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], **and Michel Besserve**[†2]

[1]University of Tübingen, Germany
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[3]University of Cambridge, Cambridge, United Kingdom
[4]Amazon Web Services, Tübingen, Germany
{patrik.reizinger,jack.brady,wieland.brendel}@uni-tuebingen.de,
{luigi.gresele,jvk,bs,gmartius,besserve}@tue.mpg.de, zietld@amazon.de

Abstract

Variational autoencoders (VAEs) are a popular framework for modeling complex data distributions; they can be efficiently trained via variational inference by maximizing the evidence lower bound (ELBO), at the expense of a gap to the exact (log-)marginal likelihood. While VAEs are commonly used for disentangled representation learning, it is unclear why ELBO maximization would yield such representations, since unregularized maximum likelihood estimation generally cannot invert the data-generating process without additional assumptions. Yet, VAEs often succeed at this task. We seek to elucidate this apparent paradox 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 refer to as *self-consistency*. Leveraging self-consistency, we show that the ELBO converges to a regularized log-likelihood. This allows VAEs to perform what has recently been termed independent mechanism analysis (IMA): it adds an inductive bias towards decoders with column-orthogonal Jacobians, which helps recovering the true latent factors. The gap between ELBO and log-likelihood is therefore welcome, since it bears unanticipated benefits for nonlinear representation learning. In experiments on synthetic and image data, we show that VAEs uncover the true latent factors when the data generating process satisfies the IMA assumption.

1 Introduction

Latent Variable Models (LVMs) allow to effectively approximate a complex data distribution and to sample from it [3, 48]. Deep LVMs employ a neural network (the *decoder* or *generator*) to parameterize the conditional distribution of the observations given latent variables, which are typically assumed to be independent. However, Maximum Likelihood Estimation (MLE) of the model parameters is computationally intractable. In *Variational Autoencoders (VAEs)* [35, 56], the exact log-likelihood is substituted with a tractable lower bound, the evidence lower bound (ELBO). This objective introduces an approximate posterior of the latents given the observations (the *encoder*) from a suitable variational distribution whose mean and covariance are parametrized by neural networks. The encoder is introduced to efficiently train a deep LVM: however, it is not explicitly designed to extract useful representations [17, 58].

[*]Equal contribution. Code available at: github.com/rpatrik96/ima-vae
[†]Senior author

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

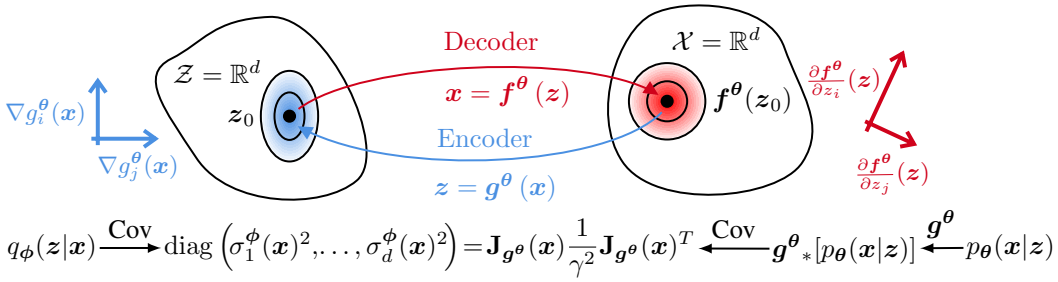

Figure 1: **Modeling choices in VAEs promote *Independent Mechanism Analysis (*IMA*)* [23].** We assume a Gaussian VAE (3), and prove that in the near-deterministic regime the mean encoder approximatetely inverts the mean decoder, $\boldsymbol{g}^{\boldsymbol{\theta}} \approx \boldsymbol{f}^{\boldsymbol{\theta}^{-1}}$ (*self-consistency*, Prop. 1). **Bottom:** Closing the gap requires matching the covariances of the variational (LHS, $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$) and the true posterior (RHS, approximated by $\boldsymbol{g}^{\boldsymbol{\theta}}{}_*[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})]$, cf. § 3.2 for details). Under self-consistency, an encoder with diagonal covariance enforces a row-orthogonal encoder Jacobian $\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{x})$—or equivalently, a column-orthogonal decoder Jacobian $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})$. This regularization was termed Independent Mechanism Analysis (IMA) [23] and shown to be beneficial for learning the true latent factors. The connection elucidates unintended benefits of using the ELBO for representation learning.

Nonetheless, VAEs and their variants are widely used in representation learning [25, 1], where they often recover semantically meaningful representations [39, 8, 34, 5]. Our understanding of this empirical success is still incomplete, since (deep) LVMs with independent latents are nonidentifiable from i.i.d. data [29, 42]; different models fitting the data equally well may yield arbitrarily different representations, thus rendering the recovery of a ground truth generative model impossible. While auxiliary variables, weak supervision [28, 31, 21, 43, 72, 19], or specific model constraints [29, 67, 68, 26, 23] can help identifiability, the mechanism through which the ELBO may enforce a useful inductive bias remains unclear, despite recent efforts [5, 57, 38, 15, 71].

In this work, we investigate the benefits of optimizing the ELBO for representation learning by analyzing VAEs in a *near-deterministic* limit for the conditional distribution parametrized by the nonlinear decoder. Our first result concerns the encoder's optimality in this regime. Previous works relied on the intuitive assumption that the encoder inverts the decoder in the optimum [50, 38, 71]; 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 previous work [50]. The regularization term allows VAEs to perform what has been termed Independent Mechanism Analysis (IMA) [23]: it encourages column orthogonality of the decoder's Jacobian. This generalizes previous findings based on linearizations or approximations of the ELBO [57, 44, 38], and allows us to characterize the gap w.r.t. the log-likelihood in the deterministic limit. Our results elucidate the gap between ELBO and exact log-likelihood as a possible mechanism through which the ELBO implements a useful inductive bias. Unlike the unregularized log-likelihood, the IMA-regularized objective can help invert the data generating process under suitable assumptions [23]. We verify this by training VAEs in experiments on synthetic and image data, showing that they can recover the ground truth factors when the IMA assumptions are met.

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

- we characterize and prove *self-consistency* of VAEs in the near-deterministic regime (i.e., when the decoder variance tends to zero), justifying its usage in previous works (§ 3.1);
- we show that under self-consistency, the ELBO converges to a regularized log-likelihood (§ 3.2), and discuss its possible role as a useful inductive bias in representation learning;
- we test the applicability of our theoretical results in experiments on synthetic and image data, and show that VAEs recover the true latent factors when the IMA assumptions are met (§ 4).

2 Background

We will connect two unsupervised learning objectives: the ELBO in VAEs and the IMA-regularized log-likelihood. Both stem from LVMs with latent variables $\boldsymbol{z}$ distributed according to a *prior* $p_0(\boldsymbol{z})$, and a mapping from $\boldsymbol{z}$ to observations $\boldsymbol{x}$ given by a conditional generative model $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$.

Variational Autoencoders. Optimizing the data likelihood $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ in deep LVMs—i.e., finding decoder parameters $\boldsymbol{\theta}$ maximizing $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p_0(\boldsymbol{z})d\boldsymbol{z}$—is intractable in general, so approximate objectives are required. Variational approximations [63] replace the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ by an approximate one, called the *variational posterior* $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$, which is a stochastic mapping $\boldsymbol{x} \mapsto \boldsymbol{z}$ with parameters $\boldsymbol{\phi}$. This allows to evaluate a tractable evidence lower bound (ELBO) [35, 56] of the model's log-likelihood that can be defined as

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

The two terms in (1) are sometimes interpreted as a reconstruction term measuring the sample quality of the decoder and a regularizer—the Kullback-Leibler Divergence (KL) between the prior and the encoder [36]. The variational approximation trades off computational efficiency with a difference w.r.t. the exact log-likelihood, which is expressed alternatively as (see [17, 36] and Appx. A)

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

where the KL between variational and true posteriors characterizes the *gap*: if the variational family of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ does not include $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$, the ELBO will be strictly smaller than $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$.

VAEs [35] rely on the variational approximation in (1) to train deep LVMs where neural networks parametrize the *encoder* $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and the *decoder* $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$. A common modeling choice constrains the variational family of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ to a factorized Gaussian with posterior means $\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})$ and variances $\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2$ for the kth factor $z_k|\boldsymbol{x}$, and with a diagonal covariance $\boldsymbol{\Sigma}_{\boldsymbol{z}|\boldsymbol{x}}^{\boldsymbol{\phi}}$; and the decoder to a factorized Gaussian, conditional on $\boldsymbol{z}$, with mean $\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})$ and an isotropic covariance in d dimensions,

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

The deterministic limit of VAEs. The stochasticity of VAEs makes it nontrivial to relate them to generative models with deterministic decoders such as Independent Component Analysis (see paragraph below), though postulating a deterministic regime (where the decoder precision γ^2 becomes infinite) is possible. Interestingly, Nielsen et al. [50] explored this deterministic limit and argued that *deterministic* VAEs optimize an exact log-likelihood, similar to normalizing flows [55, 51]. Normalizing flows model arbitrarily complex distributions with a simple base distribution $p_0(\boldsymbol{z})$ and a juxtaposition of *deterministic and invertible* transformations $\boldsymbol{f}^{\boldsymbol{\theta}}$ through the change of variables[3]

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \log p_0(\boldsymbol{z}) - \log\left|\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})\right|. \tag{4}$$

The comparison is nontrivial, since VAEs contain an encoder and a decoder, whereas normalizing flows consist of a single architecture. Nielsen et al. [50] made this analogy by resorting to what we call a *self-consistency assumption*, stating that the VAE encoder inverts the decoder. We define self-consistency in the *near-deterministic* regime: as the decoder variance goes to zero, i.e. $\gamma \to +\infty$.

Definition 1 ((Near-deterministic) self-consistency)**.** *For a fixed $\boldsymbol{\theta}$, assume that mean decoder $\boldsymbol{f}^{\boldsymbol{\theta}}$ is invertible with inverse $\boldsymbol{g}^{\boldsymbol{\theta}}$, and that a map associates each choice of decoder parameters and observation $(\boldsymbol{\theta}, \gamma, \boldsymbol{x})$ to an encoder parameter $(\boldsymbol{\theta}, \gamma, \boldsymbol{x}) \mapsto \widehat{\boldsymbol{\phi}}(\boldsymbol{\theta}, \gamma, \boldsymbol{x})$, we say the VAE is self-consistent whenever*

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) \to \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) \quad \textit{and} \quad \boldsymbol{\sigma}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 \to \mathbf{0}\ \textit{, as } \gamma \to +\infty\,. \tag{5}$$

The encoder parameter map reflects the choice of a particular encoder model for each $(\boldsymbol{\theta}, \gamma)$ pair:[4] in § 3.1, we study this problem by introducing and justifying a particular choice for $\widehat{\boldsymbol{\phi}}$ (see also § 5). This self-consistency assumption appears central to deterministic claims [50, 38], but has not yet been proven. In particular, Nielsen et al. [50] assume that taking the deterministic limit is well-behaved. However, VAEs' *near-deterministic* properties have not been investigated analytically.

Identifiability, ICA, and IMA. Independent Component Analysis (ICA) [9, 30] models observations as the *mixing* of a latent vector $\boldsymbol{z}$ with independent components through a deterministic function $\boldsymbol{f}$, i.e., $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{z}), p_0(\boldsymbol{z}) = \prod_i p_0(z_i)$.[5] In ICA the focus is on defining conditions under which the original

[3] note that in normalizing flows the change of variables is usually expressed in terms of $\boldsymbol{g}^{\boldsymbol{\theta}} = \boldsymbol{f}^{\boldsymbol{\theta}\,-1}$

[4] both the ELBO and $\widehat{\boldsymbol{\phi}}$ depends on the decoder precision γ: we will omit this in the following for simplicity

[5] the conditional distribution $p(\boldsymbol{x}|\boldsymbol{z})$ is therefore degenerate

latent variables can be recovered from observations—i.e., the model is identifiable by design [31]. The goal is to learn an unmixing $\boldsymbol{g}^{\boldsymbol{\theta}}$ such that the recovered components $\boldsymbol{y} = \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})$ are estimates of the true ones up to some ambiguities (e.g., permutation and element-wise nonlinear transformations). Unfortunately, the nonlinear problem is nonidentifiable without further constraints [16, 29]: any two observationally equivalent models can yield components which are arbitrarily entangled, thus making recovery of the ground truth factors impossible. This is typically shown by suitably constructed counterexamples [29, 42], and it was argued to imply impossibility statements for unsupervised disentanglement [42, 65]. Identifiability can be recovered when *auxiliary* variables [31, 21, 33, 19] are available, or exploiting a temporal structure in the data [28, 24].

Restrictions on the mixing function class (e.g., linear [9]) are another possibility to recover identifiability [29, 67]. Recently, Gresele et al. [23] proposed restricting the function class by taking inspiration from the *principle of independent causal mechanisms* [52], 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 $\partial \boldsymbol{f}^{\boldsymbol{\theta}} / \partial z_k$, and their non-statistical independence amounts to an orthogonality condition. While full identifiability has not been proved for this model class, it was shown to rule out classical families of spurious solutions used as counterexamples to identifiability of unconstrained non-linear ICA [23, 4]. Moroever, Buchholz et al. [4] further demonstrated local identifiability of this function class. Also, IMA constraints were empirically shown [23, 62] to help recover the ground truth through regularization of the log-likelihood in (4) with an objective $\mathcal{L}_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{z}) := \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \lambda \cdot c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{z})$, where $\lambda > 0$ and the regularization term $c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{z})$ and its expectation $C_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, p_0)$ are given by

$$c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{z}) = \sum_{k=1}^{d} \log \left\| \frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}(\boldsymbol{z}) \right\| - \log \left| \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z}) \right|; \quad C_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, p_0) = \mathbb{E}_{p_0(\boldsymbol{z})} \Big[c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{z}) \Big], \tag{6}$$

and termed *local* (resp. *global*) IMA contrast. When $\boldsymbol{f}^{\boldsymbol{\theta}}$ is in the IMA function class (i.e., $C_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, p_0)$ vanishes), the objective is equal to the log-likelihood; otherwise, it lower bounds it.

3 Theory

Our theoretical analysis assumes that all the model's defining densities ($p_0(\boldsymbol{z})$, $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$) are factorized. We also assume 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 $\boldsymbol{\theta} \mapsto \boldsymbol{f}^{\boldsymbol{\theta}}$, and the mean and standard deviation encoder classes, $\boldsymbol{\phi} \mapsto \boldsymbol{\mu}^{\boldsymbol{\phi}}$ and $\boldsymbol{\phi} \mapsto \boldsymbol{\sigma}^{\boldsymbol{\phi}}$ s.t.*

- *(i) $p_0(\boldsymbol{z}) \sim \prod_k m(z_k)$, with m being smooth and fully supported on $\mathbb{R}$, having a 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 $\boldsymbol{\theta}$, $f^{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^d$ is a bijection with inverse $\boldsymbol{g}^{\boldsymbol{\theta}}$, 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 [50] (which additionally assumes quasi-deterministic encoders), and in [44, 38]. Unlike Nielsen et al. [50], 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 [50] (see also § 5).

Due to considering $\gamma \to +\infty$, results are stated in the following "big-O" notation for an integer p:

$$f(\boldsymbol{x}, \gamma) = g(\boldsymbol{x}, \gamma) + O_{\gamma \to +\infty}(1/\gamma^p) \iff \gamma^p \| f(\boldsymbol{x}, \gamma) - g(\boldsymbol{x}, \gamma) \| \text{ is bounded as } \gamma \to +\infty\,.$$

3.1 Self-consistency

In this section, we will prove a *self-consistency* result in the near-deterministic regime. This rests on 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

$$\widehat{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{\theta}) \in \arg\max_{\boldsymbol{\phi}} \text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \arg\min_{\boldsymbol{\phi}} \text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] . \tag{7}$$

We call *self-consistent* ELBO the resulting achieved value, denoted as

$$\text{ELBO}^*(\boldsymbol{x}; \boldsymbol{\theta}) = \text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \widehat{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{\theta})) . \tag{8}$$

The expression in (7) corresponds to a problem of *information projection* [10, 48] 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 [11], 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. 7-8 in Appx. C.2), we obtain the following self-consistency result.

Proposition 1. *[Self-consistency of near-deterministic VAEs] Under Assumption 1, for all $\boldsymbol{x}$, $\boldsymbol{\theta}$, as $\gamma \to +\infty$, there exists at least one global minimum solution of (7). These solutions satisfy*

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) = \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) + O(1/\gamma) \quad \textit{and} \quad \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = O(1/\gamma^2) \,, \textit{ for all } k \,. \tag{9}$$

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.

Let us now consider the relevance of this result for training VAEs, i.e., maximizing the expectation of the ELBO for an observed distribution $p(\boldsymbol{x})$. While maximization *only* w.r.t. $\boldsymbol{\phi}$ in (7) does not match common practice—which is learning $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ *jointly*—it models this process in the limit of large-capacity encoders. Indeed, in this case, (7) can be solved for each $\boldsymbol{x}$ as a separate learning problem, which entails that the following inequality is satisfied for any parameter choice

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi})\right] &= \int p(\boldsymbol{x})\text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi}) d\boldsymbol{x} \\ &\leq \int p(\boldsymbol{x})\text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \widehat{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{\theta})) d\boldsymbol{x} =: \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\text{ELBO}^*(\boldsymbol{x}; \boldsymbol{\theta})\right] .\end{aligned} \tag{10}$$

The joint optimization of encoder and decoder parameters thus reduces to optimizing the subset of pairs $(\boldsymbol{\theta}, \widehat{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{\theta}))$, and is equivalent to optimizing the expected self-consistent ELBO, that is

$$\min_{\boldsymbol{\theta}, \boldsymbol{\phi}} \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\text{ELBO}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi})\right] \iff \min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\text{ELBO}^*(\boldsymbol{x}; \boldsymbol{\theta})\right] \tag{11}$$

This problem reduction is aligned with the original purpose of the ELBO: building a tractable but optimal likelihood approximation. Namely, (i) ELBO* depends on the same parameters as the likelihood ($\boldsymbol{x}$, γ and $\boldsymbol{\theta}$), (ii) its gap $\text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right]$ is minimal. Clearly, the problem reduction of (11) is more informative only at the critical points of the VAE loss, since it allows us to compare the optimality of different decoders and Prop. 1 allows us to study this question for near-deterministic decoders.

3.2 Self-consistent ELBO, IMA-regularized log-likelihood and identifiability of VAEs

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.

Theorem 1. *[VAEs with a near-deterministic decoder approximate the* IMA *objective] Under Assumption 1, the variational posterior satisfies*

$$\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = \left(-\frac{d^2 \log p_0}{dz_k^2}(g_k^{\boldsymbol{\theta}}(\boldsymbol{x})) + \gamma^2 \left\| \left[\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]_{:k} \right\|^2\right)^{-1} + O(1/\gamma^3) \,, \tag{12}$$

and the self-consistent ELBO (19) *approximates the* IMA*-regularized log-likelihood* (6)*:*

$$\text{ELBO}^*(\boldsymbol{x}; \boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})) + O_{\gamma\to\infty}\left(1/\gamma^2\right) . \tag{13}$$

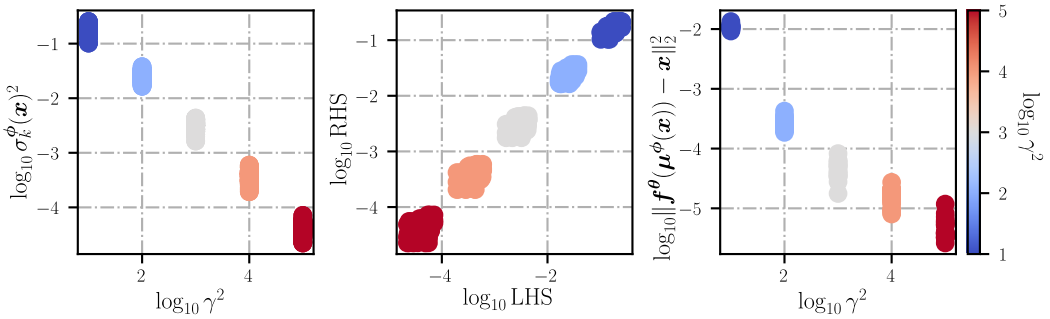

Figure 2: Self-consistency (Prop. 1) in VAE training, on a log-log plot, cf. 4.1 for details. **Left**: convergence of $\sigma_k^{\widehat{\phi}}(\boldsymbol{x})^2$ to 0; **Center:** connecting $\sigma_k^{\widehat{\phi}}(\boldsymbol{x})^2$, γ^2, and the column norms of the decoder Jacobian via LHS and RHS of (46); **Right:** convergence of $\boldsymbol{\mu}^{\widehat{\phi}}(\boldsymbol{x})$ to $\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})$

Proof is in Appx. B. Below, we provide a qualitative argument on 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 $\boldsymbol{g}^{\boldsymbol{\theta}}{}_*[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})]$ (see Appx. A for more 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 $1/\gamma^2\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}^T$. As the sufficient statistics of a Gaussian are given by its mean and covariance, the structure of the posterior covariance $\boldsymbol{\Sigma}_{\boldsymbol{z}|\boldsymbol{x}}^{\boldsymbol{\phi}}$ (which is by design diagonal, cf. (3)) is crucial for minimizing the gap in (2). Practically, this implies that 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 $\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2 = 1/\gamma^2\|\left[\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z}_0)\right]_{:k}\|^{-2}$. The self-consistent ELBO therefore converges to the IMA-regularized log-likelihood (6).

Our argument indicates that minimizing the gap between the ELBO and the log-likelihood encourages column-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 [44].

Implications for identifiability of VAEs. While previous works argued that the VAE objective favors decoders with a column-orthogonal Jacobian [57, 38], they did not exactly characterize how: our result shows that the self-consistent ELBO tends to a regularized log-likelihood, where the regularization term c_{IMA} explicitly enforces this (soft) constraint. Thus, it possibly explains why VAEs are successful in learning disentangled representations: namely, the IMA function class provably rules out certain spurious solutions for nonlinear ICA [23], and the IMA-regularized log-likelihood was empirically shown to be beneficial in recovering the true latent factors. Thus, we speak about *embracing the gap*, as its functional form equips VAEs with a useful inductive bias. While the IMA function class has not yet been shown to be identifiabile, such results exist for special cases such as conformal maps ($d = 2$ [29], generalized by the very recent work in [4]), isometries [26] and for closely-related unsupervised nonlinear ICA models [69]. In the following, we empirically corroborate that VAEs: 1) recover the ground truth sources when the mixing satisfies IMA, and thereby 2) achieve unsupervised disentanglement.

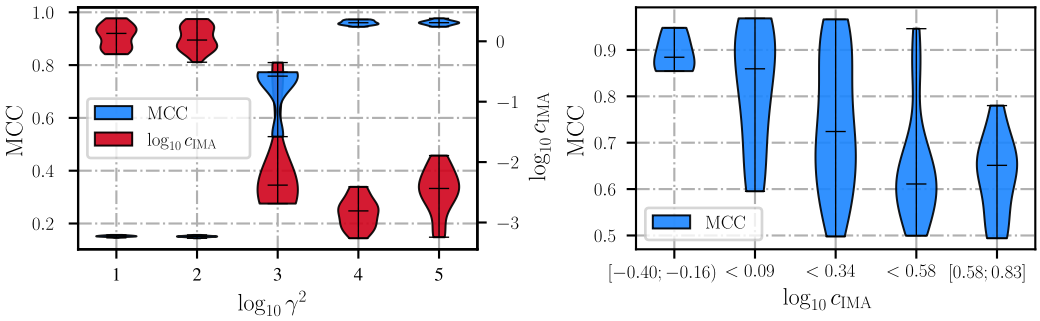

Figure 3: **Left:** c_{IMA} and Mean Correlation Coefficient (MCC) for 3-dimensional Möbius mixings **Right:** MCC depending on the *volume-preserving linear map's* c_{IMA} ($\gamma^2 = 1\text{e}5$)

4 Experiments

Our experiments serve three purposes: 1) demonstrating that self-consistency holds in practice (§ 4.1); 2) showing the relationship of the self-consistent ELBO*, the IMA-regularized and unregularized log-likelihood objectives (§ 4.2); and 3) providing empirical evidence that the connection to the IMA function class in VAEs can lead to success in learning disentangled representations (§ 4.3). More details are provided in Appx. F.

4.1 Self-consistency in practical conditions

Experimental setup. We use a 3-layer Multi-Layer Perceptron (MLP) with smooth Leaky ReLU nonlinearities [22] and orthogonal weight matrices—which intentionally does not belong to the IMA class, as our results are more general. 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 $\gamma^2 \in \{1\text{e}1; 1\text{e}2; 1\text{e}3; 1\text{e}4; 1\text{e}5\}$.
Results. Fig. 2 summarizes our results, featuring the *logarithms* on each axes. 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 (9). The **center** plot shows that the expression for $\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2$ corresponds to (46) in the optimum of the ELBO by comparing both sides of the equation. 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 (see § 5). 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 $\boldsymbol{x}$.

4.2 Relationship between ELBO*, IMA-regularized, and unregularized log-likelihoods

Experimental setup. We use an MLP $\boldsymbol{f}^{\boldsymbol{\theta}}$ with square upper-triangular weight matrices and invertible element-wise nonlinearities to construct a mixing not in the IMA class [23] and fix the VAE decoder to the ground truth such that (4) gives the true data log-likelihood. This way, we ensure that the unregularized and IMA-regularized log-likelihoods differ and make the claim of Nielsen et al. [50] comparable to ours. With a fixed decoder, the ELBO* depends only on $\boldsymbol{\phi}$, therefore we only train the encoder with γ^2 values from $[1\text{e}1; 1\text{e}5]$ (5 seeds each).
Results. Fig. 4 compares the difference of the estimate of ELBO* and the unregularized/IMA-regularized log-likelihoods after convergence over the whole dataset. As the decoder and the data are fixed, $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$ and C_{IMA} will not change during training, only the ELBO* does. The figure shows that as $\gamma \to +\infty$, ELBO* approaches $\mathcal{L}_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{z})$, as predicted by Thm. 1, and not $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$, as stated in [50]—the difference is C_{IMA}.

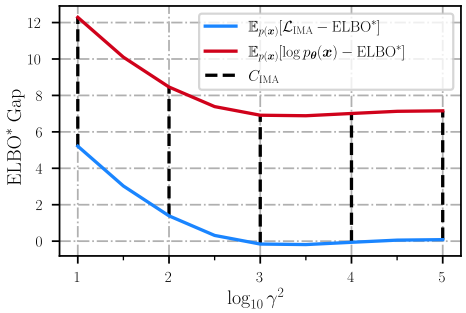

Figure 4: Comparison of the ELBO*, the IMA-regularized and unregularized log-likelihoods over different γ^2. Error bars are omitted as they are orders of magnitudes smaller

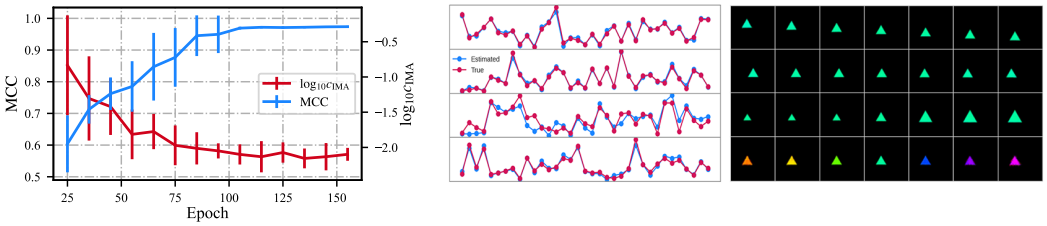

Figure 5: **Left:** c_{IMA} and MCC for Sprites [66] during training $(\gamma^2 = 1)$; **Center:** true and estimated latent factors for the best trained VAE on Sprites; **Right:** the corresponding latent interpolations and MCC values (from top to bottom): y- (0.989), x-position (0.996), scale (0.933), and color (0.989)

4.3 Connecting the IMA principle, γ^2, and disentanglement

Experimental setup (synthetic). We use 3-dimensional conformal mixings (i.e., the Möbius transform [53]) from the IMA class with *uniform* ground-truth and prior distributions. Our results quantify the relationship of the decoder Jacobian's IMA-contrast and identifiability with MCC [27] and show how this translates to disentanglement—we note that MCC was already used to quantify disentanglement [72, 37]. 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 make $c_{\text{IMA}} \neq 0$. Other parameters are the same as in § 4.1, with the exception of picking the best $\gamma^2 = 1\text{e}5$.
Results (synthetic). The **left** of Fig. 3 empirically demonstrates the benefits of optimizing the IMA-regularized log-likelihood. By increasing γ^2, MCC increases, while c_{IMA} decreases, suggesting that VAEs in the near-deterministic regime encourage disentanglement by enforcing the IMA principle. The **right** plot shows that when the mixing is outside the IMA class, MCC decreases, corroborating the benefits of such mixings for disentanglement.
Experimental setup (image). 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 [66] with a Beta ground truth distribution. Similar to [32], 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) [57, 37]. Our choice is motivated by [26, 18] showing that this data-generating process may approximately satisfy the IMA principle.
Results (image). The **left** of Fig. 5 indicates that VAEs can learn the true latent factors and MCC is *anticorrelated* with c_{IMA}, reinforcing the hypothesis that the data-generating process belongs to the IMA class. The **center** plot compares estimated and true latent factors from the best model (scaling and permutation indeterminacies are removed), whereas the **right** plot shows the corresponding latent interpolations—thus, connecting identifiability (measured by MCC) to disentanglement.

5 Limitations

The near-deterministic regime. Our theory relies on $\gamma \to +\infty$; this is the regime where posterior collapse may be avoided [44], and where calculating the reconstruction loss may be possible even without sampling [38]. However, in practice it may be unclear when γ^2 is large enough. This seems to be problem-dependent [57, 44], and possibly tied to the covariance of the observations [60, 59]. Moreover, 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 (19). 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.

Dimensionality.[**Patrik:** Identifiability for different-dimensional conformal maps is proven in [4]] The setting in § 3 requires equal dimensionality for observations $\boldsymbol{x}$ and latents $\boldsymbol{z}$, in line with work on normalizing flows [51] and nonlinear ICA [28, 31, 24] (but see, e.g., [33]). For high-dimensional images, however, it is often assumed that $\boldsymbol{x}$ lives on a lower-dimensional manifold embedded in a higher-dimensional space, where the dimensionality of $\boldsymbol{x}$ is greater than $\boldsymbol{z}$ [13]. While our theoretical results do not cover this case, we observe empirically in Fig. 5 that the predictions of our theory remain accurate when observations are high-dimensional images. Extending our theory to this setting could leverage ideas explored in, e.g., [13, 12, 7] and is left for future work.

The ELBO, the self-consistent ELBO, and amortized inference. There are in principle multiple ways to obtain self-consistency (Defn. 1). Notably, one could simply force the variational mean and variance encoder maps to behave this way; unlike [38], we model the actual behavior of VAEs trained under ELBO maximization, and obtain self-consistency as a result. For this, we assume that the optimal encoder, which minimizes the gap between ELBO and log-likelihood, can be learned. This is not guaranteed in general, since it requires universal approximation capability of the encoder. On the other hand, (19) requires *unamortized* inference to introduce ELBO*, which does not depend on $\boldsymbol{\phi}$. As in practice amortized inference may be used to efficiently estimate a single set of $\boldsymbol{\phi}$ for all $\boldsymbol{x}$ [61], it can lead to a suboptimal gap to the log-likelihood and discrepancies with our theoretical predictions.

6 Discussion

On disentanglement in unsupervised VAEs. It is widely believed that unsupervised VAEs cannot learn disentangled representations [42, 33], motivating work on models with, e.g., conditional priors [33] or sparse decoding [47]. We show that under certain assumptions, ELBO optimization can implement useful inductive biases for representation learning, yielding disentangled representations in unsupervised VAEs. However, while our results are formulated for VAEs, some of the most successful models at disentanglement are modifications thereof—e.g., β-VAEs [25, 5], with an additional parameter β multiplying the KL in (1). While they deviate from the information projection setting considered in § 3.1, their objectives are equivalent to the ELBO in a sense described in Appx. A.3, which allows us to derive convergence to the IMA-regularized likelihood objective for $\gamma/\sqrt{\beta} \to +\infty$. This encompasses the deterministic limit, and also the setting $\beta \to 0$ with constant γ described in [38]. Whether this theoretical regime matches common practice remains an open question. Overall, we stress that we uncover *one* possible mechanism through which VAEs may achieve disentanglement. By connecting to IMA [23], we discuss implications on recovering the ground truth under suitable assumptions, extending uniqueness results presented in [38]. We speculate that our success in disentanglement is probably due to selecting data sets where the mixing is in the IMA class (cf. [26, 18]), which presumably was not the case in [42].

Characterizing the ELBO gap for nonlinear models. Thm. 1 characterizes the gap between ELBO and true log-likelihood for nonlinear VAEs, and extends the linear analysis of Lucas et al. [44] and the results of Dai et al. [14] in the affine case; we also empirically characterize the gap in the deterministic limit in § 4.2. An unanticipated consequence of this result is that—consistent with [44]—VAEs optimize the IMA-regularized log-likelihood in the near-deterministic limit, and not the unregularized one, as stated in [50].

Extensions to related work. Several papers discuss the (near-)deterministic regime [50, 57, 38, 13]. For example, Nielsen et al. [50] postulate a deterministic VAE with the encoder inverting the decoder. Also Kumar and Poole [38] work in that regime, but without justifying the relationship between the encoder and decoder. Although they show that the choice of $p_0(\boldsymbol{z})$ and $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ influences uniqueness (by, e.g., ruling out rotations), this does not imply recovering the true latents. Our approach formalizes (Defn. 1), proves (Prop. 1), and demonstrates the practical feasibility of (§ 4) the near-deterministic regime. To the best of our knowledge, all previous work relied on the linear case [44] or a (linear) approximation and the evaluation of the ELBO *around a point* to show the inductive bias on the decoder Jacobian. However, our main result (Thm. 1) yields a nonlinear equation where the decoder Jacobian can be evaluated at *any point* and is equipped with a convergence bound. Moreover, the consistency of VAE estimation for identifiable models [33] requires guarantees on $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$; our result helps proving these. Dai and Wipf [13] use a non-factorized Gaussian variational posterior and prove in their Thm. 2 (including the $\dim \boldsymbol{x} = \dim \boldsymbol{z}$ case) that in the deterministic limit their κ-simple VAE can fit perfectly arbitrary observed data (barring few assumptions), while the ELBO gap tends to zero. In contrast, we use a factorized variational posterior; this prevents the ELBO gap to vanish in the deterministic limit, except in the special case of a decoder mean in the IMA class fitting the data perfectly. Dai and Wipf [13] take the limit of $\gamma \to +\infty$ (here using γ as the square root of the decoder precision and not the decoder variance as used in [13]) to relate encoder and decoder properties in this limit in their Thm. 5, similarly to Prop. 1. In contrast to our nonlinear analysis, this is derived when optimizing w.r.t. both encoder and decoder parameters, and with a non-factorized encoder assumption, leading to fundamentally different behavior of the solutions in the deterministic limit. The work done by Sliwa et al. [62], simultaneously to ours, showcases an extensive empirical study highlighting that the IMA contrast allows distinguishing true and spurious solutions for a broad range of cases and outperforms standard regularizers such as weight decay. We discuss extended connections to the literature in Appx. D and Appx. E.

Covariance structure and IMA. We have shown that specific choices for encoder and decoder covariances regularize the decoder Jacobian, such that closing the ELBO gap constrains the decoder to belong to the IMA class. Following our intuition (Fig. 1), assuming factorized $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and isotropic $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$, IMA holds only for the *decoder*; since in the other direction the pushforward of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ through $\boldsymbol{f}^{\boldsymbol{\theta}}$ has covariance $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})\,\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}|\boldsymbol{x}}\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})^T$, which cannot be simplified in the general case. Additionally, we conjecture that assuming an isotropic encoder would constrain IMA to hold in both encoding and decoding directions (as both $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})$ and $\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{x})$ need to be column-orthogonal), such that the resulting decoder mean is constrained to have orthogonal columns of equal norms, which is a defining property of conformal maps [4]. On the other hand, we conjecture that if the observation model is not isotropic, but the encoder mode is, IMA would only be enforced for the mean *encoder* Jacobian, converging to the inverse decoder mean in the deterministic limit.

Implications for recovering the true latent factors using unsupervised VAEs. Convergence of the ELBO to the IMA-regularized log-likelihood suggests that unsupervised VAEs may recover the true factors of variation according to current identifiability results of the IMA class [4]. This is based on the following reasoning: *If the ground truth generative model belongs to the IMA class, unsupervised learning of the model with an infinite capacity VAE will, in the deterministic limit, ensure a solution that perfectly fits the data and whose decoder mean is also in the IMA class (by joint optimization of both the likelihood and the regularization term). Identifiability of the IMA class implies that the VAE will learn the true decoder (up to acceptable ambiguities); then, since self-consistency guarantees that the encoder inverts the decoder, the encoder infers the ground truth generative factors associated to observations.* Although strict identifiability for all functions in the IMA class remains to be proven, three concurrent papers provide guarantees that go towards identifiability: Leemann et al. [41] proves identifiability for a subset of the IMA class in the context of concept discovery; Zheng et al. [70] shows identifiability of nonlinear ICA by assuming a specific sparsity structure of the decoder Jacobian (called *structural sparsity*); whereas Buchholz et al. [4] introduce the concept of *local identifiability* and proves that IMA is locally identifiable.

Moreover, as mentioned in the above paragraph, we suspect that closing the ELBO gap with an isotropic encoder (while the encoder in Thm. 1 is only constrained to have diagonal covariance) constrains the decoder to be a conformal map. This is an interesting constraint, as nonlinear ICA with conformal mixings are identifiable: the two-dimensional case was first addressed with some additional constraints in [29], while the general case (in arbitrary dimension) was shown to rule out certain spurious solutions for conformal mixings [23], and finally proven to be identifiable by Buchholz et al. [4] in concurrent work. Hence, we conjecture that given a ground truth generative model with a conformal map from latent to observation space, and an unsupervised VAEs with isotropic Gaussian encoders and decoders, the true latent factors can be recovered.

Conclusion. We provide a theoretical justification for 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 ELBO and true log-likelihood and reason about its role as an inductive bias for representation learning in nonlinear VAEs. We characterize a set of assumptions under which unsupervised VAEs can be expected to disentangle and we demonstrate this behavior in experiments on synthetic and image data.

Acknowledgments and Disclosure of Funding

The authors thank the anonymous reviewers for their suggestions. 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. Wieland Brendel acknowledges financial support via an Emmy Noether Grant funded by the German Research Foundation (DFG) under grant no. BR 6382/1-1. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Dominik Zietlow and Patrik Reizinger. Patrik Reizinger acknowledges his membership in the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program.

References

[1] Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken ELBO. 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 159–168. PMLR, 2018. 2

[2] 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. 22, 23, 41

[3] Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 4. Springer New York, 2006. doi: 10.1007/978-0-387-45528-0. URL https://doi.org/10.1007/978-0-387-45528-0. 1

[4] Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function Classes for Identifiable Nonlinear Independent Component Analysis. *Advances in Neural Information Processing Systems*, 36, 2022. 4, 6, 8, 10

[5] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018. 2, 9

[6] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3), jun 2011. ISSN 0004-5411. doi: 10.1145/1970392.1970395. URL https://doi.org/10.1145/1970392.1970395. 40

[7] Anthony L Caterini, Gabriel Loaiza-Ganem, Geoff Pleiss, and John P Cunningham. Rectangular flows for manifold learning. *Advances in Neural Information Processing Systems*, 34, 2021. 8

[8] 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. 2, 20

[9] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314, 1994. 3, 4

[10] 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. 5

[11] 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. 5

[12] Edmond Cunningham and Madalina Fiterau. A change of variables method for rectangular matrix-vector products. In *International Conference on Artificial Intelligence and Statistics*, pages 2755–2763. PMLR, 2021. 8

[13] Bin Dai and David Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2018. 8, 9, 20, 40

[14] Bin Dai, Yu Wang, John Aston, Gang Hua, and David Wipf. Connections with robust PCA and the role of emergent sparsity in variational autoencoder models. *The Journal of Machine Learning Research*, 19(1):1573–1614, 2018. 9

[15] Bin Dai, Ziyu Wang, and David Wipf. The usual suspects? Reassessing blame for VAE posterior collapse. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 2313–2322. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/dai20c.html. 2

[16] George Darmois. Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, page 231, 1951. 4

[17] Carl Doersch. Tutorial on Variational Autoencoders. *ArXiv preprint*, abs/1606.05908, 2016. 1, 3, 21

[18] 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`. 8, 9, 45

[19] Élisabeth Gassiat, Sylvain Le Corff, and Luc Lehéricy. Deconvolution with unknown noise distribution is possible for multivariate signals. *Ann. Statist.*, 50(1), February 2022. ISSN 0090-5364. doi: 10.1214/21-aos2106. URL `https://doi.org/10.1214/21-aos2106`. 2, 4

[20] 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. 20, 40

[21] Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The Incomplete Rosetta Stone problem: Identifiability results for Multi-view Nonlinear ICA. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 115 of *Proceedings of Machine Learning Research*, pages 217–227. PMLR, July 2019. URL `https://proceedings.mlr.press/v115/gresele20a.html`. 2, 4

[22] 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. 7, 42, 43, 44

[23] 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`. 2, 4, 6, 7, 9, 10, 22, 42

[24] Hermanni Hälvä and Aapo Hyvärinen. Hidden Markov nonlinear ICA: Unsupervised learning from nonstationary time series. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*, pages 939–948. AUAI Press, 2020. 4, 8

[25] 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. 2, 9, 20

[26] Daniella Horan, Eitan Richardson, and Yair Weiss. When is unsupervised disentanglement possible? *Advances in Neural Information Processing Systems*, 34, 2021. 2, 6, 8, 9, 45

[27] 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. 8

[28] Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In Aarti Singh and Xiaojin (Jerry) Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 460–469. PMLR, 2017. 2, 4, 8

[29] 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, 4, 6, 10, 23, 40

[30] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, Inc., New York, May 2001. ISBN 047140540X, 0471221317. doi: 10.1002/0471221317. URL https://doi.org/10.1002/0471221317. 3

[31] Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR, 2019. 2, 4, 8

[32] 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. 8, 45

[33] 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. 4, 8, 9

[34] 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. 2, 20

[35] 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, 3, 20

[36] 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. 3

[37] 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. 8, 45

[38] Abhishek Kumar and Ben Poole. On Implicit Regularization in β-VAEs. In *International Conference on Machine Learning*, pages 5480–5490. PMLR, 2020. 2, 3, 4, 6, 8, 9, 20, 22, 40, 45

[39] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2

[40] 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. 40

[41] Tobias Leemann, Michael Kirchhof, Yao Rong, Enkelejda Kasneci, and Gjergji Kasneci. Disentangling Embedding Spaces with Minimal Distributional Assumptions, June 2022. URL http://arxiv.org/abs/2206.13872. arXiv:2206.13872 [cs, stat]. 10

[42] 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, 4, 9

[43] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6348–6359. PMLR, 2020. 2

[44] 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. 2, 4, 6, 8, 9, 18, 40, 41, 45

[45] 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. 38

[46] Emile Mathieu, Tom Rainforth, N. Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. 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 4402–4412. PMLR, 2019. 40

[47] Gemma E Moran, Dhanya Sridhar, Yixin Wang, and David M Blei. Identifiable variational autoencoders via sparse decoding. *arXiv preprint arXiv:2110.10804*, 2021. 9

[48] Kevin P Murphy. *Machine learning: A probabilistic perspective*. MIT press, 2012. 1, 5

[49] Akira Nakagawa, Keizo Kato, and Taiji Suzuki. Quantitative Understanding of VAE as a Non-linearly Scaled Isometric Embedding. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7916–7926. PMLR, July 2021. ISSN: 2640-3498. 39

[50] 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. 2, 3, 4, 7, 9, 40

[51] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *ArXiv preprint*, abs/1912.02762, 2019. 3, 8

[52] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017. 4

[53] 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`. 8, 42, 44

[54] 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. 42, 44, 45

[55] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1530–1538. JMLR.org, 2015. 3

[56] 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, 3

[57] 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`. 2, 6, 8, 9, 20, 39, 40, 45

[58] Paul K. Rubenstein. Variational Autoencoders are not autoencoders, Jan 2019. URL `http://paulrubenstein.co.uk/variational-autoencoders-are-not-autoencoders/`. 1

[59] 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. 8, 21, 45

[60] Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. 2021. 8, 45

[61] Rui Shu, Hung H Bui, Shengjia Zhao, Mykel J Kochenderfer, and Stefano Ermon. Amortized inference regularization. *Advances in Neural Information Processing Systems*, 31, 2018. 9

[62] Joanna Sliwa, Shubhangi Ghosh, Vincent Stimper, Luigi Gresele, and Bernhard Schölkopf. Probing the Robustness of Independent Mechanism Analysis for Representation Learning, July 2022. URL `http://arxiv.org/abs/2207.06137`. arXiv:2207.06137 [cs, stat]. 4, 9

[63] 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`. 3

[64] 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`. 40

[65] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 4

[66] 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/`. 8, 45

[67] Kun Zhang and Laiwan Chan. Minimal nonlinear distortion principle for nonlinear Independent Component Analysis. *Journal of Machine Learning Research*, 9(Nov):2455–2487, 2008. 2, 4

[68] Kun Zhang and Aapo Hyvarinen. On the Identifiability of the Post-Nonlinear Causal Model. *arXiv:1205.2599 [cs, stat]*, 2012. arXiv: 1205.2599. 2

[69] Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA with unconditional priors. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022. 6

[70] Yujia Zheng, Ignavier Ng, and Kun Zhang. On the Identifiability of Nonlinear ICA: Sparsity and Beyond, June 2022. URL `http://arxiv.org/abs/2206.07751`. arXiv:2206.07751 [cs, stat]. 10

[71] 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. 2, 39

[72] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. 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 12979–12990. PMLR, 2021. 2, 8

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See § 3, § 4, § 6
   (b) Did you describe the limitations of your work? [Yes] See § 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appx. H
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We use no data about human subjects.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See § 3
   (b) Did you include complete proofs of all theoretical results? [Yes] See § 3, Appx. B
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the supplementary material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See § 4, Appx. F
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See § 4
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appx. F
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See § 4 and the supplementary material
   (b) Did you mention the license of the assets? [Yes] See the supplementary material
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We included the code and the experiment logs.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

Table of Contents

A Complementary notes **19**
- A.1 ELBO decompositions . . . 19
- A.2 Justification of the intuition . . . 20
- A.3 A connection between the β parameter of β-VAEs and the decoder precision γ^2 . 20
- A.4 c_{IMA} as the (scaled) left KL measure of diagonality of $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}^T \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}$ . . . 22
- A.5 Assessing the value of c_{IMA} for recovering the true latents . . . 23

B Main Theoretical Results **23**
- B.1 Proof of Proposition 1 . . . 23
- B.2 Proof of Theorem 1 . . . 30

C Auxiliary results **35**
- C.1 Squared norm statistics . . . 35
- C.2 KL divergence bounds . . . 35
- C.3 Taylor formula-based approximations . . . 38
- C.4 Variational posterior variance optimization problem . . . 39

D Related work **39**
- D.1 Implicit inductive biases in the ELBO . . . 39
- D.2 (Near)-deterministic VAEs . . . 40

E Further remarks on the the IMA–VAE connection **41**
- E.1 Linear VAE from Lucas et al. [44] . . . 41

F Experimental details **42**
- F.1 The relationship of weight matrix structures and the IMA function class . . . 42
- F.2 Self-consistency in practical conditions (§ 4.1) . . . 42
- F.3 Relationship between ELBO*, IMA-regularized, and unregularized log-likelihoods (§ 4.2) . . . 43
- F.4 Connecting the IMA principle, γ^2, and disentanglement (§ 4.3) . . . 44
- F.5 Optimality of γ^2 w.r.t. its MLE . . . 45

G Computational resources **45**

H Societal impact **46**

I Notation **46**
- Acronyms . . . 46
- Nomenclature . . . 46

A Complementary notes

A.1 ELBO 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):

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

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

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

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

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

which leads to (1) by rearranging the terms:

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

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

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

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

$$-\text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_0(\boldsymbol{z})\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log(p_0(\boldsymbol{z}))] + \frac{1}{2}\sum_{k=1}^{d} \log \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2 + \kappa_d\,. \tag{14}$$

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

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] = -\frac{\gamma^2}{2}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})\|^2\right] + d\log\gamma - \frac{d}{2}\log(2\pi)\,. \tag{15}$$

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).

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

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

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

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

$$\begin{aligned}\text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] = &-\sum_{k=1}^{d} \log \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + c(\boldsymbol{x}, \gamma) \\ &+ \frac{1}{2}\mathbb{E}_{z\sim q_{\boldsymbol{\phi}}(\cdot|\boldsymbol{x})}\left[\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})\|^2\gamma^2 - 2\sum_{k=1}^{d} \log m(z_k)\right],\end{aligned} \tag{16}$$

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 (16) 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 $\boldsymbol{\mu}^{\boldsymbol{\phi}}$ and $\boldsymbol{\sigma}^{\boldsymbol{\phi}}$. Much of the derivations to obtain our results will revolve around constructing bounds that no longer involve such expectations, but instead only depend on $\boldsymbol{\mu}^{\boldsymbol{\phi}}$ and $\boldsymbol{\sigma}^{\boldsymbol{\phi}}$.

A.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 Appx. B.

As the decoder becomes deterministic, the marginal observed density becomes the pushforward of the latent prior by $\boldsymbol{f}^{\boldsymbol{\theta}}$ [6] 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 $\boldsymbol{x} = \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z}_0)$, we get

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

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

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

Using the isotropic Gaussian decoder assumption, we get

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

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

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

with $\mathbf{G} = \mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z}_0)) = \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z}_0)^{-1}$, 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 $\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z}_0)$ (i.e. replacing the mapping by its Jacobian at that point, $\mathbf{G}$).

A.3 A connection between the β parameter of β-VAEs and the decoder precision γ^2

In the context of disentanglement, a commonly used variant of standard VAEs [35] is the β-VAE [8, 25, 34, 57, 38]. 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 [13, 20, 38, 57]. The β-VAE objective [25] can be written as

$$\mathcal{L}_\beta(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] - \beta KL\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})\|p_0(\boldsymbol{z})\right] , \quad \beta > 0 . \tag{17}$$

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

The influence of the decoder precision γ^2 and the β parameters on the objective have been related in the literature, see for example [17, § 2.4.3]—and similar observations can be found in [59, § 3.1]. Under the assumption of a Gaussian decoder, the ELBO from eq. (1) can be written as (making now explicit mention of its decoder parameter γ in parenthesis):

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

with $c(\gamma,d) = d\log\gamma - \frac{d}{2}\log(2\pi)$.

In contrast, the β-VAE objective $\mathcal{L}_\beta(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$ (also with explicit mention of γ) is expressed as:

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

We thus can link this expression to an ELBO as follows:

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

When we restrict ourselves to common practice, the optimizations of both the ELBO and the $\mathcal{L}_\beta$ are performed with a fixed value of γ (and with fixed β for $\mathcal{L}_\beta$). This entails that there is an equivalence of the solutions resulting from the optimization of each objective, as they differ only by additive and multiplicative constants. In particular, given the above expression, the following result is immediate:

Proposition 2. *Let us define the self-consistent β-VAE objective as*

$$\mathcal{L}_\beta{}^*(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi},\gamma) = \min_{\boldsymbol{\phi}} \mathcal{L}_\beta(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi},\gamma)\,. \tag{18}$$

Then the following normalized self-consistent β-VAE objective

$$\frac{1}{\beta}\left(\mathcal{L}_\beta{}^*(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi},\frac{\gamma}{\sqrt{\beta}}) - d\log\gamma + \frac{d}{2}\log(2\pi)\right) + d\log(\frac{\gamma}{\sqrt{\beta}}) - \frac{d}{2}\log(2\pi) = \text{ELBO}^*(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi},\frac{\gamma}{\sqrt{\beta}})\,. \tag{19}$$

converges to the regularized IMA objective as $\frac{\gamma}{\sqrt{\beta}} \to +\infty$ under the same conditions as in Thm. 1.

Proof. The β-VAE objective is, up to an additive constant and a strictly positive multiplicative constant, identical to the ELBO objective. As a consequence their optimum and the values of parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ at which they are achieved are identical. □

This suggests that choosing a fixed β or a β growing as, for example, $\log\gamma$ or even $\sqrt{\gamma}$ would lead to self-consistent solutions with the same properties as the vanilla VAE in the deterministic decoder limit $\gamma \to +\infty$, and notably robustness to spurious solutions.

Generalizing Kumar and Poole [38] Moreover, this connection also shows that our result generalizes previous work. Particularly, considering $\gamma^2 = 1$ when expressing the optimal variance analogous to eq. (55), we can discover the same expression as in [38].

Namely, the objective function has the form (where we now indicate explicitly the dependence on γ and emphasize that both γ and β are fixed):

$$\mathcal{L}_\beta(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi},\gamma) = -\frac{1}{2}\sum_{k=1}^{d}\Bigg[\beta\log\frac{1}{\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2} - \beta + \beta\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\left(-\frac{d^2\log p_0}{dz_k^2}(g_k^{\boldsymbol{\theta}}(\boldsymbol{x})) + \frac{\gamma^2}{\beta}\left\|\left[\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]_{:k}\right\|^2\right) - 2\beta\log(m(g_k^{\boldsymbol{\theta}}(\boldsymbol{x})))\Bigg] + d\log\gamma - \frac{d}{2}\log(2\pi). \quad (20)$$

As a consequence, the optimal $\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2$ now includes β compared to (55):

$$\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = \left(-\frac{d^2\log p_0}{dz_k^2}(g_k^{\boldsymbol{\theta}}(\boldsymbol{x})) + \frac{\gamma^2}{\beta}\left\|\left[\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]_{:k}\right\|^2\right)^{-1}. \quad (21)$$

That is, β affects the decoder Jacobian column norms as $1/\gamma^2$. They are not exactly equivalent though—note the β factor in front of the log-prior terms:

$$\mathcal{L}_\beta(\boldsymbol{x};\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}) = -\frac{1}{2}\sum_{k=1}^{d}\Bigg[\beta\log\left(-\frac{d^2\log p_0}{dz_k^2}(g_k^{\boldsymbol{\theta}}(\boldsymbol{x})) + \frac{\gamma^2}{\beta}\left\|\left[\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]_{:k}\right\|^2\right) - 2\beta\log(m(g_k^{\boldsymbol{\theta}}(\boldsymbol{x})))\Bigg] + d\log\gamma + cst. \quad (22)$$

Again, as both β and γ are fixed, the additive/multiplicative constants are irrelevant. Thus, β has a similar effect on column-orthogonality as $1/\gamma^2$. We formalize this observation in the following remark:

Remark 1. *β affects the column norms of the decoder Jacobian in the same way as $1/\gamma^2$. One can think of having $\gamma' = \gamma/\sqrt{\beta}$ and tying the tuning of β, γ'.*

Remark 2 (Generalization of Kumar and Poole [38])**.** *Assuming the same conditions as in Thm. 1 and $\gamma^2 = 1$ (i.e., a Gaussian decoder with a Hessian $\mathbf{H} = \mathbf{I}_d$ and Gaussian prior with $\mathbf{I}_d$ as second-order derivative) with the β-VAE loss, then expressing the optimal posterior covariance with $\gamma/\sqrt{\beta}$, we get Eq. (11) of Kumar and Poole [38],*

$$\boldsymbol{\Sigma}_{\boldsymbol{z}|\boldsymbol{x}}^{\boldsymbol{\phi}} = \left(\mathbf{I}_d + \frac{1}{\beta}\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})\right)\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})\right)^T\right)^{-1}$$

as a special case.

A.4 c_{IMA} as the (scaled) left KL measure of diagonality of $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}^T\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}$

In our paper, we used the original definition for c_{IMA} ([23, (8)]), which we restate here:

$$c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}},\boldsymbol{z}) = \sum_{k=1}^{d}\log\left\|\tfrac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}(\boldsymbol{z})\right\| - \log\left|\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})\right|.$$

However, an alternative formulation exists: namely, $c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}},\boldsymbol{z})$ can be thought of as the (scaled) *left KL measure of diagonality* of the *square* matrix $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})^T\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})$, as shown in [2]. That is, we can rewrite the above as [23, (23)] (with $\mathbf{A} = \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})^T\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})$):

$$c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}},\boldsymbol{z}) = \frac{1}{2}\mathrm{D}_{\mathrm{KL}}^{\text{left}}(\mathbf{A}) \quad (23)$$

$$= -\frac{1}{2}\log\left|\mathrm{diag}(\mathbf{A})^{-1/2}\mathbf{A}\mathrm{diag}(\mathbf{A})^{-1/2}\right| \quad (24)$$

$$= \frac{1}{2}\left(\log|\mathrm{diag}(\mathbf{A})| - \log|\mathbf{A}|\right) \quad (25)$$

The importance of eq. (23) is twofold: i) it provides a theoretical analysis of c_{IMA} as a measure of the column-orthogonality of $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})$ (or, equivalently, the diagonality of $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})^T \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})$); and ii) it elucidates why c_{IMA} *can be used in the* $\dim \boldsymbol{x} \neq \dim \boldsymbol{z}$ *case,* but only to measure column-orthogonality (to exploit the beneficial properties of IMA for identifiability, the theory of IMA needs to be extended to this case). We leverage ii) as a justification to use c_{IMA} in our image experiments in § 4.3.

A.5 Assessing the value of c_{IMA} for recovering the true latents

c_{IMA}, given by (6), measures the deviation of the learned decoder from the IMA function class. As it is positive and unbounded, it is practically relevant to investigate the following question: *how much violation of the* IMA *assumption is acceptable to recover the true latents?* Expressed differently, we are interested in whether a threshold can be specified for $c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{z})$ to decide whether the true (but unknown) latent factors can be recovered.

Unfortunately, our answer is negative, but this is not specific to IMA theory. Even in the case of linear ICA we cannot specify a threshold for the non-Gaussianity of the latent (source) variables to ensure that the ground-truth factors can be recovered [29].

We acknowledge that, among others, c_{IMA} being unbounded makes it harder to interpret. Thus, we go back to first principles to develop an imperfect but hopefully more practical intuition on how to assess the value of c_{IMA} to decide whether the true latents are recovered (if the true mixing is in the IMA class). As we already pointed out in Appx. A.4, c_{IMA} is the left KL measure of diagonality of $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}^T \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}$ [2]. In their analysis, the authors provide closed form solutions expressions to compare different diagonality measures.

For our purposes, [2, Fig. 1] provides an important insight: it shows that in the two-dimensional case c_{IMA} increases nonlinearly in a variable r expressing the degress of diagonality of a matrix (more precisely as $-\log\left(1-r^2\right)$, where $r=0$ denotes the identity matrix, whereas $r=1$ a matrix with parallel columns). The takeaway for us is that simply comparing two c_{IMA} values can be misleading. For the two-dimensional case, we could define an expression that is linear in r (but this requires us to accept that r is a -in some sense- suitable measure of the diagonality of a matrix), namely:

$$c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{z}) = -\frac{1}{2}\log\left(1-r^2\right) \implies r = \sqrt{1-\exp\left(-2c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{z})\right)}.$$

Unfortunately, the d-dimensional case only has a power series formulation; thus, it is nontrivial how to extend the above reasoning. However, eq. (24) expresses c_{IMA} as the negative log determinant of a scaled matrix, where the determinant of $\operatorname{diag}(\mathbf{A})^{-1/2}\mathbf{A}\operatorname{diag}(\mathbf{A})^{-1/2}$ is between 0 and 1 ($\mathbf{A} = \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})^T \mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})$). Thus, we can convert c_{IMA} to the $[0;1]$ interval as follows:

$$c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{z}) = -\frac{1}{2}\log\left|\operatorname{diag}(\mathbf{A})^{-1/2}\mathbf{A}\operatorname{diag}(\mathbf{A})^{-1/2}\right| \tag{26}$$

$$= -\frac{1}{2}\log\left|\hat{\mathbf{A}}\right| \tag{27}$$

$$0 \leq \exp\left(-2c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{z})\right) \leq 1. \tag{28}$$

The merit of the above expression is a more natural way to compare values that are normalized to the $[0;1]$ interval. Namely, the original formulation of c_{IMA} may potentially lead to problems, as, e.g., a value of t and $10t$ does not mean that one model is ten times better in recovering the true latents. Nonetheless, this can be thought (at most) as a small step towards making the analysis of the empirical value of c_{IMA} a useful tool in practical scenarios, where we do not have access to the true latent factors.

B Main Theoretical Results

B.1 Proof of Proposition 1

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

$$\boldsymbol{x}|\boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z}), \gamma^{-2}\mathbf{I}_d\right) \tag{29}$$

(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

$$z_k|\boldsymbol{x} \sim \mathcal{N}(\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x}), \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2) \tag{30}$$

(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 3 (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 $KL\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right]$.

Proof. Using Prop. 7, we have the lower bound

$$\begin{aligned}\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq -\sum_{k=1}^{d}\left[\log \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + \log m(\mu_k^{\boldsymbol{\phi}})\right] + c(\boldsymbol{x}, \gamma) \\ + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right\|^2 + \sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right] .\end{aligned} \tag{31}$$

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:

$$\begin{aligned}\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq (d-1)n(B, \gamma) - \log \sigma_{k_0}^{\boldsymbol{\phi}}(\boldsymbol{x}) \\ - d\max_t(\log m(t)) + c(\boldsymbol{x}, \gamma) + \frac{\gamma^2}{2}B^{-2}\left[\sigma_{k_0}^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right] .\end{aligned} \tag{32}$$

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. 7,

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq -\sum_{k=1}^{d}\left[\log\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + \log m(\mu_k^{\boldsymbol{\phi}})\right] + c(\boldsymbol{x},\gamma) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right\|^2 + \sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right], \quad (33)$$

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 $\boldsymbol{\mu}^{\boldsymbol{\phi}}$:

$$\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq dn(B,\gamma) - d\max_t\left(\log m(t)\right) + c(\boldsymbol{x},\gamma) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right\|^2\right]. \quad (34)$$

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 (7) 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 $(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}, \boldsymbol{\sigma}^{\widehat{\boldsymbol{\phi}}})$ 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 4 (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*

$$\forall \boldsymbol{z}, \boldsymbol{w} \in \mathbb{R}^d : \qquad \left\|\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z}) - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{w})\right\|_2 \leq C\|\boldsymbol{z} - \boldsymbol{w}\|_2, \quad (35)$$

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d : \qquad \left\|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{y})\right\|_2 \leq B\|\boldsymbol{x} - \boldsymbol{y}\|_2. \quad (36)$$

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 (7) satisfies

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) = \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) + O(1/\gamma) \quad (37)$$

$$\boldsymbol{\sigma}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = O(1/\gamma^2). \quad (38)$$

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

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

and

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

Proof. We start from the lower bound expression of Prop. 7

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

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

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

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

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

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

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

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

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

Finally, using Prop. 8, 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)

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

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

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

which simplifies to

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

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

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

such that

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

and finally

$$B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\mu}^{\widehat{\phi}}(\boldsymbol{x})\right\|^2 + (1-\nu)\sum_{k=1}^{d}\sigma_k^{\widehat{\phi}}(\boldsymbol{x})^2\right] \leq \frac{2d}{\gamma^2}\left(\frac{1}{2}(C^2-1) + E + D\left[\frac{\|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})\|^2}{d} + \frac{1}{\gamma^2}\right] + M + \frac{1}{2}\log(B^2/\nu)\right) \quad (39)$$

Taking $\nu = 1$ in (39) we get the first intended inequality

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

Alternatively, (39) implies

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

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

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

□

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

Proposition 1. *[Self-consistency of near-deterministic VAEs] Under Assumption 1, for all $\boldsymbol{x}$, $\boldsymbol{\theta}$, as $\gamma \to +\infty$, there exists at least one global minimum solution of (7). These solutions satisfy*

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) = \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) + O(1/\gamma) \quad \textit{and} \quad \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = O(1/\gamma^2)\,, \textit{ for all } k\,. \tag{9}$$

Proof. We only have to check that Assum. 1 allow fulfilling the following requirements of Prop. 4:

- the Lipschitz continuity requirements in Prop. 4 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. 4 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. 3.

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:

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

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

$$\begin{aligned}
&\int q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}) \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) \left[\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})\|^2 \gamma^2 - 2 \sum_{k=1}^{d} \log m(z_k) \right] d\boldsymbol{z} \\
&\qquad = \int q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}) z_k \left[\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})\|^2 \gamma^2 - 2 \sum_{k=1}^{d} \log m(z_k) \right] d\boldsymbol{z}\,, \forall k
\end{aligned} \tag{40}$$

Based on this expression we derive the following result.

Proposition 5. *Under Assum. 1, as $\gamma \to +\infty$*

$$\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) = \boldsymbol{x} + \frac{1}{\gamma^2}\mathbf{J}^{-T}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})} n'(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) + O(1/\gamma^3). \tag{41}$$

and

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) = \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) + \frac{1}{\gamma^2}\mathbf{J}^{-1}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})}\mathbf{J}^{-T}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})} n'(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})) + O(1/\gamma^3) \tag{42}$$

Proof. We start from the constraints of (40) that we rewrite

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

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

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

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.)

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

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 $n : \boldsymbol{z} \to \log(m(\boldsymbol{z}))$)

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

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

$$\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) = \boldsymbol{x} + \frac{1}{\gamma^2} \mathbf{J}^{-T}_{\boldsymbol{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})} n'(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) + O(1/\gamma^3). \tag{43}$$

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 (43)

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

□

B.2 Proof of Theorem 1

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 6 (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*

$$\left\| \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) - \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) \right\| = O_{\gamma \to +\infty}(1/\gamma^2), \tag{44}$$

$$\left\| \boldsymbol{\sigma}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 \right\|^2 = O_{\gamma \to +\infty}(1/\gamma^2). \tag{45}$$

then

$$\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = \left(-\frac{d^2 \log p_0}{dz_k^2}(g_k^{\boldsymbol{\theta}}(\boldsymbol{x})) + \gamma^2 \left\| \left[\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}} \left(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) \right) \right]_{:k} \right\|^2 \right)^{-1} + O(1/\gamma^3), \tag{46}$$

and the self-consistent ELBO (19) *approximates the* IMA-*regularized log-likelihood* (6)*:*

$$\text{ELBO}^*(\boldsymbol{x}; \boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})) + O_{\gamma \to \infty}(1/\gamma^2). \tag{47}$$

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

$$\text{ELBO}^*(\boldsymbol{x}; \boldsymbol{\theta}) = -\text{KL}\left[q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x}) || p_0(\boldsymbol{z}) \right] + \mathbb{E}_{q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) \right], \tag{48}$$

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 (14) for this term in the ELBO. Taking optimal encoder parameters, we get the posterior regularization term for the ELBO*

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

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

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

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 $\xi \in [\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x}), u]$ if $u \geq \mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})$, or $\xi \in [u, \mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})]$ if $u \leq \mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})$ 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[7]

$$\begin{aligned} \left|\mathbb{E}_{z_k\sim\mathcal{N}(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}),\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2)}[\log(m(z_k))] - \log(m(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))) - \frac{1}{2}n''(\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\right| \\ \leq F\mathbb{E}\left[\left|u-\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})\right|^3\right] = F\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^3\frac{2^{3/2}}{\sqrt{\pi}}\,. \end{aligned} \tag{49}$$

As the assumptions entail that optimal posterior variances $\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2$ get small for γ large (cf. (45)), 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

$$\begin{aligned} -\mathrm{KL}\left[q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})||p_0(\boldsymbol{z})\right] = \sum_{k=1}^{d}\log(m(g_k^{\boldsymbol{\theta}}(\boldsymbol{x}))) + \frac{1}{2}n''(g_k^{\boldsymbol{\theta}}(\boldsymbol{x}))\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 + \frac{1}{2}\log\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 \\ + \kappa_d + O_{\gamma\to+\infty}(1/\gamma^2)\,. \end{aligned} \tag{50}$$

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

Reconstruction term Now switching to the first (reconstruction) term of the ELBO*, adapting the decomposition of (15) by using optimal encoder parameters we get

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

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

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\left\|\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z}) - \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})) - \sum_{k=1}^{d}\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\boldsymbol{z}^o}(z_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))\right\|^2\right] \leq \frac{d^3}{4}3K^2\sum_i \sigma_i^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^4\,. \quad (51)$$

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

$$\begin{aligned}\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\left\|\sum_{k=1}^{d}\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|z^o}(z_k - \mu_k^{\boldsymbol{\phi}}(\boldsymbol{x}))\right\|^2\right] &= \text{trace}\left[\text{Cov}\left[\sum_{k=1}^{d}\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})}(z_k - \mu_k^{\boldsymbol{\phi}}(\boldsymbol{x}))\right]\right] \\ &= \sum_{k=1}^{d}\left[\left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})}\right\|^2 \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right]\,. \quad (52)\end{aligned}$$

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

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

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

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

Each term of the upper bound can be bounded for the optimum encoder parameters: using from left to right the assumption of (44) and (51), respectively, leading to

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

Getting back to the whole reconstruction term, using additionally the variance self-consistency assumption (45), 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

$$\mathbb{E}_{q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] = -\frac{\gamma^2}{2}\sum_{k=1}^{d}\left[\left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})}\right\|^2 \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2\right] + d\log\gamma - \frac{d}{2}\log(2\pi) + O_{\gamma\to+\infty}(1/\gamma^2) \quad (53)$$

ELBO* approximation As a consequence of (50) and (53) the ELBO* becomes

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

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 $\widehat{\text{ELBO}}_k(\sigma_k^2)$ for optimal $\gamma^2\sigma_k^*$ we get (see lemma 4):

$$\gamma^2\sigma_k^{*2} = \left(-\frac{1}{\gamma^2}n''({g^{\boldsymbol{\theta}}}_k(\boldsymbol{x})) + \left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|{g^{\boldsymbol{\theta}}}_k(\boldsymbol{x})}\right\|^2\right)^{-1} \tag{54}$$

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

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

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)

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

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

$$\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = \sigma_k^{*2} + O(1/\gamma^3) = \left(-n''({g^{\boldsymbol{\theta}}}_k(\boldsymbol{x})) + \gamma^2\left\|\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|{g^{\boldsymbol{\theta}}}_k(\boldsymbol{x})}\right\|^2\right)^{-1} + O(1/\gamma^3). \tag{55}$$

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*

$$\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = \left(-\frac{d^2\log p_0}{dz_k^2}(g_k^{\boldsymbol{\theta}}(\boldsymbol{x})) + \gamma^2\left\|\left[\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\left(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right]_{:k}\right\|^2\right)^{-1} + O(1/\gamma^3)\,, \tag{12}$$

and the self-consistent ELBO (19) *approximates the* IMA*-regularized log-likelihood* (6)*:*

$$\text{ELBO}^*(\boldsymbol{x};\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - c_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}},\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})) + O_{\gamma\to\infty}\left(1/\gamma^2\right). \tag{13}$$

Proof. This is just a corollary of Proposition 6 because Proposition 5 entails through (42) the required $O(1/\gamma^2)$ rate of convergence for the optimal variational mean in (44), while (45) is fulfilled through Prop. 1. □

C Auxiliary results

C.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 trace$[AB]$ = 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

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

□

C.2 KL divergence bounds

Proposition 7 (Lipschtiz continuity-based lower bound)**.** *Assume $\boldsymbol{g}^{\boldsymbol{\theta}}$ is Lipschitz continuous with Lipschitz constant $B > 0$, in the sense*

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

Then for any encoder parameter choice

$$\begin{aligned}\textit{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \geq -\sum_{k=1}^{d}\left[\log\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x}) + \log m(\mu_k^{\boldsymbol{\phi}})\right] + c(\boldsymbol{x},\gamma)\\ + \frac{\gamma^2}{2}B^{-2}\left[\left\|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right\|^2 + \sum_{k=1}^{d}\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right]\,,\end{aligned} \tag{56}$$

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 (16),

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

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

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

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

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

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

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

by reordering the terms we finally get

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

which is the stated KL lower bound. □

Proposition 8 (Optimal encoder KL divergence upper bound)**.** *Assume $\boldsymbol{f}^{\boldsymbol{\theta}}$ is Lipschitz continuous with Lipschitz constant $C>0$, in the sense that*

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

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 (7)

$$KL\left[q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \;\leq\; d\left(\frac{1}{2}C^2 + E + D\left[\frac{\|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})\|^2}{d} + \frac{1}{\gamma^2}\right]\right) - \frac{d}{2} + \log p_{\boldsymbol{\theta}}(\boldsymbol{x})\,, \quad (57)$$

and

$$\begin{aligned}\limsup_{\gamma\to+\infty} KL\left[q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] \leq & \; d\left(\frac{1}{2}C^2 + E\right) + D\|\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})\|^2 \\ & - \frac{d}{2} - \log|\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}))| + \log(p_0(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}))) \quad (58)\end{aligned}$$

Proof. Starting from the KL divergence expression (16),

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

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):

$$\boldsymbol{\mu}^{\boldsymbol{\phi}^*}(\boldsymbol{x}) = \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) \tag{59}$$

$$\boldsymbol{\sigma}^{\boldsymbol{\phi}^*}(\boldsymbol{x}) = \frac{1}{\gamma} \tag{60}$$

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

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

Using Lemma 1 we get

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

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

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

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

□

C.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 $\boldsymbol{z}_o \in \mathbb{R}^d$

$$\mathbb{E}_{\boldsymbol{z}}\left[\left\|\boldsymbol{f}(\boldsymbol{z}) - \boldsymbol{f}(\boldsymbol{z}_o) - \sum_k \frac{\partial \boldsymbol{f}}{\partial z_k}_{|\boldsymbol{z}_o}(z_k - z_k^o)\right\|^2\right] \leq \frac{d^3}{4} 3K^2 \sum_i \left(\sigma_i^{\boldsymbol{\phi}}\right)^4 \tag{61}$$

Proof. As described in [45, p. 162], for the l-th component of the function

$$\begin{aligned}f_l(\boldsymbol{z}) = f_l(\boldsymbol{z}_o) + \sum_k \frac{\partial f_l}{\partial z_k}_{|\boldsymbol{z}_o}(z_k - z_k^o) + \frac{1}{2!}\sum_{i,j}\frac{\partial f_l}{\partial z_i \partial z_j}_{|\boldsymbol{z}_o + t_{ij}(\boldsymbol{z}-\boldsymbol{z}_o)}(z_i - z_i^o)(z_j - z_j^o)\,, t_{ij} \in (0;1)\ . \\ = f_l(\boldsymbol{z}_o) + \sum_k \frac{\partial f_l}{\partial z_k}_{|\boldsymbol{z}_o}(z_k - z_k^o) + \frac{1}{2!}\sum_{i,j}(\boldsymbol{z}-\boldsymbol{z}_o)^T \mathbf{H}_k(\boldsymbol{z}-\boldsymbol{z}_o)\,,\end{aligned} \tag{62}$$

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 [8] leading to the bound

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

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

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:

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

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

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

□

C.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

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

□

D Related work

D.1 Implicit inductive biases in the ELBO

Rolinek et al. [57] 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 $\boldsymbol{\Sigma}_{\boldsymbol{z}|\boldsymbol{x}}^{\boldsymbol{\phi}}$) are only experimentally investigated. The same authors extend their work in [71], 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. Nakagawa et al. [49] builds upon the results of [57] and provides a *novel interpretation* of VAEs by introducing implicit variables to express the latent space in terms of an isometric embedding.

Lucas et al. [44] 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) [64] 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 Appx. E.1 that in the limit of a deterministic decoder linear Gaussian VAEs optimize the IMA objective with $\lambda = 1$. Dai and Wipf [13] present more general results than [44] since they use affine functions. Additionally, a connection to Robust PCA [6] is established.

Kumar and Poole [38] generalizes [57], as it admits a variational posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ with *block-diagonal covariance* with a uniqueness result for diagonal $\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}|\boldsymbol{x}}$. The authors derive a formula for the optimal $\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}|\boldsymbol{x}}$ [38, 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 [38, 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.

Dai and Wipf [13] use a different setting that turns out to be very interesting to compare to ours. The most important is that while we use a factorized Gaussian variational posterior, Dai and Wipf [13] use a non-factorized Gaussian, which leads to major differences. Broadly construed, Dai and Wipf [13] are able to show in their Theorem 2 (which includes the case of equal latent and observation dimensions matching our setting) that in the deterministic limit, their κ-simple VAE can perfectly fit arbitrary observed data (barring few assumptions), while the ELBO gap tends to zero. The way it is proven relies on a first step with the Darmois construction [29], chosing the decoder mean parameter such that its pushforward is exactly the observation distribution. Then in a second step, by an appropriate choice of variational posterior parameters, they show that asymptotically the ELBO gap (i.e., the KL divergence between true and variational posteriors) tends to zero in the deterministic limit. In contrast, our constraint of factorized variational posterior does not allow the ELBO gap to vanish in the deterministic limit (unless the decoder mean that fits the data perfectly is in the IMA class, which is a very special case; in particular, if the Darmois construction is used). For this reason, the proofs and scope of our results are very different: (i) we use information theoretic bounds to show that the encoder inverts the decoder mean (independently from the fact that this one may or may not fit the data perfectly); (ii) we obtain a rigorous convergence to the IMA regularized likelihood, which demonstrates that the gap is not eliminated in the deterministic limit.

Regarding our result for β-VAEs (Prop. 2), the approach of Mathieu et al. [46] is similar as the authors show that $\mathcal{L}_\beta$ can be expressed in terms of a *rescaled* ELBO. The difference is that Mathieu et al. [46] uses a rescaling of the parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$, whereas we only scale γ^2.

D.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 [44]) to avoiding sampling for the reconstruction loss term [38]. Several papers argued for a similar setting: Dai and Wipf [13] take the limit of $\gamma \to +\infty$ (here using γ as the square root of the decoder precision and not the decoder variance as used in [13]) to derive a result relating encoder and decoder properties in this limit in their Theorem 5, that has a similar flavor to Prop. 1. In contrast to our nonlinear analysis, this is derived when optimizing with respect to both encoder and decoder parameter, and as stated in the previous section, the non-factorized encoder assumptions leads to fundamentally different behavior of the solutions in the deterministic limit. Rolinek et al. [57] refer to the *polarized regime* (a property of which is that encoder variances are small, cf. [57, Definition 1]), Kumar and Poole [38] argue for "concentrated" variational posteriors. Ghosh et al. [20] substitute stochasticity with a regularizer on the decoder Jacobian from an intuitive, whereas Kumar et al. [40] motivate these results from an injective flow perspective. Nielsen et al. [50] 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.

E 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.

E.1 Linear VAE from Lucas et al. [44]

We restate the linear VAE model of [44]:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}\left(\mathbf{W}\boldsymbol{z} + \boldsymbol{\mu}; \frac{1}{\gamma^2}\mathbf{I}_d\right) \tag{63}$$

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\mathbf{V}\left(\boldsymbol{x} - \boldsymbol{\mu}\right); \mathbf{D}\right), \tag{64}$$

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

$$\mathbf{D}^* = \frac{1}{\gamma^2}\left(\text{diag}\left(\mathbf{W}^T\mathbf{W}\right) + \frac{1}{\gamma^2}\mathbf{I}_d\right)^{-1} \tag{65}$$

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. 9.

Proposition 9 (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 [44, Appendix C.2], it is shown that the gap between exact log-likelihood and ELBO for linear Gaussian VAEs in stationary points reduces to

$$\text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] = \frac{1}{2}\left(\log\det\tilde{\mathbf{M}} - \log\det\mathbf{M}\right) \tag{66}$$

$$\mathbf{M} = \mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_d \tag{67}$$

$$\tilde{\mathbf{M}} = \text{diag}\left(\mathbf{W}^T\mathbf{W}\right) + \frac{1}{\gamma^2}\mathbf{I}_d, \tag{68}$$

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

$$\text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] = \log\frac{\left|\text{diag}\left(\mathbf{W}^T\mathbf{W}\right) + \frac{1}{\gamma^2}\mathbf{I}_d\right|}{\left|\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_d\right|} \tag{69}$$

$$= \log\frac{\left|\text{diag}\left(\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_d\right)\right|}{\left|\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_d\right|} \tag{70}$$

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

$$\text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] = \log\frac{\left|\text{diag}\left(\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_d\right)\right|}{\left|\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_d\right|} \tag{71}$$

$$= \log\frac{\left|\text{diag}\left(\mathbf{U}\left[\boldsymbol{\Lambda} + \frac{1}{\gamma^2}\mathbf{I}_d\right]\mathbf{U}^T\right)\right|}{\left|\mathbf{U}\left[\boldsymbol{\Lambda} + \frac{1}{\gamma^2}\mathbf{I}_d\right]\mathbf{U}^T\right|}, \tag{72}$$

which is by definition the local IMA contrast c_{IMA} (cf. [23, 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. 9, especially (72), 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 (72) 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* (72) *is zero.*

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

F Experimental details

F.1 The relationship of weight matrix structures and the IMA function class

During the experiments we have used different weight matrices either to *ensure* that the mixing is within or to *exclude* it from the IMA function class. Here we summarize our choices also including the *depth* of the network as it can affect the mixing's place w.r.t. the IMA function class.

When we use *orthogonal* weight matrices (§ 4.1,§ 4.2), then a single-layer network is within the IMA class, but adding more layers with elements-wise nonlinearities will move the MLP outside the function class. When using *triangular* MLPs (§ 4.2), the network is also outside the IMA class (triangular matrices are orthogonal only when they are *diagonal*, see also [23, Lemma C.1]).

Notably, Möbius transforms [53] are conformal maps (thus, they are in the IMA class) irrespective of the structure of the weight matrix used (cf. Appx. F.4 for details).

F.2 Self-consistency in practical conditions (§ 4.1)

For the self-consistency experiments, the mixing is a 3-layer MLP with smooth Leaky ReLU nonlinearities [22] 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 $\gamma^2 \in \{1\text{e}1; 1\text{e}2; 1\text{e}3; 1\text{e}4; 1\text{e}5\}$. Additional parameters are described in Tab. 1. Training is continued until the ELBO* improves on the *validation set* (we use early stopping [54]), then all metrics are reported for the maximum ELBO* (Fig. 2).

Table 1: Hyperparameters for the self-consistency experiments (§ 4.1)

PARAMETER	VALUES
ENCODER	3-LAYER MLP
DECODER	3-LAYER MLP
ACTIVATION	SMOOTH LEAKY RELU [22]
BATCH SIZE	64
# SAMPLES (TRAIN-VAL-TEST)	$42 - 12 - 6$K
LEARNING RATE	1e-3
d	3
GROUND TRUTH	GAUSSIAN
$p_0(\boldsymbol{z})$	GAUSSIAN
$\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}\|\boldsymbol{x}}$	DIAGONAL
γ^2	{1e1; 1e2; 1e3; 1e4; 1e5}
# SEEDS	20

F.3 Relationship between ELBO*, IMA-regularized, and unregularized log-likelihoods (§ 4.2)

Table 2: Hyperparameters for the *triangular MLP* (*not* from the IMA class) ELBO*–$\mathcal{L}_{\text{IMA}}$–log-likelihood experiments (§ 4.2)

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	1e-4
d	2
GROUND TRUTH	GAUSSIAN
$p_0(\boldsymbol{z})$	GAUSSIAN
$\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}\|\boldsymbol{x}}$	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.

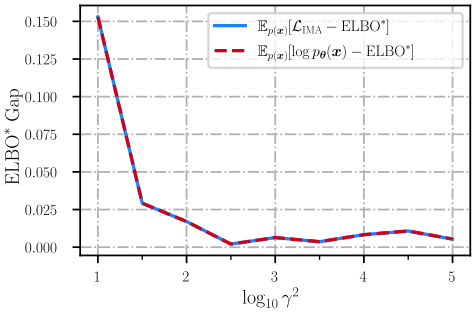

Figure 6: Comparison of the ELBO*, the IMA-regularized and unregularized log-likelihoods over different γ^2 with an IMA-class mixing

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 $1\mathrm{e}{-4}$ and the batch size 64.

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. 6 demonstrates.

Table 3: Hyperparameters for the *orthogonal MLP* (from the IMA class) ELBO*–$\mathcal{L}_{\text{IMA}}$–log-likelihood experiments (§ 4.2)

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	$1\mathrm{e}{-4}$
d	2
GROUND TRUTH	UNIFORM
$p_0(\boldsymbol{z})$	UNIFORM
$\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}\mid\boldsymbol{x}}$	DIAGONAL
γ^2	$[1\mathrm{e}1; 1\mathrm{e}5]$
C_{IMA} (MIXING)	0

F.4 Connecting the IMA principle, γ^2, and disentanglement (§ 4.3)

Synthetic data (Möbius transform) We use 3-dimensional conformal mixings (i.e., the Möbius transform [53]) 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 $\gamma^2 = 1\mathrm{e}5$ and report further parameters in Tab. 4. Training is continued until the ELBO* improves on the *validation set* (we use early stopping [54]), then all metrics are reported for the maximum ELBO* (Fig. 3).

Table 4: Hyperparameters for the *synthetic (Möbius)* IMA–disentanglement experiments (§ 4.3) with a linear map

PARAMETER	VALUES
ENCODER	3-LAYER MLP
DECODER	3-LAYER MLP
ACTIVATION	SMOOTH LEAKY RELU [22]
BATCH SIZE	64
# SAMPLES (TRAIN-VAL-TEST)	$42 - 12 - 6$K
LEARNING RATE	$1\mathrm{e}{-3}$
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 [66] with a Beta ground truth distribution. Similar to [32], 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) [57, 37]. Our choice is motivated by [26, 18] 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 [54]), then all metrics are reported for the maximum ELBO*.

Table 5: Hyperparameters for the *image (Sprites)* IMA–disentanglement experiments (§ 4.3)

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

F.5 Optimality of γ^2 w.r.t. its MLE

During our experiments, we *do not optimize* γ^2, as it is generally the case in the literature [57, 44, 38]. However, as noted by Rybkin et al. [59], doing so could lead to superior sample quality. The price we need to pay for improved sample generation is a more difficult optimization task (also noted in [60]): including γ^2 as a trainable parameter might require a careful learning rate tuning, and smaller learning rates can yield suboptimal likelihood values in the beginning [59].

During our experiments, we confirmed that making γ^2 learnable (all else being equal) yields sub-optimal results, particularly in terms of MCC. Thus, we opted for comparing our hyperparameter setting to the *maximum likelihood estimate of the decoder variance*, as proposed in [59, Eq. (8)]. Accomodating the parameter γ^2 instead of the decoder variance, we reformulate the equation as:

$$\gamma^2_{\text{MLE}} = \arg\max_{\gamma^2} \mathcal{N}\left(\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{z}\right), \frac{1}{\gamma^2}\mathbf{I}_d\right) = \frac{1}{\text{MSE}\left(\boldsymbol{x}, \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right)\right)} \tag{73}$$

$$= \left[\frac{1}{|\mathcal{X}|}\sum_{\boldsymbol{x}\in\mathcal{X}} \left\|\boldsymbol{x} - \boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right)\right\|^2\right]^{-1}, \tag{74}$$

i.e., the MLE is the mean squared error between observations $\boldsymbol{x}$ and the *decoded mean encodings* $\boldsymbol{f}^{\boldsymbol{\theta}}\left(\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})\right)$, where $|\mathcal{X}|$ denotes the number of observations. Interestingly, this is the inverse of the quantity we report on the right plot of Fig. 2.

To compare γ^2_{MLE} (calculated as eq. (73)) and the optimal value of γ^2 we found via grid search from the values $\{1\text{e}1; 1\text{e}2; 1\text{e}3; 1\text{e}4; 1\text{e}5\}$, we plot the log of both values in Fig. 7. We can observe that for all values except 1e5, γ^2_{MLE} is larger, sometimes with more than one order of magnitude. For 1e5, the mean (for the 20 seeds) lie in the range $[0.8\text{e}5; 3.8\text{e}5]$ with a mean and standard deviation of $2.3 \pm 0.77\text{e}5$, indicating that $\gamma^2 = 1\text{e}5$ and γ^2_{MLE} are in the same order of magnitude, corroborating that we used the optimal setting up to the granularity of our original grid search.

G Computational resources

The self-consistency (§ 4.1), the likelihood comparison (§ 4.2), and the synthetic experiments with the Möbius transform (§ 4.3, particularly Fig. 3) were ran on a MacBook Pro with a Quad-Core Intel

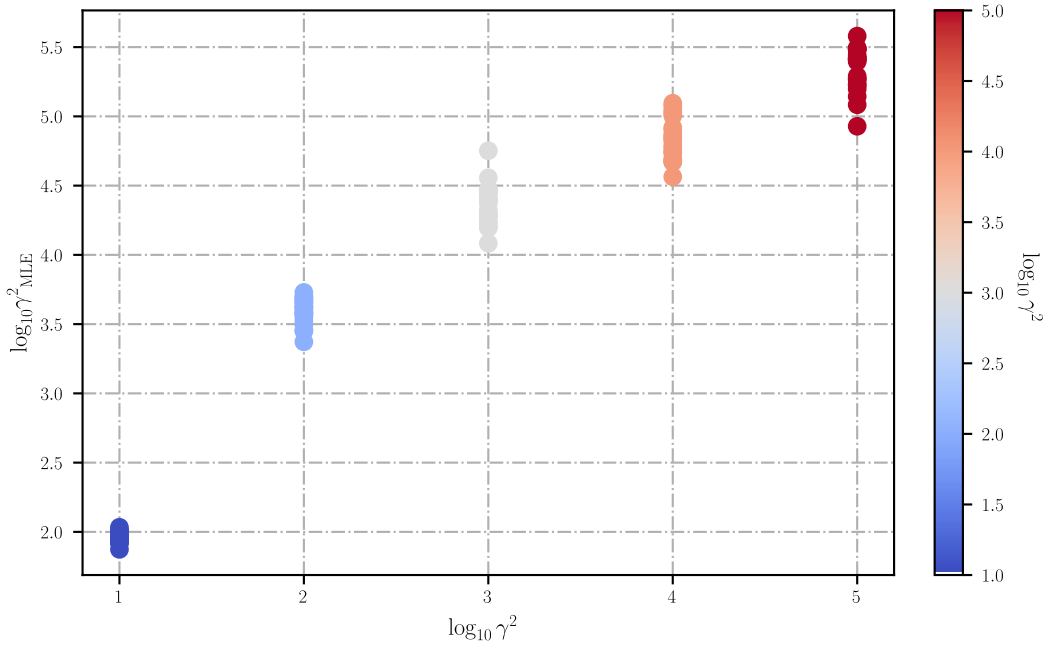

Figure 7: Comparison of γ^2_{MLE} and the optimal γ^2 we found via grid search (experimental details are the same as in § 4.1, detailed in Appx. F.2)

Core i5 CPU and required approximately nine days. The Sprites experiments (§ 4.3, particularly Fig. 5) required approximately four and a half days on an Nvidia RTX 2080 GPU.

H 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.

I 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

MLE Maximum Likelihood Estimation
MLP Multi-Layer Perceptron
MSE Mean Squared Error

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 $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$
$p(\boldsymbol{x})$ data distribution
$p_0(\boldsymbol{z})$ 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