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$. While such information projection problems are well-studied for closed convex sets where they yield a unique minimizer [5], the set projected onto in our case is not convex (convex combinations of arbitrary Gaussians are not Gaussian), making this problem of independent interest. After establishing upper and lower bounds on the KL divergence (exposed in Prop. 6-7 in § 8.2), we obtain the following self-consistency result.

Proposition 1. *[Self-consistency of near-deterministic VAEs] Under Assum. 1,* $\forall\, \boldsymbol{x}, \boldsymbol{\theta}$*, as* $\gamma \to +\infty$*, there exists at least one global minimum solution of (5), satisfying*

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) = \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) + O(1/\gamma); \; \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = O(1/\gamma^2)\,, \; \forall k. \; (7)$$

Prop. 1 states that minimizing the ELBO gap (equivalently, maximizing the ELBO) w.r.t. the encoder parameters $\boldsymbol{\phi}$ implies in the limit of large γ that the encoder's mean $\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})$ tends to $\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})$, the image of $\boldsymbol{x}$ by the *inverse* decoder. We can interpret this as the decoder "inverting" the encoder. Additionally, the variances of the encoder will converge to zero, in line with empirical observations of practitioners.

3.2 SELF-CONSISTENT ELBO AND IMA-REGULARIZED LOG-LIKELIHOOD

We want to investigate how the choice of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ implicitly regularizes the Jacobians of their means $\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})$ and $\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})$ in the near-deterministic regime. Exploiting self-consistency, we are able to precisely characterize how this happens: we formalize this in Thm. 1.

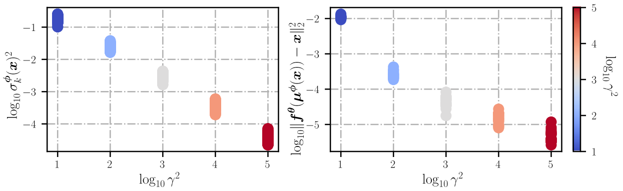


Figure 2: Self-consistency (Prop. 1) in VAE training, on a log-log plot, cf. 4 for details. **Left**: convergence of $\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2$ to 0; **Right:** convergence of $\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})$ to $\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})$

Theorem 1. *[VAEs with a near-deterministic decoder approximate the* IMA *objective] Under Assumption 1, the variational posterior satisfies (denoting* $n'' = \frac{d^2 \log p_0}{dz_k^2}$*)*

$$\begin{aligned}\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = \left[\gamma^2 \left\| \left[\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]_{:k} \right\|^2 - n''(g_k^{\boldsymbol{\theta}}(\boldsymbol{x}))\right]^{-1} \\ + O(1/\gamma^3)\,, \quad (8)\end{aligned}$$

and the self-consistent ELBO (6) *approximates the* IMA*-regularized log-likelihood [11]:*

$$\begin{aligned}\mathrm{ELBO}^*(\boldsymbol{x}; \boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - c_{\mathrm{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})) \\ + O_{\gamma \to \infty}(1/\gamma^2)\,. \quad (9)\end{aligned}$$

Proof is in § 7. We qualitatively describe the interplay between distributional assumptions in the VAE and implicit constraints on the decoder's Jacobian and its inverse.

Modeling assumptions implicitly regularize the mean decoder class $\boldsymbol{f}^{\boldsymbol{\theta}}$ under self-consistency. In the near deterministic regime, $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ gets close to the pushforward distribution of the prior by the mean decoder $\boldsymbol{f}^{\boldsymbol{\theta}}{}_*[p_0(\boldsymbol{z})]$, which can be used to show that the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p_0(\boldsymbol{z})/p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is approximately the pushforward through the inverse mean decoder $\boldsymbol{g}^{\boldsymbol{\theta}}{}_*[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})]$ (cf. § 6 for details). If we select a given latent $\boldsymbol{z}_0$ and denote its image by $\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z}_0)$, then we can locally linearize $\boldsymbol{g}^{\boldsymbol{\theta}}$ by its Jacobian $\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}} = \mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z}_0))$, yielding a Gaussian for the pushforward distribution $\boldsymbol{g}^{\boldsymbol{\theta}}{}_*[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})]$ with covariance $1/\gamma^2 \mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}} \mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}^T$ (Fig. 1). As the sufficient statistics of a Gaussian are given by its mean and covariance, the structure of the posterior covariance $\boldsymbol{\Sigma}_{\boldsymbol{z}|\boldsymbol{x}}^{\boldsymbol{\phi}}$ (by design diagonal, cf. (3)) is crucial for minimizing the gap in (2). Thus, in the zero gap limit, the covariances of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ should match, i.e., $1/\gamma^2 \mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}} \mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}^T$ will be diagonal with entries $\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2$ and therefore $\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}$ has orthogonal rows. We can express the decoder Jacobian via the inverse function theorem as $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z}_0) = \mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z}_0))^{-1}$. As the inverse of a row-orthogonal matrix has orthogonal columns, $\boldsymbol{f}^{\boldsymbol{\theta}}$ satisfies the IMA principle. Additionally, we can relate the variational posterior's variances to the column-norms of $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}$ as $\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2 = 1/\gamma^2 \left\| \left[\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z}_0)\right]_{:k} \right\|^{-2}$. The self-consistent ELBO therefore converges to the IMA-regularized log-likelihood [11].

Our argument indicates that minimizing the gap between the ELBO and the log-likelihood encourages column-

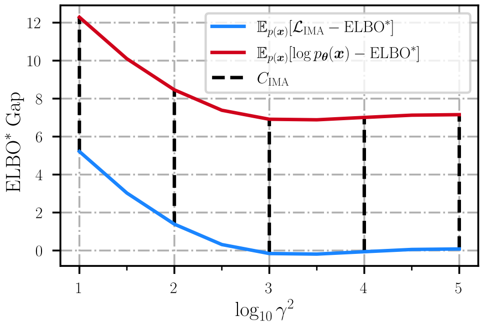


Figure 3: Comparison of the ELBO*, the IMA-regularized and unregularized log-likelihoods over different γ^2. Error bars are omitted as they are orders of magnitude smaller

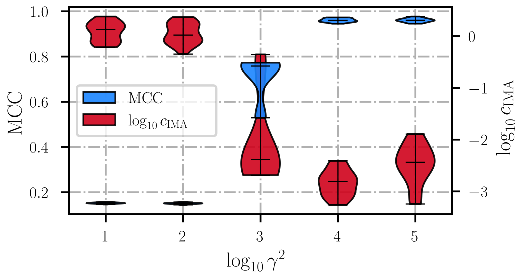


Figure 4: c_{IMA} and Mean Correlation Coefficient (MCC) for 3-dimensional Möbius mixings

orthogonality in $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}$ by matching the covariances of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and $\boldsymbol{g}^{\boldsymbol{\theta}}{}_*\,[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})]$. When $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$, the gap is closed; this is only possible if the decoder is in the IMA class, for which c_{IMA} vanishes and the ELBO *tends to an exact log-likelihood*. To the best of our knowledge, we are the first to prove this for nonlinear functions, extending related work for linear VAEs [25].

4 EXPERIMENTS

Self-consistency in practical conditions Fig. 2 empirically verifies Prop. 1 of self-consistency. We generate data by mixing latents sampled from a standard normal distribution using an MLP with smooth nonlinearities [10] and orthogonal weights—which intentionally does not belong to the IMA class, as our results are more general. We then train a Gaussian VAE (Assum. 1) with 20 seeds for each γ^2 from $\{1\mathrm{e}1; 1\mathrm{e}2; 1\mathrm{e}3; 1\mathrm{e}4; 1\mathrm{e}5\}$. The **left** plot shows that the posterior variances $\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2$ converge to zero with a $1/\gamma^2$ rate, as predicted by (7). The **right** plot shows approximate convergence of the mean encodings $\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})$ to $\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})$ with a $1/\gamma$ rate. As $\boldsymbol{f}^{\boldsymbol{\theta}}$ is not guaranteed to be invertible, we use instead the *optimal* encoder and decoder parameters to compare $\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))$ to $\boldsymbol{x}$.

Relationship between ELBO*, IMA-regularized, and unregularized log-likelihoods Fig. 3 compares the difference of the estimate of ELBO* and the unregularized/IMA-regularized log-likelihoods after convergence. We generate data by mixing latents with an invertible MLP not in the IMA class [11]. This way, we ensure that the unregularized and IMA-regularized log-likelihoods differ and make the claim of Nielsen et al. [28] comparable to ours. We then train a VAE where we fix the decoder to the ground-truth mixing. With a fixed decoder, the ELBO* depends only on $\boldsymbol{\phi}$; thus, we only train the encoder with γ^2 values from $[1\mathrm{e}1; 1\mathrm{e}5]$ (5 seeds each). As the decoder and the data are fixed, $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$ and C_{IMA} will not change during training, only ELBO* does. The figure shows that as $\gamma \to +\infty$, ELBO* approaches $\mathcal{L}_{\mathrm{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{z})$, as predicted by Thm. 1, and not $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$, as stated in [28]—the difference is C_{IMA}.

Connecting IMA, γ^2, and disentanglement Fig. 4 quantifies the relationship between the orthogonality of the decoder's Jacobian measured by the c_{IMA} and identifiability of the ground-truth latents measured by the Mean Correlation Coefficient (MCC) [14]. We use 3-dimensional Möbius transforms [29], which are in the IMA class [11], for the ground-truth mixing with *uniform* ground-truth and model prior distributions. By increasing γ^2, MCC increases, while c_{IMA} decreases, suggesting that VAEs in the near-deterministic regime promote identifiability by enforcing column-orthogonality of the decoder's Jacobian.

5 DISCUSSION

The near-deterministic regime. Our theory relies on $\gamma \to +\infty$, however, in practice large values of γ^2 may be harder to optimize due to an exploding reconstruction term in (1). This may be one explanation for the slight deviation of Fig. 2, right from our theory's predictions: while convergence of $\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})$ to $\boldsymbol{g}^{\boldsymbol{\theta}}$ matches the prediction in Prop. 1, its rate is not precisely the one predicted for the self-consistent ELBO (6). Another cause could be the encoder's finite capacity. Nonetheless, we have experimentally shown that for realistic hyperparameters, VAEs' behavior matches the predictions of our theory for the near-deterministic regime.

Characterizing the ELBO gap for nonlinear models. Thm. 1 characterizes the gap between the ELBO and true log-likelihood for nonlinear VAEs, extending the linear analysis of [25]; we also empirically characterize the gap in the deterministic limit. An unanticipated consequence of this is that (consistent with [25]) VAEs optimize the IMA-regularized log-likelihood in the near-deterministic limit, and not the unregularized one, as stated in [28].

Conclusion. We theoretically justify the widely-used self-consistency assumption in the near-deterministic regime of small decoder variance. Using this result, we show that the self-consistent ELBO converges to the IMA-regularized log-likelihood, and not to the unregularized one. Thus, we can characterize the gap between the ELBO and true log-likelihood and reason about its role as an inductive bias for representation learning in nonlinear VAEs.

Author Contributions

PR conceived the idea and managed the project. LG, JvK, and MB developed the theory with help from PR. JB and PR ran the experiments. All authors wrote the paper.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A & 01IS18039B, and by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. WB acknowledges financial support via an Emmy Noether Grant funded by the German Research Foundation (DFG) under grant no. BR 6382/1-1. DZ and PR thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support. PR acknowledges his membership in the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program.

References

[1] Khaled Alyani, Marco Congedo, and Maher Moakher. Diagonality measures of hermitian positive-definite matrices with application to the approximate joint diagonalization problem. *Linear Algebra Appl.*, 528: 290–320, September 2017. ISSN 0024-3795. doi: 10.1016/j.laa.2016.08.031. URL `https://doi.org/10.1016/j.laa.2016.08.031`. 29

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. ISSN 0162-8828, 2160-9292. doi: 10.1109/tpami.2013.50. URL `https://doi.org/10.1109/tpami.2013.50`. 2

[3] Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2615–2625, 2018. 11

[4] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley series in telecommunications. John Wiley & Sons, Inc., New York, 1991. ISBN 0471062596, 0471200611. doi: 10.1002/0471200611. URL `https://doi.org/10.1002/0471200611`. 3

[5] I. Csiszar and F. Matus. Information projections revisited. *IEEE Trans. Inf. Theory*, 49(6):1474–1490, June 2003. ISSN 0018-9448. doi: 10.1109/tit.2003.810633. URL `https://doi.org/10.1109/tit.2003.810633`. 3

[6] Bin Dai and David Wipf. Diagnosing and enhancing vae models. In *International Conference on Learning Representations*, 2018. 11

[7] Carl Doersch. Tutorial on Variational Autoencoders. *ArXiv preprint*, abs/1606.05908, 2016. 2, 11

[8] David L. Donoho and Carrie Grimes. Image Manifolds which are Isometric to Euclidean Space. *J. Math. Imaging Vis.*, 23(1):5–24, July 2005. ISSN 0924-9907, 1573-7683. doi: 10.1007/s10851-005-4965-4. URL `https://doi.org/10.1007/s10851-005-4965-4`. 32

[9] Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael J. Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 11, 28

[10] Luigi Gresele, Giancarlo Fissore, Adrián Javaloy, Bernhard Schölkopf, and Aapo Hyvärinen. Relative gradient optimization of the Jacobian term in unsupervised deep learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 4, 30, 32

[11] Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanisms analysis, a new concept? In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 28233–28248. Curran Associates, Inc., December 2021. URL `https://proceedings.neurips.cc/paper/2021/file/edc27f139c3b4e4bb29d1cdbc45663f9-Paper.pdf`. 1, 2, 3, 4, 22, 29

[12] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 11

[13] Daniella Horan, Eitan Richardson, and Yair Weiss. When is unsupervised disentanglement possible? *Advances in Neural Information Processing Systems*, 34, 2021. 32

[14] Aapo Hyvärinen and Hiroshi Morioka. Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3765–3773, 2016. 4

[15] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, April 1999. ISSN 0893-6080. doi: 10.1016/s0893-6080(98)00140-3. URL `https://doi.org/10.1016/s0893-6080(98)00140-3`. 2

[16] Jack Brady and Geoffrey Roeder. iSprites: A Dataset for Identifiable Multi-Object representation Learning. In *ICML2020: Workshop on Object-Oriented Learning*, 2020. URL `https://github.com/oolworkshop/oolworkshop.github.io/blob/master/pdf/OOL_25.pdf`. 32

[17] Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR, 2020. 2

[18] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2654–2663. PMLR, 2018. 11

[19] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 1, 2, 11

[20] Diederik P. Kingma and Max Welling. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000056. URL `https://doi.org/10.1561/2200000056`. arXiv: 1906.02691. 2

[21] David A. Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan M. Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 32

[22] Abhishek Kumar and Ben Poole. On Implicit Regularization in β-VAEs. In *International Conference on Machine Learning*, pages 5480–5490. PMLR, 2020. 1, 3, 11, 28

[23] Abhishek Kumar, Ben Poole, and Kevin Murphy. Regularized Autoencoders via Relaxed Injective Probability Flow. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 4292–4301. PMLR, 2020. 28

[24] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, 2019. 2

[25] James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. Don't blame the ELBO! a linear VAE perspective on posterior collapse. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9403–9413, 2019. 1, 3, 4, 28, 29

[26] Jerrold E. Marsden and Anthony Tromba. *Vector calculus*. W.H. Freeman and Company, New York, sixth edition edition, 2012. ISBN 978-1-4292-1508-4. 26

[27] Kevin P Murphy. *Machine learning: A probabilistic perspective*. MIT press, 2012. 3

[28] Didrik Nielsen, Priyank Jaini, Emiel Hoogeboom, Ole Winther, and Max Welling. SurVAE flows: Surjections to bridge the gap between VAEs and flows. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1, 3, 4, 28

[29] Robert Phillips. Liouville's theorem. *Pac. J. Math.*, 28 (2):397–405, February 1969. ISSN 0030-8730, 0030-8730. doi: 10.2140/pjm.1969.28.397. URL https://doi.org/10.2140/pjm.1969.28.397. 4, 32

[30] Lutz Prechelt. Early stopping - but when? In *Neural Networks: Tricks of the Trade, volume 1524 of LNCS, chapter 2*, pages 55–69. Springer-Verlag, 1997. 30, 32, 33

[31] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1278–1286. JMLR.org, 2014. 1, 2

[32] Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue PCA directions (by accident). In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12406–12415. IEEE, June 2019. doi: 10.1109/cvpr.2019.01269. URL https://doi.org/10.1109/cvpr.2019.01269. 1, 11, 28, 32

[33] Oleh Rybkin, Kostas Daniilidis, and Sergey Levine. Simple and effective VAE training with calibrated decoders. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9179–9189. PMLR, 2021. 11

[34] Michael Struwe. *Variational Methods*, volume 991. Springer Berlin Heidelberg, 2000. ISBN 9783662041963, 9783662041949. doi: 10.1007/978-3-662-04194-9. URL https://doi.org/10.1007/978-3-662-04194-9. 1

[35] Michael E. Tipping and Christopher M. Bishop. Probabilistic Principal Component Analysis. *J. R. Stat. Soc. B*, 61(3):611–622, August 1999. ISSN 1369-7412, 1467-9868. doi: 10.1111/1467-9868.00196. URL https://doi.org/10.1111/1467-9868.00196. 28

[36] Yu Wang, Bin Dai, Gang Hua, John Aston, and David Wipf. Recurrent variational autoencoders for learning nonlinear generative models in the presence of outliers. *#IEEE_J_STSP#*, 12(6):1615–1627, December 2018. ISSN 1932-4553, 1941-0484. doi: 10.1109/jstsp.2018.2876995. URL https://doi.org/10.1109/jstsp.2018.2876995. 1

[37] Nicholas Watters, Loic Matthey, Sebastian Borgeaud, Rishabh Kabra, and Alexander Lerchner. Spriteworld: A flexible, configurable reinforcement learning environment. https://github.com/deepmind/spriteworld/, 2019. URL https://github.com/deepmind/spriteworld/. 32

[38] Dominik Zietlow, Michal Rolinek, and Georg Martius. Demystifying Inductive Biases for (Beta-) VAE Based Architectures. In *International Conference on Machine Learning*, pages 12945–12954. PMLR, 2021. 1, 28

Part

Appendix

Table of Contents

6 COMPLEMENTARY NOTES

6.1 VAESELBO DECOMPOSITIONS

Connection between (1) **and** (2). Here we show how the two decompositions of the ELBO objective in (1) and (2) can be connected. We start from equation (2):

$$\mathrm{ELBO}(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\phi}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right].$$

By definition of KL-divergence, and applying Bayes rule, we get

$$\begin{aligned}\mathrm{ELBO}(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\phi}) &= \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\left(\log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) - \log p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right)d\boldsymbol{z}\\ &= \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\left(\log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) - \log\left(p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\frac{p_0(\boldsymbol{z})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})}\right)\right)d\boldsymbol{z}\,.\end{aligned}$$

We observe that the two terms involving $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ cancel, resulting in

$$\mathrm{ELBO}(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\phi}) = -\int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\left(\log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) - \log\left(p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p_0(\boldsymbol{z})\right)\right)d\boldsymbol{z},$$

which leads to (1) by rearranging the terms:

$$\mathrm{ELBO}(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] - \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_0(\boldsymbol{z})\right].$$

Expressions for the two terms in equation (1) **under Assum. 1.** The above two terms take the following form in our setting. For the second ("KL") term, we get

$$\begin{aligned}-\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_0(\boldsymbol{z})\right] &= \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\log p_0(\boldsymbol{z})d\boldsymbol{z} - \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z}\\ &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log(p_0(\boldsymbol{z}))] + H(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}))\,,\end{aligned}$$

where H denotes the entropy. Writing the expression for the entropy of univariate Gaussian variables ($1/2\log(2\pi\sigma^2)+1/2$), we have under Assum. 1

$$H(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})) = \frac{d}{2}\left(\log(2\pi)+1\right) + \frac{1}{2}\sum_{k=1}^{d}\log\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2 = \kappa_d + \frac{1}{2}\sum_{k=1}^{d}\log\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2,$$

where we introduce the dimension dependent constant $\kappa_d = \frac{d}{2}\left(\log(2\pi)+1\right)$. This leads to

$$-\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_0(\boldsymbol{z})\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log(p_0(\boldsymbol{z}))] + \frac{1}{2}\sum_{k=1}^{d}\log\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2 + \kappa_d\,. \tag{10}$$

The first ("reconstruction") term, under the isotropic Gaussian decoder of Assum. 1, takes the form

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] = -\frac{\gamma^2}{2}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\|\boldsymbol{x}-\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}\right)\|^2\right] + d\log\gamma - \frac{d}{2}\log(2\pi)\,. \tag{11}$$

Expression for the gap between ELBO and log-likelihood Let us now write the KL divergence between variational and true posteriors, which is the gap appearing in (2).

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] = -\int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{z} - H(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}))$$

Using again the expression of the entropy of Gaussian variables, this leads to

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] = -\int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{z} - \sum_{k=1}^{d}\log\sigma_k^{\boldsymbol{\phi}}(x) - \frac{d}{2}\left(\log(2\pi)+1\right),$$

such that, using the Bayes formula for the true posterior and Assum. 1, we get

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] = -\sum_{k=1}^{d} \log \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + c(\boldsymbol{x},\gamma) \\ + \frac{1}{2}\mathbb{E}_{z\sim q_{\boldsymbol{\phi}}(\cdot|\boldsymbol{x})}\left[\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}\right)\|^2\gamma^2 - 2\sum_{k=1}^{d}\log m(z_k)\right], \quad (12)$$

with additive constant $c(\boldsymbol{x},\gamma) = -\frac{d}{2}\left(\log(\gamma^2)+1\right) + \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$. Note the $\log(2\pi)$ term in the previous expression cancels with the one coming from the true log posterior.

The analysis of the optima of (12) is non-trivial due to the second term which involves taking expectations of functions of $\boldsymbol{z}$ w.r.t. its posterior distribution $q_{\boldsymbol{\phi}}$ parameterized by $\boldsymbol{\mu}^{\phi}$ and $\boldsymbol{\sigma}^{\phi}$. Much of the derivations to obtain our results will revolve around constructing bounds that no longer involve such expectations, but instead only depend on $\boldsymbol{\mu}^{\phi}$ and $\boldsymbol{\sigma}^{\phi}$.

6.2 JUSTIFICATION OF THE INTUITION

We add here more qualitative details to the statement of subsection 3.2 that the true posterior density is approximately the pushforward of $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}=\boldsymbol{z}_0)$. Note that they are not meant to replace a rigorous treatment, which is deferred to § 7.

As the decoder becomes deterministic, the marginal observed density becomes the pushforward of the latent prior by $\boldsymbol{f}^{\boldsymbol{\theta}}$ [1] such that

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx p_0\left(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})\right)\left|\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{x})\right|.$$

The true posterior is therefore approximately

$$p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p_0(\boldsymbol{z})/p_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p_0(\boldsymbol{z})/p_0\left(\boldsymbol{g}^{\boldsymbol{\theta}}(x)\right)\left|\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{x})\right|^{-1}.$$

Conditioning on a given observation $\boldsymbol{x} = \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right)$, we get

$$\begin{aligned} p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x} = \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right)) &= p_{\boldsymbol{\theta}}(\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right)|\boldsymbol{z})p_0(\boldsymbol{z})/p_{\boldsymbol{\theta}}(\boldsymbol{x} = \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right)) \\ &\approx p_{\boldsymbol{\theta}}(\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right)|\boldsymbol{z})p_0(\boldsymbol{z})/p_0\left(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right))\right)\left|\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right))\right|^{-1} \\ &\approx p_{\boldsymbol{\theta}}(\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right)|\boldsymbol{z})p_0(\boldsymbol{z})/p_0\left(\boldsymbol{z}_0\right)\left|\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right))\right|^{-1} \end{aligned}$$

Neglecting the variations of the prior relative to those of the posterior (due to near-determinism), we make the approximation $p_0(\boldsymbol{z}) \approx p_0\left(\boldsymbol{z}_0\right)$ such that the above approximation becomes

$$p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x} = \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right)) \approx p_{\boldsymbol{\theta}}(\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right)|\boldsymbol{z})\left|\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{z}_0\right)\right|.$$

Using the isotropic Gaussian decoder assumption, we get

$$p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x} = \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right)) \approx \frac{\gamma^d}{\sqrt{2\pi}^d}\exp\left(-\frac{\gamma^2}{2}\left\|\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right) - \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}\right)\right\|^2\right)\left|\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{z}_0\right)\right|.$$

In the near-deterministic regime, this posterior distribution should be concentrated in the region where $\boldsymbol{z}$ is close to $\boldsymbol{z}_0$, we can then further approximate this density using a Taylor formula

$$\begin{aligned} p_{\boldsymbol{\theta}}(z|\boldsymbol{x} = \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right)) &\approx \frac{\gamma^d}{\sqrt{2\pi}^d}\exp\left(-\frac{\gamma^2}{2}\left\|\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{z}_0\right)\left(\boldsymbol{z}_0 - \boldsymbol{z}\right)\right\|^2\right)\left|\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{z}_0\right)\right| \\ &= \frac{\sqrt{2\pi}^{-d}\gamma^d}{\sqrt{\left|\mathbf{G}\mathbf{G}^T\right|}}\exp\left(-\frac{1}{\gamma^2}\left(\boldsymbol{z}_0 - \boldsymbol{z}\right)^T\left(\mathbf{G}\mathbf{G}^T\right)^{-1}\left(\boldsymbol{z}_0 - \boldsymbol{z}\right)\right), \end{aligned}$$

with $\mathbf{G} = \mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right)) = \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{z}_0\right)^{-1}$, which is also matching the expression of the pushforward of the Gaussian density $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z} = \boldsymbol{z}_0)$ by the linearization of $\boldsymbol{g}^{\boldsymbol{\theta}}$ around $\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}_0\right)$ (i.e. replacing the mapping by its Jacobian at that point, $\mathbf{G}$).

[1] because the conditional distribution of the decoder tends to a Dirac measure at $\boldsymbol{f}^{\boldsymbol{\theta}}$

6.3 A CONNECTION BETWEEN THE β PARAMETER OF β-ZVAES AND THE DECODER PRECISION γ^2

In the context of disentanglement, a commonly used variant of standard VAEs [19] is the β-VAE [3, 12, 18, 32, 22]. In this model, an additional parameter β is added to modify the weight of the KL term in (1), whereas the decoder precision γ^2 is typically set to one [6, 9, 22, 32]. The β-VAE objective [12] can be written as

$$\mathcal{L}_\beta(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] - \beta KL\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\|p_0(\boldsymbol{z})\right]. \tag{13}$$

The influence of the decoder precision γ^2 and the β parameters on the objective have been related in the literature, see for example [7, § 2.4.3]—and similar observations can be found in [33, § 3.1]. Under the assumption of a Gaussian decoder, the ELBO from eq. (1) can be written as

$$\begin{aligned}
\text{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi}) &= -\text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_0(\boldsymbol{z})\right] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right]\\
&= -\text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_0(\boldsymbol{z})\right] - \frac{\gamma^2}{2}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\left\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})\right\|^2\right] + d\log\gamma - \frac{d}{2}\log(2\pi)\\
&= \gamma^2\left[-\frac{1}{\gamma^2}\text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_0(\boldsymbol{z})\right] - \frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\left\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})\right\|^2\right] + c(\gamma)\right]; \qquad (14)\\
c(\gamma) &:= \frac{d}{\gamma^2}\log\gamma - \frac{d}{2\gamma^2}\log(2\pi)
\end{aligned}$$

Given that usually optimization is performed with a fixed value for γ for the ELBO (and with fixed β for $\mathcal{L}_\beta$), this suggests that β and $1/\gamma^2$, play a similar role in (13) and (14)—since the γ^2 outside parenthesis only changes the objective and its gradients by a global scaling factor.

7 MAIN THEORETICAL RESULTS

7.1 PROOF OF ??

We proceed in two steps: first we prove the existence of variational parameters that achieve a global minimum of the ELBO gap, then we characterize its near-deterministic properties. We then combine these results, which rely on specific assumptions, to obtain our main text result under Assum. 1.

We initially use the following milder assumptions than in main text to prove intermediate results.

Assumption 2 (Gaussian Encoder-Gaussian Decoder VAE, minimal properties)**.** *We are given a fixed latent prior and three parameterized classes of $\mathbb{R}^d \to \mathbb{R}^d$ mappings: the mean decoder class $\boldsymbol{\theta} \mapsto \boldsymbol{f}^{\boldsymbol{\theta}}$, and the mean and standard deviation encoder classes, $\boldsymbol{\phi} \mapsto \boldsymbol{\mu}^{\boldsymbol{\phi}}$ and $\boldsymbol{\phi} \mapsto \boldsymbol{\sigma}^{\boldsymbol{\phi}}$ such that*

(i) the latent prior has a factorized independent and identically distributed (i.i.d.) density $p_0(\boldsymbol{z}) \sim \prod_k m(z_k)$, with m smooth fully supported on $\mathbb{R}$, with concave $\log m$,

(ii) conditional on the latent, the decoder has a factorized Gaussian density $p_{\boldsymbol{\theta}}$ with mean $\boldsymbol{f}^{\boldsymbol{\theta}}$ such that

$$\boldsymbol{x}|\boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z}), \gamma^{-2}\mathbf{I}_d\right) \tag{15}$$

(iii) the encoder is factorized Gaussian with posterior mean and variance maps $\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x}), \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2$ for each component k, leading to the factorized posterior density $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ such that

$$z_k|\boldsymbol{x} \sim \mathcal{N}(\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x}), \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2) \tag{16}$$

(iv) the mean and variance encoders classes can fit any function,

(v) for all possible $\boldsymbol{\theta}$, $\boldsymbol{f}^{\boldsymbol{\theta}}$ is a diffeomorphism of $\mathbb{R}^d$ with inverse $\boldsymbol{g}^{\boldsymbol{\theta}}$.

Existence of at least one global minimizer of the gap between true and variational posterior is given by the following proposition.

Proposition 2 (Existence of global minimum)**.** *Under Assumption 2. For a fixed $\boldsymbol{\theta}$ assume additionally that $\boldsymbol{g}^{\boldsymbol{\theta}}$ is Lipschitz continuous with Lipschitz constant $B > 0$, in the sense that*

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d: \qquad \left\|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{y})\right\|_2 \leq B\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

Then there exists at least one choice ($\boldsymbol{\mu}^{\boldsymbol{\phi}} \in \mathbb{R}^d$, $\boldsymbol{\sigma}^{\boldsymbol{\phi}} \in \mathbb{R}^d_{>0}$) that achieves the minimum of KL $[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})]$.

Proof. Using Prop. 6, we have the lower bound

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq -\sum_{k=1}^{d}\left[\log\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + \log m(\mu_k^{\boldsymbol{\phi}})\right] + c(\boldsymbol{x},\gamma) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right\|^2 + \sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right]. \quad (17)$$

We then notice (see lemma 4) that for all k,

$$\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) \to -\log\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + \frac{\gamma^2}{2}B^{-2}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2$$

achieves a global minimum $n(B,\gamma) = -\log(B/\gamma) + 1/2$ at $\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) = B/\gamma$.

For arbitrary k_0, we now 1) lower bound the $k \neq k_0$ terms by $n(B,\gamma)$; 2) lower bound and all the $\log m$ terms by their global maximum, which exists by Assum. 1i (log-concave prior); and 3) drop the non-negative squared norm term, leading to the following weaker lower bound:

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq (d-1)n(B,\gamma) - \log\sigma_{k_0}^{\boldsymbol{\phi}}(\boldsymbol{x}) - d\max_t\left(\log m(t)\right) + c(\boldsymbol{x},\gamma) + \frac{\gamma^2}{2}B^{-2}\left[\sigma_{k_0}^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right]. \quad (18)$$

The KL divergence is well-defined and finite for any choice of parameters in their domain, therefore it achieves a particular value $K_0 \geq 0$ at one arbitrary selected point of the domain. Since for all k, the lower bound tends to $+\infty$ for both $\sigma_k^{\boldsymbol{\phi}} \to +\infty$ (as the quadratic term dominates the $-\log$ term) and $\sigma_k^{\boldsymbol{\phi}} \to 0^+$, there exist $a > b > 0$ (possibly dependent on $(\gamma, \boldsymbol{x})$) such that $\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] > K_0$ for any $\sigma_k^{\boldsymbol{\phi}} < b$ or $\sigma_k^{\boldsymbol{\phi}} > a$.

Moreover, starting again from the lower bound from Prop. 6,

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq -\sum_{k=1}^{d}\left[\log\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + \log m(\mu_k^{\boldsymbol{\phi}})\right] + c(\boldsymbol{x},\gamma) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right\|^2 + \sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right], \quad (19)$$

we now focus on $\boldsymbol{\mu}^{\boldsymbol{\phi}}$ and lower bound all $\boldsymbol{\sigma}^{\boldsymbol{\phi}}$ terms. With this, we get the following weaker lower bound in terms of $\boldsymbol{\mu}^{\boldsymbol{\phi}}$:

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq dn(B,\gamma) - d\max_t\left(\log m(t)\right) + c(\boldsymbol{x},\gamma) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right\|^2\right]. \quad (20)$$

The lower bound also tends to $+\infty$ for $\|\boldsymbol{\mu}^{\boldsymbol{\phi}}\| \to +\infty$, so there exists a radius $R > 0$ (possibly dependent on $(\gamma, \boldsymbol{x})$) such that $\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] > K_0$ if $\|\boldsymbol{\mu}^{\boldsymbol{\phi}}\| > R$.

As a consequence, the infimum ($\leq K_0$) of the minimization problem (5) cannot be achieved outside the compact set $(\boldsymbol{\mu}^{\boldsymbol{\phi}}, \boldsymbol{\sigma}^{\boldsymbol{\phi}}) \in \{\boldsymbol{\mu}^{\boldsymbol{\phi}} \in \mathbb{R}^d : \|\boldsymbol{\mu}^{\boldsymbol{\phi}}\| \leq R\} \times [a,b]^d$. Since the divergence is continuous in $(\boldsymbol{\mu}^{\boldsymbol{\phi}}, \boldsymbol{\sigma}^{\boldsymbol{\phi}})$, there exists a value $(\widehat{\boldsymbol{\mu}^{\boldsymbol{\phi}}}, \widehat{\boldsymbol{\sigma}^{\boldsymbol{\phi}}})$ in this compact set achieving the minimum of the KL over the whole parameter domain, and all values achieving this minimum are in this compact set. □

For given $\boldsymbol{x}$, $\boldsymbol{\theta}$ and $\gamma > 0$, the variational posterior KL divergence mapping

$$(\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x}), \boldsymbol{\sigma}^{\boldsymbol{\phi}}(\boldsymbol{x})) \to \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right]$$

thus has a minimum, and by smoothness of this mapping, this minimum can be characterized by the vanishing gradient of the KL divergence with respect to the parameters. Now, let us try to characterize how this minimum behaves for large γ.

Proposition 3 (Self-consistency of the encoder in the deterministic limit)**.** *Under Assum. 2, assume additionally $\boldsymbol{f}^{\boldsymbol{\theta}}$ and $\boldsymbol{g}^{\boldsymbol{\theta}}$ are Lipschitz continuous with respective Lipschitz constants $C, B > 0$, in the sense that*

$$\forall \boldsymbol{z}, \boldsymbol{w} \in \mathbb{R}^d : \qquad \left\|\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}\right) - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{w})\right\|_2 \leq C\|\boldsymbol{z} - \boldsymbol{w}\|_2 \,, \tag{21}$$

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d : \qquad \left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{y})\right\|_2 \leq B\|\boldsymbol{x} - \boldsymbol{y}\|_2 \,. \tag{22}$$

Assume additionally that $-\log m$ is quadratically dominated, in the sense that

$$\exists D > 0, E > 0 : \qquad -\log m(u) \leq D|u|^2 + E \,, \qquad \forall u \in \mathbb{R}.$$

Then for all $\boldsymbol{x}, \boldsymbol{\theta}$, as $\gamma \to +\infty$, any global minimum of (5) satisfies

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) = \boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) + O(1/\gamma) \tag{23}$$

$$\boldsymbol{\sigma}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = O(1/\gamma^2) \,. \tag{24}$$

More precisely, for all $\boldsymbol{x} \in \mathbb{R}^d$, $\gamma > 0$

$$\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right\|^2 \leq B^2 \frac{2d}{\gamma^2}\left(\frac{1}{2}(C^2 - 1) + E + D\left[\frac{\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2}{d} + \frac{1}{\gamma^2}\right] + M + \frac{1}{2}\log(B^2)\right) \,.$$

and

$$\sum_{k=1}^{d} \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 \leq B^2 \frac{4d}{\gamma^2}\left(\frac{1}{2}(C^2 - 1) + E + D\left[\frac{\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2}{d} + \frac{1}{\gamma^2}\right] + M + \frac{1}{2}\left(\log(2B^2)\right)\right) \,.$$

Proof. We start from the lower bound expression of Prop. 6

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq -\sum_{k=1}^{d}\left[\log \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + \log m(\mu_k^{\boldsymbol{\phi}})\right] + c(\boldsymbol{x}, \gamma) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\boldsymbol{\phi}}\right\|^2 + \sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right] \,,$$

with $c(\boldsymbol{x}, \gamma) = -\frac{d}{2}\left(\log(\gamma^2) + 1\right) + \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$. For any $\nu \in (0, 1]$, we can thus write

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq \sum_{k=1}^{d}\left[-\log \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + \nu\gamma^2 B^{-2}\frac{\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2}{2} - \log m(\mu_k^{\boldsymbol{\phi}})\right] + c(\boldsymbol{x}, \gamma) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\boldsymbol{\phi}}\right\|^2 + (1 - \nu)\sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right] \,.$$

Now, from lemma 4 we get

$$\forall u > 0 : \qquad -\log u + \alpha u^2/2 \geq \frac{1}{2}\log(\alpha) + \frac{1}{2} \,.$$

We exploit this lower bound to obtain

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq \frac{d}{2}\left(\log(\nu\gamma^2 B^{-2}) + 1\right) - \sum_{k=1}^{d}\left[\log m(\mu_k^{\boldsymbol{\phi}})\right] + c(\boldsymbol{x}, \gamma) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\boldsymbol{\phi}}\right\|^2 + (1 - \nu)\sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right] \,.$$

Using the expression of $c(\boldsymbol{x},\gamma)$ we get

$$
\begin{aligned}
\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq \frac{d}{2}\left(\log(\nu B^{-2}) + \log\gamma^2 + 1\right) - \sum_{k=1}^{d}\left[\log m(\mu_k^{\boldsymbol{\phi}})\right] - \frac{d}{2}\left(\log\gamma^2 + 1\right)& \\
+ \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\boldsymbol{\phi}}\right\|^2 + (1-\nu)\sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right]&.
\end{aligned}
$$

and both the "$d\log\gamma$" as well as "$d/2$" terms cancel out such that

$$
\begin{aligned}
\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq \frac{d}{2}\left(\log(\nu B^{-2})\right) - \sum_{k=1}^{d}\left[\log m(\mu_k^{\boldsymbol{\phi}})\right] + \log p_{\boldsymbol{\theta}}(\boldsymbol{x})& \\
+ \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\boldsymbol{\phi}}\right\|^2 + (1-\nu)\sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right]&.
\end{aligned}
$$

Finally, using Prop. 7, the above right hand side is bounded from above by a constant as $\gamma \to +\infty$, and as a consequence, the positive factor of the γ^2 term must vanish (by continuity assumption and its limits note $-\log m$ is bounded from below)

$$
\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\boldsymbol{\phi}}\right\|^2 + (1-\nu)\sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2 \to 0
$$

This entails that both positive terms it comprises must vanish too.

More precisely, we get the inequality between lower and upper bounds at the optimal solution

$$
\begin{aligned}
&\frac{d}{2}\left(\log(\nu B^{-2})\right) - \sum_{k=1}^{d}\left[\log m(\mu_k^{\widehat{\boldsymbol{\phi}}})\right] + \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \\
&\qquad + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right\|^2 + (1-\nu)\sum_{k=1}^{d}\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\right] \\
&\qquad\qquad \leq d\left(\frac{1}{2}C^2 + E + D\left[\frac{\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2}{d} + \frac{1}{\gamma^2}\right]\right) - \frac{d}{2} + \log p_{\boldsymbol{\theta}}(\boldsymbol{x}),
\end{aligned}
$$

which simplifies to

$$
\begin{aligned}
&\frac{d}{2}\left(\log(\nu B^{-2})\right) - \sum_{k=1}^{d}\left[\log m(\mu_k^{\widehat{\boldsymbol{\phi}}})\right] + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right\|^2 + (1-\nu)\sum_{k=1}^{d}\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\right] \\
&\qquad\qquad \leq d\left(\frac{1}{2}C^2 + E + D\left[\frac{\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2}{d} + \frac{1}{\gamma^2}\right]\right) - \frac{d}{2}.
\end{aligned}
$$

Moreover by continuity assumption and its limits, $-\log m$ is bounded from below by $-M = -\max_t \log m(t)$, yielding

$$
\begin{aligned}
&\frac{d}{2}\left(\log(\nu B^{-2}) - 2M\right) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right\|^2 + (1-\nu)\sum_{k=1}^{d}\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\right] \\
&\qquad\qquad \leq d\left(\frac{1}{2}(C^2-1) + E + D\left[\frac{\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2}{d} + \frac{1}{\gamma^2}\right]\right)
\end{aligned}
$$

such that

$$
\begin{aligned}
&\frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right\|^2 + (1-\nu)\sum_{k=1}^{d}\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\right] \\
&\qquad\qquad \leq d\left(\frac{1}{2}(C^2-1) + E + D\left[\frac{\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2}{d} + \frac{1}{\gamma^2}\right] - \frac{1}{2}\left(\log(\nu B^{-2}) - 2M\right)\right)
\end{aligned}
$$

and finally

$$B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)-\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right\|^2+(1-\nu)\sum_{k=1}^{d}\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\right]$$
$$\leq\frac{2d}{\gamma^2}\left(\frac{1}{2}(C^2-1)+E+D\left[\frac{\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2}{d}+\frac{1}{\gamma^2}\right]+M+\frac{1}{2}\log(B^2/\nu)\right)\qquad(25)$$

Taking $\nu=1$ in (25) we get the first intended inequality

$$\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)-\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right\|^2\leq B^2\frac{2d}{\gamma^2}\left(\frac{1}{2}(C^2-1)+E+D\left[\frac{\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2}{d}+\frac{1}{\gamma^2}\right]\right.$$
$$\left.+M+\frac{1}{2}\log(B^2)\right)\,.$$

Alternatively, (25) implies

$$(1-\nu)\sum_{k=1}^{d}\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\leq B^2\frac{2d}{\gamma^2}\left(\frac{1}{2}(C^2-1)+E+D\left[\frac{\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2}{d}+\frac{1}{\gamma^2}\right]\right.$$
$$\left.+M+\frac{1}{2}\left(\log(B^2/\nu)\right)\right)$$

Taking a fixed value of ν, say $1/2$, we get the second intended inequality

$$\sum_{k=1}^{d}\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\qquad\leq\qquad B^2\frac{4d}{\gamma^2}\left(\frac{1}{2}(C^2-1)+E+D\left[\frac{\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2}{d}+\frac{1}{\gamma^2}\right]+M+\frac{1}{2}\left(\log(2B^2)\right)\right)\,.$$

□

We now restate the main text proposition and provide the proof.

Proposition 1. *[Self-consistency of near-deterministic VAEs] Under Assum. 1, $\forall\,\boldsymbol{x},\,\boldsymbol{\theta}$, as $\gamma\to+\infty$, there exists at least one global minimum solution of (5), satisfying*

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})=\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)+O(1/\gamma);\ \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2=O(1/\gamma^2)\,,\ \forall k.\qquad(7)$$

Proof. We only have to check that Assum. 1 allow fulfilling the following requirements of Prop. 3:

- the Lipschitz continuity requirements in Prop. 3 results from the boundedness of the first order derivatives of the decoder mean and of its inverse (by using the multivariate Taylor theorem),
- concavity of $\log m$, required by Assum. 2, is a direct consequence of non-positivity of the second-order logarithmic derivative of m in Assum. 1i,
- quadratic domination of $-\log m$ comes from the boundedness of the second-order logarithmic derivative of m (by integrating twice).

Then Prop. 3 follows and the $O(1/\gamma)$ convergence of the variational posterior mean of the inverse, as well as the $O(1/\gamma^2)$ convergence of the variational posterior variance. □

Finer approximation of parameter values We now derive a finer result for the convergence of the mean, that we will exploit in Thm. 1. This relies on the existence of an optimum shown by Prop. 2.

At such optimum $\widehat{\boldsymbol{\phi}}$ we thus have for all k

$$\frac{\partial}{\partial\mu_k^{\boldsymbol{\phi}}}\left[\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right]\right]_{|\widehat{\boldsymbol{\phi}}}=0\,,$$

and

$$\frac{\partial}{\partial \sigma_k^{\boldsymbol{\phi}}} \left[\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right]\right]_{|\widehat{\boldsymbol{\phi}}} = 0\,.$$

We derive the constraints entailed by the first expression:

$$\begin{aligned}\frac{\partial}{\partial \mu_k^{\boldsymbol{\phi}}} \left[\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right]\right]_{|\widehat{\boldsymbol{\phi}}} &= \frac{1}{2}\int \frac{\partial}{\partial \mu_k^{\boldsymbol{\phi}}} q_{\boldsymbol{\phi}}(\boldsymbol{z}) \left[\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}\right)\|^2\gamma^2 - 2\sum_{k=1}^{d}\log m(z_k)\right] d\boldsymbol{z} \\ &= \frac{1}{2}\int \prod_{j\neq k} q_{\boldsymbol{\phi}}^{j}(z_j)\frac{\partial q_{\boldsymbol{\phi}}^{k}(z_k)}{\partial \mu_k^{\boldsymbol{\phi}}} \left[\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}\right)\|^2\gamma^2 - 2\sum_{k=1}^{d}\log m(z_k)\right] d\boldsymbol{z}\end{aligned}$$

with

$$\frac{\partial q_{\boldsymbol{\phi}}^{k}(z_k)}{\partial \mu_k^{\boldsymbol{\phi}}} = \frac{\mu_k^{\boldsymbol{\phi}} - z_k}{{\sigma_k^{\boldsymbol{\phi}}}^2} q_{\boldsymbol{\phi}}^{k}(z_k),$$

which leads to a set of constraints at optimum

$$\begin{aligned}&\int q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z})\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) \left[\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}\right)\|^2\gamma^2 - 2\sum_{k=1}^{d}\log m(z_k)\right] d\boldsymbol{z} \\ &\qquad = \int q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}) z_k \left[\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}\right)\|^2\gamma^2 - 2\sum_{k=1}^{d}\log m(z_k)\right] d\boldsymbol{z}\,,\ \forall k \qquad (26)\end{aligned}$$

Based on this expression we derive the following result.

Proposition 4. *Under Assum. 1, as* $\gamma \to +\infty$

$$\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) = \boldsymbol{x} + \frac{1}{\gamma^2}\mathbf{J}^{-T}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})} n'(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) + O(1/\gamma^3). \qquad (27)$$

and

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) = \boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) + \frac{1}{\gamma^2}\mathbf{J}^{-1}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})}\mathbf{J}^{-T}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})} n'(\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)) + O(1/\gamma^3) \qquad (28)$$

Proof. We start from the constraints of (26) that we rewrite

$$\begin{aligned}&\int q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}) \left(z_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right) \left[\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}\right)\|^2\gamma^2\right] d\boldsymbol{z} \\ &\qquad = \int q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}) \left(z_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right) \left[2\sum_{k=1}^{d}\log m(z_k)\right] d\boldsymbol{z}\end{aligned}$$

We then proceed to approximate the left hand side using a Taylor formula. Assuming bounded Hessian components, we can upper and lower bound using third order centered absolute moments of the Gaussian as

$$\gamma^2 \int q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}) \left(z_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right) \left[\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) - \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})}(\boldsymbol{z} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))\|^2\right] d\boldsymbol{z} \qquad + \qquad O(1/\gamma),$$

which we can rewrite (by 1) expanding the norm of the sum; 2) removing constants in the bracket, which lead to zeros after multiplying the zero mean variable and taking the expectation; 3) using Gaussianity, all centered third order terms vanish.)

$$
\begin{aligned}
\gamma^2 \int q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}) \left(z_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right) \Big[&\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))\|^2 + \|\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})}(\boldsymbol{z} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))\|^2 \\
&-2 \left\langle \boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})),\, \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})}(\boldsymbol{z} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) \right\rangle \Big] d\boldsymbol{z} + O(1/\gamma) \\
&= \gamma^2 \int q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}) \left(z_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right) \Big[\|\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})}(\boldsymbol{z} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))\|^2 \\
&-2 \left\langle \boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})),\, \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})}(\boldsymbol{z} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) \right\rangle \Big] d\boldsymbol{z} + O(1/\gamma) \\
= \gamma^2 \int q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}) \left(z_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right) &\Big[(\boldsymbol{z} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))^T \mathbf{J}^T_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})} \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})}(\boldsymbol{z} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) \\
&-2 \left\langle \boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})),\, \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})}(\boldsymbol{z} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) \right\rangle \Big] d\boldsymbol{z} + O(1/\gamma) \\
= \gamma^2 \int q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}) \left(z_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right) &\Big[-2 \left\langle \boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})),\, \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})}(\boldsymbol{z} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) \right\rangle \Big] d\boldsymbol{z} + O(1/\gamma)
\end{aligned}
$$

Finally computing this integral we get the left hand side as

$$
-2\gamma^2 \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 \left\langle \boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})),\, [\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})}]_{.k} \right\rangle + O(1/\gamma)
$$

For the right hand side we get using a Taylor expansion (with notation $n : \boldsymbol{z} \to \log(m(\boldsymbol{z}))$)

$$
\begin{aligned}
&\int q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}) \left(z_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))\right) \left[2 \sum_{k=1}^{d} \log m(z_k) \right] d\boldsymbol{z} \\
&\quad = \int q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}) \left(z_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))\right) \left[2 \sum_{k=1}^{d} \log m(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) + n'(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))(z_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) \right] d\boldsymbol{z} + O(1/\gamma^2) \\
&\qquad = 2\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 n'(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) + O(1/\gamma^2).
\end{aligned}
$$

Equating the non-negligible terms of the left and right-hand sides we get for each k

$$
\gamma^2 \left\langle \boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})),\, [\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})}]_{.k} \right\rangle = -n'(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) + O(1/\gamma)
$$

such that

$$
(\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})))^T \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})} = -\frac{1}{\gamma^2} n'(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) + O(1/\gamma^3),
$$

where n' is applied component-wise. Because the Jacobian is everywhere invertible (implicit consequence of Lipschitz assumptions), we can solve for this equations and get

$$
\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) = \boldsymbol{x} + \frac{1}{\gamma^2} \mathbf{J}^{-T}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})} n'(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) + O(1/\gamma^3). \tag{29}
$$

Using again a similar Taylor approximation we get

$$
\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) = \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) + \frac{1}{\gamma^2} \mathbf{J}^{-1}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})} \mathbf{J}^{-T}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})} n'(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) + O(1/\gamma^3).
$$

This equation has the shortcoming of still referring to the posterior mean on both sides. To fix this, we first note that it implies, by boundedness of the Jacobian, that

$$
|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) - \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})| \leq \frac{1}{\gamma^2} K |n'(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))| + O(1/\gamma^3).
$$

By bounding the second-order derivative of the log prior, we get

$$|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) - \boldsymbol{g}^{\boldsymbol{\theta}}\,(\boldsymbol{x})\,| \leq \frac{1}{\gamma^2} K |n'(\boldsymbol{g}^{\boldsymbol{\theta}}\,(\boldsymbol{x})) + O(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) - \boldsymbol{g}^{\boldsymbol{\theta}}\,(\boldsymbol{x}))| + O(1/\gamma^3),$$

which implies

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) = \boldsymbol{g}^{\boldsymbol{\theta}}\,(\boldsymbol{x}) + O(1/\gamma^2)\,,$$

i.e., we obtain an improved convergence rate. Using this rate and Taylor theorem, we obtain the final equation by replacing the variational posterior mean by the inverse decoder in (29)

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) = \boldsymbol{g}^{\boldsymbol{\theta}}\,(\boldsymbol{x}) + \frac{1}{\gamma^2}\mathbf{J}^{-1}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})}\mathbf{J}^{-T}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})} n'(\boldsymbol{g}^{\boldsymbol{\theta}}\,(\boldsymbol{x})) + O(1/\gamma^3)$$

□

7.2 PROOF OF ??

This will be a corollary of the following result, that uses as a key assumption a rate of $O(1/\gamma^2)$ in the convergence of the self-consistency equation of the variational mean.

Proposition 5 (VAEs with log-concave factorized prior and close-to-deterministic decoder approximate the IMA objective)**.** *Under Assum. 1, if additionally the VAE satisfies the following self-consistency in the deterministic limit*

$$\left\|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) - \boldsymbol{g}^{\boldsymbol{\theta}}\,(\boldsymbol{x})\right\| = O_{\gamma\to+\infty}(1/\gamma^2)\,, \tag{30}$$

$$\left\|\boldsymbol{\sigma}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\right\|^2 = O_{\gamma\to+\infty}(1/\gamma^2)\,. \tag{31}$$

then

$$\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = \left(-\frac{d^2\log p_0}{dz_k^2}(g_k^{\boldsymbol{\theta}}(\boldsymbol{x})) + \gamma^2\left\|\left[\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]_{:k}\right\|^2\right)^{-1} + O(1/\gamma^3)\,, \tag{32}$$

and the self-consistent ELBO (6) *approximates the* IMA*-regularized log-likelihood* (4)*:*

$$\text{ELBO}^*(\boldsymbol{x};\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})) + O_{\gamma\to\infty}\left(1/\gamma^2\right). \tag{33}$$

Proof. We start from the self-consistent ELBO decomposition as "reconstruction error plus posterior regularization" terms:

$$\text{ELBO}^*(\boldsymbol{x};\boldsymbol{\theta}) = -\text{KL}\left[q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})||p_0(\boldsymbol{z})\right] + \mathbb{E}_{q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right], \tag{34}$$

and continue with reformulating both terms, based on Assum. 1. That is, p_0 is factorized with components i.i.d. distributed according to a fully supported **log-concave** density $z_k \sim m$.

Posterior regularization term Assum. 1 gives us the formula of (10) for this term in the ELBO. Taking optimal encoder parameters, we get the posterior regularization term for the ELBO*

$$-\text{KL}\left[q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})||p_0(\boldsymbol{z})\right] = \mathbb{E}_{q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})}[\log(p_0(\boldsymbol{z}))] + \frac{1}{2}\sum_{k=1}^{d}\left[\log\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\right] + \kappa_d\,,$$

with $\kappa_d = \frac{d}{2}\left(\log(2\pi)+1\right)$. Using the factorized Gaussian encoder and i.i.d. prior assumptions we get

$$-\quad \text{KL}\left[q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})||p_0(\boldsymbol{z})\right] \quad = \quad \sum_{k=1}^{d}\mathbb{E}_{z_k\sim\mathcal{N}(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}),\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2)}[\log(m(z_k))] \quad + \quad \frac{1}{2}\sum_{k=1}^{d}\left[\log\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\right] \quad + \quad \kappa_d\,,$$

where we rewrote the distribution p_0 as $p_0 = \prod_k m(z_k)$.

Based on the Taylor theorem, with a residual in Lagrange form of $n = \log m$, we have that for all k and u there exists $\xi \in [\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}), u]$ if $u \geq \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})$, or $\xi \in [u, \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})]$ if $u \leq \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})$ such that

$$\begin{aligned} n(u) = \log(m(u)) = \log(m(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))) + n'(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))(u - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) \\ + \frac{1}{2} n''(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))(u - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))^2 + \frac{1}{3!} n^{(3)}(\xi)(u - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))^3 \end{aligned}$$

We assumed that $|n^{(3)}|$ is bounded over $\mathbb{R}$ by F, such that

$$\begin{aligned} -F \left| u - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) \right|^3 \leq \log(m(u)) - \log(m(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))) - n'(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))(u - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) \\ - \frac{1}{2} n''(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))(u - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))^2 \leq F \left| u - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) \right|^3 . \end{aligned}$$

Taking the expectation and using the expression of centered Gaussian absolute moments[2]

$$\begin{aligned} \left| \mathbb{E}_{z_k \sim \mathcal{N}(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}), \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2)} [\log(m(z_k))] - \log(m(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))) - \frac{1}{2} n''(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 \right| \\ \leq F \mathbb{E} \left[\left| u - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) \right|^3 \right] = F \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^3 \frac{2^{3/2}}{\sqrt{\pi}} . \end{aligned} \tag{35}$$

As the assumptions entail that optimal posterior variances $\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2$ get small for γ large (cf. (31)), this implies the near-deterministic approximation

$$\mathbb{E}_{z_k \sim \mathcal{N}(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}), \sigma_k(\boldsymbol{x})^2)} [\log(m(z_k))] = \log(m(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))) + \frac{1}{2} n''(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 + O_{\gamma \to +\infty}(1/\gamma^3) .$$

In addition, using again a Taylor formula and the self-consistency assumption for the mean

$$\begin{aligned} \log(m(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))) = \log(m(g_k^{\boldsymbol{\theta}}(\boldsymbol{x}))) + n'(g_k^{\boldsymbol{\theta}}(\boldsymbol{x}))(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) - g_k^{\boldsymbol{\theta}}(\boldsymbol{x})) + O_{\gamma \to +\infty}(1/\gamma^2) \\ = \log(m(g_k^{\boldsymbol{\theta}}(\boldsymbol{x}))) + O_{\gamma \to +\infty}(1/\gamma^2). \end{aligned}$$

Moreover, using again a Taylor formula for n'' under boundedness of $n^{(3)}$ and again using the self-consistency assumption for the mean yields

$$n''(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) = n''(g_k^{\boldsymbol{\theta}}(\boldsymbol{x})) + O(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) - g_k^{\boldsymbol{\theta}}(\boldsymbol{x})) = n''(g_k^{\boldsymbol{\theta}}(\boldsymbol{x})) + O_{\gamma \to +\infty}(1/\gamma^2) .$$

Overall this leads to the approximation of the posterior regularization term

$$\begin{aligned} -\mathrm{KL} \left[q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x}) || p_0(\boldsymbol{z}) \right] = \sum_{k=1}^{d} \log(m(g_k^{\boldsymbol{\theta}}(\boldsymbol{x}))) + \frac{1}{2} n''(g_k^{\boldsymbol{\theta}}(\boldsymbol{x})) \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 + \frac{1}{2} \log \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 \\ + \kappa_d + O_{\gamma \to +\infty}(1/\gamma^2) . \end{aligned} \tag{36}$$

Reconstruction term Now switching to the first (reconstruction) term of the ELBO*, adapting the decomposition of (11) by using optimal encoder parameters we get

$$\mathbb{E}_{q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})} [\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})] = -\frac{\gamma^2}{2} \mathbb{E}_{q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})} \left[\| \boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z}) \|^2 \right] + d \log \gamma - \frac{d}{2} \log(2\pi).$$

Then in the small encoder noise limit $\sigma_k(\boldsymbol{x})^2 \ll 1, \forall k$ (justified by Prop. 1), we rely on a Taylor approximation around the posterior mean $\boldsymbol{z}^o = \boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})$ based on Lemma 3, which bounds this approximation as follows

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\left\| \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z}) - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) - \sum_{k=1}^{d} \frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\boldsymbol{z}^o} (z_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) \right\|^2 \right] \leq \frac{d^3}{4} 3K^2 \sum_i \sigma_i^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^4 . \tag{37}$$

[2] see e.g. https://arxiv.org/pdf/1209.4340

The linear term in this approximation is easily computed using successively Lemma 1 and Lemma 2 to get an expression with the squared column norms of the partial derivatives scaled by the standard deviations $\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})}$. We get

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\left\|\sum_{k=1}^{d}\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|z^o}(z_k-\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x}))\right\|^2\right] = \text{trace}\left[\text{Cov}\left[\sum_{k=1}^{d}\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})}(z_k-\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x}))\right]\right]$$
$$= \sum_{k=1}^{d}\left[\left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})}\right\|^2 \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right]. \quad (38)$$

This term can be used as an approximation for the expectation term in the reconstruction loss thanks to the following reverse triangle inequality

$$\left|\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\|\boldsymbol{x}-\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})\|^2\right] - \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\left\|\sum_{k=1}^{d}\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|z^o}(z_k-\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x}))\right\|^2\right]\right|$$
$$= \left|\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\|\boldsymbol{x}-\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})\|^2\right] - \sum_{k=1}^{d}\left[\left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})}\right\|^2 \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right]\right|$$
$$\leq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\left\|\boldsymbol{x}-\left(\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})-\sum_{k=1}^{d}\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|z^o}(z_k-\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x}))\right)\right\|^2\right],$$

such that the resulting upper bound can be itself bounded as follows

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\left\|\boldsymbol{x}-\left(\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})-\sum_{k=1}^{d}\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\boldsymbol{z}^o}(z_k-\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x}))\right)\right\|^2\right]$$
$$\leq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\left\|\boldsymbol{x}-\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x}))\right\|^2\right] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\left\|\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})-\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x}))-\sum_{k=1}^{d}\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})}(z_k-\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x}))\right\|^2\right].$$

Each term of the upper bound can be bounded for the optimum encoder parameters: using from left to right the assumption of (30) and (37), respectively, leading to

$$\left|\mathbb{E}_{q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})}\left[\|\boldsymbol{x}-\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})\|^2\right] - \sum_{k=1}^{d}\left[\left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})}\right\|^2 \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\right]\right|$$
$$\leq O_{\gamma\to+\infty}(1/\gamma^4) + \frac{d^3}{4}3K^2\sum_i \sigma_i^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^4.$$

Getting back to the whole reconstruction term, using additionally the variance self-consistency assumption (31), the above shows that we can make the approximation

$$\mathbb{E}_{q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] = -\frac{\gamma^2}{2}\sum_{k=1}^{d}\left[\left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})}\right\|^2 \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\right] + d\log\gamma - \frac{d}{2}\log(2\pi) + O_{\gamma\to+\infty}(1/\gamma^2)$$

We can further replace the dependency of the derivatives on the encoder mean using a Taylor formula for the derivative

$$\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})} = \frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})} + O(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})-\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})) = \frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})} + O(1/\gamma^2)$$

such that

$$\mathbb{E}_{q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] = -\frac{\gamma^2}{2}\sum_{k=1}^{d}\left[\left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})}\right\|^2 \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\right] + d\log\gamma - \frac{d}{2}\log(2\pi) + O_{\gamma\to+\infty}(1/\gamma^2) \tag{39}$$

ELBO* approximation As a consequence of (36) and (39) the ELBO* becomes

$$\begin{aligned}
\mathrm{ELBO}^*(\boldsymbol{x};\boldsymbol{\theta}) &= -\frac{1}{2}\sum_{k=1}^{d}\Bigg[\log\frac{1}{\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2} + \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\left(-n''(g^{\boldsymbol{\theta}}{}_k(\boldsymbol{x})) + \gamma^2\left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|g^{\boldsymbol{\theta}}{}_k}\right\|^2\right) \\
&\qquad - 2\log(m(g^{\boldsymbol{\theta}}{}_k(\boldsymbol{x})))\Bigg] + d\log\gamma + \kappa_d - \frac{d}{2}\log(2\pi) + O_{\gamma\to\infty}(1/\gamma^2) \\
&= -\frac{1}{2}\sum_{k=1}^{d}\Bigg[\log\frac{1}{\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2} - 1 + \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\left(-n''(g^{\boldsymbol{\theta}}{}_k(\boldsymbol{x})) + \gamma^2\left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})}\right\|^2\right) \\
&\qquad - 2\log(m(g^{\boldsymbol{\theta}}{}_k(\boldsymbol{x})))\Bigg] + d\log\gamma + O_{\gamma\to\infty}(1/\gamma^2) \\
&= \widehat{\mathrm{ELBO}}(\boldsymbol{\sigma}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2;\boldsymbol{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}) + \sum_{k=1}^{d}\log(m(g^{\boldsymbol{\theta}}{}_k(\boldsymbol{x}))) + O_{\gamma\to\infty}(1/\gamma^2)\,,
\end{aligned}$$

where we isolated the terms that depend on parameters $\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2$ and γ in the approximate objective $\widehat{\mathrm{ELBO}}(\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2;\boldsymbol{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}})$ that we define for arbitrary $\boldsymbol{\sigma}^2$.

$$\begin{aligned}
\widehat{\mathrm{ELBO}}(\boldsymbol{\sigma}^2;\boldsymbol{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}) &= -\frac{1}{2}\sum_{k=1}^{d}\left[\log\frac{1}{\gamma^2\sigma_k^2} - 1 + \sigma_k^2\left(-n''(g^{\boldsymbol{\theta}}{}_k(\boldsymbol{x})) + \gamma^2\left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})}\right\|^2\right)\right] \\
&= \sum_{k=1}^{d}\widehat{\mathrm{ELBO}}_k(\sigma_k^2;\boldsymbol{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}})
\end{aligned}$$

Where we further break this objective in d components $\widehat{\mathrm{ELBO}}_k(\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2;\boldsymbol{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}})$ according to the terms of the sum as follows

$$\widehat{\mathrm{ELBO}}_k(\sigma_k^2;\boldsymbol{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}) = -\frac{1}{2}\left[\log\frac{1}{\gamma^2\sigma_k^2} - 1 + \gamma^2\sigma_k^2\left(-\frac{1}{\gamma^2}n''(g^{\boldsymbol{\theta}}{}_k(\boldsymbol{x})) + \left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})}\right\|^2\right)\right]$$

and where we note that $-n'' \geq 0$ due to the log-concavity assumption.

Solving term in k $\widehat{\mathrm{ELBO}}_k(\sigma_k^2)$ for optimal $\gamma^2\sigma_k^*$ we get (see lemma 4):

$$\gamma^2\sigma_k^{*2} = \left(-\frac{1}{\gamma^2}n''(g^{\boldsymbol{\theta}}{}_k(\boldsymbol{x})) + \left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|g^{\boldsymbol{\theta}}{}_k(\boldsymbol{x})}\right\|^2\right)^{-1} \tag{40}$$

and the resulting optimal value $\widehat{\mathrm{ELBO}}_k^*(\boldsymbol{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}) = \widehat{\mathrm{ELBO}}_k(\sigma_k^{*2};\boldsymbol{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}})$ is

$$\widehat{\mathrm{ELBO}}_k^*(\boldsymbol{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}})^* = -\frac{1}{2}\log\left(-\frac{1}{\gamma^2}n''(g^{\boldsymbol{\theta}}{}_k(\boldsymbol{x})) + \left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|g^{\boldsymbol{\theta}}{}_k(\boldsymbol{x})}\right\|^2\right)$$

A Taylor formula around this optimum leads, for some value $\xi_\gamma(\boldsymbol{x})$ lying between σ_k^{*2} and σ_k^2 to (note the first order derivative vanishes, and the second order derivative is upper bounded hence the second line)

$$\begin{aligned}\widehat{\text{ELBO}}_k(\sigma_k^2;\boldsymbol{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}) = \widehat{\text{ELBO}}_k^*(\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}) + \frac{d\widehat{\text{ELBO}}_k(\boldsymbol{x};\boldsymbol{\theta},\widehat{\boldsymbol{\phi}})}{d\gamma^2\sigma_k^2}_{|\sigma_k^{*2}}&(\gamma^2\sigma_k^2-\gamma^2\sigma_k^{*2})\\ + \frac{d^2\widehat{\text{ELBO}}_k(\boldsymbol{x};\boldsymbol{\theta},\widehat{\boldsymbol{\phi}})}{d(\gamma^2\sigma_k^2)^2}_{|\xi_\gamma(\boldsymbol{x})}&(\gamma^2\sigma_k^2-\gamma^2\sigma_k^{*2})^2\\ \leq \widehat{\text{ELBO}}_k^*(\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}) - \frac{1}{2}\left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})}\right\|^2 &(\gamma^2\sigma_k^2-\gamma^2\sigma_k^{*2})^2\end{aligned}$$

as a consequence the non-approximate solution for the true optimal ELBO*, as γ grows, must achieve a value below this quadratic function, up to a term in $O(1/\gamma^2)$, and at the same time above $\widehat{\text{ELBO}}^*$, also up to a term in $O(1/\gamma^2)$. This entails that it is restricted to a smaller and smaller domain near the approximate solution and we get

$$\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = \sigma_k^{*2} + O(1/\gamma^3) = \left(-n''({g^{\boldsymbol{\theta}}}_k(\boldsymbol{x})) + \gamma^2\left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|{g^{\boldsymbol{\theta}}}_k(\boldsymbol{x})}\right\|^2\right)^{-1} + O(1/\gamma^3). \tag{41}$$

Leading to the approximation of the true objective

$$\text{ELBO}^*(\boldsymbol{x};\boldsymbol{\theta}) = -\frac{1}{2}\sum_{k=1}^{d}\left[\log\left(-\frac{1}{\gamma^2}n''(\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})) + \left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})}\right\|^2\right) - 2\log(m(\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})))\right] + O(1/\gamma^2),$$

which reduces to

$$\text{ELBO}^*(\boldsymbol{x};\boldsymbol{\theta}) = \log p_0(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})) - \frac{1}{2}\sum_{k=1}^{d}\left[\log\left\|\left[\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}))\right]_{:k}\right\|^2\right] + O(1/\gamma^2),$$

which is the IMA objective.

□

We now restate the main text theorem and provide its proof.

Theorem 1. *[VAEs with a near-deterministic decoder approximate the* IMA *objective] Under Assumption 1, the variational posterior satisfies (denoting $n'' = \frac{d^2\log p_0}{dz_k^2}$)*

$$\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = [\gamma^2\|\left[\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]_{:k}\|^2 - n''(g_k^{\boldsymbol{\theta}}(\boldsymbol{x}))]^{-1} + O(1/\gamma^3), \tag{8}$$

and the self-consistent ELBO (6) *approximates the* IMA*-regularized log-likelihood [11]:*

$$\text{ELBO}^*(\boldsymbol{x};\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}},\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})) + O_{\gamma\to\infty}(1/\gamma^2). \tag{9}$$

Proof. This is just a corollary of Proposition 5 because Proposition 4 entails through (28) the required $O(1/\gamma^2)$ rate of convergence for the optimal variational mean in (30), while (31) is fulfilled through Prop. 1. □

8 AUXILIARY RESULTS

8.1 SQUARED NORM STATISTICS

Lemma 1 (Squared norm variance decomposition)**.** *For multivariate RV X with mean m*

$$\mathbb{E}\left[\|X\|^2\right] = \textit{trace}\left[\textit{Cov}(X)\right] + \|m\|^2$$

Proof.

$$\mathbb{E}\|X-m\|^2 = \mathbb{E}\left\langle X-m,\, X-m\right\rangle = \mathbb{E}\left[\langle X,\, X\rangle - 2\mathbb{E}\left\langle m,\, X\right\rangle + \langle m,\, m\rangle\right]$$

hence

$$\mathbb{E}\|X-m\|^2 = \mathbb{E}\left[\|X\|^2\right] - \|m\|^2$$

This leads to (using that the trace of a scalar is the scalar itself)

$$\mathbb{E}\left[\|X\|^2\right] = \mathbb{E}\left[\text{trace}\left[\|X-m\|^2\right]\right] + \|m\|^2 = \text{trace}\left[\mathbb{E}\left[(X-m)^T(X-m)\right]\right] + \|m\|^2$$

because $\text{trace}[AB] = \text{trace}[BA]$ we get

$$\mathbb{E}\left[\|X\|^2\right] = \text{trace}\left[\mathbb{E}\left[(X-m)(X-m)^T\right]\right] + \|m\|^2 = \text{trace}\left[\text{Cov}(X)\right] + \|m\|^2$$

□

Lemma 2 (Trace of transformed unit covariance)**.** *When the covariance matrix $Cov(\boldsymbol{\epsilon})$ is the identity, then*

$$\textit{trace}[Cov(A\boldsymbol{\epsilon})] = \sum_k \|[A]_{.k}\|^2\,,$$

Proof. For arbitrary matrix A, $Cov(A\boldsymbol{\epsilon}) = ACov(\boldsymbol{\epsilon})A^T$ and thus

$$\text{trace}[Cov(A\boldsymbol{\epsilon})] = \text{trace}[ACov(\boldsymbol{\epsilon})A^T] = \text{trace}[A^TACov(\boldsymbol{\epsilon})]\,.$$

Moreover, in our case $Cov(\boldsymbol{\epsilon})$ is the identity such that

$$\text{trace}[Cov(A\boldsymbol{\epsilon})] = \text{trace}[A^TA] = \sum_k \|[A]_{.k}\|^2\,,$$

□

8.2 KL DIVERGENCE BOUNDS

Proposition 6 (Lipschtiz continuity-based lower bound)**.** *Assume $\boldsymbol{g^\theta}$ is Lipschitz continuous with Lipschitz constant $B>0$, in the sense*

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d,\, \left\|\boldsymbol{g^\theta}(\boldsymbol{x}) - \boldsymbol{g^\theta}(\boldsymbol{y})\right\|_2 \leq B\|\boldsymbol{x}-\boldsymbol{y}\|_2\,.$$

Then for any encoder parameter choice

$$\textit{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq -\sum_{k=1}^{d}\left[\log\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + \log m(\mu_k^{\boldsymbol{\phi}})\right] + c(\boldsymbol{x},\gamma)$$
$$+\frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g^\theta}(\boldsymbol{x}) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right\|^2 + \sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right]\,, \quad (42)$$

with $c(\boldsymbol{x},\gamma) = -\frac{d}{2}\left(\log(\gamma^2)+1\right) + \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$.

Proof. Starting from the KL divergence expression (12),

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] = -\sum_{k=1}^{d} \log \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}}\left[\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}\right)\|^2\gamma^2 - 2\sum_{k=1}^{d}\log m(z_k)\right] + c(\boldsymbol{x},\gamma)$$

with additive constant $c(\boldsymbol{x},\gamma) = -\frac{d}{2}\left(\log(\gamma^2)+1\right) + \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$. By Lipschitz continuity

$$\begin{aligned}\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq & -\sum_{k=1}^{d} \log \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) \\ & + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}}\left[B^{-2}\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{z}\|^2\gamma^2 - 2\sum_{k=1}^{d}\log m(z_k)\right] + c(\boldsymbol{x},\gamma)\,.\end{aligned}$$

using Lemma 1 applied to $\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{z}$, $\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ we get

$$\begin{aligned}\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] &\geq -\sum_{k=1}^{d} \log \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right\|^2 + \text{trace}\left[Cov\left[\boldsymbol{z}\right]\right]\right] \\ &\quad - \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}}\left[\sum_{k=1}^{d}\log m(z_k)\right] + c(\boldsymbol{x},\gamma) \\ &\geq -\sum_{k=1}^{d} \log \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right\|^2 + \sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right] \\ &\quad - \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}}\left[\sum_{k=1}^{d}\log m(z_k)\right] + c(\boldsymbol{x},\gamma)\,.\end{aligned}$$

Using Jensen's inequality for $-\log m$ (convex by Assum. 1(i)), we get

$$\begin{aligned}\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq & -\sum_{k=1}^{d}\left[\log \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})\right] + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right\|^2 + \sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right] \\ & - \sum_{k=1}^{d}\left[\log m(\mu_k^{\boldsymbol{\phi}})\right] + c(\boldsymbol{x},\gamma)\end{aligned}$$

by reordering the terms we finally get

$$\begin{aligned}\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq & -\sum_{k=1}^{d}\left[\log \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + \log m(\mu_k^{\boldsymbol{\phi}})\right] + c(\boldsymbol{x},\gamma) \\ & + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right\|^2 + \sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right]\end{aligned}$$

which is the stated KL lower bound. □

Proposition 7 (Optimal encoder KL divergence upper bound)**.** *Assume $\boldsymbol{f}^{\boldsymbol{\theta}}$ is Lipschitz continuous with Lipschitz constant $C>0$, in the sense that*

$$\forall \boldsymbol{z},\boldsymbol{w}\in\mathbb{R}^d: \qquad \left\|\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}\right) - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{w})\right\|_2 \leq C\|\boldsymbol{z}-\boldsymbol{w}\|_2\,.$$

Assume, $-\log m$ is quadratically dominated, in the sense that

$$\exists D>0, E>0, \forall u\in\mathbb{R}, -\log m(u) \leq D|u|^2 + E\,.$$

Then for the optimal encoder solution of (5)

$$KL\left[q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \quad\leq\quad d\left(\frac{1}{2}C^2 + E + D\left[\frac{\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2}{d} + \frac{1}{\gamma^2}\right]\right) \quad-\quad \frac{d}{2} \quad+\quad \log p_{\boldsymbol{\theta}}(\boldsymbol{x})\,, \tag{43}$$

and

$$\limsup_{\gamma\to+\infty} KL\left[q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \leq d\left(\frac{1}{2}C^2+E\right)+D\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2$$
$$-\frac{d}{2}-\log|\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right))|+\log(p_0(\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right))) \quad (44)$$

Proof. Starting from the KL divergence expression (12),

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] = -\sum_{k=1}^{d}\log\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})+\frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}}\left[\|\boldsymbol{x}-\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}\right)\|^2\gamma^2-2\sum_{k=1}^{d}\log m(z_k)\right]$$
$$+c(\boldsymbol{x},\gamma)$$

with additive constant $c(\boldsymbol{x},\gamma)=-\frac{d}{2}\left(\log(\gamma^2)+1\right)+\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$.

Let us choose the following posterior (by universal approximation capabilities of the encoder):

$$\boldsymbol{\mu}^{\boldsymbol{\phi}^*}(\boldsymbol{x})=\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) \quad (45)$$
$$\boldsymbol{\sigma}^{\boldsymbol{\phi}^*}(\boldsymbol{x})=\frac{1}{\gamma} \quad (46)$$

Using Lipschitz continuity we get

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}^*}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \leq -\sum_{k=1}^{d}\log\sigma_k^{\boldsymbol{\phi}^*}(\boldsymbol{x})+\frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}^*}}\left[C^2\|\boldsymbol{\mu}^{\boldsymbol{\phi}^*}(\boldsymbol{x})-\boldsymbol{z}\|^2\gamma^2-2\sum_{k=1}^{d}\log m(z_k)\right]$$
$$+c(\boldsymbol{x},\gamma)$$

then, using

$$\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}^*}}\left[\|\boldsymbol{\mu}^{\boldsymbol{\phi}^*}(\boldsymbol{x})-\boldsymbol{z}\|^2\right]=\sum_{k=1}^{d}\mathbb{E}_{z_k\sim\mathcal{N}(\mu_k^{\boldsymbol{\phi}^*}(\boldsymbol{x}),\sigma_k^{\boldsymbol{\phi}^*}(\boldsymbol{x})^2)}\left[|\mu_k^{\boldsymbol{\phi}^*}(\boldsymbol{x})-z_k|^2\right]=\sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}^*}(\boldsymbol{x})^2\,,$$

we get

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}^*}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \leq \sum_{k=1}^{d}\left(-\log\sigma_k^{\boldsymbol{\phi}^*}(\boldsymbol{x})+\frac{1}{2}C^2\sigma_k^{\boldsymbol{\phi}^*}(\boldsymbol{x})^2\gamma^2\right)$$
$$-\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}^*}}\left[\sum_{k=1}^{d}\log m(z_k)\right]+c(\boldsymbol{x},\gamma)$$

using quadratic domination

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}^*}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \leq \sum_{k=1}^{d}\left(-\log\sigma_k^{\boldsymbol{\phi}^*}(\boldsymbol{x})+\frac{1}{2}C^2\sigma_k^{\boldsymbol{\phi}^*}(\boldsymbol{x})^2\gamma^2\right)$$
$$+\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}^*}}\left[dE+\sum_{k=1}^{d}D|z_k|^2\right]+c(\boldsymbol{x},\gamma)$$
$$\leq \sum_{k=1}^{d}\left(-\log\sigma_k^{\boldsymbol{\phi}^*}(\boldsymbol{x})+\frac{1}{2}C^2\sigma_k^{\boldsymbol{\phi}^*}(\boldsymbol{x})^2\gamma^2\right)$$
$$+dE+D\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}^*}}\left[|z_k|^2\right]+c(\boldsymbol{x},\gamma)$$

Using Lemma 1 we get

$$\begin{aligned}
\mathrm{KL}\left[(q_{\boldsymbol{\phi}^*}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \leq & \sum_{k=1}^{d}\left(-\log\sigma_k^{\boldsymbol{\phi}^*}(\boldsymbol{x})+\frac{1}{2}C^2\sigma_k^{\boldsymbol{\phi}^*}(\boldsymbol{x})^2\gamma^2\right)\\
& + dE + D\left[\|\boldsymbol{\mu}^{\boldsymbol{\phi}^*}(\boldsymbol{x})\|^2+\|\boldsymbol{\sigma}^{\boldsymbol{\phi}^*}(\boldsymbol{x})\|^2\right]+c(\boldsymbol{x},\gamma)\\
\leq & \, d\left(\log\gamma+\frac{1}{2}C^2\right)+dE+D\left[\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2+\frac{d}{\gamma^2}\right]-\frac{d}{2}\left(\log(\gamma^2)+1\right)+\log p_{\boldsymbol{\theta}}(\boldsymbol{x})
\end{aligned}$$

hence for a parameter $\widehat{\boldsymbol{\phi}}$ achieving the minimum divergence we get

$$\begin{aligned}
\mathrm{KL}\left[q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \leq & \,\mathrm{KL}\left[q_{\boldsymbol{\phi}^*}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \leq d\left(\log\gamma+\frac{1}{2}C^2\right)\\
& + dE + D\left[\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2+\frac{d}{\gamma^2}\right]-\frac{d}{2}\left(\log(\gamma^2)+1\right)+\log p_{\boldsymbol{\theta}}(\boldsymbol{x})\\
& \leq d\left(\frac{1}{2}C^2+E+D\left[\frac{\|\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\|^2}{d}+\frac{1}{\gamma^2}\right]\right)-\frac{d}{2}+\log p_{\boldsymbol{\theta}}(\boldsymbol{x})
\end{aligned}$$

As $\gamma \to +\infty$, $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \to |\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right))|^{-1}p_0(\boldsymbol{g}^{\boldsymbol{\theta}}\left(\boldsymbol{x}\right))$ such that the KL divergence for the optimal solutions is upper bounded by a finite number.

□

8.3 TAYLOR FORMULA-BASED APPROXIMATIONS

Lemma 3 (Bound on expectation of multivariate Taylor expansion)**.** *Assume $\boldsymbol{f}:\mathbb{R}^d\to\mathbb{R}$ is C^2 and assume $\boldsymbol{z}$ is a multivariate RV on $\mathbb{R}^d$ with indepedent Gaussian components such that*

$$z_k \sim \mathcal{N}(\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x}),\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2)$$

then for all $\boldsymbol{z}_o\in\mathbb{R}^d$

$$\mathbb{E}_{\boldsymbol{z}}\left[\left\|\boldsymbol{f}(\boldsymbol{z})-\boldsymbol{f}(\boldsymbol{z}_o)-\sum_k\frac{\partial\boldsymbol{f}}{\partial z_k}_{|\boldsymbol{z}_o}(z_k-z_k^o)\right\|^2\right]\leq\frac{d^3}{4}3K^2\sum_i\left(\sigma_i^{\boldsymbol{\phi}}\right)^4 \tag{47}$$

Proof. As described in [26, p. 162], for the l-th component of the function

$$\begin{aligned}
f_l(\boldsymbol{z}) = f_l(\boldsymbol{z}_o)+\sum_k\frac{\partial f_l}{\partial z_k}_{|\boldsymbol{z}_o}(z_k-z_k^o)+\frac{1}{2!}\sum_{i,j}\frac{\partial f_l}{\partial z_i\partial z_j}_{|\boldsymbol{z}_o+t_{ij}(\boldsymbol{z}-\boldsymbol{z}_o)}(z_i-z_i^o)(z_j-z_j^o)\,,t_{ij}\in(0;1)\ .\\
= f_l(\boldsymbol{z}_o)+\sum_k\frac{\partial f_l}{\partial z_k}_{|\boldsymbol{z}_o}(z_k-z_k^o)+\frac{1}{2!}\sum_{i,j}(\boldsymbol{z}-\boldsymbol{z}_o)^T\mathbf{H}_k(\boldsymbol{z}-\boldsymbol{z}_o)\,, \quad (48)
\end{aligned}$$

where the second line puts 1/2 of the partial derivatives in matrix form (note it is not exactly the Hessian as derivatives are taken at different points). As a consequence

$$\begin{aligned}
\left(f_l(\boldsymbol{z})-f_l(\boldsymbol{z}_o)-\sum_k\frac{\partial f_l}{\partial z_k}_{|\boldsymbol{z}_o}(z_k-z_k^o)\right)^2 &= \left((\boldsymbol{z}-\boldsymbol{z}_o)^T\mathbf{H}_k(\boldsymbol{z}-\boldsymbol{z}_o)\right)^2,\\
&\leq \|\mathbf{H}_k\|_2^2\,\|\boldsymbol{z}-\boldsymbol{z}_o\|^4\\
&\leq \|\mathbf{H}_k\|_F^2\,\|\boldsymbol{z}-\boldsymbol{z}_o\|^4
\end{aligned}$$

where $\|\mathbf{H}_k\|_2$ is the spectral norm of the matrix and $\|\mathbf{H}_k\|_F$ is the Frobenious norm [3] leading to the bound

$$\left(f_l(\boldsymbol{z}) - f_l(\boldsymbol{z}_o) - \sum_k \frac{\partial f_l}{\partial z_k}_{|\boldsymbol{z}_o} (z_k - z_k^o)\right)^2 \leq \frac{d^2}{4} K^2 \|\boldsymbol{z} - \boldsymbol{z}_o\|^4 ,$$

where K is an upper bound on the absolute second order derivatives. We have $(z_k - z_k^o) = \sigma_k^{\boldsymbol{\phi}}(x)\epsilon_k$, with ϵ multivariate normal, so taking the expectation of the above simplifies to:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Z}}\left(f_l(\boldsymbol{z}) - f_l(\boldsymbol{z}_o) - \sum_k \frac{\partial f_l}{\partial z_k}_{|\boldsymbol{z}_o} (z_k - z_k^o)\right)^2 &\leq \frac{d^2}{4} K^2 \mathbb{E}_{\boldsymbol{Z}} \|\boldsymbol{z} - \boldsymbol{z}_o\|^4 ,\\ &= \frac{d^2}{4} K^2 \mathbb{E}_{\boldsymbol{Z}} \sum_{i,j} \left\|z_i - z_j^o\right\|^2 \left\|z_i - z_j^o\right\|^2 \\ &= \frac{d^2}{4} K^2 \sum_i \mathbb{E}_{\boldsymbol{Z}} \|z_i - z_i^o\|^4 \\ &= \frac{d^2}{4} 3K^2 \sum_i \left(\sigma_i^{\boldsymbol{\phi}}\right)^4 .\end{aligned}$$

Now gathering all components f_l to get the squared norm yields:

$$\mathbb{E}_{\boldsymbol{Z}}\left[\left\|\boldsymbol{f}(\boldsymbol{z}) - \boldsymbol{f}(\boldsymbol{z}_o) - \sum_k \frac{\partial \boldsymbol{f}}{\partial z_k}_{|\boldsymbol{z}_o} (z_k - z_k^o)\right\|^2\right] \leq \frac{d^3}{4} 3K^2 \sum_i \left(\sigma_i^{\boldsymbol{\phi}}\right)^4 .$$

□

8.4 VARIATIONAL POSTERIOR VARIANCE OPTIMIZATION PROBLEM

Lemma 4. *For $\alpha > 0$, the function*

$$\begin{aligned} h_\alpha : &\mathbb{R}_{>0} \to \mathbb{R} \\ &u \mapsto -\log u - \frac{1}{2} + \alpha u^2/2 = \frac{1}{2}\log\frac{1}{u^2} - \frac{1}{2} + \alpha u^2/2 \end{aligned}$$

is strictly convex and achieves its global minimum $\min h_\alpha = \frac{1}{2}\log\alpha$ *for* $u^* = \frac{1}{\sqrt{\alpha}}$.

Proof. Function h_α is stricly convex as a sum of two stricly convex functions. Its derivative,

$$\frac{dh_\alpha}{du}(u) = -\frac{1}{u} + \alpha u,$$

thus vanishes only at the minimum for $u^* = \frac{1}{\sqrt{\alpha}}$. We then get that

$$\min h_\alpha = h_\alpha(u^*) = \frac{1}{2}\log\alpha\,.$$

□

[3] first inequality comes from Cauchy-Schwartz: $< x, Ax > \leq \|x\|\|Ax\| \leq \|x\|\|A\|_2\|x\|$, second is a classical inequality between norms

9 RELATED WORK

9.1 IMPLICIT INDUCTIVE BIASES IN THE ELBO

Rolinek et al. [32] reason about the connection to Principal Component Analysis (PCA) in the context of nonlinear Gaussian VAEs with an isotropic prior and assume that the variational posterior has *diagonal covariance with distinct singular values*. The authors make it explicit that they investigate the consequences of optimizing the ELBO. They locally linearize the decoder to show the inductive bias in VAEs that promotes decoder orthogonality. Their results hold for β-VAEs, where β should be in the range of satisfying the polarized regime assumption (i.e., when the VAE is close to partial posterior collapse). The validity of the assumptions (polarized regime and distinct singular values in $\mathbf{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}|\boldsymbol{x}}$) are only experimentally investigated. The same authors extend their work in [38], completing the connection to PCA for *linear* models. Their experiments, inspired by the connection to PCA for linear models, show that local perturbations in the data prohibit disentanglement for non-linear models.

Lucas et al. [25] prove that *linear Gaussian* VAEs with an isotropic prior give rise to a *column-orthogonal decoder* and therefore uniquely recover the PCA coordinate axes (not just the correct subspace, as Probabilistic Principal Component Analysis (PPCA) [35] does), yielding identifiability for Gaussian models—but only when the eigenvalues of the data covariance are distinct. In their work, the decoder variance is shown to be small when avoiding posterior collapse. More interestingly, the authors derive a formula for the ELBO gap in the linear case that is remarkably similar to the IMA objective. We show in § 10.1 that in the limit of a deterministic decoder linear Gaussian VAEs optimize the IMA objective with $\lambda = 1$.

Kumar and Poole [22] generalizes [32], as it admits a variational posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ with *block-diagonal covariance* with a uniqueness result for diagonal $\mathbf{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}|\boldsymbol{x}}$. The authors derive a formula for the optimal $\mathbf{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}|\boldsymbol{x}}$ [22, Eq. 12], showing that when the decoder Hessian $\mathbf{H}$ is diagonal, the decoder Jacobian will be column-orthogonal even for *non-Gaussian* decoders. Their analysis relies on a "concentrated" $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ (i.e., they work in what we term the near-deterministic regime) and sufficiently small values of β—this relationship can be read off from [22, Eq. 12]. Interestingly, the authors also show that rotations of the latents can be ruled out, though they do not connect the decoder structure (especially, column-orthogonality of its Jacobian) to any specific generative model for the data, or to considerations on identifiability of the ground truth sources.

9.2 (NEAR)-DETERMINISTIC VAES

Recent work was inspired by the normalizing flow literature and the shortcomings of the stochastic VAE architecture to propose designs that are (near)-deterministic. Arguments for this regime range from avoiding posterior collapse (as demonstrated in [25]) to avoiding sampling for the reconstruction loss term [22]. Several papers argued for a similar setting: Rolinek et al. [32] refer to the *polarized regime* (a property of which is that encoder variances are small, cf. [32, Definition 1]), Kumar and Poole [22] argue for "concentrated" variational posteriors. Ghosh et al. [9] substitute stochasticity with a regularizer on the decoder Jacobian from an intuitive, whereas Kumar et al. [23] motivate these results from an injective flow perspective. Nielsen et al. [28] also take a normalizing flow perspective to connect VAEs to deterministic models. Besides benefits of avoiding posterior collapse or possible improvements during optimization, this regime serves as a potential connection to the identifiability literature.

10 FURTHER REMARKS ON THE THE IMA–VAE CONNECTION

In this section, we elaborate on the connection between VAEs and IMA, by showing that previous work on linear VAEs can be directly connected to optimizing $\mathcal{L}_{\text{IMA}}$. Our intent with this analysis is to provide additional insights about the role of γ in a simpler setting.

10.1 LINEAR VAE FROM LUCAS ET AL.

We restate the linear VAE model of [25]:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}\left(\mathbf{W}\boldsymbol{z} + \boldsymbol{\mu}; \frac{1}{\gamma^2}\mathbf{I}_d\right) \tag{49}$$

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\mathbf{V}\left(\boldsymbol{x} - \boldsymbol{\mu}\right); \mathbf{D}\right), \tag{50}$$

where $\mathbf{D}$ is a diagonal matrix, $\mathbf{W}$ the decoder and $\mathbf{V}$ the encoder weights, $\boldsymbol{\mu}$ the mean latent representation.

The authors show that in stationary points, the optimal value for $\mathbf{D}$ is

$$\mathbf{D}^* = \frac{1}{\gamma^2}\left(\operatorname{diag}\left(\mathbf{W}^T\mathbf{W}\right) + \frac{1}{\gamma^2}\mathbf{I}_d\right)^{-1} \tag{51}$$

If we substitute this expression into the ELBO gap (i.e., the KL between the variational and true posteriors), we get a similar expression to c_{IMA}—as formalized in Prop. 8.

Proposition 8 (The ELBO converges to $\mathcal{L}_{\text{IMA}}$ for linear Gaussian VAEs if $\gamma \to +\infty$)**.** *For linear Gaussian VAEs, in the limit of $\gamma \to \infty$, the* ELBO *equals the* IMA*-regularized log-likelihood in stationary points with $\lambda = 1$.*

Proof. In [25, Appendix C.2], it is shown that the gap between exact log-likelihood and ELBO for linear Gaussian VAEs in stationary points reduces to

$$\begin{aligned} \text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] &= \frac{1}{2}\left(\log\det\tilde{\mathbf{M}} - \log\det\mathbf{M}\right) &(52)\\ \mathbf{M} &= \mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_d &(53)\\ \tilde{\mathbf{M}} &= \operatorname{diag}\left(\mathbf{W}^T\mathbf{W}\right) + \frac{1}{\gamma^2}\mathbf{I}_d, &(54)\end{aligned}$$

where $\mathbf{W}$ is the decoder weight matrix. Reformulating the above expression, we arrive at :

$$\begin{aligned} \text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] &= \log\frac{\left|\operatorname{diag}\left(\mathbf{W}^T\mathbf{W}\right) + \frac{1}{\gamma^2}\mathbf{I}_d\right|}{\left|\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_d\right|} &(55)\\ &= \log\frac{\left|\operatorname{diag}\left(\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_d\right)\right|}{\left|\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_d\right|} &(56)\end{aligned}$$

Noting that $\mathbf{W}^T\mathbf{W}$ is symmetric with a Singular Value Decomposition (SVD) of $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ ($\mathbf{U}$ is orthogonal, $\boldsymbol{\Lambda}_{ii} = \|[\mathbf{W}]_{:k}\|^2$), and $\mathbf{I}_d = \mathbf{U}\mathbf{U}^T$; thus:

$$\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_d = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T + \frac{1}{\gamma^2}\mathbf{U}\mathbf{U}^T = \mathbf{U}\left[\boldsymbol{\Lambda} + \frac{1}{\gamma^2}\mathbf{I}_d\right]\mathbf{U}^T$$

Therefore, (56) can be reformulated as the left KL-measure of diagonality [1] of the matrix $\mathbf{U}\left[\boldsymbol{\Lambda} + 1/\gamma^2\mathbf{I}_d\right]\mathbf{U}^T$:

$$\begin{aligned} \text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] &= \log\frac{\left|\operatorname{diag}\left(\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_d\right)\right|}{\left|\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_d\right|} &(57)\\ &= \log\frac{\left|\operatorname{diag}\left(\mathbf{U}\left[\boldsymbol{\Lambda} + \frac{1}{\gamma^2}\mathbf{I}_d\right]\mathbf{U}^T\right)\right|}{\left|\mathbf{U}\left[\boldsymbol{\Lambda} + \frac{1}{\gamma^2}\mathbf{I}_d\right]\mathbf{U}^T\right|}, &(58)\end{aligned}$$

which is by definition the local IMA contrast c_{IMA} (cf. [11, Appendix C.1]). When $\gamma \to +\infty$, the above expression converges to the left KL-measure of diagonality for $\mathbf{W}^T\mathbf{W}$, i.e., the local IMA contrast for the decoder.

$\gamma \to +\infty$ thus means that the ELBO converges to the IMA regularized log-likelihood $\mathcal{L}_{\text{IMA}}$ with $\lambda = 1$:

$$\begin{aligned} \text{ELBO} &= \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right]\\ &= \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - c_{\text{IMA}}(\mathbf{W}, \boldsymbol{z}),\end{aligned}$$

which concludes the proof. □

Prop. 8, especially (58), gives us intuitive understanding on why and how γ influences how much the orthogonality of $\mathbf{W}$ is enforced.

1. Small γ (high observation noise) means that there is no reason to promote the orthogonality of the decoder, as the high noise level (i.e., low-quality fit of $\boldsymbol{x}$) will drive (58) towards diagonality via $1/\gamma^2$.

2. On the other hand, when $\gamma \to +\infty$, then the orthogonality of the decoder is promoted. That is, the decoder precision γ^2 acts akin to a weighting factor influencing how strong the IMA principle should be enforced.

We can observe that the ELBO recovers the exact log-likelihood for column-orthogonal $\mathbf{W}$:

Corollary 1 (For column-orthogonal $\mathbf{W}$ the ELBO equals the exact log-likelihood)**.** *When* $\mathbf{W}$ *is in the form* $\mathbf{W} = \mathbf{OD}$, *then* $\operatorname{diag}\left(\mathbf{W}^T\mathbf{W}\right) = \mathbf{W}^T\mathbf{W} = \mathbf{DO}^T\mathbf{OD} = \mathbf{D}^2$, *i.e. the* ELBO *corresponds to the exact log-likelihood since* (58) *is zero.*

Corollary 1 also implies that γ does not affect the gap between ELBO and exact log-likelihood for column-orthogonal $\mathbf{W}$.

11 EXPERIMENTAL DETAILS

11.1 SELF-CONSISTENCY IN PRACTICAL CONDITIONS (??)

For the self-consistency experiments, the mixing is a 3-layer MultiLayer Perceptron (MLP) with smooth Leaky ReLU nonlinearities [10] and orthogonal weight matrices—which intentionally does not belong to the IMA class, since our self-consistency result is not constrained to the IMA class. The 60,000 source samples are drawn from a standard normal distribution and fed into a VAE composed of a 3-layer MLP encoder and decoder with a Gaussian prior. We use 20 seeds for each $\gamma^2 \in \{1\mathrm{e}1; 1\mathrm{e}2; 1\mathrm{e}3; 1\mathrm{e}4; 1\mathrm{e}5\}$. Additional parameters are described in Tab. 1. Training is continued until the ELBO* improves on the *validation set* (we use early stopping [30]), then all metrics are reported for the maximum ELBO* (Fig. 2).

Table 1: Hyperparameters for the self-consistency experiments (§ 4)

Parameter	Values
Encoder	3-layer MLP
Decoder	3-layer MLP
Activation	smooth Leaky ReLU [10]
Batch size	64
# Samples (train-val-test)	$42 - 12 - 6$k
Learning rate	$1\mathrm{e}{-3}$
d	3
Ground truth	Gaussian
$p_0(\boldsymbol{z})$	Gaussian
$\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}\|\boldsymbol{x}}$	Diagonal
γ^2	$\{1\mathrm{e}1; 1\mathrm{e}2; 1\mathrm{e}3; 1\mathrm{e}4; 1\mathrm{e}5\}$
# Seeds	20

11.2 RELATIONSHIP BETWEEN ELBO*, IMA-REGULARIZED, AND UNREGULARIZED LOG-LIKELIHOODS (??)

Table 2: Hyperparameters for the *triangular MLP* (*not* from the IMA class) ELBO*–$\mathcal{L}_{\text{IMA}}$–log-likelihood experiments (**??**)

PARAMETER	VALUES
ENCODER	3-LAYER MLP
DECODER	2-LAYER TRIANGULAR MLP (GROUND TRUTH)
ACTIVATION	SIGMOID
BATCH SIZE	64
# SAMPLES (TRAIN-VAL-TEST)	$100 - 30 - 15$K
LEARNING RATE	$1\text{e}{-4}$
d	2
GROUND TRUTH	GAUSSIAN
$p_0(\boldsymbol{z})$	GAUSSIAN
$\boldsymbol{\Sigma}^{\phi}_{\boldsymbol{z}\|\boldsymbol{x}}$	DIAGONAL
γ^2	[1e1; 1e5]
# SEEDS	5
C_{IMA} (MIXING)	7.072

For the experiments comparing the ELBO*, IMA-regularized, and unregularized log-likelihoods, data is generated by mixing points from a standard Gaussian prior using an invertible neural network. When the mixing is not in the IMA-class (Tab. 2), we use a two-layer neural network with sigmoid nonlinearites and triangular weight matrices. When the mixing is from the IMA-class (Tab. 3), we use a one-layer neural network with orthogonal weight matrices. The data dimensionality in both cases is two.

Training is carried out using a VAE with a decoder fixed to the ground-truth and separate encoder models for the means and variances of the approximate posterior. The encoder comprises two three-layer neural networks with ReLU non-linearities and a hidden layer size of 50. Due to training instabilities when using a large γ, we train the model by first fixing the mean encoder to the ground-truth inverse of the mixing for the first 30 epochs; thus, only training the variances. We then train both for the remaining epochs. Training is stopped after the ELBO* plateaus on the *validation set*. A training set of 100,000 samples is used, with a validation set and test set of 30,000 and 15,000 samples, respectively. The learning rate is $1\text{e}{-4}$ and the batch size 64.

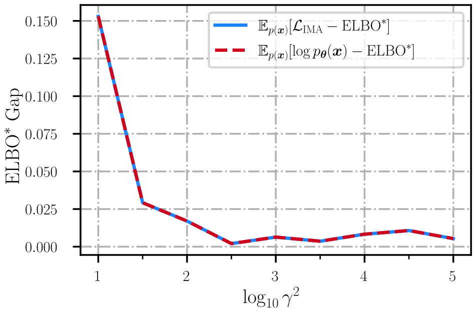


Figure 5: Comparison of the ELBO*, the IMA-regularized and unregularized log-likelihoods over different γ^2 with an IMA-class mixing

We provide additional results when the mixing is from the IMA class (Tab. 3): as C_{IMA} is zero, we expect that both $\mathcal{L}_{\text{IMA}}$ and the unregularized log-likelihood match. Indeed, this is what Fig. 5 demonstrates.

Table 3: Hyperparameters for the *orthogonal MLP* (from the IMA class) ELBO*–$\mathcal{L}_{\text{IMA}}$–log-likelihood experiments (**??**)

PARAMETER	VALUES
ENCODER	3-LAYER MLP
DECODER	1-LAYER ORTHOGONAL MLP (GROUND TRUTH)
ACTIVATION	SIGMOID
BATCH SIZE	64
# SAMPLES (TRAIN-VAL-TEST)	$100 - 30 - 15$K
LEARNING RATE	1e-4
d	2
GROUND TRUTH	UNIFORM
$p_0(\boldsymbol{z})$	UNIFORM
$\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}\mid\boldsymbol{x}}$	DIAGONAL
γ^2	$[1\text{e}1; 1\text{e}5]$
C_{IMA} (MIXING)	0

11.3 CONNECTING THE IMA PRINCIPLE, γ^2, AND DISENTANGLEMENT (??)

Synthetic data (Möbius transform) We use 3-dimensional conformal mixings (i.e., the Möbius transform [29]) from the IMA class with the functional form:

$$\boldsymbol{x} = \boldsymbol{t} + \alpha \frac{\mathbf{W}\left(\boldsymbol{z} - \boldsymbol{b}\right)}{\left\|\boldsymbol{z} - \boldsymbol{b}\right\|^{\epsilon}},$$

where $\boldsymbol{t}, \boldsymbol{b} \in \mathbb{R}^d$, $\mathbf{W} \in \mathbb{R}^{d\times d}$, $\alpha \in \mathbb{R}$, and $\epsilon = 2$ (to ensure nonlinearity) with $d = 3$. Both ground-truth and prior distributions are *uniform* to avoid the singularity when $\boldsymbol{z} = \boldsymbol{b}$.

To determine whether a mixing from the IMA class is beneficial for disentanglement, we apply a volume-preserving linear map after the Möbius transform (using 100 seeds) to construct a mixing outside of the IMA class. We fix $\gamma^2 = 1\text{e}5$ and report further parameters in Tab. 4. Training is continued until the ELBO* improves on the *validation set* (we use early stopping [30]), then all metrics are reported for the maximum ELBO* (Fig. 4).

Table 4: Hyperparameters for the *synthetic (Möbius)* IMA–disentanglement experiments (**??**) with a linear map

PARAMETER	VALUES
ENCODER	3-LAYER MLP
DECODER	3-LAYER MLP
ACTIVATION	SMOOTH LEAKY RELU [10]
BATCH SIZE	64
# SAMPLES (TRAIN-VAL-TEST)	$42 - 12 - 6$K
LEARNING RATE	1e-3
d	3
GROUND TRUTH	UNIFORM
$p_0(\boldsymbol{z})$	UNIFORM
$\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}\mid\boldsymbol{x}}$	DIAGONAL
γ^2	1e5
# SEEDS	100
C_{IMA} (MIXING)	$[0.398; 6.761]$

Image data (Sprites) We train a VAE (not β-VAE) with a factorized Gaussian posterior and Beta prior on a Sprites image dataset generated using the spriteworld renderer [37] with a Beta ground truth distribution. Similar to [16], we use four latent factors, namely, *x- and y-position, color and size*, and omit factors that can be problematic, such as shape (as it is discrete) and rotation (due to symmetries) [32, 21]. Our choice is motivated by [13, 8] showing that the data-generating process presumably is in the IMA class. The architecture both for encoder and decoder consists of four convolutional and three linear layers with ReLU nonlinearities (Tab. 5). Training is continued until the ELBO* improves on the *validation set* (we use early stopping [30]), then all metrics are reported for the maximum ELBO*.

Table 5: Hyperparameters for the *image (Sprites)* IMA–disentanglement experiments (**??**)

Parameter	Values
Encoder	4-layer Conv2D + 3-layer MLP
Decoder	4-layer Conv2D + 3-layer MLP
Activation	ReLU
Batch size	64
# Samples (train-val-test)	$42 - 12 - 6$k
Learning rate	1e-5
d	3
Ground truth	Beta
$p_0(\boldsymbol{z})$	Beta
$\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}\vert\boldsymbol{x}}$	Diagonal
γ^2	1e0
# Seeds	10

12 COMPUTATIONAL RESOURCES

The self-consistency (§ 4), the likelihood comparison (**??**), and the synthetic experiments with the Möbius transform (**??**, particularly Fig. 4) were ran on a MacBook Pro with a Quad-Core Intel Core i5 CPU and required approximately nine days. The Sprites experiments (**??**, particularly **??**) required approximately four and a half days on an Nvidia RTX 2080 GPU.

13 SOCIETAL IMPACT

Our paper presents basic research and is mainly theoretical, though the lack of direct connection to a specific application does not mean that our results could not be used for malevolent purposes. We acknowledge that providing a possible mechanism for why unsupervised VAEs can learn disentangled representations can inform specific actors that unsupervised VAEs might be used to extract the true generating factors. Since no auxiliary variables, labels, or conditional distributions are required, this might lead to a broader use of unsupervised VAEs for trying to learn the true generating factors—including applications with potentially negative societal impact such as extracting features from images, video, or text for personal identification; thus, possibly violating the desire of those who intend to remain anonymous.

14 NOTATION

ACRONYMS

ELBO evidence lower bound
IMA Independent Mechanism Analysis

i.i.d. independent and identically distributed
ICA Independent Component Analysis

KL Kullback-Leibler Divergence

LVM Latent Variable Model

MCC Mean Correlation Coefficient
MLP MultiLayer Perceptron

PCA Principal Component Analysis
PPCA Probabilistic Principal Component Analysis

SVD Singular Value Decomposition

VAE Variational Autoencoder

NOMENCLATURE

Independent Mechanism Analysis

$C_{\mathbf{IMA}}$ global IMA contrast
α scalar field
$\mathbf{D}$ general diagonal matrix
$\mathbf{O}$ orthogonal matrix
$\boldsymbol{y}$ reconstructed sources

$\mathcal{L}_{\mathbf{IMA}}$ IMA loss function

$c_{\mathbf{IMA}}$ local IMA contrast

Variational Autoencoder

$\mathbf{V}$ weight matrix of a linear encoder

$\mathbf{W}$ weight matrix of a linear decoder

$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})$ optimal mean of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$

$\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})$ mean of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$

$\boldsymbol{\phi}$ parameters of the variational posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$

$\boldsymbol{\sigma}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2$ optimal variance of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$

$\boldsymbol{\theta}$ parameters of the decoder $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$

γ square root of the precision of the VAE decoder

$\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}|\boldsymbol{x}}$ covariance matrix of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$

$\mathcal{L}_{\beta}$ β-VAE loss function

$\boldsymbol{f}^{\boldsymbol{\theta}}$ decoder

$\boldsymbol{g}^{\boldsymbol{\theta}}$ inverse decoder

$\widehat{\boldsymbol{\phi}}$ optimal parameters of the variational posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$

$p_0(\boldsymbol{z})$ latent prior distribution

$p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ true posterior distribution of the decoded samples of the VAE, mapping $\boldsymbol{x} \mapsto \boldsymbol{z}$, parametrized by $\boldsymbol{\theta}$

$p_{\boldsymbol{\theta}}(\boldsymbol{x})$ marginal likelihood

$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ conditional distribution of the decoded samples of the VAE, mapping $\boldsymbol{z} \mapsto \boldsymbol{x}$, parametrized by $\boldsymbol{\theta}$

$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ variational posterior of the VAE, mapping $\boldsymbol{x} \mapsto \boldsymbol{z}$ parametrized by $\boldsymbol{\phi}$

$q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})$ optimal variational posterior of the VAE, mapping $\boldsymbol{x} \mapsto \boldsymbol{z}$ parametrized by $\boldsymbol{\phi}$

$\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})$ optimal mean of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ in dimension k

$\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})$ mean of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ in dimension k

$\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2$ optimal variance of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ in dimension k

$\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2$ variance of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ in dimension k

$g^{\boldsymbol{\theta}}$ inverse decoder component

$\mathbf{H}$ Hessian matrix

$\mathbf{I}_d$ d-dimensional identity matrix

$\mathbf{J}$ Jacobian matrix

$\boldsymbol{\Sigma}$ covariance matrix

$\boldsymbol{x}$ observation vector

$\boldsymbol{z}$ latent vector

$\mathcal{X}$ observation space

d dimensionality of the observation space $\mathcal{X}$

z latent single component