Adversarial robustness of VAEs through the lens of local geometry

Asif Khan
School of Informatics,
University of Edinburgh, U.K.

Amos Storkey
School of Informatics,
University of Edinburgh, U.K.

Abstract

In an unsupervised attack on variational autoencoders (VAEs), an adversary finds a small perturbation in an input sample that significantly changes its latent space encoding, thereby compromising the reconstruction for a fixed decoder. A known reason for such vulnerability is the distortions in the latent space resulting from a mismatch between approximated latent posterior and a prior distribution. Consequently, a slight change in an input sample can move its encoding to a low/zero density region in the latent space resulting in an unconstrained generation. This paper demonstrates that an optimal way for an adversary to attack VAEs is to exploit a directional bias of a stochastic pullback metric tensor induced by the encoder and decoder networks. The pullback metric tensor of an encoder measures the change in infinitesimal latent volume from an input to a latent space. Thus, it can be viewed as a lens to analyse the effect of input perturbations leading to latent space distortions. We propose robustness evaluation scores using the eigenspectrum of a pullback metric tensor. Moreover, we empirically show that the scores correlate with the robustness parameter β of the $\beta-$VAE. Since increasing β also degrades reconstruction quality, we demonstrate a simple alternative using *mixup* training to fill the empty regions in the latent space, thus improving robustness with improved reconstruction.

1 INTRODUCTION

Variational autoencoders (VAEs) belong to a class of deep generative models with a stochastic encoder-decoder network (Kingma and Welling, 2014). The encoder parame-

terises the variational distribution over latent variables conditioned on data samples, and the decoder estimates the data distribution through the latent distribution. Thus, VAEs serve a dual purpose of estimating data density and providing a rich representation space with uncertainty quantification. Recently several works have shown the application of VAEs to high-fidelity image generation (Vahdat and Kautz, 2020), music generation (Roberts et al., 2017), video generation (Wu et al., 2021), and many others. However, like other machine learning models, VAEs are also susceptible to adversarial attacks, as demonstrated in several recent works (Tabacof et al., 2016; Gondim-Ribeiro et al., 2018; Willetts et al., 2021; Kos et al., 2017; Kuzina et al., 2021). In a typical setup, an adversary can attack a VAE by learning a small perturbation to an input that will lead to a large change in its latent encoding. This mechanism takes the form of an optimisation problem (introduced later in Equation 5), which is generally solved using stochastic gradient methods (Chakraborty et al., 2021; Szegedy et al., 2014). For a more comprehensive overview of attacks on VAEs and other generative models, we refer readers to (Sun et al., 2021).

The primary reason for the vulnerability of VAEs to attacks is the distortion in the latent space resulting from the mismatch between approximated posterior and latent space prior, also known as the *posterior-prior gap*. Hence, the latent space is non-smooth, and the representations of similar inputs tend to be significantly distant in the space under the Euclidean metric. Methods such as (Mathieu et al., 2019; Willetts et al., 2021) have emphasised the importance of reducing the *posterior-prior* for learning disentangled latent space that also improves the robustness of VAEs. $\beta-$VAE (Higgins et al., 2017) formulation introduces a parameter β to directly control the gap. Other methods (Chen et al., 2018b; Kim and Mnih, 2018; Esmaeili et al., 2019) utilise the total correlation (TC) term to disentangle the latent coordinates of VAEs that also smooths the latent space proving helpful in improving the robustness. The limitation of these approaches is that they cause over-smoothing of the reconstructed samples and require a careful training mechanism to balance the regularisation term. Willetts et al. (2021) address this problem by introducing a TC term in hierarchical VAEs that provide robustness along with sharp reconstruction. However, most robustness methods use reg-

ularisation terms, which do not provide meaningful insights for quantifying robustness. Therefore, the notion of a small change in the input to the large change in latent space is not well established. Moreover, comparing these schemes relies on the visual inspection of distorted images at different magnitudes of adversarial loss.

Much recently Camuto et al. (2021) proposed a theoretical framework that considers the uncertainty of encoder for studying the robustness of VAEs. However, the attacks in the input space do not consider the effect of geometry induced by an encoder or decoder maps. In another recent work Kuzina et al. (2021) proposed an asymmetric KL term to capture the difference between the latent representation of input and its perturbation. They obtain the ϵ using the Jacobian of the mean latent code evaluated at the input perturbation. Nevertheless, their optimisation objective does not provide any geometrical insights. Likewise, they do not consider the contribution of the standard deviation term when computing Jacobian. Thus, do not account for the uncertainty in the representation space. Our paper shows that the geometry induced by the stochastic nature of encoder mapping provides the intuition behind the sensitivity of VAEs that can be a valuable tool for understanding robustness.

The central theme of our paper is to view the adversarial attack problem through the lens of manifold geometry. Unlike existing approaches treating the input space as Euclidean, we propose to utilise the stochastic pullback-metric tensor induced by the encoder map to measure the distance in the input space. We demonstrate that the distortion in the latent space results in a directional bias appearing in the form of an anisotropic metric tensor. We show that an optimal way for an adversary to design an attack is by moving along the dominant eigendirection of the induced metric tensor. Moreover, we propose scores to quantify robustness using the eigenspectrum of the pullback metric tensor. We hypothesise that methods reducing *posterior-prior gap* improve robustness, directly influencing the induced metric tensors. To this end, we demonstrate that the proposed scores correlate with the β parameter of $\beta-$VAE used to control robustness. To our knowledge, such a geometric view of the robustness of VAEs has not been previously investigated.

$\beta-$VAE is known to generate over-smooth samples for a high value of β. We introduce a simple mechanism using a mixup training (Zhang et al., 2018) that improves the robustness of VAE without much effect on reconstruction quality. Specifically, we introduce a regularisation loss term that forces the encoder to fill empty regions with linear interpolation in data space and ensures the decoder generates the respective interpolations, thus avoiding the issue of an unconstrained generation. We empirically demonstrate that such a training scheme improves the robustness measured by the proposed scores.

2 BACKGROUND

2.1 $\beta-$Variational autoencoder

$\beta-$VAE (Higgins et al., 2017) is a probabilistic encoder-decoder framework that simultaneously parameterises the latent distribution and emission distribution using deep neural networks. Consider a sample $\mathbf{x} \in \mathbf{X} = \mathbb{R}^N$ drawn from unknown data distribution $p(\mathbf{x})$, VAE learns an approximate posterior distribution $\mathsf{q}_\phi(\mathbf{z}|\mathbf{x})$ over latent variables $\mathbf{z} \in \mathbf{Z} = \mathbb{R}^{d_z}$ using a stochastic encoder network, and an emission distribution $p_\theta(\mathbf{x}|\mathbf{z})$ using a stochastic decoder network. The parameters θ of an encoder network and ϕ of a decoder network are learned by maximising the evidence lower bound (ELBO),

$$\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta KL[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \quad (1)$$

where KL stands for Kullback-Leibler divergence (Kullback and Leibler, 1951), and a parameter β controls the smoothness of latent distribution, setting $\beta = 1$ is equivalent to a standard VAE (Kingma and Welling, 2014).

2.2 Adversarial attacks on VAEs

The adversarial attacks on VAEs assume access to a pre-trained encoder-decoder network. The adversary aims to exploit the capacity of VAE by finding small perturbations in an input sample that lead to a large change in its latent encoding or reconstruction. In a supervised scenario, the adversary starts with input and finds a minimal change that can match the reconstruction to the known target. In an unsupervised setting, the aim is to maximise the distance in latent codes, which will corrupt the reconstruction. Several recent developments have proposed mechanisms for designing an adversary and evaluating the robustness of existing deep generative models (Tabacof et al., 2016; Willetts et al., 2021). Our paper takes a geometrical viewpoint of an unsupervised attack by analysing the metric tensor induced by the stochastic encoder network. We first introduce the unsupervised variational attack problem and later present our approach in Section 2.3.

For an encoder neural network $f_\theta : \mathbf{X} \to \mathbf{Z}$ the unsupervised latent space attack optimises the objective (Gondim-Ribeiro et al., 2018),

$$\max_{\boldsymbol{\eta}:||\boldsymbol{\eta}||_2 \leq \eta_0} d(f_\theta(\mathbf{x}), f_\theta(\mathbf{x} + \boldsymbol{\eta})) \quad (2)$$

where η_0 is a small constant that decides the severity of the attack, and $d(.,.)$ is a distance function that measures the proximity in the latent space. A common approach to finding a corruption $\boldsymbol{\eta}$ is to use stochastic gradient methods (Sun et al., 2021; Willetts et al., 2021).
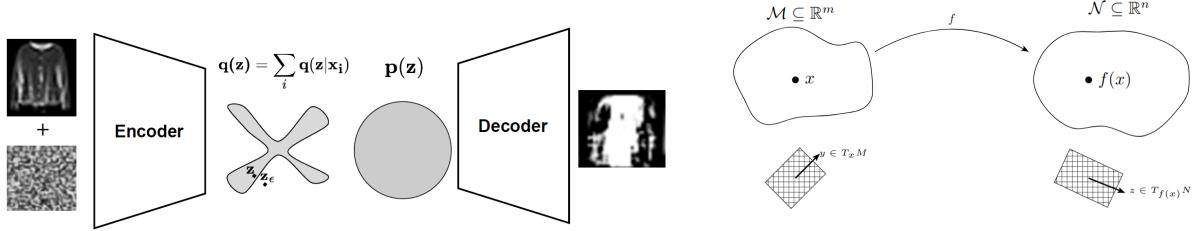
Figure 1: *Left:* Illustration that adversarial examples find non-smooth change in the latent encodings. A small perturbation in the input sample exploits a direction that maximally changes latent encoding by moving from high density to low/zero density region in the latent space. In this paper, we show an optimal perturbation can be found by moving along the eigendirection of the local pullback metric tensor of a data point. *Right:* A smooth mapping f from the data manifold $\mathcal{M}$ to the latent manifold $\mathcal{N}$ induces pullback metrics on $\mathcal{M}$. The jacobian $\mathbf{J}_{f(x)} = \frac{\partial f}{\partial \mathbf{x}}$ is a linear map from a tangent vector $y \in T_x M$ to a tangent vector $z \in T_{f(x)} N$ that induces a pullback Riemannian metric tensor $\mathbf{G_x} = \mathbf{J}_{f(x)}^T \mathbf{J}_{f(x)}$. The determinant of metric tensor $\mathbf{G_x}$ represents the change in infinitesimal volume element when projected to the latent space.

2.3 Latent space geometry

In this section, we introduce definitions from Riemannian geometry relevant to the context of our work.

Definition 2.1. A $n-$dimensional manifold $\mathcal{M}$ is a topological space where for every $\mathbf{x} \in \mathcal{M}$ there exist a neighborhood region $\mathcal{V}_\mathbf{x}$ homeomorphic to $\mathcal{R}^n$ (Lee, 2006).

Definition 2.2. A Riemannian metric for a smooth manifold $\mathcal{M}$ is a bilinear, symmetric, positive definite map $\mathbf{G}_x : \mathcal{T}_x \mathcal{M} \times \mathcal{T}_\mathbf{x} \mathcal{M} \to \mathbb{R}$ for all $\mathbf{x} \in \mathcal{M}$, where $T_\mathbf{x} \mathcal{M}$ is a tangent plane at point x on the manifold (Lee, 2006).

Definition 2.3. A smooth manifold $\mathcal{M}$ with a Riemannian metric $\mathbf{G}$ defined on every point of a manifold is called a Riemannian Manifold (Lee, 2006).

Definition 2.4. Given a mapping $f : \mathcal{M} \to \mathcal{N}$ from smooth manifold $\mathcal{M}$ to $\mathcal{N}$, for any $\mathbf{x} \in \mathcal{M}$ the pull-back metric $\mathbf{G_x}$ induced by the mapping f is given as $\mathbf{G_x} = \mathbf{J}_{f(\mathbf{x})}^T \mathbf{G}_{f(\mathbf{x})} \mathbf{J}_{f(\mathbf{x})}$, where $\mathbf{J}_{f(\mathbf{x})} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$.

2.4 Latent space distortion

To understand the latent distortion, we use an alternative derivation of ELBO (Makhzani et al., 2016),

$$- RE[\log p_\theta(\mathbf{x}|\mathbf{z})] - CE[q_\phi(\mathbf{z})||p(\mathbf{z})] + H[q_\phi(\mathbf{z}|\mathbf{x})] \quad (3)$$

where RE is the reconstruction error, CE is the cross entropy, and H is the entropy. For a Gaussian encoder, the H term maximises the variance of latent encoding, the RE term forces the distribution to be peakier (low variance), and the CE term is to force the approximated posterior to match with the prior distribution. Optimizing the full objective means the encoder has to match a fixed prior and simultaneously have a peaky distribution. Due to this tradeoff, the encoder fails to match the fixed prior, resulting in the *posterior-prior gap*. As a result, latent space is distorted with regions of low or zero density for which the decoder is

unconstrained. For the derivation of Equation 3 with a more comprehensive discussion, we refer readers to (Makhzani et al., 2016; Tomczak, 2022).

The *posterior-prior gap* implies small changes in the input sample can result in significant changes in the latent encoding leading to abrupt changes in the reconstruction. Figure 1 on the left demonstrates an example of a vulnerability of VAEs.

2.5 Related work

The distortions in the latent space of VAEs have been previously shown to result in limited generalisation capacity of VAEs (Rezende and Viola, 2018; Chen et al., 2020). Rezende and Viola (2018) further proposed a constrained optimisation to control the model performance. In our work, we take a geometric view of such distortion and use it to investigate the robustness of VAEs.

The use of encoder Jacobian matrices has previously appeared in approximating a tangent space of a data manifold, which captures the data points' sensitivity to its latent encoding (Rifai et al., 2011b,a). However, they don't consider the stochasticity of the encoder and decoder mappings. Several recent works treat the decoder mapping of VAEs as a smooth immersion and use the pullback as an induced metric in the latent space. The computation of such metrics has been helpful in various applications such as drawing on manifold samples, latent space interpolation, clustering, motion planning and many more (Arvanitidis et al., 2018; Yang et al., 2018; Hauberg, 2018; Chen et al., 2018a; Shao et al., 2017; Arvanitidis et al., 2020; Mohammadi et al., 2022). (Chen et al., 2020) propose to flatten the pullback metric tensor in the latent space induced by the decoder, which allows them to use standard Euclidean distance as a metric in latent space. The computation of a pullback in Chen et al. (2018a); Yang et al. (2018); Chen et al. (2020) does not consider the contribution of the uncertainty in the

MNIST: (a) Original ↓ Reconstruction   (b) Corrupted δ_1 ↓ Reconstruction   (c) Corrupted δ_2 ↓ Reconstruction

FMNIST: (d) Original ↓ Reconstruction   (e) Corrupted δ_1 ↓ Reconstruction   (f) Corrupted δ_2 ↓ Reconstruction
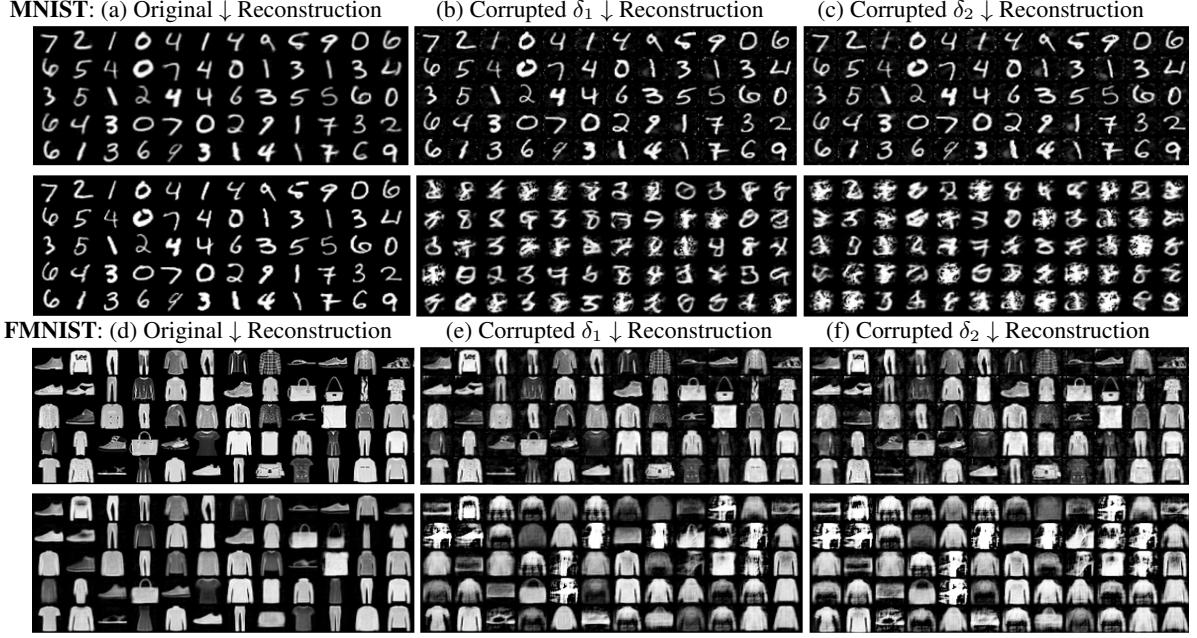
Figure 2: Illustration of adversarial attack along the dominant eigenvector of a stochastic pullback metric tensor. The first two rows are the results of MNIST data, and the bottom two are on the FashionMNIST dataset. We evaluate the reconstruction for original images and its two corrupted versions with different step sizes $\delta_1 = 0.5233$ and $\delta_2 = 0.7443$. Moving along eigendirection doesn't affect the input image but significantly changes its reconstruction.

decoder mapping. It is, therefore, limited in its ability to capture the topological properties of the manifold. Arvanitidis et al. (2018); Hauberg (2018) proposed to consider the uncertainty of the decoder by treating the Gaussian decoder as a random projection of a deterministic manifold. This viewpoint allows them to treat the reconstruction space as a random manifold and the pullback metric tensors as stochastic, proving helpful in handling topological holes and low-density regions.

We want to remark much of the existing methods focus on improving the sampling in the latent space or improving the performance on tasks. Here, we show such distortions can be exploited by an adversary and discuss the importance of local geometry for evaluating and training robust VAEs. Previously Zhao et al. (2019); Sun et al. (2019); Martin and Elster (2020) studied the spectrum of Fisher information (pullback from probability simplex to the input space) of a classifier to investigate the robustness of adversarial perturbations. However, there is no such study for generative models to our knowledge. Also, the metric tensor considered in our paper considers the effect of uncertainty in the latent space, which is vital for understanding the latent distortions.

3   ADVERSARIAL ATTACK EXPLOIT LOCAL GEOMETRY

Consider an encoder f_ϕ as a smooth immersion from a data manifold to a latent manifold. We can then utilise the pullback metric tensor induced by f_ϕ to express the infinitesimal distance in the input space in terms of the local metric tensor of the latent space. Thus, unlike the existing methods that rely on Euclidean distance in the input space, we use the pullback metric tensor to measure the infinitesimal distance.

Here, we first express the adversarial optimisation problem in terms of the pullback metric tensor. Next, we show that the adversary can exploit the directional bias of a metric tensor to design optimal attacks.

Remark 3.1. The infinitesimal distance between the representation of any data point $\mathbf{x}$ and its perturbation $\mathbf{x}_\eta = \mathbf{x} + \boldsymbol{\eta}$ under parametric encoder f_θ for $\boldsymbol{\eta}$ small in l_2 is approximated as,

i) $d(f_\theta(\mathbf{x}), f_\theta(\mathbf{x} + \boldsymbol{\eta})) = \boldsymbol{\eta}^T \mathbf{G_x} \boldsymbol{\eta}$, where $\mathbf{G_x} = \mathbf{J}_{f_\theta(\mathbf{x})}^T \mathbf{J}_{f_\theta(\mathbf{x})}$ for locally flat latent manifold,

ii) $d(f_\theta(\mathbf{x}), f_\theta(\mathbf{x} + \boldsymbol{\eta})) = \boldsymbol{\eta}^T \mathbf{G_x} \boldsymbol{\eta}$, where $\mathbf{G_x} = \mathbf{J}_{f_\theta(\mathbf{x})}^T \mathbf{G_z} \mathbf{J}_{f_\theta(\mathbf{x})}$ for a latent Riemannian manifold equipped with metric tensor $\mathbf{G_z}$. Here $\mathbf{G_z} = \mathbf{J}_{g_\phi(\mathbf{z})}^T \mathbf{J}_{g_\phi(\mathbf{z})}$ is a pullback under the parametric decoder mapping g_ϕ.

where $\mathbf{J}_{f_\theta(\mathbf{x})} \in \mathbb{R}^{d_z \times N}$ is a Jacobian matrix, d_z is the dimensionality of $\mathbf{Z}$ and N is the dimensionality of $\mathbf{X}$. The

matrix $\mathbf{G_x}$ is a symmetric, positive definite matrix known as a pullback metric tensor under the mapping f_θ. We can use it to measure the local inner product for every $\mathbf{x}$ in the input space $\mathbf{x} \in \mathbf{X}$.

3.1 Stochastic pull-back metric tensor

Eklund and Hauberg (2019); Arvanitidis et al. (2018) proposed an expected pullback metric tensor for measuring distances in the latent space of deep generative models. Here we utilise the pullback of an encoder mapping to evaluate the robustness of representations. Consider a stochastic encoder mapping f_θ expressed as a combination of mean $\boldsymbol{\mu}_\theta(\mathbf{x})$ and standard deviation $\boldsymbol{\sigma}_\theta(\mathbf{x})$ parameterisations: $f_\theta(\mathbf{x}) = \boldsymbol{\mu}_\theta(\mathbf{x}) + \boldsymbol{\epsilon} \odot \boldsymbol{\sigma}_\theta(\mathbf{x})$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_d)$, then a Jacobian of f_θ with respect to input $\mathbf{x}$ can be expressed as $\mathbf{J}_{f_\theta} = \mathbf{J}_{\boldsymbol{\mu(x)}} + \boldsymbol{\epsilon} \odot \mathbf{J}_{\boldsymbol{\sigma(x)}}$ and the pullback matrix $\mathbf{G_x}$ as,

$$\mathbf{G_x} = (\mathbf{J}_{\boldsymbol{\mu(x)}} + \boldsymbol{\epsilon} \odot \mathbf{J}_{\boldsymbol{\sigma(x)}})^T (\mathbf{J}_{\boldsymbol{\mu(x)}} + \boldsymbol{\epsilon} \odot \mathbf{J}_{\boldsymbol{\sigma(x)}})$$
$$= \mathbf{J}_{\boldsymbol{\mu(x)}}^T \mathbf{J}_{\boldsymbol{\mu(x)}} + \mathbf{J}_{\boldsymbol{\mu(x)}}^T \boldsymbol{\epsilon} \mathbf{J}_{\boldsymbol{\mu(x)}} + \mathbf{J}_{\boldsymbol{\sigma(x)}}^T \boldsymbol{\epsilon} \mathbf{J}_{\boldsymbol{\mu(x)}} + \mathbf{J}_{\boldsymbol{\sigma(x)}}^T \boldsymbol{\epsilon}^2 \mathbf{J}_{\boldsymbol{\sigma(x)}}$$

We can view the latent space as a random projection of a deterministic manifold. Under the assumption the sample paths from f_θ are smooth, we can treat the metric tensor as a stochastic matrix. The metric of the random latent manifold can be estimated in expectation as $\hat{\mathbf{G}}_\mathbf{x} = \mathbb{E}_{\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})}[\mathbf{G_x}]$. Since, $\boldsymbol{\epsilon}$ is a zero mean and unit covariance the $\mathbb{E}[\boldsymbol{\epsilon}] = 0$ and $\mathbb{E}[\boldsymbol{\epsilon}^2] = 1$ the final expected metric tensor is,

$$\hat{\mathbf{G}}_\mathbf{x} = \mathbf{J}_{\boldsymbol{\mu(x)}}^T \mathbf{J}_{\boldsymbol{\mu(x)}} + \mathbf{J}_{\boldsymbol{\sigma(x)}}^T \mathbf{J}_{\boldsymbol{\sigma(x)}} \quad (4)$$

Theorem 3.2. *Given a stochastic encoder mapping f_θ, for an arbitrary data point $\mathbf{x} \in \mathbf{X}$ the adversarial perturbation $\mathbf{x}_\eta$ under l_2 norm is optimal when moving along the eigendirection of a stochastic pullback metric tensor induced by f_θ at $\mathbf{x}$.*

Proof. By using the stochastic pull back metric tensor given in Equation 4, we reformulate the adversarial attack optimisation of Equation 5 as,

$$\max_{\boldsymbol{\eta}} \quad \boldsymbol{\eta}^T \hat{\mathbf{G}}_\mathbf{x} \boldsymbol{\eta}$$
$$\text{subject to} \quad ||\boldsymbol{\eta}||_2 = \eta_0 \quad (5)$$

Next, to solve the problem, we combine the constraints by introducing Lagrange multiplier $\boldsymbol{\lambda}$,

$$\max_{\boldsymbol{\eta}} \quad \boldsymbol{\eta}^T \hat{\mathbf{G}}_\mathbf{x} \boldsymbol{\eta} + \boldsymbol{\lambda}(\eta_0 - ||\boldsymbol{\eta}||_2^2) \quad (6)$$

The closed-form solution of the above optimisation takes the form:

$$\hat{\mathbf{G}}_\mathbf{x} \boldsymbol{\eta} = \boldsymbol{\lambda} \boldsymbol{\eta} \quad (7)$$

thus a pair $(\boldsymbol{\lambda}, \boldsymbol{\eta})$ represents the eigenvalue and eigenvector of the stochastic pullback metric tensor, the eigenvector with the largest eigenvalue corresponds to the direction of maximal change. $\square$


(a) $\beta = 0.01$ (b) $\beta = 1.0$ (c) $\beta = 7.5$
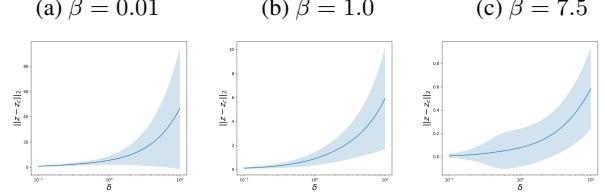
Figure 3: The plot shows the change in the latent encoding of $\beta-$VAE (in terms of Euclidean distance) for different values of β when moving along the dominant eigendirection of a pullback metric tensor $\hat{\mathbf{G}}_\mathbf{x}$ with different step size δ. We can see for small β, the changes are of much higher magnitude compared to larger β, demonstrating that increasing the β makes the latent space more smooth.

Thus, for any given input $\mathbf{x}$ an adversary can design an optimal attack by taking a step along the eigendirection of $\hat{\mathbf{G}}_\mathbf{x}$ as $\mathbf{x}_c = \mathbf{x} + \delta\lambda\boldsymbol{\eta}$, where δ is a step size. To compromise the reconstruction the step size δ can be chosen such that $||\mathbf{x} - \hat{\mathbf{x}}_c||_2 > ||\mathbf{x} - \hat{\mathbf{x}}||_2$, where $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}_c$ are reconstructions of original input and its corrupted version.

3.2 Robustness evaluation

A robustness method should suppress the maximum eigenvalue of the pullback metric tensor. Moreover, it should eliminate the directional bias resulting from the anisotropic distribution of eigenvalues of a pullback metric tensor. To quantify these two effects, we report the following two scores,

Spectral Radius for a matrix $\mathbf{G}$ is defined as,

$$\rho(\mathbf{G}) = \max\{|\boldsymbol{\lambda}|, \boldsymbol{\lambda} \text{ is an eigenvalue of } \mathbf{G}\} \quad (8)$$

A robust model will have a low value of spectral radius.

Von Neumann Entropy (Bengtsson et al., 2008) $\mathbf{S}$ of a metric tensor $\mathbf{G}$ is given by the Shannon entropy of its eigenvalues $\mathbf{S} = -\sum_k \boldsymbol{\lambda}_k \log \boldsymbol{\lambda}_k$. The high value would imply the metric tensor is anisotropic, resulting in a directional bias. Thus, a robust model will have a low value of $\mathbf{S}$.

4 RESULTS AND DISCUSSION

As pointed out in Remark (3.1), we can treat the latent space as (i) locally flat or (ii) a Riemannian manifold equipped with a metric tensor at all points. In the latter case, similar to the encoder, the stochastic pull-back metric tensor in the latent space can be computed as $\mathbf{G_z} = \mathbf{J}_{g_\omega(\mathbf{z})}^T \mathbf{J}_{g_\omega(\mathbf{z})}$ where g_θ is a decoder network. The main paper presents case (i) results where latent space is $\mathbf{G_z} = \mathbf{I}$. We also analyze the case $\mathbf{G_z}$ is a pullback induced by the decoder network; the results for this one are reported in Supplementary B.

We used PyTorch (Paszke et al., 2019) for the implementation of our work. The experimental details, including

CelebA: (a) Original ↓ Reconstruction   (b) Corrupted δ_1 ↓ Reconstruction   (c) Corrupted δ_2 ↓ Reconstruction

Figure 4: Illustration of adversarial attack along the dominant eigenvector of a stochastic pullback metric tensor on CelebA dataset. We evaluate the reconstruction for original images and its two corrupted versions with different step sizes $\delta_1 = 0.5233$ and $\delta_2 = 0.7443$.

the training procedure, are discussed in Supplementary Section A. The implementation is publicly available on GitHub[1]. Here, we first empirically demonstrate the vulnerability of a VAE using a one-step attack along the dominant eigendirection of input samples. Next, we investigate the robustness of $\beta-$VAE and discuss a simple mixup strategy that fills a *posterior-prior gap* in the latent space, flattens the latent space and ensures the decoder generated valid samples.

4.1 Adversarial attack

Figure 2 demonstrates the two instances of corruption along the dominant eigenvector of $\beta = 1$ VAE on MNIST (Lecun et al., 1998) and FashionMNIST (Xiao et al., 2017) datasets. For each dataset, the three columns in the first row are a set of original images and their corrupted version with a step size of $\delta = 0.5223$ and $\delta = 0.7443$. In the second row, we report their respective reconstructions. We observe that with $\delta = 0.5223$, the reconstruction significantly differs from the original images, and for $\delta = 0.7443$ gets much more severe, exposing the capacity of VAE. This result proves an attacker can exploit the metric tensor's directional bias to design a one-shot attack. Figure 4 further demonstrates a similar finding on the CelebA (Liu et al., 2015) dataset.

4.2 Robustness evaluation

In this section, we first investigate the connection between our proposed scores and the robustness of $\beta-$VAE and later discuss an alternative robustness scheme using a mixup training loss.

4.2.1 $\beta-$VAE

The β parameter in a $\beta-$VAE controls the gap between an approximated posterior and a prior distribution. The high value of β reduces the gap, thus eliminating the latent distortions which an adversary can exploit. We now investigate

the effect of increasing values of β on the two scores. We sample 50 values of β with a logarithmic spacing between $[0.01, 10]$. We trained the encoder-decoder model for each parameter and computed the two scores $\rho(\hat{\mathbf{G}})$ and $\mathbf{S}$ for every sample point. Due to the high computational cost of training 50 different models per dataset, in this section, we limit the experiments to MNIST and FashionMNIST datasets.

Figure 5 first and the second column in row one (MNIST) and in row three (FashionMNIST) reports the histogram of the scores for four different values of β. We observe that the higher value of β suppresses the spectral radius. Similarly, the von Neumann entropy is decreased, demonstrating that the local directions get isotropic. Importantly, this indicates the adversary cannot exploit the directional bias for high values of β with $\boldsymbol{\eta}$ small in the norm. In the third and fourth columns of rows one and three, we report the mean and standard deviation of the scores computed for fifty increasing values of β. The results demonstrate that by reducing the KL gap, parameter β prevents distortion in the latent space eliminating the directional bias exploited by an adversary.

Next, we examine the connection between the step size δ and the strength of attacks under different values of β. We report the mean squared error (MSE) between an original image and its reconstruction under varying corruption rates along five dominant eigendirections. We generated 40 logarithmic spacing steps in the range $[0.01, 10]$. Figure 5 second row (MNIST) and fourth row (FashionMNIST) demonstrate the MSE vs step size averaged across all test samples for four different values of β. We observe that for the small β, all five directions tend to get high MSE, and as β increases, it requires a larger step size to get a significant change in MSE.

Further, to analyse the latent distortions, we generate perturbations of increasing magnitude for each sample $\mathbf{x}$ by increasing step size δ along the dominant eigendirection of the pullback metric tensor. In Figure 3, we report the

---

[1]https://github.com/MdAsifKhan/RobustnessVAE/
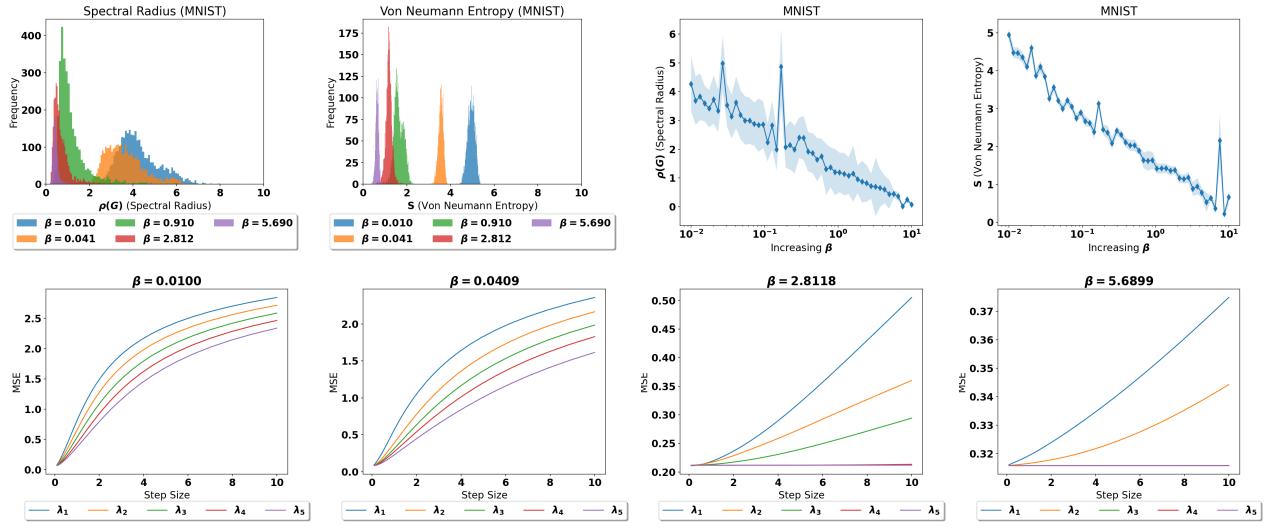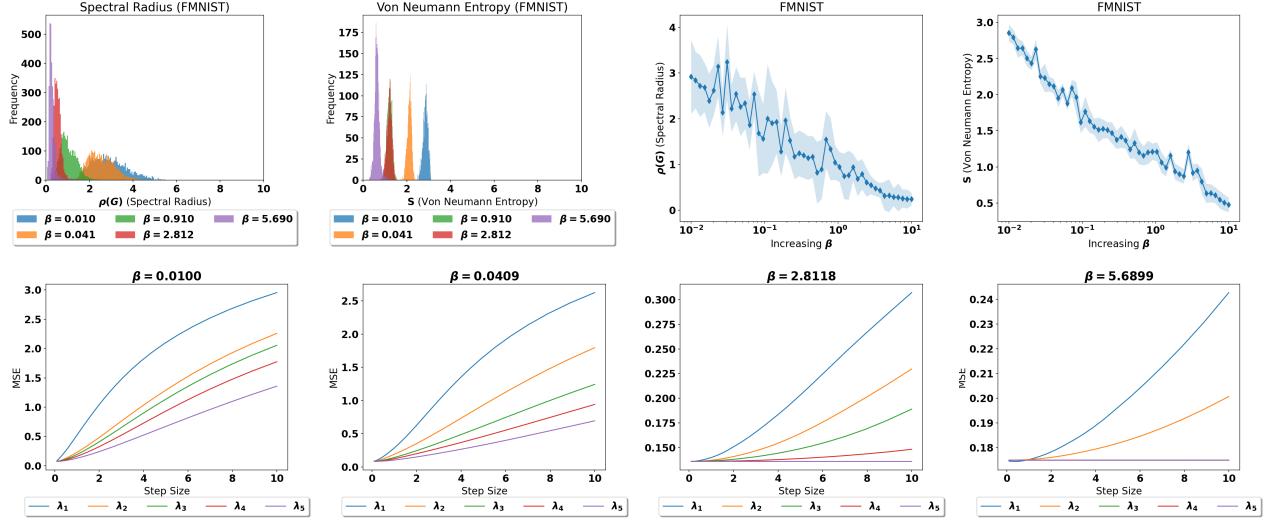
(a) Robustness evaluation of $\beta-$VAE on MNIST.

(b) Robustness evaluation of $\beta-$VAE on FashionMNIST.

Figure 5: Figure (a), on the left, we report the histogram of spectral radius and Von Neumann entropy (on test samples) for different values of β in β-VAE. On the right, we report the average of two scores across test samples for an increasing value of β. We observe that increasing the value of β suppresses the metric tensor's maximum eigenvalue, and the eigenspectrum distribution gets more isotropic. In the second row, we corrupt the test images along the top five eigendirections (denoted by $\lambda_1, \lambda_2, \lambda_3, \lambda_4,$ and λ_5) with an increasing step size for different values of β. The plots describe the average MSE across test samples. We observe that the average step size increases for a higher value of β. Increasing the value of β reduces the *posterior-prior gap*, minimising distortion in the latent space. Figure (b) demonstrates similar observations on the FashionMNIST dataset.

distance between the latent encoding of the original input and its perturbations averaged across data samples.

Increasing the β improves the contribution of the uncertainty term, showcasing that the probabilistic encoder-decoder model is more robust than its deterministic counterpart. The stochastic metric tensor comprises two terms resulting from the latent space's variational distribution. For small values of β, the encoder does not do well in quantifying the uncertainty and fails to match the prior; as an outcome, the latent

space is more distorted, resulting in empty and low-density regions given by dominating eigendirections of the metric tensor. The distortions are reduced for a high value of β; accordingly, the second term can account for the uncertainty in latent space preventing the eigenvalues from getting large.

In Supplementary Section B, we report the results for the case where latent space is equipped with a metric tensor $\mathbf{G_z}$ induced by the decoder network. We observe a higher value of spectral radius when incorporating the geometry of the

Robustness evaluation of *mixup* on MNIST and FashionMNIST.
(a) MNIST   (b) FashionMNIST
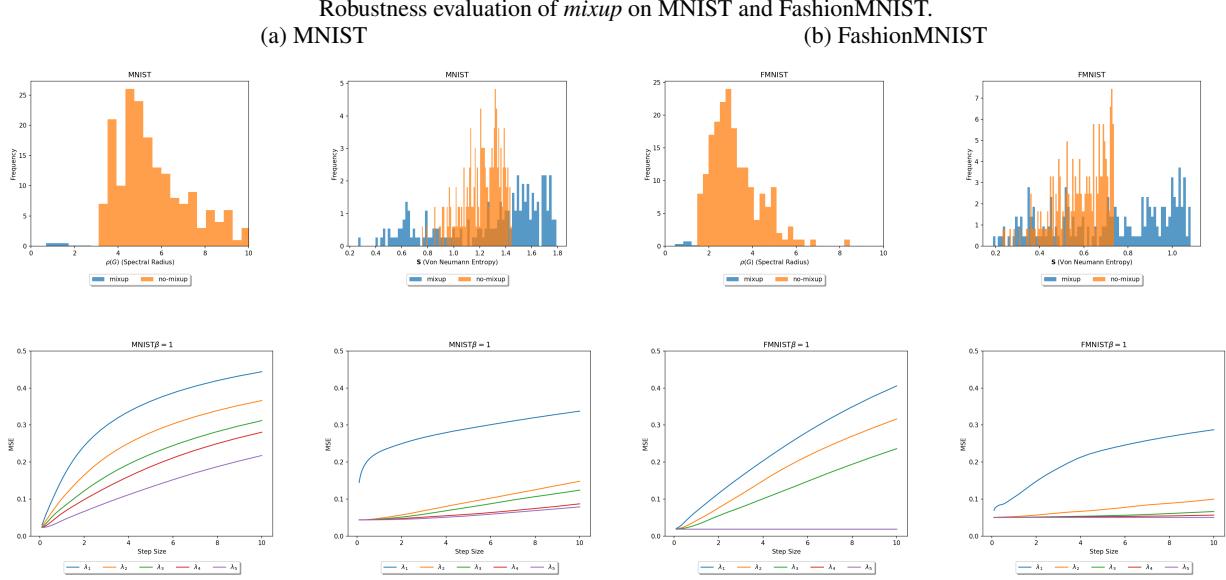
Figure 6: Figure (a), first row, two columns are histograms of spectral radius and Von Neumann entropy (on test samples) with and without mixup regularisation. Second row, we corrupt the test images along the top five eigendirections (denoted by $\lambda_1, \lambda_2, \lambda_3, \lambda_4,$ and λ_5) with an increasing step size for different values of β. The plots describe the average MSE across test samples. We observe mixup suppresses the spectral radius, and the eigenspectrum distribution gets more isotropic. Figure (b) demonstrates similar observations on the FashionMNIST dataset.

decoder, which implies the latent space is locally curved.

4.2.2 Latent *mixup*

In $\beta-$VAE, a higher value of β reduces the *posterior-prior gap* but degrades the reconstruction quality. We propose filling up the low or zero-density region in the latent space as an alternative solution to the latent distortion problem. To this end, we utilise a *mixup* training strategy. Mixup is a simple and powerful data augmentation technique that has been shown to improve the robustness and generalisation capabilities of classification models (Zhang et al., 2018; Lamb et al., 2019; Verma et al., 2019). For a given pair of distinct data points $\mathbf{x}_i$ and $\mathbf{x}_j$, we use linear interpolation between them to draw samples in the low-density region; likewise, we take the linear interpolation between the encodings $\mathbf{z}_i$ and $\mathbf{z}_j$, of the respective samples,

$$\mathbf{z}_m = \alpha\mathbf{z}_i + (1-\alpha)\mathbf{z}_j, \quad \mathbf{x}_m = \alpha\mathbf{x}_i + (1-\alpha)\mathbf{x}_j \quad (9)$$

where α is sampled from a Beta distribution $\mathcal{B}(a,b)$ with shape parameters a and b that we set to 0.5. We now introduce the following regularisation penalty to a VAE objective,

$$C = ||\mathbf{z}_m - f_\phi(\mathbf{x}_m)||_2 + ||g_\theta(\mathbf{z}_m) - \mathbf{x}_m||_2 \quad (10)$$

where the first term is to force the encoder to match mixing in the input space, and the second term is to force the decoder to match mixing in the latent space to mixing in the

reconstruction space. Combining the two loss terms lets the encoder fill the empty region of latent space such that the linear interpolation in a latent space corresponds to linear interpolation in input space. Thus reducing the gap between the posterior and a prior and preventing the decoder from generating unconstrained output. We report the robustness scores for increasing step size δ.

Figure 6 shows the comparison of robustness scores with mixup training on MNIST and FashionMNIST datasets. We observe mixup suppresses the spectral radius for two datasets, thus improving the robustness. Figure 7 further demonstrates the qualitative performance of VAE trained with a *mixup* loss.
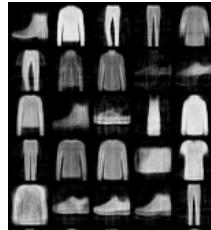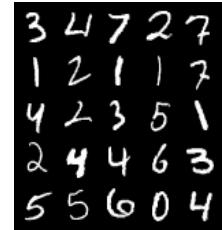
5 CONCLUSION, LIMITATIONS AND FUTURE SCOPE

We have presented a geometrical perspective of adversarial attacks and introduced scores for measuring the robustness of VAEs. We have shown that the sensitivity of the encoder at a given input depends on eigendirections of stochastic pullback metric tensor that an adversary can exploit to design an attack. We proposed evaluation scores using the spectral radius and Von Neumann's entropy of a pullback metric tensor. Moreover, we demonstrate the scores correlate with parameter β of $\beta-$VAE, providing geometrical insights into the robustness due to parameter β of $\beta-$VAE.

A caveat with $\beta-$VAE is that increasing the β trades off the

FashionMNIST: (a) Original (b) Reconstruction **MNIST**: (c) Original (d) Reconstruction
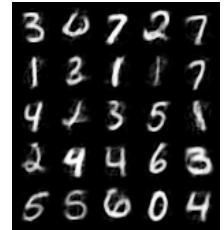

Figure 7: Here, we report a qualitative evaluation of training VAE with a mixup loss. We report the input samples and their respective reconstructions. On the left are the results of the FashionMNIST and the rights of MNIST.

capacity of representations with the quality of reconstruction resulting in over-smooth reconstructions for a large value of β, which further implies a tradeoff between the spectral radius and robustness. To circumvent the above issue, we utilised a *mixup* training scheme that fills the empty region of latent space and improves the robustness measured in terms of proposed scores. We want to remark *mixup* does not guarantee that all empty or low-density regions are filled in the latent space. Here coverage depends on the regularisation hyperparameter as well as the size of the training dataset. Thus, it is still possible the regions of zero-low density exist. However, the overall training scheme reduces latent distortions.

Recently few mechanisms have been proposed to reduce the distortion in latent space and improve the robustness of VAEs (Willetts et al., 2021; Kuzina et al., 2021). We hypothesise that the benefits of such robustness measures can be better established geometrically by investigating their pullback metric tensors. Another recent work (Kuzina et al., 2022) shows the adversarial attacks on VAEs target to move latent encoding to the region of low or zero density. Moreover, they proposed using a Markov Chain Monte Carlo (MCMC) as a correction scheme during the inference time. We wish to investigate the applicability of the proposed scores to the above methods in future work.

A limitation of our current work is that we only consider the unsupervised attack when a target sample is unknown. Moreover, there can be different forms of attack by replacing l_2-norm with more general $p-$norms. We wish to study these in future work.

Acknowledgements

References

Arvanitidis, G., Hansen, L. K., and Hauberg, S. (2018). Latent Space Oddity: on the Curvature of Deep Generative Models. In *International Conference on Learning Representations*.

Arvanitidis, G., Hauberg, S., and Schölkopf, B. (2020). Geometrically Enriched Latent Spaces. In *International Conference on Artificial Intelligence and Statistics*.

Bengtsson, I., Zyczkowski, K., and Milburn, G. J. (2008). Geometry of quantum states: an introduction to quantum entanglement by ingemar bengtsson and karol zyczkowski. *Quantum Inf. Comput.*, 8(8):860.

Camuto, A., Willetts, M., Roberts, S. J., Holmes, C. C., and Rainforth, T. (2021). Towards a Theoretical Understanding of the Robustness of Variational Autoencoders. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021,*, volume 130 of *Proceedings of Machine Learning Research*, pages 3565–3573. PMLR.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D. (2021). A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45.

Chen, N., Klushyn, A., Ferroni, F., Bayer, J., and Van Der Smagt, P. (2020). Learning Flat Latent Manifolds with VAEs. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Chen, N., Klushyn, A., Kurle, R., Jiang, X., Bayer, J., and Smagt, P. (2018a). Metrics for Deep Generative Models. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1550. PMLR.

Chen, R. T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. (2018b). Isolating Sources of Disentanglement in Variational Autoencoders. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Eklund, D. and Hauberg, S. (2019). Expected path length on random manifolds. *ArXiv*, abs/1908.07377.

Esmaeili, B., Wu, H., Jain, S., Bozkurt, A., Siddharth, N., Paige, B., Brooks, D. H., Dy, J., and van de Meent, J.-W. (2019). Structured Disentangled Representations. In

Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2525–2534. PMLR.

Gondim-Ribeiro, G., Tabacof, P., and Valle, E. (2018). Adversarial Attacks on Variational Autoencoders. *arXiv preprint arXiv:1806.04646*.

Hauberg, S. (2018). Only Bayes should learn a manifold (on the estimation of differential geometric structure from data). *ArXiv*, abs/1806.04994.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. (2017). 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.

Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France. PMLR.

Kim, H. and Mnih, A. (2018). Disentangling by Factorising. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kos, J., Fischer, I., and Song, D. X. (2017). Adversarial Examples for Generative Models. *2018 IEEE Security and Privacy Workshops (SPW)*, pages 36–42.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86.

Kuzina, A., Welling, M., and Tomczak, J. M. (2021). Diagnosing Vulnerability of Variational Auto-Encoders to Adversarial Attacks. *RobustML Workshop@ICLR 2021*.

Kuzina, A., Welling, M., and Tomczak, J. M. (2022). Alleviating Adversarial Attacks on Variational Autoencoders with MCMC. In *Advances in Neural Information Processing Systems*.

Lamb, A., Verma, V., Kannala, J., and Bengio, Y. (2019). Interpolated Adversarial Training: Achieving Robust Neu-

ral Networks Without Sacrificing Too Much Accuracy. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, page 95–103, New York, NY, USA. Association for Computing Machinery.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, J. M. (2006). *Riemannian manifolds: An Introduction to Curvature*, volume 176. Springer Science & Business Media.

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

Makhzani, A., Shlens, J., Jaitly, N., and Goodfellow, I. (2016). Adversarial Autoencoders. In *International Conference on Learning Representations*.

Martin, J. and Elster, C. (2020). Inspecting adversarial examples using the fisher information. *Neurocomputing*, 382:80–86.

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. (2019). Disentangling Disentanglement in Variational Autoencoders. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4402–4412. PMLR.

Mohammadi, H. B., Hauberg, S., Arvanitidis, G., Neumann, G., and Rozo, L. D. (2022). Reactive Motion Generation on Learned Riemannian Manifolds. *arxiv*, abs/2203.07761.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Rezende, D. J. and Viola, F. (2018). Taming VAEs. *arXiv preprint arXiv:1810.00597*.

Rifai, S., Dauphin, Y. N., Vincent, P., Bengio, Y., and Muller, X. (2011a). The Manifold Tangent Classifier. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011b). Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 833–840, Madison, WI, USA. Omnipress.

Roberts, A., Engel, J., and Eck, D. (2017). Hierarchical Variational Autoencoders for Music. In *Workshop on Machine Learning for Creativity and Design, NIPS*.

Shao, H., Kumar, A., and Fletcher, P. T. (2017). The Riemannian Geometry of Deep Generative Models. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 428–4288.

Sun, H., Zhu, T., Zhang, Z., Jin, D., Xiong, P., and Zhou, W. (2021). Adversarial Attacks Against Deep Generative Models on Data: A Survey. *IEEE Transactions on Knowledge & Data Engineering*, (01):1–1.

Sun, K., Koniusz, P., and Wang, Z. (2019). Fisher-Bures Adversary Graph Convolutional Networks. In *Uncertainty in Artificial Intelligence*, page 161. arXiv:1903.04154 [cs.LG].

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Tabacof, P., Tavares, J., and Valle, E. (2016). Adversarial Images for Variational Autoencoders. *Adversarial Training Workshop, NIPS*.

Tomczak, J. M. (2022). *Deep Generative Modeling*. Springer.

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

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., and Bengio, Y. (2019). Manifold mixup: Better representations by interpolating hidden states. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6438–6447. PMLR.

Willetts, M. J., Camuto, A., Rainforth, T., Roberts, S., and Holmes, C. C. (2021). Improving {vae}s' Robustness to Adversarial Attack. In *International Conference on Learning Representations*.

Wu, B., Nair, S., Martín-Martín, R., Fei-Fei, L., and Finn, C. (2021). Greedy Hierarchical Variational Autoencoders for Large-Scale Video Prediction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2318–2328.

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

Yang, T., Arvanitidis, G., Fu, D., Li, X., and Hauberg, S. (2018). Geodesic Clustering in Deep Generative Models. *ArXiv*, abs/1809.04747.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhao, C., Fletcher, P. T., Yu, M., Peng, Y., Zhang, G., and Shent, C. (2019). The Adversarial Attack and Detection under the Fisher Information Metric. AAAI Press.

Supplementary Materials

A IMPLEMENTATION DETAILS

We use PyTorch Paszke et al. (2019) to implement our work, and all networks are trained on a single 11GB Nvidia RTX 2080 GPU. For MNIST and FashionMNIST datasets, we use the same architecture across all the experiments. The encoder network is a four-layer multi-layered perceptron (MLP) with 256, 256, 512 and 32 hidden units. The latent space distribution is a multivariate Gaussian with mean and standard deviation parameterised by two 32×32 linear mappings. We use the standard zero mean and unit covariance prior on the latent space. The decoder network is the inverse of an encoder with 32, 512, 256 and 256 hidden units. We use tanh as an activation function and batch-normalisation Ioffe and Szegedy (2015) before all activations.

For the CelebA dataset, the encoder is a convolutional neural network (CNN) with four convolution layers with an increasing number of filters 32, 64, 128, 256, and 512 followed by a dense layer that maps to a latent space. The latent distribution is a multivariate Gaussian with mean and standard deviation given by 128×128 linear mappings. The decoder is an inverse of encoder architecture with a dense layer that maps to 1024 hidden units followed by transpose convolution layers with decreasing number of filters 512, 256, 128, 64 and 32. All convolution layers are followed with a tanh activation function and batch-normalisation layer. For all models, we use Adam optimiser Kingma and Ba (2015) with a learning rate of 0.003.

B EXTENDED RESULTS

We also conducted experiments for the case when latent space is locally curved that is $\mathbf{G_z} = \mathbf{J}_{g_\phi(\mathbf{z})}^T \mathbf{J}_{g_\phi(\mathbf{z})}$ where g_ϕ is a decoder network with parameters ϕ. For a stochastic decoder mapping $g_\phi = \boldsymbol{\mu}(\mathbf{z}) + \boldsymbol{\epsilon} \odot \boldsymbol{\sigma}_\phi(\mathbf{z})$ the metric tensor $\mathbf{G_z}$ can further be expressed as $\mathbf{G_z} = \mathbf{J}_{\boldsymbol{\mu}_\phi(\mathbf{z})}^T \mathbf{J}_{\boldsymbol{\mu}_\phi(\mathbf{z})} + \mathbf{J}_{\boldsymbol{\sigma}_\phi(\mathbf{z})}^T \mathbf{J}_{\boldsymbol{\sigma}_\phi(\mathbf{z})}$. We use this expression of pullback metric tensor in the latent space to measure infinitesimal distance which results in the following expression of combined pullback metric tensor,

$$\hat{\mathbf{G}}_{\mathbf{x}} = \mathbf{J}_{\boldsymbol{\mu}_\theta(\mathbf{x})}^T \mathbf{G_z} \mathbf{J}_{\boldsymbol{\mu}_\theta(\mathbf{x})} + \mathbf{J}_{\boldsymbol{\sigma}_\theta(\mathbf{x})}^T \mathbf{G_z} \mathbf{J}_{\boldsymbol{\sigma}_\theta(\mathbf{x})} \tag{11}$$

Figure 8-9 show the results on MNIST and FashionMNIST. Figure 10 further discusses the analysis of robustness scores.
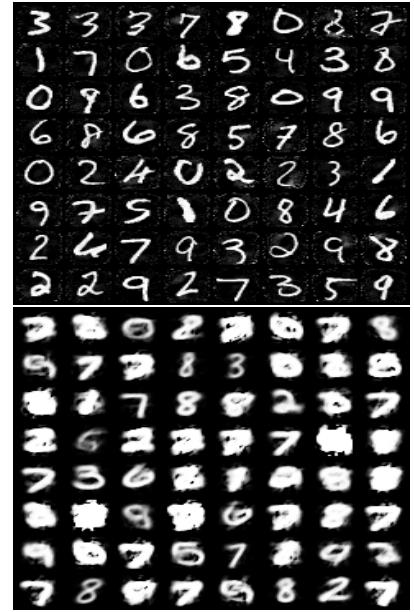
MNIST: (a) Original ↓ Reconstruction   (b) Corrupted δ_1 ↓ Reconstruction   (c) Corrupted δ_2 ↓ Reconstruction


Figure 8: Illustration of adversarial attack along the dominant eigenvector of a stochastic pullback metric tensor given by Equation 11. We evaluate the reconstruction for original images and its two corrupted versions with different step sizes $\delta_1 = 0.5233$ and $\delta_2 = 0.7443$.
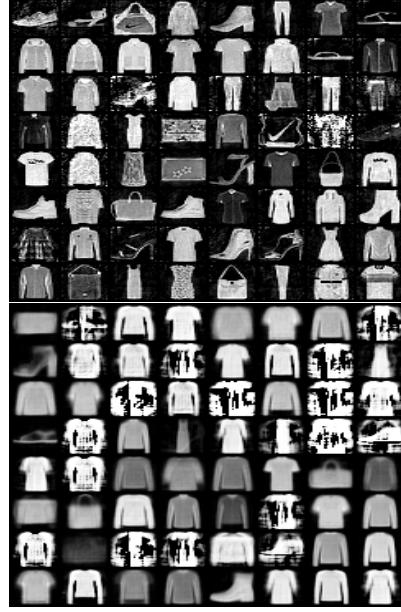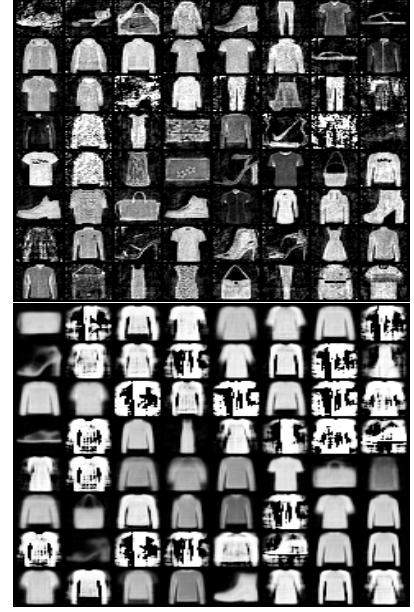
FMNIST: (a) Original ↓ Reconstruction   (b) Corrupted δ_1 ↓ Reconstruction   (c) Corrupted δ_2 ↓ Reconstruction

Figure 9: Illustration of adversarial attack along the dominant eigenvector of a stochastic pullback metric tensor given by Equation 11. We evaluate the reconstruction for original images and its two corrupted versions with different step sizes $\delta_1 = 0.5233$ and $\delta_2 = 0.7443$.

(a) Robustness evaluation of $\beta-$VAE on MNIST.

(b) Robustness evaluation of $\beta-$VAE on FashionMNIST.
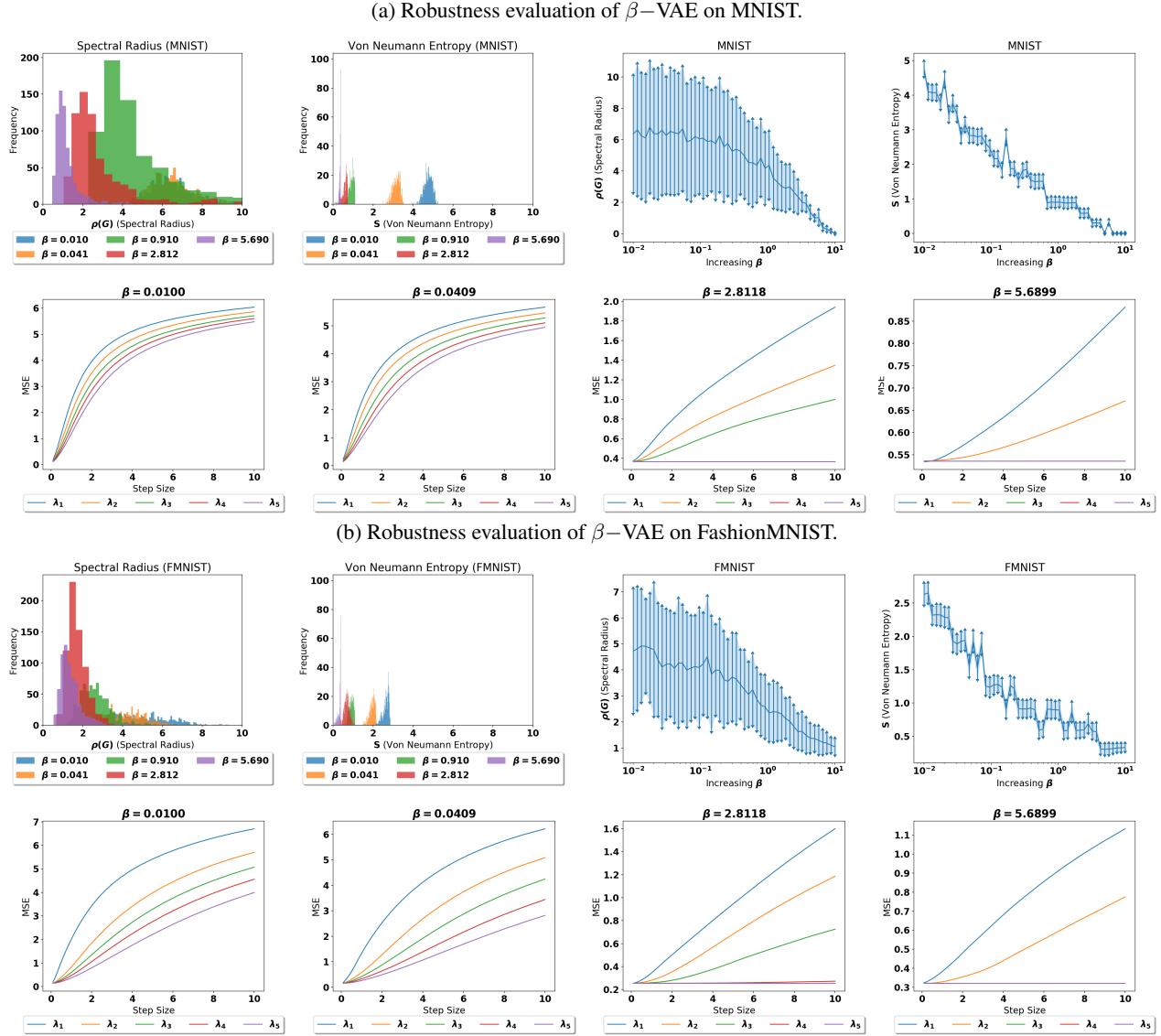
Figure 10: Evaluation using stochastic pullback metric tensor given by Equation 11. Figure (a), on the left, we report the histogram of spectral radius and Von Neumann entropy (on test samples) for different values of β in β-VAE. On the right, we report the average of two scores across test samples for an increasing value of β. Increasing the value of β suppresses the metric tensor's maximum eigenvalue, and the eigenspectrum distribution gets more isotropic. In the second row, we corrupt the test images along the top five eigendirections (denoted by $\lambda_1, \lambda_2, \lambda_3, \lambda_4,$ and λ_5) with an increasing step size for different values of β. The plots describe the average MSE across test samples. We observe for a higher value of β, the average step size increases. Increasing the value of β reduces the *posterior-prior gap*, minimising distortion in the latent space. Figure (b) demonstrates similar observations on the FashionMNIST dataset.