UNSCENTED AUTOENCODER

Anonymous authors
Paper under double-blind review

ABSTRACT

The Variational Autoencoder (VAE) is a seminal approach in deep generative modeling with latent variables. It performs posterior inference by parameterizing a distribution of latent variables in the stochastic encoder (while penalizing the disparity to an assumed standard normal prior), and achieves sample reconstruction via a deterministic decoder. In our work, we start from a simple interpretation of the reconstruction process: a nonlinear transformation of the stochastic encoder. We apply the Unscented Transform (UT) from the field of filtering and control – a well-known distribution approximation used in the Unscented Kalman Filter (UKF). A finite set of statistics called sigma points that are sampled deterministically provides a more informative and lower-variance posterior representation than the ubiquitous noise-scaling of the reparameterization trick. Inspired by the unscented transform, we derive a novel deterministic flavor of the VAE, the Unscented Autoencoder (UAE), trained purely with regularization-like terms on the per-sample, full-covariance posterior. A key ingredient for the good performance is the Wasserstein distribution metric in place of the Kullback-Leibler (KL) divergence, effectively performing covariance matrix regularization while allowing for a sharper posterior, which especially benefits reconstruction. Nevertheless, our results are consistent with recent findings showing that deterministic models can ensure good sample quality and smooth interpolation in the latent space. We empirically show superior performance in Fréchet Inception Distance (FID) scores over closely-related models, in addition to a lower training variance than the VAE.

1 INTRODUCTION

VAEs (Kingma et al., 2015; Rezende et al., 2014) are a seminal 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 application domains such as image (Vahdat & Kautz, 2020), language (Kusner et al., 2017; Bowman et al., 2015), 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; Tripp et al., 2020; Townsend et al., 2019; 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 quality. 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 (Dai et al., 2020; Chen et al., 2017; Lucas et al., 2019), 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 several 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 function. To account for sampling only at the sigma points, we add a regularizer for decoder smoothness around these points, similar to Ghosh et al. (2019). The unscented transform naturally supports using a non-diagonal posterior distribution, which is very common in the filtering literature. While a non-diagonal posterior can have negative effects on the structure of the latent space (Zietlow et al., 2021; Rolinek et al., 2019), it can have a positive effect on the optimization of the variational objective (Dai et al., 2018) and the reconstruction quality. Finally, 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). We conduct rigorous experiments on several standard image datasets to compare our modifications against the closely-related model from (Ghosh et al., 2019) as well as the VAE as baselines.

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 Regularized Autoencoder (RAE) in Ghosh et al. (2019); our model can be considered as part of this class. As previously mentioned, we employ multiple modifications to the VAE, namely the **Wasserstein metric**, **full-covariance** representation, **decoder regularization**, and the **Unscented Transform**; we outline the section accordingly.

The **Wasserstein distance** is used in Tolstikhin et al. (2018); Patrini et al. (2020) to regularize the aggregated posterior $q_{\text{agg}}(\mathbf{z}) = \mathbb{E}_{p(\mathbf{x})}\left[q(\mathbf{z}|\mathbf{x})\right]$ 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 post-hoc density estimation. From a theoretical point of view, we do not fix the prior but learn the manifold; the distribution is learned by ex-post density estimation.

The **full-covariance** representation 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. In this context, the UT is an efficient way to train models with full covariance posteriors. We are not aware of a work successfully employing a full covariance posterior.

Our work incorporates several ideas from the recently published RAE (Ghosh et al., 2019). We also incorporate a **decoder regularization** term in our loss, which promotes smoothness of the latent space. It is based on the decoder Jacobian and accounts for sampling only at specific sigma points. In contrast to the RAE however, we do not assume a deterministic encoder as not every datapoint might be encoded with the same fidelity. Furthermore, we employ 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.

Finally, we employ 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 with samples for propagation and update have emerged. These approaches can be clustered according to the employed sampling method – random as in 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[1] aims to achieve a good coverage of the distribution approximated by the mean and covariance, which usually better reflects the nonlinearities of the functions that are applied to the distribution. For a more comprehensive overview of the unscented transform as well as the UKF, we refer the reader to Menegaz et al. (2015).

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 one obtains the well-known ELBO $\mathcal{L}$

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

which is maximized w.r.t. θ and ϕ. The first term accounts for the quality of reconstructed samples and the KL-divergence 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.

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

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[2], 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.

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 $\boldsymbol{\Sigma} = \mathbf{0}$, 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 $\{\boldsymbol{\chi}_i\}_{i=0}^{2n}$, $\boldsymbol{\chi}_i \in \mathbb{R}^n$ of $2n+1$ symmetric points centered around the mean (incl. the mean), e.g. for $1 \leq i \leq n$,

$$\boldsymbol{\chi}_0 = \boldsymbol{\mu}, \quad \boldsymbol{\chi}_i = \boldsymbol{\mu} + \sqrt{(\kappa + n)\boldsymbol{\Sigma}}\big|_i, \quad \boldsymbol{\chi}_{i+n} = \boldsymbol{\mu} - \sqrt{(\kappa + n)\boldsymbol{\Sigma}}\big|_i \,, \tag{5}$$

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 $\hat{\boldsymbol{\chi}}_i = \boldsymbol{f}(\boldsymbol{\chi}_i)$ and estimate the mean and covariance of the $\boldsymbol{f}$-transformed distribution

$$\hat{\boldsymbol{\mu}} = \textstyle\sum_{i=0}^{2n} \hat{\boldsymbol{\chi}}_i, \quad \hat{\boldsymbol{\Sigma}} = \textstyle\sum_{i=0}^{2n} (\hat{\boldsymbol{\chi}}_i - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\chi}}_i - \hat{\boldsymbol{\mu}})^T \,. \tag{6}$$

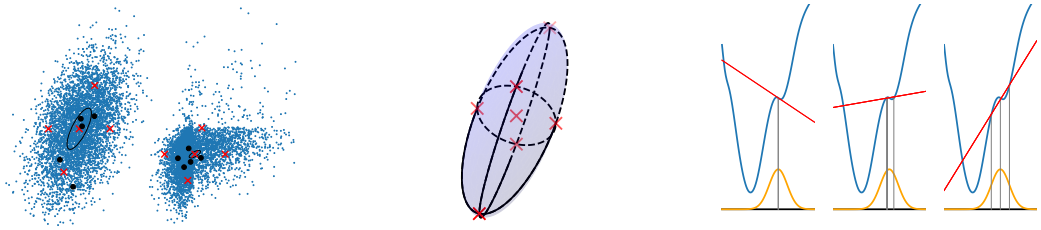
A visualization of the sigma points and their transformation is depicted in Fig. 1a. The procedure in Eq. (5) and (6) effectively applies the fully-known function $\boldsymbol{f}$ to an approximating set of points whose mean and covariance equal the original distribution's. Therefore, the mean and covariance of the transformed sigma points will be closer to the true transformed mean and covariance than 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.

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 $\mathcal{N}(\mathbf{0}, \mathbf{I})$ 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 since the samples lie on the border of a hyperellipsoid induced by the covariance matrix $\boldsymbol{\Sigma}$ (example in Fig. 1b). Thus, while computing the gradients of the loss function (which is 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 1c. Further analytical and empirical arguments validating the lower gradient variance claim are provided in Appendix B.

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



(a) (b) (c)

Figure 1: (a) **(transforming 2D sigma points)** (best viewed in color) 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. It is clear that 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 three random samples drawn from the standard normal, potentially far away from the true gradient. Right: gradient of the loss function at the three sigma points; although the estimate is potentially biased, 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.

The sigma-sampling of the unscented transform 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 unscented transform 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 models in filtering and control. Guided by the success of the method, we hypothesize that applying the unscented transform in the VAE setting has the potential to provide the best-possible approximation of any given learned posterior. With these insights, we develop the novel UAE model presented in the next section.

5 Unscented Auto-Encoder (UAE)

The UAE is a deterministic 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 unscented transform of the posterior $q_\phi(\mathbf{z}|\mathbf{x})$ parameterized by the encoder $E_\phi(\mathbf{x}) = \{\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\Sigma}_\phi(\mathbf{x})\}$. The latent features $\mathbf{z}$ can be obtained by deterministically sampling a single or multiple sigma points, resulting in a lower variance sampling than of the reparameterization trick in Eq. (4). The key ingredient enabling the good performance of the overall model as well as the unscented transform in this context is replacing the vanilla KL divergence with the Wasserstein distribution metric, which effectively performs a regularization of the posterior moments. It is employed in conjunction with a full covariance posterior commonly used in filtering applications. The decoder regularization further applies a smoothing effect on the latent space – it is formally derived in Sec. 5.2. The full training objective consists of optimizing $\phi, \theta \leftarrow \arg\min_{\phi,\theta} \mathcal{L}_{\text{UAE}}$,

$$\mathcal{L}_{\text{UAE}} = E_{\mathbf{x}\sim p_{\text{data}}} \mathcal{L}_{\text{REC}} + \beta \mathcal{L}_{\text{KL}} + \gamma \mathcal{L}_{D_\theta \text{REG}} \,, \tag{7}$$

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

Due to a deterministic decoder, the **reconstruction term** $\mathcal{L}_{\text{REC}}$ is a straightforward L_2 loss

$$\mathcal{L}_{\text{REC}} = \|\mathbf{x} - D_\theta(\mathbf{z})\|_2^2 \,, \quad \mathbf{z} \sim \{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi, \boldsymbol{\Sigma}_\phi)\}_{i=0}^{2n} \,, \tag{8}$$

with a caveat that the n-dimensional vector $\mathbf{z}$ is sampled uniformly from the set of sigma points. For simplicity, we sample a single sigma point but explore multiple sigma point sampling in Appendix D.

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

Table 1: A comparison of the VAE, RAE-GP (employing a Gradient Penalty (GP) on the decoder), and UAE loss functions, including the intermediate models VAE*, UT-VAE*, UT-VAE*-full-$\boldsymbol{\Sigma}_\phi$ (weights omitted for clarity). The posterior terms $\mathbf{z}$, $\boldsymbol{\mu}_\phi$, $\boldsymbol{\sigma}_\phi$, and $\boldsymbol{\Sigma}_\phi$ are realized given the sample $\mathbf{x}$.

The **KL divergence term** $\mathcal{L}_{\text{KL}}$ of the posterior against the multivariate normal prior comes from its definition for two multivariate Gaussians[3]

$$\mathcal{L}_{\text{KL}} = \|\boldsymbol{\mu}_\phi\|_2^2 + \text{tr}(\boldsymbol{\Sigma}_\phi) - n - \log\det\boldsymbol{\Sigma}_\phi = \|\boldsymbol{\mu}_\phi\|_2^2 + \text{tr}(\boldsymbol{\Sigma}_\phi) - n - 2\text{tr}(\log\boldsymbol{L}_\phi) \tag{9}$$

and allows for a full covariance matrix. However, due to favorable optimization properties and higher-quality reconstruction, we use the Wasserstein metric between distributions in place of the KL divergence. 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 Frobenius norm of the mismatch between the lower triangular matrix and the identity

$$\|\boldsymbol{L}_\phi - \mathbf{I}\|_F^2 = \text{tr}(\boldsymbol{\Sigma}_\phi) - 2\text{tr}(\boldsymbol{L}_\phi)\ . \tag{10}$$

It differs from the original objective in Eq. (9) 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 harshly penalized by the logarithm in Eq. (9)), yielding a sharper posterior. Furthermore, it favors a diagonal $\boldsymbol{\Sigma}_\phi$, but importantly, does not assert such shape as in the case of the VAE and implicitly RAE.

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

$$\mathcal{L}_{D_\theta\text{REG}} = \lambda_{\max}(\boldsymbol{\Sigma}_\phi)\|\nabla_\mathbf{z} D_\theta(\mathbf{z})\|_2^2\ . \tag{11}$$

We approximate the $\lambda_{\max}(\boldsymbol{\Sigma}_\phi)$ by the largest diagonal, which is correct for the diagonal $\boldsymbol{\Sigma}_\phi$. Furthermore, Eq. (10) anyway pushes the $\boldsymbol{\Sigma}_\phi$ to be diagonal.

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. These are the VAE*, using the Wasserstein metric from Eq. (10) on a standard diagonal VAE posterior, UT-VAE*, with the unscented transform in place of the reparameterization trick, and full covariance UT-VAE*-full $\boldsymbol{\Sigma}_\phi$, differing from the given UAE only in the lack of a decoder regularization term. Additional models employing different combinations of the loss function components are provided in Appendix E, Tab. 5.

5.1 Sampling from the deterministic 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), 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 feature vectors $\mathbf{z}$ (a single sigma point for each training example) sampled in Eq. (8) 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 generate new samples. For a fair comparison, we utilize this procedure when sampling all the models.

[3] $D_{\text{KL}}(\mathcal{N}_0,\mathcal{N}_1) = \frac{1}{2}(\text{tr}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0) - n + (\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0) + \log(\frac{\det\boldsymbol{\Sigma}_1}{\det\boldsymbol{\Sigma}_0}))$

5.2 ELBO DERIVATION

In the following, we analytically derive the UAE model from Eq. (7). 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

$$\underset{\phi,\theta}{\arg\min} \quad E_{x\sim p_{\text{data}}}\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{KL}} \tag{12}$$

$$\text{s.t.} \quad \|D_\theta(\mathbf{z}_1) - D_\theta(\mathbf{z}_2)\|_p < \epsilon, \quad \mathbf{z}_1, \mathbf{z}_2 \sim q_\phi(\mathbf{z}|\mathbf{x}),\ \forall \mathbf{x} \sim p_{\text{data}}. \tag{13}$$

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.

Ghosh et al. (2019) showed that the constraint in Eq. (13) can be reformulated as

$$\sup\{\|\nabla_{\mathbf{z}} D_\theta(\mathbf{z})\|_p\} \cdot \sup\{\|\mathbf{z}_1 - \mathbf{z}_2\|_p\} < \epsilon\,. \tag{14}$$

We provide the full derivation in Appendix C. In Eq. (14), $\nabla_{\mathbf{z}} D_\theta(\mathbf{z})$ 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. (14) into Eq. (12) via the Lagrange multiplier γ, obtaining the following RAE loss function[4]

$$\mathcal{L}_{\text{RAE}} = \|\mathbf{x} - D_\theta(\mathbf{z})\|_2^2 + \beta\|\mathbf{z}\|_2^2 + \gamma\|\nabla_{\mathbf{z}} D_\theta(\mathbf{z})\|_2^2\,. \tag{15}$$

Here, the KL-term from Eq. (12) 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. (14) 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 $\boldsymbol{\Sigma}_\phi$. Furthermore, the UAE does not require the constant variance assumption; the KL-term can be provided for a full covariance. Thus, we arrive at the following analytical UAE loss function from Eq. (7)

$$\begin{aligned}\mathcal{L}_{\text{UAE}} = \|\mathbf{x} - D_\theta(\mathbf{z})\|_2^2 + \|\boldsymbol{\mu}_\phi\|_2^2 - n + \text{tr}(\boldsymbol{\Sigma}_\phi) - \log\det\boldsymbol{\Sigma}_\phi + \\ + \gamma\lambda_{\max}(\boldsymbol{\Sigma}_\phi)\|\nabla_{\mathbf{z}} D_\theta(\mathbf{z})\|_2^2\,, \quad \mathbf{z} \sim \{\boldsymbol{\chi}_i(\boldsymbol{\mu}_\phi, \boldsymbol{\Sigma}_\phi)\}_{i=0}^{2n}\,.\end{aligned} \tag{16}$$

In practice, we replace the logarithm of the KL-term with a linear term, see Eq. (10). This completes the derivation of the UAE loss function in Eq. (7) from the ELBO condition in Eq. (12).

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 fitting 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. 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

In practice, the training of VAEs can be sensitive to the weighting of the KL term, which can lead to posterior collapse (Dai et al., 2020). The main factor is the strong variance regularization of the KL with its $\log$ term, which can be written as

$$\mathcal{L}_{\text{KL}} = \|\boldsymbol{\mu}_\phi\|_2^2 + \text{tr}(\boldsymbol{\Sigma}_\phi) - n - \log\det\boldsymbol{\Sigma}_\phi = \|\boldsymbol{\mu}_\phi\|_2^2 + \text{tr}(\boldsymbol{\Sigma}_\phi) - n - 2\textstyle\sum_i \log L_{\phi,ii} \tag{17}$$

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

[4] In Ghosh et al. (2019), the decoder gradient penalty from Eq. (15) 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.

	Fashion-MNIST			CIFAR10			CelebA		
	Rec.	Sample	Interp.	Rec.	Sample	Interp.	Rec.	Sample	Interp.
VAE	47.38	51.74	66.04	160.05	173.45	170.33	65.86	67.66	68.08
RAE-no reg.	34.60	41.61	57.0	140.70	158.08	154.33	40.35	47.62	50.10
RAE-L2	33.80	41.46	60.97	142.08	159.68	154.97	39.03	46.37	50.65
RAE-GP	**32.40**	**39.01**	**56.28**	139.83	158.17	153.83	39.87	46.38	**46.46**
VAE*	33.72	40.39	59.76	136.90	156.83	151.41	45.15	50.29	53.23
UT-VAE*	33.57	40.31	57.21	134.21	153.09	147.32	48.29	56.14	54.14
UT-VAE*-full $\boldsymbol{\Sigma}_\phi$	35.78	42.97	66.75	126.00	149.52	141.67	43.00	53.39	51.22
UAE	33.30	40.81	60.13	**120.95**	**147.07**	**137.39**	**37.93**	**44.59**	**46.46**

Table 2: Comparison of the architectures from Tab. 1: VAE, RAE, intermediate models, and the UAE. VAE* employs the Wasserstein metric in place of the KL divergence. The UT variants use the unscented transform and sample a single sigma point given $\boldsymbol{\mu}_\phi$ and (diagonal or full) $\boldsymbol{\Sigma}_\phi$. Three RAE variants are provided: RAE-no reg. without decoder regularization, RAE-GP with the Gradient Penalty (GP) from Eq. (15), and RAE-L2 with decoder weight decay.

posteriors (results in Appendix E). Despite these problems the KL divergence is theoretically sound. Hoffman & Johnson (2016) showed that $D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ can be reformulated into two terms, one that weakly pushes toward overlapping 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, 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 also do not require a certain aggregated posterior to sample from.

The Wasserstein-2 metric for Gaussians $\mathcal{N}_1 = \mathcal{N}(\boldsymbol{\mu}_\phi, \boldsymbol{\Sigma}_\phi)$ and $\mathcal{N}_2 = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ can be written as

$$W_2(\mathcal{N}_1, \mathcal{N}_2) = \|\boldsymbol{\mu}_\phi\|_2^2 + \mathrm{tr}(\boldsymbol{\Sigma}_\phi) + n - 2\mathrm{tr}(\boldsymbol{\Sigma}_\phi^{1/2}) = \|\boldsymbol{\mu}_\phi\|_2^2 + \mathrm{tr}(\boldsymbol{\Sigma}_\phi) + n - 2\mathrm{tr}(\boldsymbol{L}_\phi)\ . \tag{18}$$

The last three terms can be reformulated into the term used in Eq. (10)

$$\mathrm{tr}(\boldsymbol{\Sigma}_\phi) + n - 2\mathrm{tr}(\boldsymbol{L}_\phi) = \mathrm{tr}(\boldsymbol{L}_\phi^T \boldsymbol{L}_\phi - 2\boldsymbol{L}_\phi + \boldsymbol{I}) = \mathrm{tr}((\boldsymbol{L}_\phi - \mathbf{I})^T(\boldsymbol{L}_\phi - \mathbf{I})) = \|\boldsymbol{L}_\phi - \mathbf{I}\|_F^2\ . \tag{19}$$

Disregarding the constant terms, it is clear that Eq. (17) and Eq. (18) 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 analysis provided in Appendix F. Overall, our empirical analysis shows that replacing $D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ with the Wasserstein-2 distance $W_2(q_\phi(\mathbf{z}|\mathbf{x}), p(\mathbf{z}))$ results in better performance.

6 Results

In the following, we present quantitative and qualitative results of the UAE model 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) datasets. We aim to delineate the effects of the Wasserstein metric, the unscented transform compared to the reparameterization trick (in the diagonal and full-covariance context), and the decoder regularization of the latent space. Furthermore, we investigate whether sampling multiple sigma points is beneficial, similarly to drawing many samples from the prior in the VAE. These results are provided in Appendix D. In addition to evaluating the reconstruction and random sample quality (using a fitted mixture for all models, see Sec. 5.1), we investigate if sampling only at the sigma points in training preserves the latent space structure, e.g. does not create "holes". Here, we evaluate the quality of interpolated samples. The metric for evaluating reconstruction, sampling, and interpolation quality is the popular FID (Heusel et al., 2017), which quantifies the distance between two distributions by their samples. Detailed information about the network architecture, training procedure, and the choice of FID computation datasets is given in Appendix A.

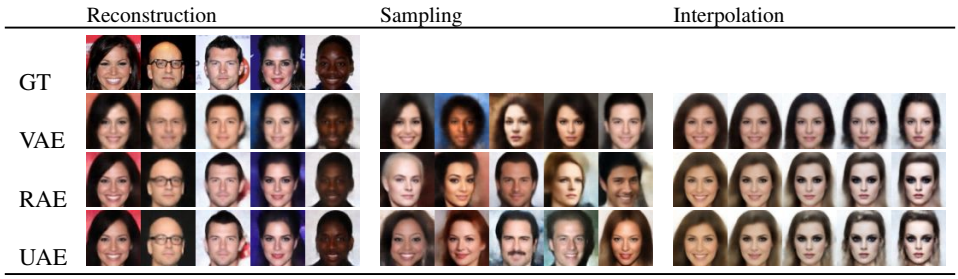


Figure 2: Qualitative results on the CelebA dataset of the VAE, RAE, and UAE models. The VAE and UAE models use 8 samples or sigma points, which in the case of the UAE corresponds to an approximately 10% improvement in FID scores (Appendix D, Tab. 4) over the closest RAE baseline.

The main results are provided in Tab. 2. The table can be interpreted as an ablation study, where, starting from the basic VAE model, we successively apply the four components of the UAE model: the Wasserstein metric from Eq. (10) in place of the KL divergence, unscented transform sampling, full covariance matrix in the posterior, and the decoder regularization based on the maximum eigenvalue of $\mathbf{\Sigma}_\phi$. An ablation study of additional model combinations is provided in Appendix E.

Among the results in Tab. 2, the deterministic RAE baseline sets the context with a significantly higher performance over the vanilla VAE. Likewise, the Wasserstein-metric VAE* preserves the latent space regularization (in spirit of the RAE) but extends it to a probabilistic, non-constant variance setting. The model can be considered on-par with the non-regularized RAE: outperforms it on CIFAR10, is on-par on Fashion-MNIST, but behind on CelebA. More importantly, it can be seen that the VAE* model 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 is an important result indicating that the rigidity of the KL divergence w.r.t. the 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 applied on a diagonal covariance in the VAE* context brings mixed results: minor improvements over VAE* on Fashion-MNIST and CIFAR10, and a minor regression on CelebA. Nevertheless, the UT-VAE*-full $\mathbf{\Sigma}_\phi$ row shows that allowing a full covariance representation (as common in unscented transform applications) enables the UT sampling to boost all metrics on CIFAR10 and CelebA. However, Fashion-MNIST appears to not benefit from modeling correlations in the posterior; we posit that it is due to the lower dimensionality of the input space. Finally, the eigenvalue-weighted decoder regularization in the UAE row (essentially UT-VAE*-full $\mathbf{\Sigma}_\phi$-GP) further applies a strong smoothing effect to compensate for sampling at fixed points. It is especially helpful on CelebA, where the UAE model brings a major gain of more than 40% over the VAE baseline in reconstruction. Overall, compared to the RAE, the UAE achieves large improvements on CIFAR10 and a minor improvement on CelebA, while being slightly behind on Fashion-MNIST. However, compared to the vanilla VAE, the UAE achieves very large improvements on all datasets in all metrics.

The given UAE model in Tab. 2 samples only a single sigma point. Therefore, additional performance gains can be reached simply by using multiple sigmas in training. Tab. 4 in Appendix D shows such results, where metrics further improve at the expense of an observed approximately linear scaling of the training time. Qualitative results on CelebA are shown in Fig. 2 and confirm the superior FID scores: the UAE samples appear sharper than the RAE and significantly more realistic than the VAE. Qualitative results on Fashion-MNIST and CIFAR10 are provided in Appendix G.

7 Conclusion

In this paper, we introduced a novel VAE architecture employing the Unscented Transform – a lower-variance alternative to the reparameterization trick. Through the Wasserstein metric, we established a framework enabling the good performance of the UT. By breaking the rigidity of the KL divergence w.r.t. posterior variance, we unlocked performance improvements brought on by sharper, full-covariance posteriors that still preserve a smooth latent space. Our work contributes an important step toward establishing competitive deterministic generative models.

REFERENCES

L Accardi. De finetti theorem. *Hazewinkel, Michiel, Encyclopaedia of Mathematics, Kluwer Academic Publishers*, 2001.

Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 66–75. PMLR, 2019.

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 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.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In Yoshua Bengio and Yann LeCun (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.

Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. 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.

Bin Dai and David Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1e0X3C9tQ.

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

Bin Dai, Ziyu Wang, and David Wipf. The usual suspects? reassessing blame for vae posterior collapse. In *International Conference on Machine Learning*, pp. 2313–2322. PMLR, 2020.

A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Oxford Handbook of Nonlinear Filtering*, 2011.

Partha Ghosh, Mehdi SM Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436*, 2019.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *CoRR*, abs/1406.2661, 2014. URL http://arxiv.org/abs/1406.2661.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *arXiv preprint arXiv:1706.08500*, 12(1), 2017.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. 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.

Matthew D Hoffman and Matthew J Johnson. Elbo surgery: Yet another way to carve up the evidence lower bound. In *Proc. Workshop Adv. Approx. Bayesian Inference*, pp. 2, 2016.

Simon Julier, Jeffrey Uhlmann, and Hugh F Durrant-Whyte. 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.

Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick Van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. *arXiv preprint arXiv:1605.06432*, 2016.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In Doina Precup and Yee Whye Teh (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.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. 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.

Emile Mathieu, Tom Rainforth, N Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In Kamalika Chaudhuri and Ruslan Salakhutdinov (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.

Henrique MT Menegaz, João Y Ishihara, Geovany A Borges, and Alessandro N Vargas. A systematization of the unscented kalman filter theory. *IEEE Transactions on automatic control*, 60(10): 2583–2598, 2015.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Giorgio Patrini, Rianne van den Berg, Patrick Forre, Marcello Carioni, Samarth Bhargav, Max Welling, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. In *Uncertainty in Artificial Intelligence*, pp. 733–743. PMLR, 2020.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 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.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12406–12415, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 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.

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

Ilya O. Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. 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.

James Townsend, Thomas Bird, and David Barber. 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.

Austin Tripp, Erik A. Daxberger, and José Miguel Hernández-Lobato. Sample-efficient optimization in the latent space of deep generative models via weighted retraining. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/81e3225c6ad49623167a4309eb4b2e75-Abstract.html.

Jeffrey Uhlmann. *Dynamic map building and localization: new theoretical foundations.* PhD thesis, University of Oxford, 1995.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *Neural Information Processing Systems (NeurIPS)*, 2020.

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

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

Wu Zhang, Min Liu, and Zong-gui Zhao. 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.

Dominik Zietlow, Michal Rolinek, and Georg Martius. Demystifying inductive biases for (beta-) vae based architectures. In *International Conference on Machine Learning*, pp. 12945–12954. PMLR, 2021.