Unscented Autoencoder

Faris Janjoš [1]   **Lars Rosenbaum** [1]   **Maxim Dolgov** [1]   **J. Marius Zöllner** [2]

Abstract

The Variational Autoencoder (VAE) is a seminal approach in deep generative modeling with latent variables. Interpreting its reconstruction process as a nonlinear transformation of samples from the latent posterior distribution, we apply the Unscented Transform (UT) – a well-known distribution approximation used in the Unscented Kalman Filter (UKF) from the field of filtering. A finite set of statistics called sigma points, sampled deterministically, provides a more informative and lower-variance posterior representation than the ubiquitous noise-scaling of the reparameterization trick, while ensuring higher-quality reconstruction. We further boost the performance by replacing the Kullback-Leibler (KL) divergence with the Wasserstein distribution metric that allows for a sharper posterior. Inspired by the two components, we derive a novel, deterministic-sampling flavor of the VAE, the Unscented Autoencoder (UAE), trained purely with regularization-like terms on the per-sample posterior. We empirically show competitive performance in Fréchet Inception Distance (FID) scores over closely-related models, in addition to a lower training variance than the VAE[1].

1. Introduction

The Variational Autoencoder (VAE) (Rezende et al., 2014; Kingma et al., 2015) is a widely used method for learning deep latent variable models via maximization of the data likelihood using a reparametrized version of the Evidence Lower Bound (ELBO). Deep latent variable models are used as generative models in a variety of applica-
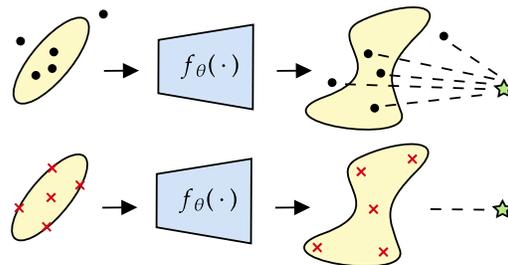
Figure 1: The VAE decoder $f_\theta(\cdot)$ can be interpreted as a nonlinear mapping of the Gaussian posterior distribution generated by the encoder, resulting in a non-Gaussian output distribution. The standard VAE (top) samples randomly from the posterior (black points) and matches each decoded sample to the ground truth (green star). Our model (bottom) samples and transforms fixed posterior sigma points (red) instead. By matching the mean of the transformed points, we push the entire output distribution to resemble the ground truth.

tion domains such as image (Vahdat & Kautz, 2020), language (Bowman et al., 2015; Kusner et al., 2017), and dynamics modeling (Karl et al., 2016). A good generative model requires the VAE to produce high-quality samples from the prior latent variable distribution and a disentangled latent representation is desired to control the generation process (Higgins et al., 2017). Another important application of deep latent variable models is representation learning, where the goal is to induce a latent representation facilitating downstream tasks (Bengio et al., 2013; Townsend et al., 2019; Tripp et al., 2020; Rombach et al., 2022). In many of these tasks a good sample quality, as well as a 'well-behaved' latent representation with a high reconstruction accuracy is desired.

Since their introduction, VAEs have been one of the methods of choice in generative modeling due to their comparatively easy training and the ability to map data to a lower dimensional representation as opposed to generative adversarial networks (Goodfellow et al., 2014). However, despite their popularity there are still open challenges in VAE training addressed by recent works. A major problem of VAEs is their tendency to have a trade-off between the quality of samples from the prior and the reconstruction qual-

---

[1]Robert Bosch GmbH, Corporate Research, 71272 Renningen, Germany [2]Research Center for Information Technology (FZI), 76131 Karlsruhe, Germany. Correspondence to: <first-name.last-name@de.bosch.com>.

[1]Code available at: https://github.com/boschresearch/unscented-autoencoder

ity. This trade-off can be attributed to overly simplistic priors (Bauer & Mnih, 2019), encoder/decoder variance (Dai & Wipf, 2019), weighting of the KL divergence regularization (Higgins et al., 2017; Tolstikhin et al., 2018), or the aggregated posterior not matching the prior (Tolstikhin et al., 2018; Ghosh et al., 2019). Furthermore, the VAE objective can be prone to spurious local maxima leading to posterior collapse (Chen et al., 2017; Lucas et al., 2019; Dai et al., 2020), which is characterized by the latent posterior (partially) reducing to an uninformative prior. Finally, the variational objective requires approximations of expectations by sampling, which causes increased gradient variance (Burda et al., 2016) and makes the training sensitive to several hyperparameters (Bowman et al., 2015; Higgins et al., 2017).

Our main technical contributions are two modifications to the original VAE objective resulting in an improved sample and reconstruction quality. We propose to use a well-known algorithm from the filtering and control literature, the Unscented Transform (UT) (Uhlmann, 1995), to obtain lower-variance, albeit potentially biased gradient estimates for the optimization of the variational objective. A lower variance is achieved by only sampling at the sigma points of the variational posterior and transforming these points with a deterministic decoder. In this context, we show that reconstructing the entire posterior distribution via its sigma points (visualized in Fig. 1) is superior in resulting image quality to reconstructing individual random samples. Furthermore, we observe that the regularization toward a standard normal prior using a KL divergence often harshly penalizes low variance along some components even though the low variance is usually beneficial for reconstruction. Thus, we use a different regularization based on the Wasserstein metric (Patrini et al., 2020). To account for resulting sharper posteriors, we add a regularizer for decoder smoothness around the mean encoded value, similar to (Ghosh et al., 2019). We conduct rigorous experiments on several standard image datasets to compare our modifications against the VAE baseline, the closely-related Regularized Autoencoder (RAE) (Ghosh et al., 2019), the Importance-Weighted Autoencoder (IWAE) (Burda et al., 2016), as well as the Wasserstein Autoencoder (WAE) (Tolstikhin et al., 2018).

2. Related Work

Many recent works on VAEs focus on understanding and addressing still existing problems like undesired posterior collapse (Dai et al., 2020), trade-off between sample and reconstruction quality (Tolstikhin et al., 2018; Bauer & Mnih, 2019), or non-interpretable latent representations (Rolinek et al., 2019; Higgins et al., 2017). Other recent works suggest to move from the probabilistic VAE

models to deterministic models, such as the RAE in (Ghosh et al., 2019); our model can be considered as part of this class. As previously mentioned, we employ two major modifications to the VAE, namely the **Unscented Transform** and the **Wasserstein metric**, as well as **decoder regularization**; we outline the section accordingly.

We use the **Unscented Transform** (Uhlmann, 1995) from the field of nonlinear filtering within signal processing. In this context, the signal state estimate is often assumed to be Gaussian in order to maintain tractability. However, nonlinear prediction and measurement models always invalidate this assumption at each time step so that a re-approximation becomes necessary. A commonly used approach is the Extended Kalman Filter (EKF), where a linearization of the models is employed so that the Gaussian state remains Gaussian during filtering. In contrast, alternative approaches that represent the Gaussian state (assuming application in the context of the VAE posterior) with samples for propagation and update have emerged. These approaches can be clustered according to the employed sampling method – random as in (Gaussian) particle filters (Doucet & Johansen, 2011) or deterministic, e.g. in the UKF (Julier et al., 2000). In the UKF, the n-dimensional Gaussian is approximated with $2n + 1$ deterministic samples, which can be propagated through the nonlinearities and are sufficient for computing the statistics of a Gaussian distribution, i.e. its mean and covariance. This procedure is referred to as the Unscented Transform (UT).

The use of deterministic sampling[2] aims to achieve a good coverage of the distribution represented with the mean and covariance. Although this approach produces biased estimates of the involved expectations compared to random sampling due to non-i.i.d. samples, it often captures well the nonlinearities applied to the distribution, for a finite, small set of samples in the filtering context. This observation can transfer to neural networks due to their Lipschitz continuity (Khromov & Singh, 2023). Our UT experiments empirically underline this expectation. For a more comprehensive overview of the UT and the UKF, we refer the reader to (Menegaz et al., 2015).

The UT uses several samples to get an estimate of the moments of a nonlinearly transformed probability distribution. Along those lines, our method also relates to the IWAE (Burda et al., 2016) and some of its extensions (Tucker et al., 2018). IWAE uses importance weighting of K posterior samples to obtain a variational distribution closer to the true posterior (Cremer et al., 2017). The method is known to have a diminishing gradient signal for the inference network (Rainforth et al., 2018) if no additional improvements are used (Tucker et al., 2018). Using the Wasserstein metric, the inference distribution is sharp,

---

[2] Sampling from a set of points at fixed locations in the domain.

so practically there is not much gain in a more complex distribution. However, multiple samples can help to obtain lower variance gradient estimates, which also applies to the IWAE by taking a multiple of K samples. Sampling only at the sigma points reduces this variance even more and is known to empirically work well in filtering and control.

The **Wasserstein metric** is used in (Tolstikhin et al., 2018; Patrini et al., 2020) to regularize the aggregated posterior $q_{\mathrm{agg}}(\mathbf{z}) = \mathbb{E}_{p(\mathbf{x})}[q(\mathbf{z}|\mathbf{x})]$ toward the standard normal prior. The authors also show that such an objective is an upper bound to the Wasserstein distance between the sampling distribution of the generative model and the data distribution if the regularization is scaled by the Lipschitz constant of the generator. In contrast, we do not regularize the aggregated posterior, but use the Wasserstein distance to weakly regularize the mean and variance of the encoder, such that neither explodes and we can do ex-post density estimation. From a theoretical point of view, we do not fix the prior but learn the manifold; the aggregated posterior is learned by fitting a mixture to the encoded data points.

Finally, our work incorporates several ideas from the recently published RAE (Ghosh et al., 2019). We also use a **decoder regularization** term based on the decoder Jacobian in our loss, which promotes smoothness of the latent space. In contrast to the RAE however, we generalize the term from a deterministic to a stochastic encoder as not every data point might be encoded with the same fidelity. Furthermore, we employ ex-post density estimation as we do not explicitly regularize the aggregated posterior toward a prior. Conceptually, the UAE can be placed between the VAE, characterized by significant sampling variance, and the purely deterministic RAE.

3. Problem Description

Most generative models take a max-likelihood approach to model a real-world distribution $p(\mathbf{x})$ via the θ-parameterized probabilistic generator model $p_\theta(\mathbf{x})$

$$\theta \leftarrow \arg\max_\theta \quad \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})}[\log p_\theta(\mathbf{x})] . \tag{1}$$

In this setting, latent variable generative approaches assume an underlying structure in $p(\mathbf{x})$ not directly observable from the data and model this structure with a latent variable $\mathbf{z}$, which is well-motivated by de Finetti's theorem (Accardi, 2001). As a result, the distribution $p(\mathbf{x})$ can be represented as a product of tractable distributions. However, directly incorporating $\mathbf{z}$ via an integral $\int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$ is intractable; thus, one introduces an amortized variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$ (Zhang et al., 2018) and obtains

$$\log p_\theta(\mathbf{x}) = \log \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] . \tag{2}$$

This model assumption is the basis of variational inference. Applying Jensen's inequality yields the well-known ELBO, denoted by $\mathcal{L}$

$$\begin{aligned}\log p_\theta(\mathbf{x}) \geq \mathcal{L} = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]- \\ - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})) ,\end{aligned} \tag{3}$$

which is maximized w.r.t. θ and ϕ. The first term accounts for the quality of reconstructed samples and the $D_{\mathrm{KL}}(...)$ term pushes the approximate posterior to mimic the prior, i.e. it enforces a $p(\mathbf{z})$-like structure to the latent space.

Training on $\mathcal{L}$ in Eq. (3) requires computing gradients w.r.t. θ and ϕ. This is relatively straightforward for the generator parameters, however, requiring a high-variance policy gradient for the posterior parameters. To avoid this issue in practice, the reparameterization trick (Kingma et al., 2015) is used to simplify the sampling of the approximate posterior by means of an easy-to-sample distribution. Assuming a Gaussian posterior $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can sample a multivariate normal and obtain the latent feature vector via the deterministic transformation

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T . \tag{4}$$

With the help of the reparameterization trick, the VAE (Kingma & Welling, 2013) provides a framework for optimizing the loss function from the condition in Eq. (3) via an encoder–decoder generative latent variable model. The encoder $E_\phi(\mathbf{x}) = \{\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\Sigma}_\phi(\mathbf{x})\}$ parameterizes a multivariate Gaussian $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\Sigma}_\phi(\mathbf{x}))$, where $\boldsymbol{\Sigma}_\phi$ is usually a diagonal matrix, $\boldsymbol{\Sigma}_\phi = \mathrm{diag}(\boldsymbol{\sigma}_\phi)$. The decoder $D_\theta(\mathbf{z}) = \boldsymbol{\mu}_\theta(\mathbf{z})$ is in practice rendered deterministic: $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\theta(\mathbf{z}), \mathbf{0})$, reducing the reconstruction term in Eq. (3) to a simple mean-squared error under the expectation of the posterior $\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})}\|\mathbf{x} - \boldsymbol{\mu}_\theta(\mathbf{z})\|_2^2$. The VAE uses the reparameterization trick for efficient sampling from the posterior q_ϕ (in practice providing only a single sample to the decoder), which enables a lower-variance gradient backpropagation through the encoder.

The deterministic decoder and the reparameterization trick allow for a slightly different interpretation of the reconstruction/generation process: a (highly) nonlinear transformation of an input distribution, represented (usually) only by a single stochastic sample. The sample is white noise[3], scaled and shifted by the posterior moments. This interpretation serves as the basis for our work, where the unscented transform of the input distribution serves as an alternative to the single-stochastic-sample representation. In the next section, we outline the unscented transform representation of the input to the decoder via a set of deterministically computed and sampled sigma points.

---

[3]The white-noise interpretation is used in (Ghosh et al., 2019) to justify regularization as an alternative to the noise sampling.

4. Unscented Transform of the Posterior

4.1. Background

The unscented transform (Uhlmann, 1995) is a method to evaluate a nonlinear transformation of a distribution characterized by its first two moments. Assume a known deterministic function $\boldsymbol{f}$ applied to a distribution $P(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean and covariance $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$. If $\boldsymbol{f}$ is a linear transformation, one can describe the distribution $Q(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ at the output via $\hat{\boldsymbol{\mu}} = \boldsymbol{f}\boldsymbol{\mu}$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{f}\boldsymbol{\Sigma}\boldsymbol{f}^T$. Similarly, for a nonlinear transformation $\boldsymbol{f}$ but a zero covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{0}$, the mean of the transformed distribution is $\hat{\boldsymbol{\mu}} = \boldsymbol{f}(\boldsymbol{\mu})$. However, in the general case it is not possible to determine $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ of the $\boldsymbol{f}$-transformed distribution given $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ since the result depends on higher-order moments. Thus, the unscented transform is useful; it provides a mechanism to obtain this result via an approximation of the input distribution while assuming full knowledge of $\boldsymbol{f}$.

In computing the unscented transform, first a set of sigma points characterizing the input $P(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is chosen. The most common approach (Menegaz et al., 2015) is to take a set $\{\chi_i\}_{i=0}^{2n}$, $\chi_i \in \mathbb{R}^n$ of $2n+1$ symmetric points centered around the mean (incl. the mean), e.g. for $1 \leq i \leq n$,

$$\chi_0 = \boldsymbol{\mu} \,,$$
$$\chi_i = \boldsymbol{\mu} + \sqrt{(\kappa+n)\boldsymbol{\Sigma}}\big|_i \,, \qquad (5)$$
$$\chi_{i+n} = \boldsymbol{\mu} - \sqrt{(\kappa+n)\boldsymbol{\Sigma}}\big|_i \,,$$

where $\kappa > -n$ is a real constant and $\big|_i$ denotes the i-th column. The approximation in Eq. (5) is unbiased; the mean and covariance of the sigma points are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Thus, one can compute the transformation $\hat{\chi}_i = \boldsymbol{f}(\chi_i)$ and estimate the mean and covariance of the $\boldsymbol{f}$-transformed distribution

$$\hat{\boldsymbol{\mu}} = \tfrac{1}{2n+1} \sum_{i=0}^{2n} \hat{\chi}_i \,, \qquad (6)$$
$$\hat{\boldsymbol{\Sigma}} = \tfrac{1}{2n+1} \sum_{i=0}^{2n} (\hat{\chi}_i - \hat{\boldsymbol{\mu}})(\hat{\chi}_i - \hat{\boldsymbol{\mu}})^T \,. \qquad (7)$$

A visualization of the sigma points and their transformation is depicted in Fig. 2a. The procedure in Eq. (5-7) effectively applies the fully-known function $\boldsymbol{f}$ to an approximating set of points whose mean and covariance equal the original distribution's. Therefore, in the context of the commonly used VAE decoder nonlinearities, the mean and covariance of the transformed sigma points can be closer to the true transformed mean and covariance compared to the ones computed by propagating the same number of random samples from the original distribution.

4.2. Unscented Transform in the VAE

In an ELBO maximization setting from Eq. (3), the nonlinear transformation of the posterior in the decoder lends itself straightforwardly to the unscented transform approximation. Given any posterior defined by $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we can compute the sigma points (for example according to Eq. (5)) and provide them to the decoder. In a VAE, the sigma points provide a deterministic-sampling alternative to the reparameterization-trick-computed random samples of the latent space. Furthermore, computing the average reconstruction of the sigma points at the output of the decoder provides an approximation of the mean of the entire transformed posterior distribution in Eq. (6), while implicitly taking into account the variance in Eq. (7), as opposed to the per-sample reconstructions.

The choice of the number of sigma points provided to the decoder is similar to the sampling in Eq. (4), where one can realize a single latent vector with a single sample from $\mathcal{N}(\boldsymbol{0}, \mathbf{I})$ or multiple latents, resulting in a trade-off between reconstruction quality and computation demands (Ghosh et al., 2019). However, taking a single or few random samples in the VAE setting can produce instances very far from the mean, especially in high dimensional spaces. In contrast, sampling sigma points produces a more controlled overall estimate of the posterior (as well as producing a more accurate transformed posterior, see Eq. (6-7)) since the samples lie on the border of a hyperellipsoid induced by the covariance matrix $\boldsymbol{\Sigma}$ (example in Fig. 2b). Thus, while computing the loss function gradients (which are a function of the samples), the sigma-sampling has the potential to bring a more accurate and lower-variance estimate when all the sigma points are considered. This is illustrated in Fig 2c. Further empirical arguments validating the lower gradient variance claim are provided in Appendix B.

The sigma-sampling of the UT can be applied to any learned posterior described by its first two moments (as common in generative models), not only the VAE standard normal. With this description, the sigma points cannot be the uniquely optimal representation of the distribution since there is an infinite number of distributions that share the first two moments. However, the UT has shown superior empirical performance over other representations in extensive experiments in (Julier et al., 2000) and (Zhang et al., 2009), under various distributions and nonlinear functions, and especially for the case of differentiable functions. This has led to the UKF, built on this paradigm, being one of the major algorithms in filtering and control. Guided by the success of the method, we hypothesize that applying the UT in the VAE setting has the potential to, for a finite set of samples, provide a better approximation of the learned two-moment Gaussian posterior than the ubiquitous independent random sampling and reconstruction. With these insights, we develop the UAE model presented in the next section.
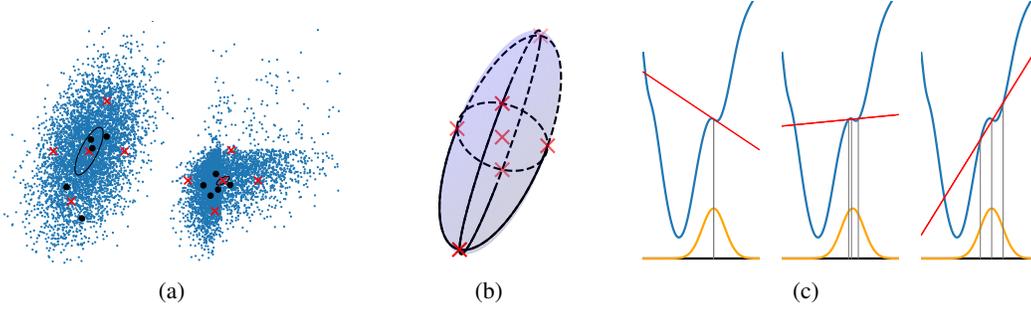
Figure 2: (best viewed in color) (a) **(transforming 2D sigma points)** Left: a Gaussian with its Monte Carlo approximation (blue), sigma points computed according to Eq. (5) (red), and five random samples (black points). Right: nonlinear RReLU activation (Xu et al., 2015) applied to the distribution, sigma points, and the random samples. In this example, the five sigma points provide a better approximation of the transformed distribution than the five random samples.
(b) **(3D sigma points)** Sigma points (red) on an ellipsoid spanned by a 3×3 covariance matrix, consisting of a central sigma point and a pair of sigma points on each axis.
(c) **(gradient variance)** Left: loss function (blue) at a sample (gray) corresponding to the standard normal (yellow) mean. The gradient of the loss function (red) at the mean is not representative of the true gradient. Middle: a high-variance gradient computed from the gradients at the three random samples drawn from the standard normal, potentially far away from the true gradient. Right: gradient of the loss function computed from the gradients at the three sigma points; although the estimate is potentially biased due to the applied nonlinear transformation, it has lower variance than if computed from the random points. The three provided examples can be interpreted as the RAE-(Ghosh et al., 2019), VAE-, and UAE-like sampling procedures.

5. Unscented Autoencoder (UAE)

The UAE is a deterministic-sampling autoencoder model maximizing the ELBO. It addresses the maximum likelihood optimization problem from Sec. 3, namely the $\mathcal{L}$ maximization from Eq. (3), by computing the UT of the posterior $q_\phi(\mathbf{z}|\mathbf{x})$ parameterized by the encoder $E_\phi(\mathbf{x}) = \{\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\Sigma}_\phi(\mathbf{x})\}$ (see Eq. (5-7)). The latent features $\mathbf{z}$ can be obtained by deterministically sampling multiple sigma points, resulting in a lower variance sampling than of the reparameterization trick in Eq. (4). Good performance of the model is further boosted by replacing the vanilla KL divergence with the Wasserstein distribution metric, which effectively performs a regularization of the posterior moments. The decoder regularization applies an additional smoothing effect on the latent space – it is formally derived in Sec. 5.2. The full training objective consists of optimizing $\phi, \theta \leftarrow \arg\min_{\phi,\theta} \mathcal{L}_{\text{UAE}}$,

$$\mathcal{L}_{\text{UAE}} = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} \mathcal{L}_{\text{REC}} + \beta \mathcal{L}_W + \gamma \mathcal{L}_{D_\theta \text{REG}} , \quad (8)$$

where β (from the β-VAE (Higgins et al., 2017)) and γ are weights.

The **reconstruction term** $\mathcal{L}_{\text{REC}}$ is an L_2 loss function incorporating the average of decoded sigma points

$$\mathcal{L}_{\text{REC}} = \|\mathbf{x} - \frac{1}{K} \sum_{k=1}^{K} D_\theta(\mathbf{z}_k)\|_2^2 , \\ \mathbf{z}_k \sim \{\chi_i(\boldsymbol{\mu}_\phi, \boldsymbol{\Sigma}_\phi)\}_{i=0}^{2n} , \quad (9)$$

where K n-dimensional vectors $\mathbf{z}_k$ are sampled from the set of sigma points, $K \leq 2n + 1$. Various sampling

heuristics are investigated in Appendix C. Note that this reconstruction loss function differs from the commonly used $\frac{1}{K} \sum_{k=1}^{K} \|\mathbf{x} - D_\theta(\mathbf{z}_k)\|_2^2$, where each decoded sample is matched to the ground truth. This strategy, employed in the standard multi-sample VAE, aims at getting the same output image for different samples thus demanding a certain attenuation property from the deterministic decoder. In contrast, Eq. (9) is motivated by the application of the UT in filtering where after propagating the sigma points through a nonlinear function a Gaussian is fit to the posterior (see Eq. (6-7)). By applying the loss to the mean output image, we essentially maintain a probability distribution at the output.

We use the **Wasserstein metric term** $\mathcal{L}_W$ as an alternative to the KL divergence. For a multivariate posterior and a multivariate normal prior, the KL divergence is defined as

$$\mathcal{L}_{\text{KL}} = \|\boldsymbol{\mu}_\phi\|_2^2 + \text{tr}(\boldsymbol{\Sigma}_\phi) - n - 2\text{tr}(\log \boldsymbol{L}_\phi) , \quad (10)$$

in the general case[4] of a full-covariance matrix $\boldsymbol{\Sigma}_\phi = \boldsymbol{L}_\phi \boldsymbol{L}_\phi^T$. Instead, due to favorable optimization properties and higher-quality reconstruction, we use the Wasserstein metric between distributions. This metric effectively replaces the covariance part of the KL term, $\text{tr}(\boldsymbol{\Sigma}_\phi) - 2\text{tr}(\log \boldsymbol{L}_\phi)$, with the squared Frobenius norm of the mismatch between the lower triangular matrix and the identity

---

[4]Derived from $D_{\text{KL}}(\mathcal{N}_0 \| \mathcal{N}_1) = \frac{1}{2}(\text{tr}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0) - n + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) + \log(\frac{\det \boldsymbol{\Sigma}_1}{\det \boldsymbol{\Sigma}_0}))$ for $\mathcal{N}_1(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$.

$$\mathcal{L}_W = \|\boldsymbol{L}_\phi - \mathbf{I}\|_F^2 = \mathrm{tr}(\boldsymbol{\Sigma}_\phi) - 2\mathrm{tr}(\boldsymbol{L}_\phi)\,. \quad (11)$$

It differs from the original objective in Eq. (10) only in the lack of a logarithm while sharing the same global minimum. Further details are provided in Sec. 5.3. Such a loss function allows the variance to approach zero (which is instead strongly penalized by the logarithm in Eq. (10)), yielding a sharper posterior.

The **decoder regularization term** $\mathcal{L}_{D_\theta\mathrm{REG}}$ is a generalization of the gradient penalty term in (Ghosh et al., 2019), accounting for a fully probabilistic formulation. It can be realized as a penalty on the input–output gradient of the posterior mean, weighted by the largest eigenvalue of the covariance matrix

$$\mathcal{L}_{D_\theta\mathrm{REG}} = \lambda_{\max}(\boldsymbol{\Sigma}_\phi)\|\nabla_{\boldsymbol{\mu}_\phi} D_\theta(\boldsymbol{\mu}_\phi)\|_2^2\,. \quad (12)$$

We approximate $\lambda_{\max}(\boldsymbol{\Sigma}_\phi)$ by the largest diagonal, which is correct for a diagonal $\boldsymbol{\Sigma}_\phi$.

We provide an overview of the VAE, RAE, and UAE loss functions in Tab. 1, together with the models that are conceptually between the VAE and UAE. Additional models employing different combinations of the loss function components are provided in Appendix D, Tab. 7.

5.1. Sampling From the Prior-Less UAE

Since the UAE model doesn't regularize the aggregated posterior toward the prior using the KL divergence (Hoffman & Johnson, 2016) or the Wasserstein metric (Patrini et al., 2020) (we use the per-posterior Wasserstein metric), it is not equipped with an easy-to-use sampling procedure as the VAE. To remedy this, we use the straightforward ex-post density estimation procedure described in (Ghosh et al., 2019) for the deterministic RAE model. We fit the latent means $\boldsymbol{\mu}_\phi$ for each input sample $\mathbf{x}$ to a 10-component Gaussian Mixture Model (GMM) (which has shown good performance and generalization ability in the experiments of (Ghosh et al., 2019) even for VAE models) and use the mixture to sample from the latent space. For a fair comparison, we utilize this procedure in all models.

5.2. ELBO Derivation

In the following, we analytically derive the UAE model in Eq. (8). The derivation is largely inspired from (Ghosh et al., 2019), with a few crucial differences allowing for greater generalizability and less restrictive assumptions. We start with the general ELBO minimization formulation in Eq. (3), augmented with a constraint

$$\arg\min_{\phi,\theta} \quad E_{x\sim p_{\mathrm{data}}}\mathcal{L}_{\mathrm{REC}} + \mathcal{L}_{\mathrm{KL}} \quad (13)$$

$$\mathrm{s.t.} \quad \|D_\theta(\mathbf{z}_1) - D_\theta(\mathbf{z}_2)\|_p < \epsilon,$$
$$\mathbf{z}_1, \mathbf{z}_2 \sim q_\phi(\mathbf{z}|\mathbf{x}),\ \forall \mathbf{x} \sim p_{\mathrm{data}}. \quad (14)$$

Here, the decoder outputs given any two latent vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ (any two draws from the posterior $q_\phi(\mathbf{z}|\mathbf{x})$) are bounded via their p-norm difference, for a deterministic decoder D_θ. It was shown in (Ghosh et al., 2019) that the constraint in Eq. (14) can be reformulated as

$$\sup\{\|\nabla_{\mathbf{z}} D_\theta(\mathbf{z})\|_p\} \cdot \sup\{\|\mathbf{z}_1 - \mathbf{z}_2\|_p\} < \epsilon\,. \quad (15)$$

We provide the full derivation in Appendix E. In Eq. (15), $\nabla_{\mathbf{z}} D_\theta(\mathbf{z})$ is the derivative of the decoder output w.r.t. its input (not the parameterization θ). The second term in the product depends on the parameterization of the posterior $q_\phi(\mathbf{z}|\mathbf{x})$. For a Gaussian, $\sup\{\|\mathbf{z}_1 - \mathbf{z}_2\|_p\}$ becomes a functional r of the posterior entropy, $r(\mathbb{H}(q_\phi(\mathbf{z}|\mathbf{x})))$. At this point, the RAE derivation from (Ghosh et al., 2019) takes a strong simplifying assumption of constant entropy for all samples $\mathbf{x}$, effectively asserting constant variance in the posterior. This allows to incorporate a simplified version of Eq. (15) into Eq. (13) via the Lagrange multiplier γ, obtaining the following RAE loss function[5]

$$\mathcal{L}_{\mathrm{RAE}} = \|\mathbf{x} - D_\theta(\mathbf{z})\|_2^2 + \beta\|\mathbf{z}\|_2^2 + \gamma\|\nabla_{\mathbf{z}} D_\theta(\mathbf{z})\|_2^2\,. \quad (16)$$

Here, the KL-term from Eq. (13) is approximated by $\|\mathbf{z}\|_2^2$ due to the constant variance assumption.

In the UAE formulation, the samples $\mathbf{z}_1$ and $\mathbf{z}_2$ in Eq. (15) simply correspond to the sigma points of $q_\phi(\mathbf{z}|\mathbf{x})$ parameterized by $E_\phi(\mathbf{x}) = \{\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\Sigma}_\phi(\mathbf{x})\}$. Therefore, the term $\sup\{\|\mathbf{z}_1 - \mathbf{z}_2\|_p\}$ can be computed analytically as the largest eigenvalue $\lambda_{\max}$ of the covariance matrix $\boldsymbol{\Sigma}_\phi$. We regularize the decoder in an RAE-manner around the posterior mean with $\|\nabla_{\boldsymbol{\mu}_\phi} D_\theta(\boldsymbol{\mu}_\phi)\|_p$ to enforce smoothness. Finally, the UAE does not require the constant variance assumption; we can incorporate a posterior KL-term or the Wasserstein metric used in Eq. (8). Thus, we arrive at the following analytical UAE loss function from Eq. (8)

$$\mathcal{L}_{\mathrm{UAE}} = E_{\mathbf{x}\sim p_{\mathrm{data}}}\mathcal{L}_{\mathrm{REC}} + \beta\mathcal{L}_{\mathrm{W}} + {} \\ + \gamma\lambda_{\max}(\boldsymbol{\Sigma}_\phi)\|\nabla_{\boldsymbol{\mu}_\phi} D_\theta(\boldsymbol{\mu}_\phi)\|_p\,, \quad (17)$$

where a more general form of the Eq. (15) constraint is used than in Eq. (16).

It follows from the derivation that the major difference between the RAE on the one hand and VAE and UAE on the other is that the RAE assumes constant variance in mapping the training data distribution into the latent space, thus not including any variance-compensating terms in the loss function. In effect, the RAE considers all the dimensions equally and cannot take into account that the encoder might have different uncertainty per dimension and data point.

---

[5]In (Ghosh et al., 2019), the decoder gradient penalty from Eq. (16) is the analytically derived regularization; alternatives such as weight decay and spectral norm are offered as well and can also be used in the UAE.

Table 1: A comparison of the VAE, RAE-GP (employing a Gradient Penalty (GP) on the decoder, a less general version of Eq. (12)), and UAE loss functions, including the intermediate models UT-VAE, VAE*, UT-VAE*, (weights omitted for clarity). UT-VAE uses the unscented transform in the VAE, VAE* uses the Wasserstein metric from Eq. (11), and UT-VAE* differs from the UAE only in the lack of a decoder regularization term. All models use a diagonal posterior representation (except RAE, which does not model uncertainty). The terms $\mathbf{z}$, $\boldsymbol{\mu}_\phi$, and $\boldsymbol{\sigma}_\phi$ are realized given the sample $\mathbf{x}$.

	Loss function	Posterior sampling
$\mathcal{L}_{\text{VAE}}$	$\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{x}-D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2-n+\sum_i\sigma_{\phi,i}^2-2\log\sigma_{\phi,i}$	$\mathbf{z}_k=\boldsymbol{\mu}_\phi+\boldsymbol{\sigma}_\phi\odot\boldsymbol{\epsilon}_k,\ \boldsymbol{\epsilon}_k\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
$\mathcal{L}_{\text{UT-VAE}}$	$\|\mathbf{x}-\frac{1}{K}\sum_{k=1}^{K}D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2-n+\sum_i\sigma_{\phi,i}^2-2\log\sigma_{\phi,i}$	$\mathbf{z}_k\sim\{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi,\text{diag}(\boldsymbol{\sigma}_\phi^2))\}_{i=0}^{2n}$
$\mathcal{L}_{\text{RAE-GP}}$	$\|\mathbf{x}-D_\theta(\mathbf{z})\|_2^2+\|\mathbf{z}\|_2^2+\|\nabla_\mathbf{z}D_\theta(\mathbf{z})\|_2^2$	None, $\mathbf{z}=\boldsymbol{\mu}_\phi$
$\mathcal{L}_{\text{VAE*}}$	$\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{x}-D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\|\text{diag}(\boldsymbol{\sigma}_\phi^2)-\mathbf{I}\|_F^2$	$\mathbf{z}_k=\boldsymbol{\mu}_\phi+\boldsymbol{\sigma}_\phi\odot\boldsymbol{\epsilon}_k,\ \boldsymbol{\epsilon}_k\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
$\mathcal{L}_{\text{UT-VAE*}}$	$\|\mathbf{x}-\frac{1}{K}\sum_{k=1}^{K}D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\|\text{diag}(\boldsymbol{\sigma}_\phi^2)-\mathbf{I}\|_F^2$	$\mathbf{z}_k\sim\{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi,\text{diag}(\boldsymbol{\sigma}_\phi^2))\}_{i=0}^{2n}$
$\mathcal{L}_{\text{UAE}}$	$\|\mathbf{x}-\frac{1}{K}\sum_{k=1}^{K}D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\|\text{diag}(\boldsymbol{\sigma}_\phi^2)-\mathbf{I}\|_F^2+\max(\boldsymbol{\sigma}_\phi^2)\|\nabla_{\boldsymbol{\mu}_\phi}D_\theta(\boldsymbol{\mu}_\phi)\|_2^2$	$\mathbf{z}_k\sim\{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi,\text{diag}(\boldsymbol{\sigma}_\phi^2))\}_{i=0}^{2n}$

Additionally, the difference between VAE and UAE is that the VAE incorporates a sampling procedure with higher variance than the deterministic sigma-point sampling used in the unscented transform. Therefore, loss function-wise, the UAE can be regarded as a middle-ground between the VAE and RAE – deterministic and lower-variance in training than the VAE, but with greater generalization capabilities than the RAE due to the probabilistic formulation.

5.3. Posterior Regularization via the Wasserstein Metric

The usage of the Wasserstein metric is motivated by practical properties of VAE model optimization. The training can be sensitive to the weighting of the KL divergence term, which can lead to posterior collapse (Dai et al., 2020). The main factor is the strong variance regularization of the KL divergence with its $\log$ term, which can be written as

$$\mathcal{L}_{\text{KL}} = \|\boldsymbol{\mu}_\phi\|_2^2 + \text{tr}(\boldsymbol{\Sigma}_\phi) - n - 2\sum_i \log L_{\phi,ii} \quad (18)$$

If the posterior gets more peaked, which might be necessary for good reconstructions, the divergence quickly grows toward infinity. We observed such problems in particular with full-covariance posteriors (see Appendix F).

Despite these problems the KL divergence is theoretically sound. It was shown in (Hoffman & Johnson, 2016) that $D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ can be reformulated into two terms, one that weakly pushes toward overlapping per-sample posterior distributions and a KL divergence between the aggregated posterior and the prior. The latter is required if samples are drawn from the prior and the former prevents the latent encoding from becoming a lookup table (Mathieu et al., 2019). Replacing the KL divergence with the Wasserstein-2 metric preserves the tendency toward overlapping posteriors, but does not match the aggregated posterior to a predefined prior. However, a simple connection can be found to such models, see Appendix G. Neverthe-

less, this matching is not required in our setup due to the ex-post density estimation. Furthermore, successful practical approaches like Stable Diffusion (Rombach et al., 2022) only require correctly learning the manifold and therefore do not need a certain aggregated posterior to sample from.

We use the Wasserstein-2 metric between two Gaussian distributions. Mathematically, it can be written as

$$\begin{aligned} W_2(\mathcal{N}_1,\mathcal{N}_2) &= \|\boldsymbol{\mu}_\phi\|_2^2 + \text{tr}(\boldsymbol{\Sigma}_\phi) + n - 2\text{tr}(\boldsymbol{\Sigma}_\phi^{1/2}) \\ &= \|\boldsymbol{\mu}_\phi\|_2^2 + \text{tr}(\boldsymbol{\Sigma}_\phi) + n - 2\text{tr}(\boldsymbol{L}_\phi) , \end{aligned} \quad (19)$$

for $\mathcal{N}_1 = \mathcal{N}(\boldsymbol{\mu}_\phi,\boldsymbol{\Sigma}_\phi)$ and $\mathcal{N}_2 = \mathcal{N}(\mathbf{0},\boldsymbol{I})$. The last three terms can be reformulated into Eq. (11)

$$\begin{aligned} \text{tr}(\boldsymbol{\Sigma}_\phi) + n - 2\text{tr}(\boldsymbol{L}_\phi) &= \text{tr}(\boldsymbol{L}_\phi^T\boldsymbol{L}_\phi - 2\boldsymbol{L}_\phi + \boldsymbol{I}) = \\ &= \text{tr}((\boldsymbol{L}_\phi - \mathbf{I})^T(\boldsymbol{L}_\phi - \mathbf{I})) = \|\boldsymbol{L}_\phi - \mathbf{I}\|_F^2 . \end{aligned} \quad (20)$$

Disregarding the constant terms, it is clear that Eq. (18) and Eq. (19) differ in the lack of the $\log$ term that infinitely penalizes zero-variance latents. In contrast, the Wasserstein metric even allows the posterior variance to approach zero if it helps to significantly reduce the reconstruction loss. This is evidenced in the aggregated posterior visualization of our model provided in Appendix H.

Naturally, the reduced reconstruction losses brought on by the *per-sample* Wasserstein metric in place of the KL divergence come at the cost of losing the ELBO formulation of the overall optimization problem. Furthermore, the Wasserstein distance between the *aggregated* posterior and the standard normal prior (Patrini et al., 2020) is not optimized either. Nevertheless, our empirical analysis shows that replacing the KL divergence with a Wasserstein metric regularization of the per-sample posterior results in significantly better reconstruction performance.

6. Results

In the following, we present quantitative and qualitative results of the UAE and its precursors compared to the VAE and RAE baselines on Fashion-MNIST (Xiao et al., 2017), CIFAR10 (Krizhevsky et al., 2009), and CelebA (Liu et al., 2015). We aim to delineate the effects of the UT (along with the reconstruction loss in Eq. (9)), Wasserstein metric, and the decoder regularization. Furthermore, we investigate multi-sampling and various sigma-point heuristics in Appendix C and ablate the entire loss function from Eq. (8) in Appendix D. In addition to evaluating the reconstruction and sampling quality (using a mixture for all models, see Sec. 5.1), we investigate if sampling only at the sigmas in training preserves the latent space structure (e.g. does not create 'holes') by evaluating interpolated samples. The metric is the widely-used FID (Heusel et al., 2017), which quantifies the distance between two distributions of images. Detailed information about the network architecture, training, and the choice of FID datasets is given in Appendix A.

The main results are provided in Tab. 2. The table is divided into three parts: the first part shows the effects of applying the Unscented Transform to the vanilla VAE model; the second part shows the baseline results of the RAE, while the third part shows the results of Wasserstein metric models. In the UT-VAE row of Tab. 2, *we tweak the VAE sampling to select instances at the sigma points while averaging the resulting images in the reconstruction loss, as consistent with the definition in Eq. (5-6). This simple change brings a remarkable near 40% improvement on Fashion-MNIST on average, near 15% on CIFAR10, and near 30% on CelebA.* It provides strong evidence that a higher-quality, lower-variance representation of the posterior distribution results in higher-quality decoded images.

The deterministic baseline RAE model in Tab. 2 sets the context with a significantly higher performance than the vanilla VAE. The Wasserstein metric of the VAE*, which preserves the latent space regularization in spirit of the RAE but extends it to a probabilistic, non-constant variance setting, can be considered close to the non-regularized RAE: outperforms it on CIFAR10 while being behind on Fashion-MNIST and CelebA. More importantly, the VAE* model also achieves a large improvement over the classical VAE in all metrics and on all datasets, achieved effectively only by replacing the logarithm term with a linear term. This indicates that the rigidity of the KL divergence w.r.t. posterior variance potentially harms the quality of decoded samples, particularly on the richer CIFAR10 and CelebA.

Observing the UT-VAE* row in Tab. 2, it can be seen that the unscented transform (UT) sampling in the VAE* context gives a further, albeit lesser boost in most metrics than with the KL divergence. Due to the Wasserstein metric's ability to shrink the posterior variance while approaching

convergence, the effect of any sampling is reduced. Nevertheless, it provides a considerable, approximately 10% boost on CelebA and Fashion-MNIST as well as a larger relative improvement with multiple samples than in VAE* (see Tab. 5, 6 in Appendix C). Finally, the generalized decoder regularization from Eq. (12) of the UAE applies a strong smoothing effect and further boosts the performance on CelebA and especially CIFAR10. Surprisingly, it yields a regression on Fashion-MNIST; similar effect of the gradient penalty harming the RAE performance compared to no-regularization is observable in (Ghosh et al., 2019) MNIST experiments. Overall, compared to the RAE, the UAE achieves significant improvements on CIFAR10 and a minor improvement on CelebA, while interestingly, the best model on Fashion-MNIST can be considered the UT-VAE.

In Tab. 3, we take a deeper look at the performance of the UT reconstruction loss term from Eq. (9). We empirically compare two strategies for designing the loss function: (i) use the mean reconstruction loss of images for each selected sample from the posterior (consistent with the standard VAE reconstruction loss) and (ii) apply the reconstruction loss to the mean image of samples from the posterior. Quantitative results in Tab. 3 consistently show the advantages of strategy (ii) for both the VAE and UT-VAE models using random samples and sigma points, respectively.

CelebA qualitative results are shown in Fig. 3 and reflect the FID scores: the UAE images appear similar to the RAE but significantly more realistic than the VAE. Fashion-MNIST and CIFAR10 images are provided in Appendix I.

7. Conclusion

In this paper, we introduced a novel VAE architecture employing the Unscented Transform, a lower-variance alternative to the reparameterization trick. We have challenged one of the core components of the VAE by showing that a sigma-point transform of the posterior significantly outperforms propagating random samples through the decoder. This was empirically shown for a small number of sigma points (2, 4, and 8) while taking more becomes impractical due to computationally-intensive training. Additionally, we proposed to use the Wasserstein metric, which does not optimize the ELBO. Although it can be considered as the main theoretical limitation of our model, it is a sound practical alternative to the KL divergence. By breaking its rigidity w.r.t. posterior variance, we unlocked performance improvements brought on by sharper posteriors that preserve a smooth latent space. Our work contributes an important step toward establishing competitive deterministic and deterministic-sampling generative models. Future work will thus focus on expanding the classes of supported generative models and on evaluation of further deterministic and quasi-deterministic sampling methods.

Table 2: Comparison of the architectures from Tab. 1. In all sampling instances, we select 8 random samples or sigma points. In the unscented transform models (UT-VAE, UT-VAE*, UAE), we select random sigma points on all datasets apart from CIFAR10, where pairs of sigma points along the largest eigenvalue axes are selected (see Appendix C). All RAE variants from (Ghosh et al., 2019) are provided: RAE-no-reg. without decoder regularization, RAE-GP with the Gradient Penalty (GP) from Eq. (16), RAE-L2 with decoder weight decay, and RAE-SN with spectral normalization.

	Fashion-MNIST			CIFAR10			CelebA		
	Rec.	Sample	Interp.	Rec.	Sample	Interp.	Rec.	Sample	Interp.
VAE_{8x}	44.29	48.73	61.99	110.0	120.6	118.3	65.86	68.53	68.75
UT-VAE_{8x}	27.79	**30.39**	**39.92**	91.04	111.7	104.3	50.11	54.15	54.32
RAE-no-reg.	21.56	34.79	50.27	86.79	102.1	96.80	40.79	47.88	49.97
RAE-GP	22.91	33.80	50.74	85.70	100.7	96.06	39.89	46.67	46.18
RAE-L2	**20.28**	32.06	48.52	84.27	99.26	94.23	38.78	46.44	50.33
RAE-SN	21.40	33.50	49.60	85.75	101.1	96.48	41.23	48.39	50.23
VAE*_{8x}	27.36	36.63	52.61	82.22	99.11	92.84	45.02	50.81	53.64
UT-VAE*_{8x}	23.64	31.51	48.06	81.12	100.6	93.80	40.18	47.39	49.62
UAE_{8x}	25.07	35.19	54.24	**71.97**	**89.91**	**83.50**	**38.48**	**45.60**	**45.88**

Table 3: Comparison of a VAE model using the reconstruction loss of the mean image of random samples from the posterior: $\|\mathbf{x} - \frac{1}{K}\sum_{k=1}^{K} D_\theta(\mathbf{z}_k)\|_2^2$, $\mathbf{z}_k = \boldsymbol{\mu}_\phi + \boldsymbol{\sigma}_\phi \odot \boldsymbol{\epsilon}_k$, $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, denoted by $\text{VAE}_{2x}^\dagger$, and a model with the mean reconstruction loss of sigma points from the posterior: $\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{x} - D_\theta(\mathbf{z}_k)\|_2^2$, $\mathbf{z}_k \sim \{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi, \text{diag}(\boldsymbol{\sigma}_\phi^2))\}_{i=0}^{2n}$, denoted by $\text{UT-VAE}_{2x}^\ddagger$. The UT-VAE_{2x} uses the full unscented transform with the reconstruction loss of the mean image of sigma points from the posterior: $\|\mathbf{x} - \frac{1}{K}\sum_{k=1}^{K} D_\theta(\mathbf{z}_k)\|_2^2$, $\mathbf{z}_k \sim \{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi, \text{diag}(\boldsymbol{\sigma}_\phi^2))\}_{i=0}^{2n}$, as consistent with the Unscented Transform in Eq. (5-6). In the sigma-point variants of $\text{UT-VAE}_{2x}^\ddagger$ and UT-VAE_{2x}, random sigma points are selected for Fashion-MNIST and CelebA, while largest-eigenvalue pairs are used in CIFAR10.

	Fashion-MNIST			CIFAR10			CelebA		
	Rec.	Sample	Interp.	Rec.	Sample	Interp.	Rec.	Sample	Interp.
VAE_{2x}	43.66	49.01	61.03	112.7	123.2	120.6	67.29	69.92	70.00
$\text{VAE}_{2x}^\dagger$	42.22	47.33	59.47	110.0	121.6	118.6	61.71	65.77	65.29
$\text{UT-VAE}_{2x}^\ddagger$	46.79	52.87	74.11	115.2	128.2	124.7	54.61	61.03	59.49
UT-VAE_{2x}	**36.25**	**40.30**	**53.10**	**95.70**	**115.4**	**107.3**	**51.61**	**57.42**	**56.56**
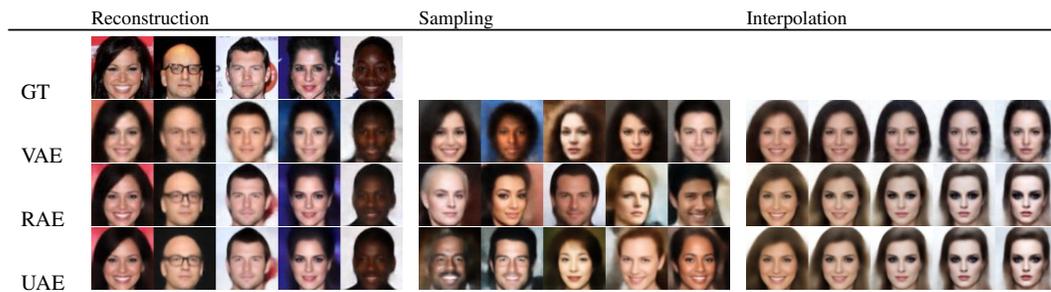
Figure 3: Qualitative results on the CelebA dataset of the VAE_{8x}, RAE-L2, and UAE_{8x} models.

References

Accardi, L. De Finetti Theorem. *Hazewinkel, Michiel, Encyclopaedia of Mathematics, Kluwer Academic Publishers*, 2001.

Bauer, M. and Mnih, A. Resampled Priors for Variational Autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 66–75. PMLR, 2019.

Bengio, Y., Courville, A. C., and Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50. URL https://doi.org/10.1109/TPAMI.2013.50.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. Generating Sentences From a Continuous Space. *arXiv preprint arXiv:1511.06349*, 2015.

Burda, Y., Grosse, R. B., and Salakhutdinov, R. Importance Weighted Autoencoders. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1509.00519.

Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., and Abbeel, P. Variational Lossy Autoencoder. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=BysvGP5ee.

Cremer, C., Morris, Q., and Duvenaud, D. Reinterpreting Importance-Weighted Autoencoders. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Syw2ZgrFx.

Dai, B. and Wipf, D. Diagnosing and Enhancing VAE Models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1e0X3C9tQ.

Dai, B., Wang, Y., Aston, J., Hua, G., and Wipf, D. 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.

Dai, B., Wang, Z., and Wipf, D. The Usual Suspects? Reassessing Blame for VAE Posterior Collapse. In *International Conference on Machine Learning*, pp. 2313–2322. PMLR, 2020.

Doucet, A. and Johansen, A. M. A Tutorial On Particle Filtering and Smoothing: Fifteen Years Later. *Oxford Handbook of Nonlinear Filtering*, 2011.

Ghosh, P., Sajjadi, M. S., Vergari, A., Black, M., and Schölkopf, B. From Variational to Deterministic Autoencoders. *arXiv preprint arXiv:1903.12436*, 2019.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative Adversarial Networks. *CoRR*, abs/1406.2661, 2014. URL http://arxiv.org/abs/1406.2661.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., and Hochreiter, S. GANs Trained by a Two Time-scale Update Rule Converge to a Nash Equilibrium. *arXiv preprint arXiv:1706.08500*, 12(1), 2017.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. Beta-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. URL https://openreview.net/forum?id=Sy2fzU9gl.

Hoffman, M. D. and Johnson, M. J. ELBO Surgery: Yet Another Way to Carve Up the Evidence Lower Bound. In *Proc. Workshop Adv. Approx. Bayesian Inference*, pp. 2, 2016.

Julier, S., Uhlmann, J., and Durrant-Whyte, H. F. A New Method for the Nonlinear Transformation of Means and Covariances In Filters and Estimators. *IEEE Transactions on automatic control*, 45(3):477–482, 2000.

Karl, M., Soelch, M., Bayer, J., and Van der Smagt, P. Deep Variational Bayes Filters: Unsupervised Learning of State Space Models From Raw Data. *arXiv preprint arXiv:1605.06432*, 2016.

Khromov, G. and Singh, S. P. Some fundamental aspects about lipschitz continuity of neural network functions, 2023.

Kingma, D. P. and Welling, M. Auto-encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kingma, D. P., Salimans, T., and Welling, M. Variational Dropout and the Local Reparameterization Trick. *Advances in neural information processing systems*, 28, 2015.

Krizhevsky, A., Hinton, G., et al. Learning Multiple Layers of Features From Tiny Images. 2009.

Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. Grammar Variational Autoencoder. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1945–1954. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/kusner17a.html.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep Learning Face Attributes In the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Lucas, J., Tucker, G., Grosse, R. B., and Norouzi, M. Understanding Posterior Collapse In Generative Latent Variable Models. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=r1xaVLUYuE.

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. Disentangling Disentanglement In Variational Autoencoders. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4402–4412. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/mathieu19a.html.

Menegaz, H. M., Ishihara, J. Y., Borges, G. A., and Vargas, A. N. A Systematization of the Unscented Kalman Filter Theory. *IEEE Transactions on automatic control*, 60 (10):2583–2598, 2015.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An Imperative Style, High-performance Deep Learning Library. *Advances in neural information processing systems*, 32, 2019.

Patrini, G., van den Berg, R., Forre, P., Carioni, M., Bhargav, S., Welling, M., Genewein, T., and Nielsen, F. Sinkhorn Autoencoders. In *Uncertainty in Artificial Intelligence*, pp. 733–743. PMLR, 2020.

Rainforth, T., Kosiorek, A., Le, T. A., Maddison, C., Igl, M., Wood, F., and Teh, Y. W. Tighter Variational Bounds Are Not Necessarily Better. In *International Conference on Machine Learning*, pp. 4277–4285. PMLR, 2018.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic Backpropagation and Approximate Inference In Deep Generative Models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. II–1278–II–1286. JMLR.org, 2014.

Rolinek, M., Zietlow, D., and Martius, G. Variational Autoencoders Pursue PCA Directions (by Accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12406–12415, 2019.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Seitzer, M. Pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.2.1.

Tolstikhin, I. O., Bousquet, O., Gelly, S., and Schölkopf, B. Wasserstein Auto-Encoders. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=HkL7n1-0b.

Townsend, J., Bird, T., and Barber, D. Practical Lossless Compression with Latent Variables Using Bits Back Coding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=ryE98iR5tm.

Tripp, A., Daxberger, E. A., and Hernández-Lobato, J. M. Sample-Efficient Optimization In the Latent Space of Deep Generative Models Via Weighted Retraining. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Tucker, G., Lawson, D., Gu, S., and Maddison, C. J. Doubly Reparameterized Gradient Estimators for Monte Carlo Objectives. *arXiv preprint arXiv:1810.04152*, 2018.

Uhlmann, J. *Dynamic Map Building and Localization: New Theoretical Foundations*. PhD thesis, University of Oxford, 1995.

Vahdat, A. and Kautz, J. NVAE: A Deep Hierarchical Variational Autoencoder. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xu, B., Wang, N., Chen, T., and Li, M. Empirical Evaluation of Rectified Activations In Convolutional Network. *arXiv preprint arXiv:1505.00853*, 2015.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. Advances In Variational Inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

Zhang, W., Liu, M., and Zhao, Z.-g. Accuracy Analysis of Unscented Transformation of Several Sampling Strategies. In *2009 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing*, pp. 377–380. IEEE, 2009.

Zietlow, D., Rolinek, M., and Martius, G. Demystifying Inductive Biases for (Beta-) VAE Based Architectures. In *International Conference on Machine Learning*, pp. 12945–12954. PMLR, 2021.

Appendix

A. Network Architecture and Training

Table 4: Network architectures of the implemented VAE, RAE, and UAE models. Batch dimensions omitted for clarity.

VAE, UAE: $\mathbf{x}_{C \times W \times H} \to \text{ENCODER} \to \{\text{FC}_{1024 \times n} : \boldsymbol{\mu}_\phi, \text{FC}_{1024 \times n} : \log \boldsymbol{\sigma}_\phi^2\} \to \mathbf{z} \to \text{DECODER} \to \hat{\mathbf{x}}$
RAE: $\mathbf{x}_{C \times W \times H} \to \text{ENCODER} \to \{\text{FC}_{1024 \times n} : \mathbf{z}_\phi\} \to \text{DECODER} \to \hat{\mathbf{x}}$
ENCODER: $\text{CONV}_{32 \times 64} \to \text{CONV}_{64 \times 128} \to \text{CONV}_{128 \times 256} \to \text{CONV}_{256 \times 512} \to \text{CONV}_{512 \times 1024} \to \text{FLATTEN}$
DECODER: $\text{FC}_{n \times 1024 \cdot 8 \cdot 8} \to \text{TCONV}_{1024 \times 512} \to \text{TCONV}_{512 \times 256} [\to \text{TCONV}_{256 \times 128}]^{\text{CelebA}} \to \text{TCONV}_{256 \text{ or } 128 \times C}$
MNIST: $C = 1, W = H = 32, n = 64$ CIFAR10: $C = 3, W = H = 32, n = 128$ CELEBA: $C = 3, W = H = 64, n = 64$

Network architectures are given in Tab. 4 and largely follow the architecture in (Ghosh et al., 2019). For consistency, all models share the same encoder/decoder structure. All encoder 2D convolution blocks contain 3×3 kernels, stride 2, and padding 1, followed by a 2D batch normalization and a Leaky-ReLU activation. The decoder transposed convolutions share the same parameters as the encoder convolutions apart from using a 4×4 kernel. The last transposed convolution (mapping to channel dimension) however has a 3×3 kernel and is followed by a tanh activation (without batch normalization).

The dataset preprocessing procedure is the following. The Fashion-MNIST images are scaled from 28×28 to 32×32. For the training dataset, we use 50k out of the 60k provided examples, leaving the remaining 10k for the validation dataset. For the test dataset, we use the provided examples. In CIFAR10, we perform a random horizontal flip on the training data followed by a normalization for all dataset subsets. We use the same training/validation/test split method as in Fashion-MNIST. In CelebA, we perform a 148×148 center crop and resize the images to 64×64. We use the provided training/validation/testing subsets.

All models are implemented in PyTorch (Paszke et al., 2019) and use the library provided in (Seitzer, 2020) for FID computation. The models are trained for 100 epochs, starting with a 0.005 learning rate that is then halved after every five epochs without improvement. The weights used in the loss functions are the following: KL-divergence (or the Wasserstein metric) terms are weighted with $\beta = 2.5e^{-4}$ in the case of VAE and UAE and $\beta = 1e^{-4}$ for the RAE. The decoder regularization terms are weighted with $\gamma = 1e^{-6}$ for both RAE and UAE. We performed minimal hyperparameter search over the weights.

In computing the FID scores, we follow the same procedure as in (Ghosh et al., 2019). In the three cases of reconstruction, sampling, and interpolation, we evaluate the FID to the test set image reconstructions as the ground-truth. In the reconstruction metric, we use the validation set image reconstructions. In sampling, we fit the training dataset latent features to a GMM (see Sec. 5.1) and sample and reconstruct the same number of elements as in the test set. In interpolation, we apply mid-point spherical interpolation between a random pair of validation set embeddings. In all cases, we generate a single image per input; this image corresponds to the posterior mean of the latent distribution. This mean latent feature vector is also used in sampling and interpolation while fitting a mixture ex-post or interpolating the latent space vectors. Thus, the resulting number of generated images for FID computation is the same regardless of the number of sigma points or samples used in training. In all experiments, the average FID score of three runs is reported, while observing a similar variation between scores of individual runs among the models employing the UT compared to the vanilla VAE. In contrast, the scores of RAE and VAE* modes were significantly more consistent.

The network architectures largely follow the structure adopted by (Ghosh et al., 2019), with the difference of the added first two encoder layers. Nevertheless, in Tab. 2, we did not manage to reproduce the FID values reported in (Ghosh et al., 2019) on CelebA and CIFAR10, even observing that removing the first two encoder layers reduces the overall performance. We suspect that it is due to the differing Tensorflow and PyTorch model implementations as well as the FID computation libraries. However, in most cases, our implementation of the RAE attains a significantly larger performance gain over the VAE than reported in (Ghosh et al., 2019).
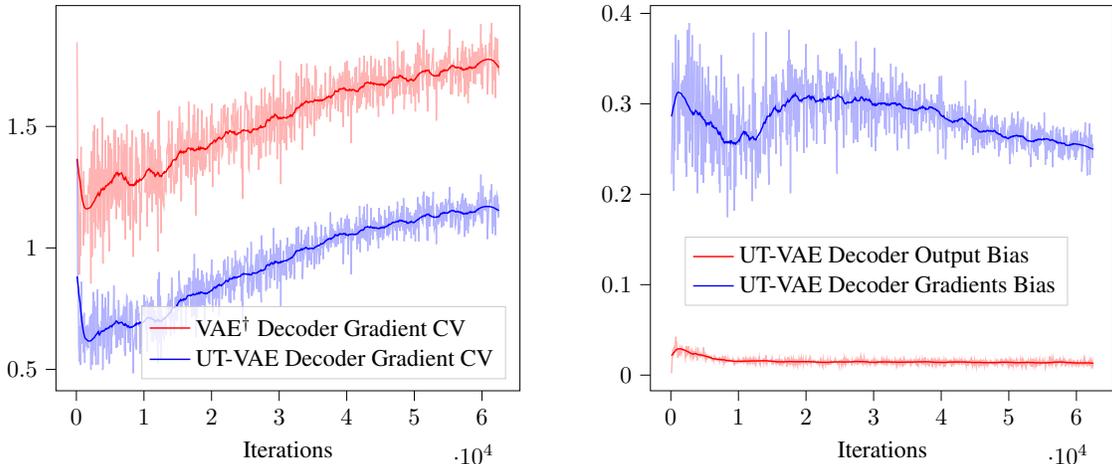
(a) Median of the decoder gradient CV for UT-VAE and VAE†

(b) Median relative bias based on an estimate of the true gradient using 200 random samples

Figure 4: Comparison of the variance and bias trade-off for the VAE† (employing the decoder output mean instead of the sample mean, see Tab. 3) and UT-VAE across approx. 60k training steps (100 epochs) on the CIFAR10 dataset. The data is based on a single training of an UT-VAE where every 50th epoch the gradient variance and bias was estimated using different sampling schemes. In case of VAE†, two random points are sampled (in accordance with the reparameterization trick), while in case of UT-VAE, a single sigma point pair is sampled.

B. Gradient Variance and Bias

In this section, we investigate the gradient variance and bias of the proposed base UT-VAE model. Compared to random sampling of the reparameterization trick, using a different integration scheme like sampling sigma points can be biased. It can nevertheless achieve lower variance depending on the nonlinear function of the decoder. Thus, for our decoder setup, we compare the gradient variance and bias of the UT-VAE (with random sigma pair sampling) and the VAE† (with random sampling) employing the decoder output mean instead of the sample mean[6] (see Tab. 3 for a performance comparison) in order to isolate the effect of sampling sigma points.

We train both models and estimate the gradient variance and bias every 50th iteration. For UT-VAE we independently sample 50 sigma point pairs, pass them through the decoder, and calculate the gradients' mean m_j and standard deviation σ_j. For VAE† we draw 2 random samples 200 times and perform the same steps to obtain m'_j and σ'_j. We calculate the median Coefficient of Variation (CV) of the gradients for both models, assuming that m'_j computed with 200 random samples is a good enough estimate of the true gradient. Furthermore, we compute the median relative bias b_{rel} for the decoder gradients and output of the UT-VAE. The CV (for UT-VAE) and b_{rel} (for decoder gradients bias) are computed as follows

$$\text{CV} = \text{median}\left\{\frac{\sigma_j}{|m_j|}\right\} \qquad b_{rel} = \text{median}\left\{\frac{|m_j - m'_j|}{|m'_j|}\right\}. \tag{21}$$

The gradient variance results are depicted in Fig. 4a. The variance of the sigma pair sampling of the UT-VAE is consistently lower than the gradient variance of the random sampling within VAE†. Interestingly, for the VAE† the standard deviation of the gradients is on average larger than the magnitude of the gradient during the whole training, whereas for the UT-VAE this is only the case at the end of the training. Fig. 4b shows the relative decoder output bias as well as the relative gradient bias of the UT-VAE at the same iterations. Whereas the relative bias at the decoder output is below 3% throughout the whole training, the bias of the gradients is around 30% of their magnitude. It is unclear whether such a substantial gradient bias is behind the good performance of the UT-VAE or if there is a performance trade-off between variance and bias. Nevertheless, our experiments show that, under a common decoder architecture, integration schemes like the UT can exhibit lower variance and higher bias while outperforming the standard VAE sampling scheme. Thus, investigating alternative integration schemes for VAEs can be a promising research direction.

---

[6]Reconstruction loss function of the VAE†: $\|\mathbf{x} - \frac{1}{K}\sum_{k=1}^{K} D_\theta(\mathbf{z}_k)\|_2^2$, $\mathbf{z}_k = \boldsymbol{\mu}_\phi + \boldsymbol{\sigma}_\phi \odot \boldsymbol{\epsilon}_k$, $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

Table 5: Analysis of the number of sampled sigma points and different heuristics, where the mean image of multiple sigma points is matched to the ground truth in the reconstruction loss. The three investigated heuristics are sampling random sigma points, random pairs of sigma points along an axis, and pairs of sigma points along axes with largest eigenvalues.

	Fashion-MNIST			CIFAR10			CelebA		
	Rec.	Samp.	Interp.	Rec.	Samp.	Interp.	Rec.	Samp.	Interp.
UT-VAE$_{1x,rand.}$	47.27	52.10	67.16	119.9	129.8	127.9	55.93	62.13	60.54
UT-VAE$_{2x,rand.}$	36.25	40.30	53.10	111.5	124.7	121.0	51.61	57.42	56.56
UT-VAE$_{4x,rand.}$	32.13	36.41	47.30	105.9	119.8	115.9	50.85	55.82	55.99
UT-VAE$_{8x,rand.}$	27.79	**30.39**	**39.92**	95.40	110.8	106.4	50.11	54.15	44.32
UT-VAE*$_{2x,rand.}$	28.26	36.36	50.69	85.88	103.7	96.90	44.32	50.33	52.40
UT-VAE*$_{4x,rand.}$	24.38	32.75	49.40	81.99	100.6	93.52	42.52	49.21	51.35
UT-VAE*$_{8x,rand.}$	**23.64**	31.51	48.06	81.10	99.87	92.48	**40.18**	**47.39**	**49.62**
UT-VAE$_{2x,rand. pairs}$	102.1	115.1	112.8	102.3	119.6	114.0	150.0	150.4	151.3
UT-VAE$_{4x,rand. pairs}$	96.85	110.1	107.3	101.0	119.5	113.4	224.3	225.0	225.4
UT-VAE$_{8x,rand. pairs}$	90.14	103.6	101.5	100.3	119.2	113.2	173.2	175.4	175.8
UT-VAE*$_{2x,rand. pairs}$	32.66	38.68	58.72	85.64	102.3	97.00	45.96	53.16	51.49
UT-VAE*$_{4x,rand. pairs}$	32.85	38.58	57.70	84.62	102.2	96.14	252.9	254.8	253.8
UT-VAE*$_{8x,rand. pairs}$	30.65	36.88	56.42	**80.51**	**98.40**	**91.96**	141.9	144.3	147.4
UT-VAE$_{2x,larg. \lambda pairs}$	106.6	118.6	115.7	95.70	115.4	107.3	54.02	60.29	60.26
UT-VAE$_{4x,larg. \lambda pairs}$	108.3	120.1	117.2	92.56	111.6	104.2	46.37	53.53	52.62
UT-VAE$_{8x,larg. \lambda pairs}$	115.5	128.8	126.3	91.04	111.7	104.3	48.59	55.22	55.29
UT-VAE*$_{2x,larg. \lambda pairs}$	33.49	42.63	61.57	82.17	100.7	93.80	55.57	61.42	61.53
UT-VAE*$_{4x,larg. \lambda pairs}$	34.94	43.18	67.65	81.61	101.3	94.11	48.41	54.70	54.80
UT-VAE*$_{8x,larg. \lambda pairs}$	31.08	41.06	64.58	81.12	100.6	93.80	45.08	51.45	52.05

C. Additional Results: Multi-Sigma Heuristics and Multi-Sample Models

The UT-VAE loss function defined in Tab. 1 samples K sigma points in the reconstruction term. Increasing the number of sigma points (up to $2n + 1$) improves the estimate of the transformed posterior distribution and thus the resulting reconstruction quality, at the expense of an approximately linear increase in training time. We observed this in most cases when training on 2, 4, and 8 sigma points, see Tab. 5. However, a much larger number of sigma points might not result in expected additional performance improvement due to significantly larger batch size, which could be mitigated by constructing approaches to select and train on a fixed, smaller batch size.

For K selected sigma points, various strategies can be used instead of sampling a discrete uniform distribution. For example, only pairs of sigma points along an axis can be chosen, conveying the width of the posterior distribution in the given dimension. This strategy can be adapted to select pairs along axes with largest eigenvalues. Tab. 5 also explores different sampling heuristics in the case of UT-VAE and UT-VAE*. We have observed that models trained with KL divergence exhibit larger variation in results w.r.t. the sampling heuristic, which is reasonable since the Wasserstein metric's posterior variance suppression diminishes the effect of sampling. The choice of the sigma-point selection heuristic turns out to have a large effect on the overall performance given a dataset. We have observed that a random selection of sigma points performs consistently well across all datasets while selecting random pairs generates reasonable results only in the case of CIFAR10. Interestingly, random-pairs performs very poorly on Fashion-MNIST and CelebA while largest eigenvalue pairs shows very good performance in the UT-VAE case on CIFAR10. In the main experiments of Tab. 2, we used a random selection for the Fashion-MNIST and CelebA models and largest-eigenvalue pairs for CIFAR10, due to its superior performance in the UT-VAE case.

Tab. 6 analyzes models using multiple samples in training. We compare the VAE* and the UAE with the classical VAE and the IWAE (Burda et al., 2016) as a baseline where multiple importance-weighted posterior samples help achieve a tighter lower bound. Observing the results, it is clear that models employing the Wasserstein metric can benefit from increasing the number of samples in training despite their ability to reduce the latent space variance, while significantly outperforming the baselines.

Table 6: Comparison of models employing multiple samples in training. The UAE uses random sigma points on Fashion-MNIST and CelebA and largest-eigenvalue pairs on CIFAR10.

	Fashion-MNIST			CIFAR10			CelebA		
	Rec.	Sample	Interp.	Rec.	Sample	Interp.	Rec.	Sample	Interp.
VAE_{1x}	45.64	49.99	61.33	116.4	126.8	124.2	68.32	71.05	71.16
VAE_{2x}	43.66	49.01	61.03	112.7	123.2	120.6	67.29	69.92	70.00
VAE_{4x}	44.94	49.51	62.29	111.7	121.3	119.5	66.32	68.87	69.06
VAE_{8x}	44.29	48.73	61.99	110.0	120.6	118.3	65.86	68.53	68.75
$IWAE_{1x}$	49.27	53.71	64.50	111.7	121.6	119.6	68.28	71.16	71.17
$IWAE_{2x}$	48.21	53.11	65.69	112.1	122.4	119.8	66.85	69.81	69.74
$IWAE_{4x}$	47.40	51.77	64.10	110.6	120.6	118.2	66.01	68.82	68.90
$IWAE_{8x}$	46.16	50.91	63.68	108.9	118.9	116.9	64.83	67.96	67.86
$VAE*_{1x}$	31.62	38.44	52.33	83.49	101.5	94.56	44.69	50.55	53.18
$VAE*_{2x}$	30.07	37.92	**52.15**	84.57	102.2	95.61	45.18	50.97	53.73
$VAE*_{4x}$	28.98	41.35	52.17	84.64	102.3	95.96	45.03	50.59	53.32
$VAE*_{8x}$	27.36	36.63	52.61	82.22	99.11	92.84	45.02	50.81	53.64
UAE_{2x}	29.29	37.59	53.69	77.71	96.37	89.71	40.07	47.28	50.51
UAE_{4x}	27.11	38.03	53.11	75.63	93.02	86.41	39.48	46.35	50.94
UAE_{8x}	**25.07**	**35.19**	54.24	**71.97**	**89.91**	**83.50**	**38.48**	**45.60**	**45.88**

D. Additional Results: Ablation Study of the Loss Components

This section provides an additional ablation study of the loss components used in the UAE model. The loss functions considered are provided in the upper half of Tab. 7 and the obtained results are in Tab. 8. There are three dimensions along which the results can be interpreted: Wasserstein metric, unscented transform, and the generalized decoder regularization (gradient penalty).

Tab. 8 is divided into two parts: the top part models use the analytical form of the KL divergence in Eq. (10) while the bottom part use the Frobenius norm mismatch derived from the Wasserstein metric in Eq. (11). It is clearly visible that the latter models strongly outperform the former, in all datasets and configurations. The loss function allows for a sharper posterior and thus larger expressiveness of the model (see Appendix H).

Similarly, the unscented transform models UT-VAE and UT-VAE* clearly outperform the random sampling and per-sample reconstruction counterparts of VAE and VAE*. In the latter case, the differences are smaller due to the sharper posterior of the VAE*. An ablation study of the unscented transform components can be found in Tab. 3.

Considering the gradient penalty models, interesting interplays can be noticed. Applying the decoder regularization on the vanilla VAE and the VAE* (this model can be considered closest to the RAE-GP) brings only minor improvements in the case of CIFAR10 and CelebA for each of the models respectively. The strong smoothing of the latent space however seems detrimental when combined with the unscented transform and the KL divergence training. One can conclude that only the latent space regularization models (such as the Wasserstein metric VAE* or the deterministic RAE) can benefit from decoder regularization. Furthermore, the effect appears to be dataset-dependent since the Fashion-MNIST VAE* and UT-VAE* slightly regress when augmented with decoder regularization.

Table 7: The loss functions used for the models in Tab. 8 and Tab. 9. The upper and lower half of the table contain diagonal and full-covariance posterior models, respectively.

	Loss function	Posterior sampling
$\mathcal{L}_{\text{VAE}}$	$\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{x}-D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2-n+\sum_i\sigma_{\phi,i}^2-2\log\sigma_{\phi,i}$	$\mathbf{z}_k=\boldsymbol{\mu}_\phi+\boldsymbol{\sigma}_\phi\odot\boldsymbol{\epsilon}_k,\ \boldsymbol{\epsilon}_k\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
$\mathcal{L}_{\text{VAE-GP}}$	$\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{x}-D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\sum_i\sigma_{\phi,i}^2-2\log\sigma_{\phi,i}+\max(\boldsymbol{\sigma}_\phi)\|\nabla_{\boldsymbol{\mu}_\phi}D_\theta(\boldsymbol{\mu}_\phi)\|_2^2$	$\mathbf{z}_k=\boldsymbol{\mu}_\phi+\boldsymbol{\sigma}_\phi\odot\boldsymbol{\epsilon}_k,\ \boldsymbol{\epsilon}_k\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
$\mathcal{L}_{\text{UT-VAE}}$	$\|\mathbf{x}-\frac{1}{K}\sum_{k=1}^{K}D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\sum_i\sigma_{\phi,i}^2-2\log\sigma_{\phi,i}$	$\mathbf{z}_k\sim\{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi,\text{diag}(\boldsymbol{\sigma}_\phi^2))\}_{i=0}^{2n}$
$\mathcal{L}_{\text{UT-VAE-GP}}$	$\|\mathbf{x}-\frac{1}{K}\sum_{k=1}^{K}D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\sum_i\sigma_{\phi,i}^2-2\log\sigma_{\phi,i}+\max(\boldsymbol{\sigma}_\phi)\|\nabla_{\boldsymbol{\mu}_\phi}D_\theta(\boldsymbol{\mu}_\phi)\|_2^2$	$\mathbf{z}_k\sim\{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi,\text{diag}(\boldsymbol{\sigma}_\phi^2))\}_{i=0}^{2n}$
$\mathcal{L}_{\text{VAE*}}$	$\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{x}-D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\|\text{diag}(\boldsymbol{\sigma}_\phi^2)-\mathbf{I}\|_F^2$	$\mathbf{z}_k=\boldsymbol{\mu}_\phi+\boldsymbol{\sigma}_\phi\odot\boldsymbol{\epsilon}_k,\ \boldsymbol{\epsilon}_k\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
$\mathcal{L}_{\text{VAE*-GP}}$	$\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{x}-D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\|\text{diag}(\boldsymbol{\sigma}_\phi^2)-\mathbf{I}\|_F^2+\max(\boldsymbol{\sigma}_\phi)\|\nabla_{\boldsymbol{\mu}_\phi}D_\theta(\boldsymbol{\mu}_\phi)\|_2^2$	$\mathbf{z}_k=\boldsymbol{\mu}_\phi+\boldsymbol{\sigma}_\phi\odot\boldsymbol{\epsilon}_k,\ \boldsymbol{\epsilon}_k\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
$\mathcal{L}_{\text{UT-VAE*}}$	$\|\mathbf{x}-\frac{1}{K}\sum_{k=1}^{K}D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\|\text{diag}(\boldsymbol{\sigma}_\phi^2)-\mathbf{I}\|_F^2$	$\mathbf{z}_k\sim\{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi,\text{diag}(\boldsymbol{\sigma}_\phi^2))\}_{i=0}^{2n}$
$\mathcal{L}_{\text{UT-VAE*-GP}}$	$\|\mathbf{x}-\frac{1}{K}\sum_{k=1}^{K}D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\|\text{diag}(\boldsymbol{\sigma}_\phi^2)-\mathbf{I}\|_F^2+\max(\boldsymbol{\sigma}_\phi)\|\nabla_{\boldsymbol{\mu}_\phi}D_\theta(\boldsymbol{\mu}_\phi)\|_2^2$	$\mathbf{z}_k\sim\{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi,\text{diag}(\boldsymbol{\sigma}_\phi^2))\}_{i=0}^{2n}$
$\mathcal{L}_{\text{VAE-full}\boldsymbol{\Sigma}_\phi}$	$\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{x}-D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\text{tr}(\boldsymbol{\Sigma}_\phi)-2\text{tr}(\log\boldsymbol{L}_\phi)$	$\mathbf{z}_k=\boldsymbol{\mu}_\phi+\boldsymbol{L}_\phi\boldsymbol{\epsilon}_k,\ \ \boldsymbol{\epsilon}_k\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
$\mathcal{L}_{\text{VAE-full}\boldsymbol{\Sigma}_\phi\text{-GP}}$	$\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{x}-D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\text{tr}(\boldsymbol{\Sigma}_\phi)-2\text{tr}(\log\boldsymbol{L}_\phi)+\lambda_{\max}(\boldsymbol{\Sigma}_\phi)\|\nabla_{\boldsymbol{\mu}_\phi}D_\theta(\boldsymbol{\mu}_\phi)\|_2^2$	$\mathbf{z}_k=\boldsymbol{\mu}_\phi+\boldsymbol{L}_\phi\boldsymbol{\epsilon}_k,\ \ \boldsymbol{\epsilon}_k\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
$\mathcal{L}_{\text{UT-VAE-full}\boldsymbol{\Sigma}_\phi}$	$\|\mathbf{x}-\frac{1}{K}\sum_{k=1}^{K}D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\text{tr}(\boldsymbol{\Sigma}_\phi)-2\text{tr}(\log\boldsymbol{L}_\phi)$	$\mathbf{z}_k\sim\{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi,\boldsymbol{\Sigma}_\phi)\}_{i=0}^{2n}$
$\mathcal{L}_{\text{UT-VAE-full}\boldsymbol{\Sigma}_\phi\text{-GP}}$	$\|\mathbf{x}-\frac{1}{K}\sum_{k=1}^{K}D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\text{tr}(\boldsymbol{\Sigma}_\phi)-2\text{tr}(\log\boldsymbol{L}_\phi)+\lambda_{\max}(\boldsymbol{\Sigma}_\phi)\|\nabla_{\boldsymbol{\mu}_\phi}D_\theta(\boldsymbol{\mu}_\phi)\|_2^2$	$\mathbf{z}_k\sim\{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi,\boldsymbol{\Sigma}_\phi)\}_{i=0}^{2n}$
$\mathcal{L}_{\text{VAE*-full}\boldsymbol{\Sigma}_\phi}$	$\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{x}-D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\|\boldsymbol{L}_\phi-\mathbf{I}\|_F^2$	$\mathbf{z}_k=\boldsymbol{\mu}_\phi+\boldsymbol{L}_\phi\boldsymbol{\epsilon}_k,\ \ \boldsymbol{\epsilon}_k\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
$\mathcal{L}_{\text{VAE*-full}\boldsymbol{\Sigma}_\phi\text{-GP}}$	$\frac{1}{K}\sum_{k=1}^{K}\|\mathbf{x}-D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\|\boldsymbol{L}_\phi-\mathbf{I}\|_F^2+\lambda_{\max}(\boldsymbol{\Sigma}_\phi)\|\nabla_{\boldsymbol{\mu}_\phi}D_\theta(\boldsymbol{\mu}_\phi)\|_2^2$	$\mathbf{z}_k=\boldsymbol{\mu}_\phi+\boldsymbol{L}_\phi\boldsymbol{\epsilon}_k,\ \ \boldsymbol{\epsilon}_k\sim\mathcal{N}(\mathbf{0},\mathbf{I})$
$\mathcal{L}_{\text{UT-VAE*-full}\boldsymbol{\Sigma}_\phi}$	$\|\mathbf{x}-\frac{1}{K}\sum_{k=1}^{K}D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\|\boldsymbol{L}_\phi-\mathbf{I}\|_F^2$	$\mathbf{z}_k\sim\{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi,\boldsymbol{\Sigma}_\phi)\}_{i=0}^{2n}$
$\mathcal{L}_{\text{UT-VAE*-full}\boldsymbol{\Sigma}_\phi\text{-GP}}$	$\|\mathbf{x}-\frac{1}{K}\sum_{k=1}^{K}D_\theta(\mathbf{z}_k)\|_2^2+\|\boldsymbol{\mu}_\phi\|_2^2+\|\boldsymbol{L}_\phi-\mathbf{I}\|_F^2+\lambda_{\max}(\boldsymbol{\Sigma}_\phi)\|\nabla_{\boldsymbol{\mu}_\phi}D_\theta(\boldsymbol{\mu}_\phi)\|_2^2$	$\mathbf{z}_k\sim\{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi,\boldsymbol{\Sigma}_\phi)\}_{i=0}^{2n}$

Table 8: Full ablation study of the models between the VAE and UAE (in the UT-VAE*-GP row), using the Wasserstein metric denoted by *, unscented transform (UT), and the decoder gradient penalty (GP) components. See the upper half Tab. 7 for the loss function definitions.

	Fashion-MNIST			CIFAR10			CelebA		
	Rec.	Sample	Interp.	Rec.	Sample	Interp.	Rec.	Sample	Interp.
$\text{VAE}_{2\text{x}}$	43.66	49.01	61.03	112.7	123.2	120.6	67.29	69.92	70.00
$\text{VAE-GP}_{2\text{x}}$	44.17	48.63	59.58	108.9	120.3	117.5	66.94	70.16	69.77
$\text{UT-VAE}_{2\text{x}}$	36.25	40.30	53.10	95.70	115.4	107.4	51.61	57.42	56.56
$\text{UT-VAE-GP}_{2\text{x}}$	47.77	65.24	72.43	102.6	118.6	113.1	100.4	102.2	100.3
$\text{VAE*}_{2\text{x}}$	30.07	37.92	52.15	84.57	102.2	95.61	45.18	50.97	53.73
$\text{VAE*-GP}_{2\text{x}}$	29.40	38.53	53.88	85.19	103.7	96.66	41.69	48.77	51.29
$\text{UT-VAE*}_{2\text{x}}$	**28.26**	**36.36**	**50.69**	82.17	100.7	93.80	44.32	50.33	52.40
$\text{UT-VAE* -GP}_{2\text{x}}$	29.29	37.59	53.69	**77.71**	**96.37**	**89.71**	**40.07**	**47.28**	**50.51**

E. ELBO Constraint Derivation

In this section, we complete the derivation of the constraint in Eq. (14) to the reformulated version in Eq. (15). The constraint in Eq. (14) can be bounded by the maximum of the decoder output in a single dimension i, multiplied by the number of dimensions

$$\|D_\theta(\mathbf{z}_1) - D_\theta(\mathbf{z}_2)\|_p \le \dim(\mathbf{x}) \cdot \sup_i \{\|d_i(\mathbf{z}_1) - d_i(\mathbf{z}_2)\|_p\} < \epsilon \,. \tag{22}$$

Using the mean value theorem, the term $\sup_i \{\|d_i(\mathbf{z}_1) - d_i(\mathbf{z}_2)\|_p\}$ can be reduced to

$$\sup_i \{\|\nabla_t d_i((1-t)\mathbf{z}_1 + t\mathbf{z}_2)\|_p \cdot \|\mathbf{z}_1 - \mathbf{z}_2\|_p\} < \epsilon \,, \tag{23}$$

Since $\mathbf{z}_1$ and $\mathbf{z}_2$ are arbitrary, the first part can be simplified and generalized over all dimensions while separating the overall product using the Cauchy-Schwarz inequality

$$\sup_i \{\|\nabla_{\mathbf{z}} d_i(\mathbf{z})\|_p \cdot \|\mathbf{z}_1 - \mathbf{z}_2\|_p\} < \epsilon \tag{24}$$

$$\sup\{\|\nabla_{\mathbf{z}} D_\theta(\mathbf{z})\|_p\} \cdot \sup\{\|\mathbf{z}_1 - \mathbf{z}_2\|_p\} < \epsilon \,, \tag{25}$$

obtaining the form in Eq. (15).

F. Full-Covariance Posterior

In this section, we aim to investigate the performance of full-covariance posterior models. The non-diagonal posterior representation is naturally supported by the unscented transform and common in filtering. However, it is seldom in VAEs – one of the key ingredients of the standard VAE model is its diagonal Gaussian posterior approximation. The induced orthogonality can implicitly have positive effects on the structure of the latent space and the decoder (Zietlow et al., 2021; Rolinek et al., 2019), but such effects highly depend on implicit biases present in the dataset (Zietlow et al., 2021). Furthermore, the diagonal posterior together with the KL regularization allows for pruning unnecessary latent dimensions, also known as desired posterior collapse (Dai et al., 2020). A full-covariance posterior does not have such implicit biases and pruning properties, but it can have a positive effect on the optimization of the variational objective, as it connects otherwise disconnected global optima (Dai et al., 2018). Furthermore, it allows for modeling correlations in the posterior. We are not aware of a work successfully employing a full-covariance posterior.

The full-covariance representation can be practically realized by predicting n-dimensional standard deviations $\boldsymbol{\sigma}_\phi$ as well as $n(n-1)/2$-dimensional correlation factors $\boldsymbol{r}_\phi$ (followed by a $\tanh$ projection into the valid $[-1, 1]$ range), and building the lower triangular covariance matrix[7] $\boldsymbol{L}_\phi$. In this way, the full-covariance matrix $\boldsymbol{\Sigma}_\phi = \boldsymbol{L}_\phi \boldsymbol{L}_\phi^T$ is ensured to be symmetric and positive semi-definite.

The results of the full-covariance models are shown in the bottom half of Tab. 9. In all KL divergence instances, the performance of the models regresses significantly compared to their counterparts in Tab. 8. This indicates that, despite its theoretical potential to connect disconnected global optima of the optimization objective, a non-diagonal latent space is nevertheless difficult to train with KL divergence, regardless of the sampling method. However, the Wasserstein metric models receive a surprising performance boost. In some cases, they significantly outperform the models from Tab. 8 on Fashion-MNIST and CelebA while achieving similar results on CIFAR10, which has less structure in its input data. It is evident that the Wasserstein metric and potentially its lower posterior variance can enable a successful utilization of correlations in the posterior.

---

[7]In the 3-dimensional case: $\boldsymbol{L}_\phi = [\sigma_1 \quad 0 \quad 0; \quad r_1\sigma_2\sigma_1 \quad \sigma_2 \quad 0; \quad r_2\sigma_3\sigma_1 \quad r_3\sigma_3\sigma_2 \quad \sigma_3]$.

Table 9: Ablation study of the models in Tab. 8 in a full-covariance setting. See Tab. 7 for the loss function definitions.

	Fashion-MNIST			CIFAR10			CelebA		
	Rec.	Sample	Interp.	Rec.	Sample	Interp.	Rec.	Sample	Interp.
VAE-full $\boldsymbol{\Sigma}_{\phi 2x}$	79.01	83.15	91.01	123.8	132.6	130.2	99.72	100.9	99.96
VAE-full $\boldsymbol{\Sigma}_{\phi}$-GP$_{2x}$	180.0	181.5	184.4	158.3	165.8	164.0	244.2	244.6	241.8
UT-VAE-full $\boldsymbol{\Sigma}_{\phi 2x}$	57.93	58.87	64.86	129.6	141.2	138.2	132.1	132.4	136.0
UT-VAE-full $\boldsymbol{\Sigma}_{\phi}$-GP$_{2x}$	133.6	136.7	136.9	208.9	217.7	212.2	303.5	304.5	303.3
VAE*-full $\boldsymbol{\Sigma}_{\phi 2x}$	31.16	40.99	54.73	85.47	103.9	96.55	42.07	48.59	50.72
VAE*-full $\boldsymbol{\Sigma}_{\phi}$-GP$_{2x}$	**19.86**	**32.71**	48.84	84.19	102.9	95.63	39.69	46.76	49.70
UT-VAE*-full $\boldsymbol{\Sigma}_{\phi 2x}$	21.96	34.17	**48.32**	**79.51**	**98.32**	**91.82**	41.54	48.32	50.29
UT-VAE*-full $\boldsymbol{\Sigma}_{\phi}$-GP$_{2x}$	24.37	34.43	51.58	82.15	100.9	94.65	**39.48**	**46.60**	**48.97**

G. Connection to Wasserstein Autoencoders

Wasserstein-distance autoencoders (Patrini et al., 2020; Tolstikhin et al., 2018) use the Wasserstein distance $W_p(q_{\text{agg}}(\mathbf{z}), p(\mathbf{z}))$ to regularize the aggregated posterior $q_{\text{agg}}(\mathbf{z})$ toward the prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Instead, we use the Wasserstein distance as a simple regularization of the per-sample posterior. However, there is a simple connection of our posterior regularization to the aggregated posterior regularization. Assuming standard normal posteriors, the aggregated posterior can be represented as a mixture

$$q_{\text{agg}}(\mathbf{z}) = \frac{1}{N} \sum_n q(\mathbf{z}|\mathbf{x}_n) = \frac{1}{N} \sum_n \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n). \tag{26}$$

In the one-dimensional case (generalizable to multiple dimensions) the mean and variance of the mixture are

$$\mathcal{N}(\mu_n, \sigma_n^2) \overset{i.d.}{=} \mathcal{N}\left(\frac{1}{N} \sum_n \mu_n, \frac{1}{N} \sum_n \left(\sigma_n^2 + \mu_n^2\right) - \left(\frac{1}{N} \sum_n \mu_n\right)^2\right). \tag{27}$$

Thus, the aggregated posterior Wasserstein metric can be represented as

$$W_2(q_{\text{agg}}(\mathbf{z}), p(\mathbf{z})) = \left(\frac{1}{N} \sum_n \mu_n\right)^2 + \frac{1}{N} \sum_n \left(\sigma_n^2 + \mu_n^2\right) - \left(\frac{1}{N} \sum_n \mu_n\right)^2 - 2\sqrt{\frac{1}{N} \sum_n \left(\sigma_n^2 + \mu_n^2\right) - \left(\frac{1}{N} \sum_n \mu_n\right)^2} =$$

$$= \frac{1}{N} \sum_n \left(\sigma_n^2 + \mu_n^2\right) - 2\sqrt{\frac{1}{N} \sum_n \left(\sigma_n^2 + \mu_n^2\right) - \left(\frac{1}{N} \sum_n \mu_n\right)^2}, \tag{28}$$

in the case $p = 2$ and while discarding constants. Similarly, the average per-sample posterior metric is

$$\frac{1}{N} \sum_n W_2(q_{\text{pp}}(\mathbf{z}|\mathbf{x}), p(\mathbf{z})) = \frac{1}{N} \sum_n \left(\mu_n^2 + \sigma_n^2 - 2\sigma_n\right) = \frac{1}{N} \sum_n \mu_n^2 + \frac{1}{N} \sum_n \sigma_n^2 - 2\frac{1}{N} \sum_n \sigma_n. \tag{29}$$

Table 10: Comparison of the Wasserstein autoencoder that utilizes the aggregated posterior Wasserstein metric, and the VAE*, utilizing the per-sample posterior Wasserstein metric in the loss.

	Fashion-MNIST			CIFAR10			CelebA		
	Rec.	Sample	Interp.	Rec.	Sample	Interp.	Rec.	Sample	Interp.
WAE-MMD	47.58	62.44	73.94	88.31	**100.35**	94.78	67.54	75.92	73.21
VAE*$_{1x}$	**31.62**	**38.44**	**52.33**	**83.49**	101.5	**94.56**	**44.69**	**50.55**	**53.18**

Comparing the aggregated posterior metric with the average per-sample posterior metric yields

$$\frac{1}{N}\sum_n\left(\sigma_n^2+\mu_n^2\right)-2\sqrt{\frac{1}{N}\sum_n\left(\sigma_n^2+\mu_n^2\right)-\left(\frac{1}{N}\sum_n\mu_n\right)^2}\leq\frac{1}{N}\sum_n\mu_n^2+\frac{1}{N}\sum_n\sigma_n^2-2\frac{1}{N}\sum_n\sigma_n \tag{30}$$

$$-2\sqrt{\frac{1}{N}\sum_n\left(\sigma_n^2+\mu_n^2\right)-\left(\frac{1}{N}\sum_n\mu_n\right)^2}\leq-2\frac{1}{N}\sum_n\sigma_n \tag{31}$$

$$\sqrt{\frac{1}{N}\sum_n\left(\sigma_n^2+\mu_n^2\right)-\left(\frac{1}{N}\sum_n\mu_n\right)^2}\geq\frac{1}{N}\sum_n\sigma_n \tag{32}$$

$$\frac{1}{N}\sum_n\left(\sigma_n^2+\mu_n^2\right)-\left(\frac{1}{N}\sum_n\mu_n\right)^2\geq\left(\frac{1}{N}\sum_n\sigma_n\right)^2 \tag{33}$$

$$\frac{1}{N}\sum_n\left(\sigma_n^2+\mu_n^2\right)\geq\left(\frac{1}{N}\sum_n\mu_n\right)^2+\left(\frac{1}{N}\sum_n\sigma_n\right)^2. \tag{34}$$

Eq. (34) can be regarded as two Jensen's inequalities $f(\mathbb{E}[x])\leq\mathbb{E}[f(x)]$, where $f(x)=x^2$, and $\mathbb{E}[x]=\frac{1}{N}\sum_n x_n$. Thus, the initial inequality holds. It shows that the per-sample posterior Wasserstein metric is an upper bound to the aggregated posterior Wasserstein metric, commonly used in the WAE (Tolstikhin et al., 2018). Therefore, we can guarantee that the Wasserstein distance of the aggregated posterior to the assumed standard normal prior will not be larger than than the average distance of per-sample posteriors.

In addition to the theoretical argument, in Tab. 10 we offer an empirical comparison of the VAE* with the WAE-MMD model from (Tolstikhin et al., 2018) with aggregated posterior weight $\lambda=10$. We observed that the per-sample posterior regularization significantly outperforms the WAE on Fashion-MNIST and CelebA, while being on par on CIFAR10.

H. Wasserstein Metric Aggregated Posterior Visualization

In Fig. 5 we present detailed plots on the posterior distributions of VAE and VAE* for the first 16 dimensions. The VAE clearly shows signs of posterior collapse (so-called *polarized regime* (Rolinek et al., 2019)); we have observed that more than half of the 128 dimensions are nearly equal to the prior. This considerably hurts the generative power of the VAE model. In contrast, the VAE* model has very low variance in all dimensions, which reflects a nearly deterministic encoder at the end of the training.
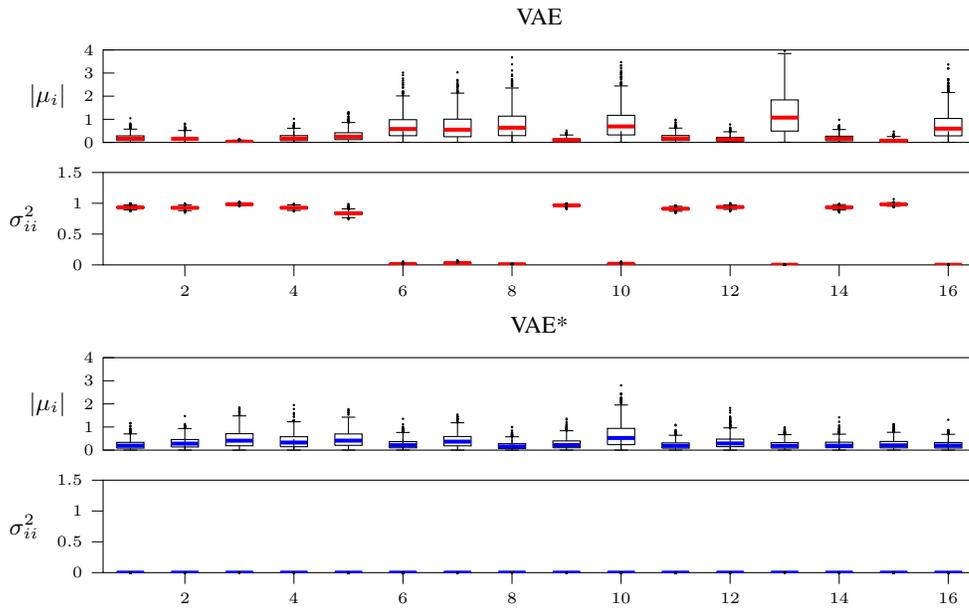
Figure 5: Comparison of the distribution of absolute means and variances of 1000 posterior samples for the VAE_{1x} and the $VAE*_{1x}$ models trained with 100 epochs on the CIFAR10 dataset. Top rows show the absolute means and the lower rows the variances of the first 16 dimensions. For the $VAE*_{1x}$ all means differ from zero while the variances are close to zero, whereas for the VAE_{1x}, 10 of 16 dimensions are effectively deactivated.

I. Qualitative Results on Fashion-MNIST and CIFAR10

Qualitative results on Fashion-MNIST and CIFAR10 are provided in Fig. 6 and Fig. 7. The same setup as in Fig. 3 is employed. It can be seen that the CIFAR10 images appear considerably richer and sharper, consistent with the results in Tab. 2 and Tab. 6.
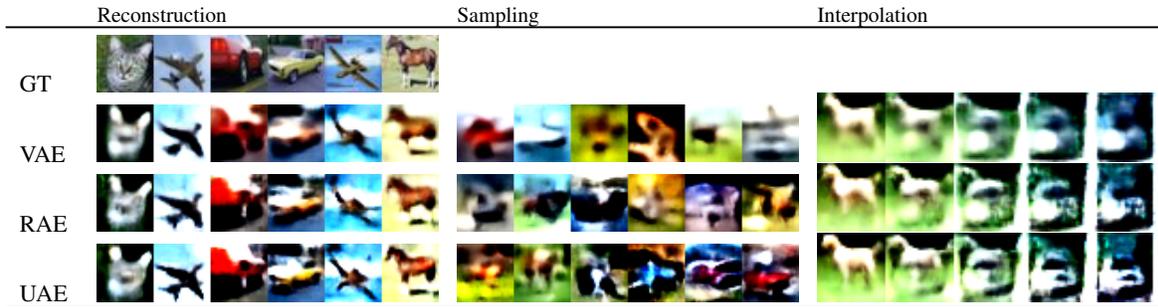
Figure 6: Qualitative results on the CIFAR10 dataset.
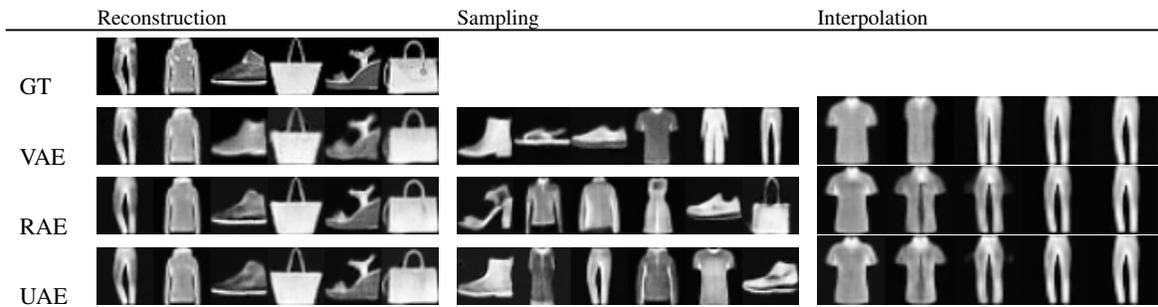
Figure 7: Qualitative results on the Fashion-MNIST dataset.