Towards a Theoretical Understanding of the Robustness of Variational Autoencoders

Alexander Camuto[1,2] **Matthew Willetts**[1,2]
Stephen Roberts[1,2] **Chris Holmes**[1,2] **Tom Rainforth**[1]
[1]University of Oxford [2]Alan Turing Institute

Abstract

We make inroads into understanding the robustness of Variational Autoencoders (VAEs) to adversarial attacks and other input perturbations. While previous work has developed algorithmic approaches to attacking and defending VAEs, there remains a lack of formalization for what it means for a VAE to be robust. To address this, we develop a novel criterion for robustness in probabilistic models: r-robustness. We then use this to construct the first theoretical results for the robustness of VAEs, deriving margins in the input space for which we can provide guarantees about the resulting reconstruction. Informally, we are able to define a region within which any perturbation will produce a reconstruction that is similar to the original reconstruction. To support our analysis, we show that VAEs trained using disentangling methods not only score well under our robustness metrics, but that the reasons for this can be interpreted through our theoretical results.

1 Introduction

Variational Autoencoders (VAEs) (Rezende et al., 2014; Kingma & Welling, 2014) have been found to be more robust to input perturbations than their deterministic counterparts, particularly those originating from adversarial attacks (Willetts et al., 2019; Schott et al., 2019; Ghosh et al., 2018). This trait has made them useful in protecting downstream tasks (Willetts et al., 2019; Schott et al., 2019; Ghosh et al., 2018).

Nevertheless, they are still not completely impervious to attack (Tabacof et al., 2016; Gondim-Ribeiro et al., 2018a; Kos et al., 2018a): a hypothetical adversary can attack a VAE by applying small input perturbations to invoke meaningful changes in the encoding. Typically this is done by trying to find perturbations which produce reconstructions close to a distinct target datapoint chosen by the adversary, rather than being representative of the original input. Such attacks have been shown to be successful in a wide range of scenarios (Tabacof et al., 2016; Gondim-Ribeiro et al., 2018a; Kos et al., 2018a; Willetts et al., 2019). Recent work has made progress towards defending against them from an empirical and algorithmic perspective (Willetts et al., 2019), by repurposing approaches designed to learn disentangled latent representations (Burgess et al., 2018; Chen et al., 2018; Mathieu et al., 2019).

However, a deeper understanding of the mechanisms underpinning the robustness of VAEs and their derivatives is still lacking. Furthermore, there are currently no theoretical foundations for this robustness or even any frameworks or formalizations for exactly what it means for a VAE to be "robust." In other words, what would it mean to have a certifiably–robust VAE? Moreover, are there scenarios where we might be able to provide theoretical guarantees of such robustness?

As a first step to addressing these questions, we develop the first metric with which to evaluate the robustness of VAEs: r-robustness. Informally, for a given input, a VAE is r-robust to a given perturbation if it is more likely that its reconstruction will fall within a ball of radius r around the undistorted maximum likelihood reconstruction, than outside it. The smaller the value of r for which we can confirm r-robustness, the more robust we can guarantee the VAE to be. Through r-robustness, we provide theoretical foundations to understand the source of VAEs' robustness and provide insights into what can cause them to more or less robust.

Using this, we next develop a *margin* of robustness, $R^r_{\mathcal{X}}(\mathbf{x})$, such that the VAE is r-robust to *any* possible perturbations of the input $\mathbf{x}$ within this margin. This, in turn, allows us to provide a notion of a certifiably–robust reconstruction as it forms a guarantee that no

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

Figure 1: Reconstructions under attack for robust and non-robust VAEs. Each subfigure shows from left to right: the original input, a perturbed input made by an adversarial attack, and the reconstruction of the perturbed input. We show results for VAEs that are robust ($R^r_{\mathcal{X}}(\mathbf{x}) \geq \|\boldsymbol{\delta}_x\|_2$) and non-robust ($R^r_{\mathcal{X}}(\mathbf{x}) < \|\boldsymbol{\delta}_x\|_2$) for a given point $\mathbf{x}$ and adversarially selected perturbation $\boldsymbol{\delta}_x$. We see that the robust VAE reconstructions are visually closer to the original input.

attack limited to the margin can reliably undermine it. An example of this is shown in Fig 1, where we demonstrate that large $R^r_{\mathcal{X}}(\mathbf{x})$ are associated with model–input pairs that are robust to adversarially generated input perturbations. Analogously to the concept of an adversarial risk (Uesato et al., 2018), $R^r_{\mathcal{X}}(\mathbf{x})$ can further be converted to a metric for the *overall* robustness of a VAE, by taking its expectation over the data generating distribution.

To make inroads towards imposing apriori constraints on a VAE that ensure that it is certifiably robust, we further derive a theoretical bound for $R^r_{\mathcal{X}}(\mathbf{x})$ as a function of the encoder variance and Jacobian. This provides insights into the characteristics of VAEs that contribute to robustness. Building on this result, we show empirically that VAEs with larger encoder variances and smaller Jacobians typically produce larger margins $R^r_{\mathcal{X}}(\mathbf{x})$ and are thus more robust to perturbations. We further demonstrate how these beneficial characteristics can be induced using methods introduced to learn disentangled representations, deriving new results for how these methods can be interpreted.

To summarize, our core contributions are that we first define a robustness metric, r-robustness, that is tailored to probabilistic generative models. We develop a margin $R^r_{\mathcal{X}}(\mathbf{x})$ on a VAE's input space within which it is r-robust to perturbations. Finally, we offer theoretical and empirical analysis—based on $R^r_{\mathcal{X}}(\mathbf{x})$ and existing disentanglement methods—that can aid the construction of robust VAEs.

2 Background

2.1 Variational Autoencoders

VAEs (Rezende et al., 2014; Kingma & Welling, 2014), and the models they have inspired (Alemi et al., 2017; Higgins et al., 2017a; Chen et al., 2018; Willetts et al., 2019), are deep latent variable models. Using $\mathbf{x} \in \mathcal{X}$ to denote data and $\mathbf{z} \in \mathcal{Z}$ to denote the latents with associated prior $p(\mathbf{z})$, a VAE simultaneously learns both a forward generative model, $p_\theta(\mathbf{x}|\mathbf{z})$, and an amortised approximate posterior distribution, $q_\phi(\mathbf{z}|\mathbf{x})$, where θ and ϕ correspond to their respective parameters, typically taking the form of neural networks. These are referred to as the decoder and encoder respectively, and a VAE can be thought of as a deep stochastic autoencoder. Under this autoencoder framework, one typically takes the reconstructions as deterministic, corresponding to the mean of the decoder, namely $g_\theta(\mathbf{z}) := \mathbb{E}_{p_\theta(\mathbf{x}|\mathbf{z})}[\mathbf{x}]$, a convention we adopt.

A VAE is trained by maximizing the evidence lower bound (ELBO) $\mathcal{L} = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}[\mathcal{L}(\mathbf{x})]$, where

$$\mathcal{L}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

and $p_{\mathcal{D}(\mathbf{x})}$ represents the empirical data distribution. The optimization is carried out using stochastic gradient descent with Monte Carlo samples, typically employing the reparameterization trick (Kingma & Welling, 2014). For example, for a Gaussian $q_\phi(\mathbf{z}|\mathbf{x})$, we draw samples as $\mathbf{z} = \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\eta} \circ \boldsymbol{\sigma}_\phi(\mathbf{x}), \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\circ$ is the element-wise product.

2.2 Adversarial Attacks for VAEs

In adversarial settings, an agent is trying to alter the behavior of a model towards a specific goal. This could involve, in the case of classification, adding a very small perturbation to an input so as to alter the model's predicted class. For many deep learning models, small changes to data imperceptible to the human eye, can drastically change a model's output.

Proposed attacks for VAEs aim to produce reconstructions close to a chosen target image by applying small distortions to the input of a VAE (Tabacof et al., 2016; Gondim-Ribeiro et al., 2018b; Kos et al., 2018b). The adversary optimizes this perturbation to minimize some measure of distance between the reconstruction and the target image *or* the distance between the embedding of the distorted image and the embedding of the target image.

2.3 Disentangled VAEs

Learning *disentangled* representations (Bengio et al., 2013) involves training a probabilistic generative model in a manner that encourages a one-to-one correspondence between dimensions of the learnt latent

space and some interpretable aspect of the data (Higgins et al., 2017a; Alemi et al., 2017; Burgess et al., 2018; Chen et al., 2018; Mathieu et al., 2019). For instance, there might be some axis in latent space encoding 'hair color' for images of people. One such method is the β-VAE (Burgess et al., 2018; Higgins et al., 2017a), which upweights the KL in the ELBO with a penalization factor β:

$$\mathcal{L}_\beta(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta\,\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))]$$

Disentangling can be difficult to achieve in practice, requiring careful hyperparameter tuning (Locatello et al., 2019; Mathieu et al., 2019; Rolinek et al., 2019).

Nevertheless, models trained under disentangling objectives have other beneficial properties. In particular, β-TC regularization (Chen et al., 2018), another method proposed for disentangling, has been shown to induce models that are more robust to adversarial attacks (Willetts et al., 2019). The encoders of β-VAEs have also been used as the perceptual part of Deep RL models to create more robust agents (Higgins et al., 2017b). Thus, regardless of the presence of disentangled generative factors, these regularization methods can induce models that are more robust to attack.

3 Robustness of VAEs

3.1 A Probabilistic Metric of Robustness

Deep learning models can be brittle. Some of the most sophisticated deep learning classifiers can be broken by simply adding small pertubations to their inputs (Szegedy et al., 2014; Shamir et al., 2019; Goodfellow et al., 2015; Papernot et al., 2016; Moosavi-Dezfooli et al., 2016). Here, perturbations that would not fool a human break neural network predictions. A model's weakness to such perturbations is called its *sensitivity*. For classifiers, we can straightforwardly define an associated *sensitivity margin*: it is the radius of the largest metric ball centered on an input $\mathbf{x}$ for which a classifier's original prediction holds for all possible perturbations within that ball.

Defining such a margin for VAEs is conceptually more difficult as, in general, the reconstructions are continuous rather than discrete. To put it another way, there is no step-change in VAE reconstructions that is akin to a change of a predicted class in classifiers; *any* perturbation in the input space will result in a change in the VAE output. To complicate matters further, a VAE's latent space is stochastic: the same input can result in different reconstructions.

As a first step to deriving robustness margins for VAEs, we now introduce a criterion for measuring robustness in probabilistic models: r-robustness. We start by presenting it in the general setting, before linking it to the specific case of VAEs.

Definition 3.1. *A model, f, operating on a point $\mathbf{x}$, that outputs a continuous random variable is r-robust for $r \in \mathbb{R}^+$, to a perturbation $\boldsymbol{\delta}$ and for an arbitrary norm $||\cdot||$ iff*

$$p(||f(\mathbf{x}+\boldsymbol{\delta}) - f(\mathbf{x})|| \leq r) > p(||f(\mathbf{x}+\boldsymbol{\delta}) - f(\mathbf{x})|| > r).$$

We will assume from now on that the norm is taken to be the 2-norm $||\cdot||_2$, such that r-robustness determines a bound for which changes in the output $f(\mathbf{x})$ induced by the perturbation $\boldsymbol{\delta}$ are more likely to fall within the hyper-sphere of radius r, than not. As r decreases, the criterion for model robustness becomes stricter. We note that r-robustness can be viewed as a probabilistic analog to the criterion for regression models presented by Nguyen & Raff (2019). We also note that r-robustness can be generalized to $p(||f(\mathbf{x}+\boldsymbol{\delta}) - f(\mathbf{x})||_2 \leq r) > m$, where $1-m$ is the allowable risk (with $m = 0.5$ in Definition 3.1).

Because this criterion is applicable to probabilistic models with continuous output spaces, it is directly relevant for ascertaining robustness in VAEs. By considering the smallest r for which the criterion holds, we can think of it as a metric that provides a probabilistic measure of the *extent* to which outputs are altered given a corrupted input: the smaller the value of r for which we can confirm r-robustness, the more robust the model.

3.2 A Robustness Margin for VAEs

We want to define a margin in a VAE's input space for which it is robust to perturbations of a given input. Perturbations that fall within this margin should not break our criterion for robustness. Formally, we want a margin in $\mathcal{X}$, $R^r_{\mathcal{X}}(\mathbf{x})$, for which any distorted input $\mathbf{x} + \boldsymbol{\delta}_x$, where $||\boldsymbol{\delta}_x||_2 < R^r_{\mathcal{X}}(\mathbf{x})$ is the perturbation, satisfies r-robustness when reconstructed.

However, to consider the robustness of VAEs, we must not only take into account the perturbation $\boldsymbol{\delta}_x$, but also the stochasticity of encoder. We can think of the decoder as taking in noisy inputs because of this stochasticity. Naturally, this noise can itself potentially cause issues in the robustness of VAE: if the level of noise is too high, we will not achieve reliable reconstructions even without perturbing the original inputs. As such, before even considering perturbations, we first need to adapt our r-robustness framework to deal with this stochasticity.

3.2.1 r-robustness for VAEs

Given an input $\mathbf{x}$, r-robustness dictates that we want to define some region in the reconstruction space,

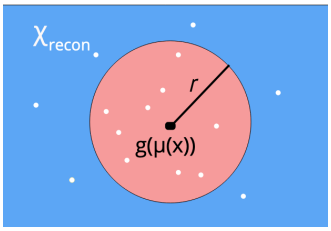


Figure 2: Illustration of r-robustness in a VAE. White dots represent possible reconstructions, with the diversity originating from the encoder stochasticity. For r-robustness to hold, the probability of our reconstruction falling within the red area—a hypersphere of radius r centered on $g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))$—needs to be greater than or equal to the probability of falling outside.

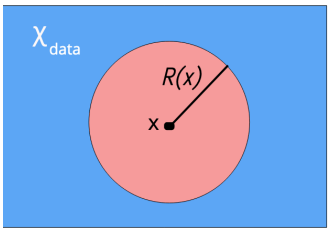


Figure 3: Illustration of the margin $R^r_{\mathcal{X}}(\mathbf{x})$, which is defined in the **input** space $\mathcal{X}$. Red represents represents the subspace where the model is r-robust, such that $p(||\Delta(\mathbf{x}, \boldsymbol{\delta}_x)||_2 \leq r) > p(||\Delta(\mathbf{x}, \boldsymbol{\delta}_x)||_2 > r)$ holds for all $\mathbf{x} + \boldsymbol{\delta}_x$ falling in this region, that is all $\boldsymbol{\delta}_x : \|\boldsymbol{\delta}_x\|_2 \leq R^r_{\mathcal{X}}(\mathbf{x})$.

$\mathcal{X}_{\text{recon}}$, within which most of the decoded samples from the latent embedding $\mathbf{z}$ will fall. We will assume that the encoder is a Gaussian as this is standard practice. Denoting $g_\theta(\mathbf{z})$ as the deterministic mapping induced by the VAE's decoder network and $\boldsymbol{\mu}_\phi(\mathbf{x})$ as the mean embedding of the encoder, we can define $g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))$ to be the "maximum likelihood" reconstruction, noting this is a deterministic function. Our aim is now to find a hyper-sphere of radius r centered on $g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))$ within which most of the possible VAE outputs for a given point $\mathbf{x}$ lie. Larger r are indicative of a greater variance in the encoding process, and as such are likely to be associated with poorer quality reconstructions.

Denoting $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as the reparameterized stochasticity of the encoder, we define the distance from the maximum likelihood reconstruction, induced by this sampling as

$$\Delta(\mathbf{x}) = g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\eta} \circ \boldsymbol{\sigma}_\phi(\mathbf{x})) - g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x})) \quad (1)$$

Using this, we see that a VAE is r-robust to the stochasticity of the encoder iff (see also Fig 2)

$$p(||\Delta(\mathbf{x})||_2 \leq r) > p(||\Delta(\mathbf{x})||_2 > r). \quad (2)$$

Informally, we want it to be more probable for reconstructions to fall within this radius r than not.

3.2.2 Robustness to distortions in data-space

Given that we have established conditions for r in Eq (2) that take into account latent space sampling, we can now return to our original objective, which was to determine a margin in the data-space $\mathcal{X}$ for which a VAE is robust to perturbations on its input. Recall that this implicitly means that we want to define a bound for robustness given two sources of perturbations: the stochasticity of the encoder, and a hypothetical input perturbation $\boldsymbol{\delta}_x$.

For simplicity of analysis, we consider the case where the perturbation is applied only to the encoder mean input and not the encoder variance input, noting that the latter is typically stable across inputs and so is less of a concern. In Fig 4 we demonstrate that adversarial attacks on VAE encoders are dominated by the perturbation to the embedding mean that is induced, thereby justifying this assumption. We note also that one can usually also simply fix the encoder variance to a constant for all datapoints without incurring substantial performance drops (Ghosh et al., 2020), thereby providing a means to ensure this assumption holds exactly if needed.

We define the distance from the maximum likelihood reconstruction, $g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))$, induced by the stochasticity of the encoder *and* an input perturbation $\boldsymbol{\delta}_x$ as

$$\Delta(\mathbf{x}, \boldsymbol{\delta}_x) = g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x} + \boldsymbol{\delta}_x) + \boldsymbol{\eta}\sigma_\phi(\mathbf{x})) - g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x})).$$

We can now define the condition for which r-robustness is satisfied on the VAE output given the two sources of perturbation as

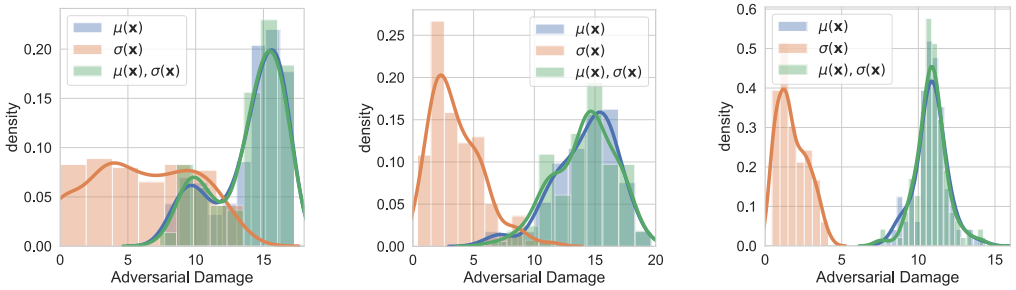
$$||\boldsymbol{\delta}_x||_2 < R^r_{\mathcal{X}}(\mathbf{x}) \;\Leftrightarrow\; p(||\Delta(\mathbf{x}, \boldsymbol{\delta}_x)||_2 \leq r) > 0.5 \quad (3)$$

Thus, $R^r_{\mathcal{X}}(\mathbf{x})$ is the margin of robustness of the VAE such that $\forall \boldsymbol{\delta}_x : ||\boldsymbol{\delta}_x||_2 < R^r_{\mathcal{X}}(\mathbf{x})$, $\mathbf{x} + \boldsymbol{\delta}_x$ is more likely than not to be reconstructed within a radius r of the maximum likelihood reconstruction $g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))$. A high level illustration of this is given in Fig 3, and Fig 5 shows a simple empirical demonstration of how $R^r_{\mathcal{X}}(\mathbf{x})$ relates to the probability of producing a good reconstruction under random input perturbations.

We note that, analogously to the concept of an adversarial risk (Uesato et al., 2018), $R^r_{\mathcal{X}}(\mathbf{x})$ can further be converted to a metric for the *overall* robustness of a VAE, by taking its expectation over the data generating distribution, namely $R^r_{\mathcal{X}} = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}[R^r_{\mathcal{X}}(\mathbf{x})]$.

3.3 Characterizing the Margin

Given this definition, we now wish to try and characterize $R^r_{\mathcal{X}}(\mathbf{x})$. In particular, we would like to understand what characteristics of the VAE are likely to



(a) MNIST (b) fMNIST (c) CIFAR10

Figure 4: *Maximum damage* adversarial attacks (see Eq (5)) on multiple VAEs trained on MNIST (a), fashion-MNIST (b), and CIFAR10 (c). We attack 25 datapoints for each VAE and propagate the attacks to the encoder mean ($\mu(\mathbf{x})$), the encoder standard deviation ($\sigma(\mathbf{x})$), or both ($\mu(\mathbf{x}), \sigma(\mathbf{x})$). Attack norms are capped to 10. Shown are distribution plots of the adversarial damage, i.e. the L_2 distance between the reconstruction resulting from the attack and the maximum likelihood reconstruction $g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))$. Clearly attacks on $\mu(\mathbf{x})$ are more harmful than on $\sigma(\mathbf{x})$, and most of the damage from attacks on both $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ stems from the attack on $\mu(\mathbf{x})$.

make it relatively larger or smaller. Ideally, we also want to establish scenarios where we might be able to provide guarantees of a minimum size for $R^r_{\mathcal{X}}(\mathbf{x})$, such that we might be able to make inroads into how one might apriori construct a certifiably-robust VAE.

A perturbation in $\mathcal{X}$, $\boldsymbol{\delta}_x$, induces a perturbation in $\mathcal{Z}$, $\boldsymbol{\delta}_z$. To determine the margins for robustness in $\mathcal{X}$, we first apply the Neyman-Pearson lemma (Neyman & Pearson, 1933; Cohen et al., 2019), assuming a "worst-case" decoder. This decoder has subspaces in $\mathcal{Z}$, where it is is either robust or non-robust, that are divided by a boundary that is normal to both the induced perturbation $\boldsymbol{\delta}_z$ and to the dimension of minimal variance in $\mathcal{Z}$, $\min_i \boldsymbol{\sigma}_\phi(\mathbf{x})_i$. We then determine the minimum perturbation norm in $\mathcal{X}$ which induces a perturbation in $\mathcal{Z}$ that crosses this boundary.

Theorem 1. *Consider a VAE with a diagonal-variance Gaussian encoder, an input* $\mathbf{x}$*, and an output margin* $r \in \mathbb{R}$ *such that the VAE is* r*-robust to the stochasticity of the encoder when the* $\mathbf{x}$ *is unperturbed as per* (2)*. Assuming standard regularity assumptions (discussed in the proof) hold for* $\boldsymbol{\mu}_\phi(\mathbf{x})$*, then*

$$R^r_{\mathcal{X}}(\mathbf{x}) \geq \frac{(\min_i \boldsymbol{\sigma}_\phi(\mathbf{x})_i)\Phi^{-1}(p(||\Delta(\mathbf{x})||_2 \leq r))}{||\mathbf{J}^{\mu}_{\phi}(\mathbf{x})||_F} + \mathcal{O}(\boldsymbol{\varepsilon}) \tag{4}$$

where $\mathcal{O}(\boldsymbol{\varepsilon})$ *represents higher order dominated terms that disappear in the limit of small perturbations,* Φ^{-1} *is the probit function,* $\mathbf{J}^{\mu}_{\phi}(\mathbf{x})_{i,j} = \partial \boldsymbol{\mu}_\phi(\mathbf{x})_i / \partial \mathbf{x}_j$ *is the Jacobian of* $\boldsymbol{\mu}_\phi(\mathbf{x})$*, and* $||\cdot||_F$ *is the Frobenius norm.*

The proof is provided in Appendix B. This bound is based on a first order approximation of $\boldsymbol{\mu}_\phi(\mathbf{x} + \boldsymbol{\delta}_x)$ around the original input $\mathbf{x}$; the impact of $\mathcal{O}(\boldsymbol{\varepsilon})$ thus depends on how well this approximation holds. As such, the result is particularly applicable to networks

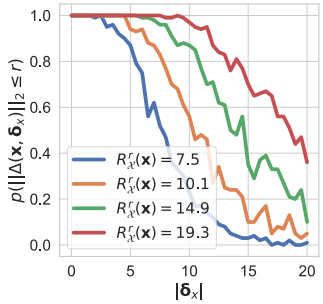
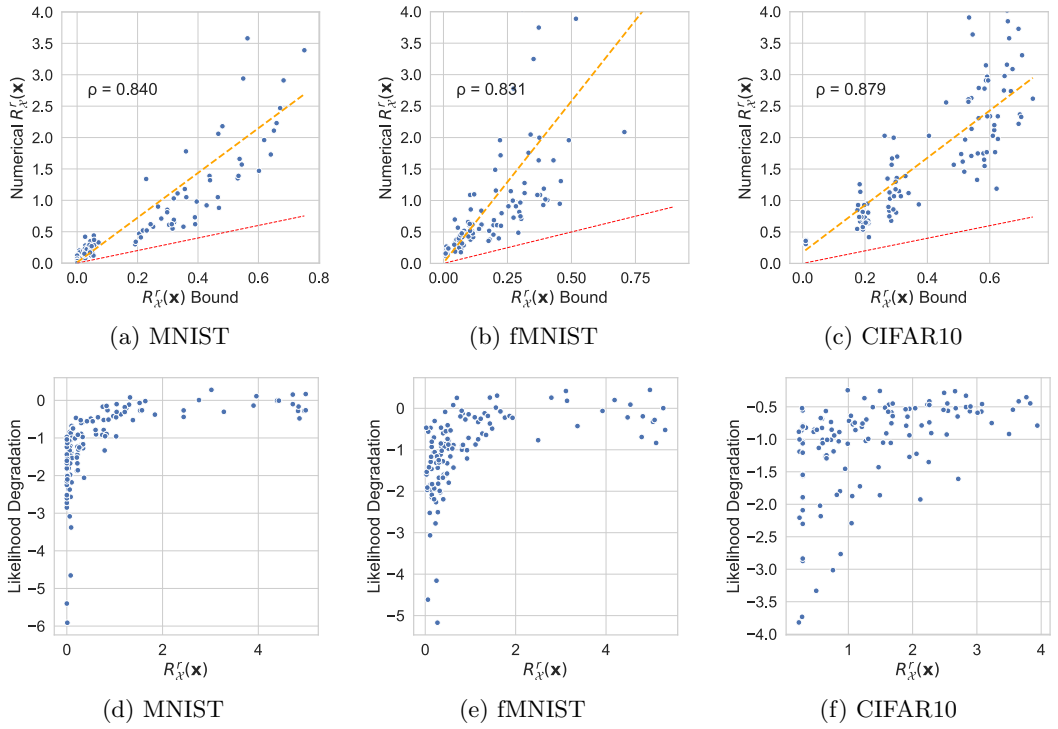


Figure 5: $R^r_{\mathcal{X}}(\mathbf{x})$ for four VAEs of varying robustness trained on MNIST. We fix the input $\mathbf{x}$ and perturbation direction $\boldsymbol{\delta}_x / \|\boldsymbol{\delta}_x\|_2$, but vary the perturbation size $\|\boldsymbol{\delta}_x\|_2$. We assess the proportion of samples which fall within $r = 4$ of the maximum likelihood reconstruction.

with piecewise linear activation functions such as the ReLU, which are locally linear and are among the most widely used activation functions. For these activation functions this bound is locally exact: $\mathcal{O}(\boldsymbol{\varepsilon})$ is exactly zero if the size of the bound is smaller than what is required to go outside the locally linear region.

This gives us margins for which VAEs are certifiably robust, up to first order expansions, to adversarial perturbations on their inputs; they have similar forms to the sensitivity margins for classifiers defined by Sokolić et al. (2017); Jakubovitz & Giryes (2018) in that both scale *inversely* with the network Jacobian. More generally, these results provide insights into the features which lead to robust VAEs. As shown in Figure 6, the bound seems to be relatively tight in practice, even when attacking both the encoder mean *and* variance. It also has a near-linear relationship with the empirically estimated robustness, such that it forms a powerful and convenient robustness metric in its own right.

Examining the bound, we see that for a given r, $R^r_{\mathcal{X}}(\mathbf{x})$



(a) MNIST (b) fMNIST (c) CIFAR10

(d) MNIST (e) fMNIST (f) CIFAR10

Figure 6: (a-c) show the empirically estimated $R^r_{\mathcal{X}}(\mathbf{x})$ against the bound for $R^r_{\mathcal{X}}(\mathbf{x})$ defined in Theorem 1, ignoring higher order terms. Each dot represents a network–input pair, with 5 separately trained networks and 25 distinct inputs considered. We show the line of best fit (in orange), the correlation coefficient ρ, and the line $y = x$ (in red) representing the theoretical bound itself. (d-f) show the relative log likelihood degradation resulting from a 'maximum-damage' adversarial attack against the numerically estimated $R^r_{\mathcal{X}}(\mathbf{x})$ for these same VAEs and inputs (see Section 4.1).

increases as the stochasticity of the encoder, i.e $\boldsymbol{\sigma}_\phi(\mathbf{x})$, increases, provided that this does not overly affect $\Phi^{-1}(p(||\Delta(\mathbf{x})||_2 \leq r))$ (see below). As $\boldsymbol{\sigma}_\phi(\mathbf{x})$ tends to 0 we recover the deterministic setting, which confers no additional protection to attack and as $\boldsymbol{\sigma}_\phi(\mathbf{x})$ increases we obtain increased protection. However, $\boldsymbol{\sigma}_\phi(\mathbf{x})$ can also have a knock–on effect on $\Phi^{-1}(p(||\Delta(\mathbf{x})||_2 \leq r))$. When $\boldsymbol{\sigma}_\phi(\mathbf{x})$ is small, this knock–on effect will typically be small relative to the direct effect of changing $\boldsymbol{\sigma}_\phi(\mathbf{x})$, but as it becomes large there is always a point where this knock–on effect will take over. Namely, our reconstructions will become increasingly poor and $\Phi^{-1}(p(||\Delta(\mathbf{x})||_2 \leq r))$ will eventually become negative, such that r-robustness does not hold even without perturbation. We can quantify this by noting that there is always a minimum r for r-robustness to be satisfied in (2). We derive a bound characterizing the minimum r for which we can confirm robustness in Appendix A.

4 Empirical Investigations

We now consider a series of empirical investigations to back up our frameworks and theoretical results. We start by assessing whether the concept of r-robustness corresponds to more commonly used measures of model robustness. Here we estimate $R^r_{\mathcal{X}}(\mathbf{x})$ numerically as in Appendix D and as demonstrated in Fig 5. Using these empirical estimations, we establish connections between r-robustness and other performance metrics during adversarial attack, confirming that larger $R^r_{\mathcal{X}}(\mathbf{x})$ correspond to model–input pairs that are more robust to adversarial attacks.

4.1 r-robustness and Adversarial Settings

We begin by evaluating our metrics in adversarial settings. We want to find the most damaging perturbations $\boldsymbol{\delta}_x$ that challenge the robustness metrics we have derived. We consider an adversary trying to distort the input data to maximally disrupt a VAE's reconstruction. Our adversary maximizes, wrt $\boldsymbol{\delta}_x$, the distance between the VAE reconstruction and the original datapoint $\mathbf{x}$, a novel adversarial attack we call *maximum damage*. We attack the encoder mean *and* variance:

$$\boldsymbol{\delta}_x^* = \arg\max_{\boldsymbol{\delta}_x} \big(||g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}+\boldsymbol{\delta}_x) + \boldsymbol{\eta}\sigma_\phi(\mathbf{x}+\boldsymbol{\delta}_x)) - g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))||_2\big). \quad (5)$$

We evaluate the success of an attack as follows. Given an embedding $\mathbf{z}^*$ formed from the mean encoding of $\mathbf{x} + \boldsymbol{\delta}_x$, we measure the likelihood of the origi-

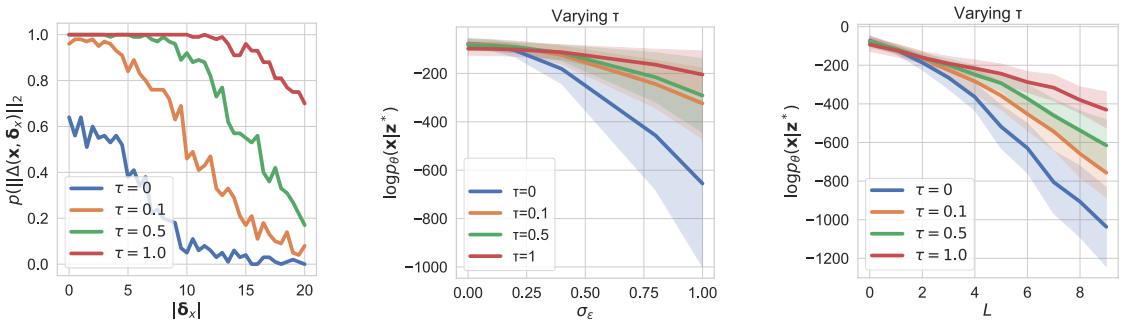


Figure 7: Ablation study on the bounds defined by Theorem 1. We train models on MNIST with $\sigma_\phi(\mathbf{x})$ offset by a constant $\tau \in [0, 0.1, 0.5, 1]$. [Left] probability that reconstructions in $\mathcal{X}_{\text{recon}}$ fall within a radius $r = 4$ centered on the 'maximum likelihood' reconstruction, $p(||\Delta(\mathbf{x}, \boldsymbol{\delta}_x)||_2 \leq r)$, as a function of $|\boldsymbol{\delta}_x|$, the magnitude of perturbations. $R^r_{\mathcal{X}}(\mathbf{x})$ is the radius $|\boldsymbol{\delta}_x|$ for which $p(||\Delta(\mathbf{x}, \boldsymbol{\delta}_x)||_2 \leq 4) > 0.5$ and clearly increases with τ. [Center] we add noise $\sim \mathcal{N}(0, \sigma^2_\epsilon)$ to a point $\mathbf{x}$ forming a noisy $\mathbf{x}^*$ and $\mathbf{z}^*$, and measure the likelihood of the original point $\mathbf{x}$ under this noisy embedding. [Right] we show the same plot where the perturbations are *maximum damage* attacks, Eq (5), where L is the maximum allowed magnitude of the attack distortion. Large τ VAEs have high likelihoods for the original point $\mathbf{x}$ as L and σ^2_ϵ increase: they are robust to attack and effective denoising models. Confidence intervals are the standard deviations of values over the entire MNIST dataset.

nal point $\mathbf{x}$ and quantify the degradation in model performance as the relative log likelihood degradation $(|(\log p(\mathbf{x}|\mathbf{z}^*) - \log p(\mathbf{x}|\mathbf{z}))|/\log p(\mathbf{x}|\mathbf{z}))$, where $\mathbf{z}$ is the embedding of $\mathbf{x}$. Fig 6 (d-f) shows that as $R^r_{\mathcal{X}}(\mathbf{x})$ increases this degradation lessens, indicating less damaging attacks. As such, larger margins for r-robustness correspond to models that are more robust to attack.

4.2 Evaluating the derived bounds

Using Theorem 1 we can gain insights into which characteristics of a VAE contribute to robustness. The encoder variance plays a prominent role and is a parameter that is easy to control. The encoder Jacobian is also present, but we found that controlling such values directly can be difficult. Penalizing the norm of this Jacobian in the VAE training objective degrades VAE generative performance, making it difficult to compare models. As such we restrict our experiments to varying the encoder variance. We do so by training models that have $\boldsymbol{\sigma}_\phi(\mathbf{x})$ offset by a constant τ, such that we artificially increase the encoder variance minimum. In Fig 7 we show that as τ increases, the numerically estimated $R^r_{\mathcal{X}}(\mathbf{x})$ also increases, supporting our claim that models with larger encoder variances have larger margins of robustness. This figure also shows that likelihood of reconstructing the original input x increases as τ increases in both an adversarial attack setting and a noisy perturbation setting. We thus see that larger τ also provides more effective denoising properties.

5 Robustness of Disentangled VAEs

We now apply our analysis to disentangling methods, which have empirically been shown to be more robust to adversarial attacks and noisy data (Willetts et al., 2019). First, we demonstrate this visually in Fig 8 where we see that β-VAEs are more resilient to attack as β increases, and thus implicitly latent space overlap increases (Mathieu et al., 2019). Second, we provide analysis to show that disentangling methods induce models with smaller encoder Jacobian norms *and* larger posterior variances, implying that they have larger margins $R^r_{\mathcal{X}}(\mathbf{x})$ by Theorem 1, a result we confirm empirically.

Disentangling increases encoder variance Empirically, increasing $\beta > 1$ in a β-VAE increases the variance of the trained encoder, saturating at the variance of the prior for large β (Mathieu et al., 2019; Locatello et al., 2019). We can shed light on this behavior by finding the optimum forms of the posterior distribution under these objective functions using calculus of variations. We find that the optimal posterior has the form of a tempered or fractional posterior (Holmes & Walker, 2017; Wenzel et al., 2020; Miller & Dunson, 2019), with an exponent $1/\beta$ on the likelihood:

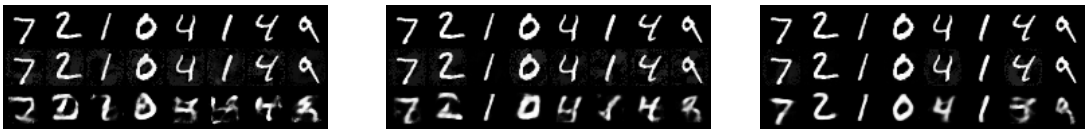
Theorem 2. *For a β-VAE, the optimum posterior is:*

$$q_\phi(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z})^{1/\beta}$$

The proof is given in Appendix C.

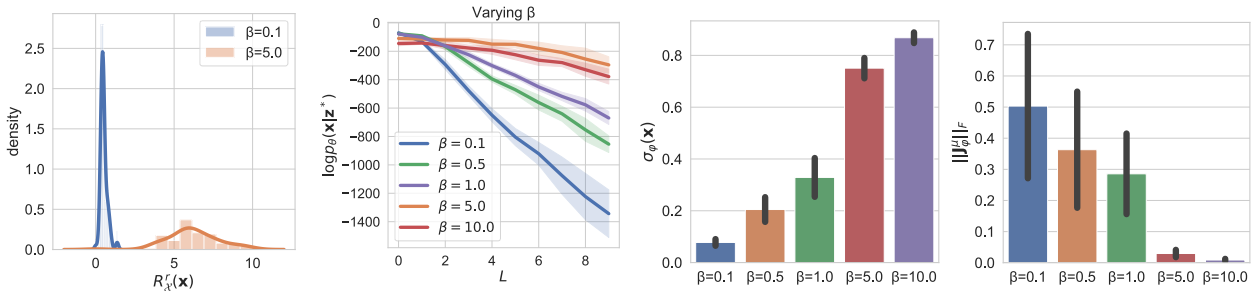
This result gives the optimal posterior as a function of β. It also tells us the β-VAE's optimal posterior in the limit of large β is the prior, as we would expect. Because the prior variance is naturally larger than that of the encoder, the encoder variance increases with β.

Disentangling penalizes Jacobian norm Assuming a Gaussian $p_\theta(\mathbf{x}|\mathbf{z})$, an encoder covariance optimal to first order, and activation functions that are



(a) $\beta = 0.1$ (b) $\beta = 1$ (c) $\beta = 10$

Figure 8: For β-VAEs trained with $\beta \in \{0.1, 1, 10\}$ we show in consequentive rows first the original data point, a perturbed version made by maximum damage adversarial attacks, and then the reconstruction given by the model. As β increases the models become more robust to attack.



(a) Sampled $R^r_{\mathcal{X}}(\mathbf{x})$ (b) Adv Attack (c) Encoder Variance (d) Encoder Jacobian

Figure 9: (a) distribution of the numerically estimated $R^r_{\mathcal{X}}(\mathbf{x})$ ($m = 0.5$) across the MNIST dataset. We see that $R^r_{\mathcal{X}}(\mathbf{x})$ increases dataset-wide for larger β. (b) likelihood of the original input given a maximum damage adversarial attack as in Eq (5). L is the maximum allowed norm of the attack. Large β models retain high likelihoods even for large L, meaning they are robust to attack. (c) and (d) show that the encoder variance increases and the encoder Jacobian norm ($||\mathbf{J}^\mu_\phi(\mathbf{x})||_F$) decreases as β increases, supporting our analysis that the changes in these values underpin the robustness observed. Confidence intervals for all plots are the standard deviation of values over the entire MNIST dataset. See Appendix E for similar experiments on other datasets.

piecewise-linear, the β–VAE objective can be approximated as (Kumar & Poole, 2020)

$$\min_{\phi,\theta} \frac{1}{2}||\mathbf{x} - g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))||^2 + \frac{\beta}{2}||\mathbf{J}^\mu_\phi(\mathbf{x})\mathbf{x}||^2_F + \frac{\beta}{2}\log|\mathbf{I} + \frac{1}{\beta}\mathbf{J}_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))\mathbf{J}^T_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))|, \quad (6)$$

As β increases $||\mathbf{J}^\mu_\phi(\mathbf{x})||^2_F$ is more penalised and we expect to learn encoders with smaller Jacobians.

Taking these results together, we expect two things to occur as β increases: the encoder variance should increase by Theorem 2 and the norm of the encoder Jacobian should decrease. We confirm this empirically in Fig 9(c,d). By Theorem 1 these two effects of increasing β should increase $R^r_{\mathcal{X}}(\mathbf{x})$ in tandem. In Fig 9(a) we confirm that the numerical estimate for $R^r_{\mathcal{X}}(\mathbf{x})$ increases *dataset-wide* for large β, that is we get a larger value for $R^r_{\mathcal{X}}$, or metric for the overall robustness of the VAE. In Appendix E, we also show that the distribution of the *bound* for $R^r_{\mathcal{X}}(\mathbf{x})$ from Theorem 1 increases dataset-wide with increasing β. In both cases it is noticeable that $R^r_{\mathcal{X}}(\mathbf{x})$ is quite a well-behaved distribution, with reasonably low variance and skew. This suggests that $R^r_{\mathcal{X}}$ can be reliably estimated in practice.

In Fig 9b we further show that these larger $R^r_{\mathcal{X}}$ values translate into larger likelihoods of the original input under adversarial attack, while in Appendix E we confirm that model sensitivity to noise is improved for larger β. We note, however, that having a β that is too large will completely undermine reconstructions (Higgins et al., 2017a; Chen et al., 2018) and lead to VAEs that are never robust because we cannot confirm r-robustness, even without input perturbations (Eq (2)). See Appendix D.1.1 for results on this.

6 Conclusion

We have defined a novel robustness metric tailored to probabilistic generative models, r-robustness, which can be used to assess the robustness of VAEs to adversarial attack. We defined a margin on a VAE's input space within which it is r-robust to perturbations and show that small norms of the encoder Jacobian and larger encoder variances are core contributors to robustness. Further, we offered theoretical and empirical analysis based on this margin, demonstrating that existing disentangling methods increase robustness by altering the optimal encoder variance and the norm of the encoder Jacobian.

Acknowledgments

This research was directly funded by the Alan Turing Institute under Engineering and Physical Sciences Research Council (EPSRC) grant EP/N510129/1. AC was supported by an EPSRC Studentship. MW was supported by EPSRC grant EP/G03706X/1. SR gratefully acknowledges support from the UK Royal Academy of Engineering and the Oxford-Man Institute. CH was supported by the Medical Research Council, the Engineering and Physical Sciences Research Council, Health Data Research UK, and the Li Ka Shing Foundation

We thank Tomas Lazauskas, Jim Madge and Oscar Giles from the Alan Turing Institute's Research Engineering team for their help and support.

References

Alemi, A. A., Fischer, I., Dillon, J. V., & Murphy, K. (2017). Deep variational information bottleneck. In *ICLR*.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1798–1828.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., & Lerchner, A. (2018). Understanding disentangling in β-VAE. In *NeurIPS*.

Chen, T. Q., Li, X., Grosse, R., & Duvenaud, D. (2018). Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*.

Cohen, J., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. In *ICML*.

Ghosh, P., Losalka, A., & Black, M. J. (2018). Resisting Adversarial Attacks using Gaussian Mixture Variational Autoencoders. (2014).

Ghosh, P., Sajjadi, M. S. M., Vergari, A., Black, M., & Schölkopf, B. (2020). From Variational to Deterministic Autoencoders. In *ICLR*.

Gondim-Ribeiro, G., Tabacof, P., & Valle, E. (2018a). Adversarial Attacks on Variational Autoencoders.

Gondim-Ribeiro, G., Tabacof, P., & Valle, E. (2018b). Adversarial Attacks on Variational Autoencoders.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *ICML*.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017a). β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework.

Higgins, I., Pal, A., Rusu, A., Matthey, L., Burgess, C., Pritzel, A., Botvinick, M., Blundell, C., & Lerchner, L. (2017b). DARLA: Improving zero-shot transfer in reinforcement learning. In *ICML*.

Holmes, C. C. & Walker, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2), 497–503.

Jakubovitz, D. & Giryes, R. (2018). Improving DNN robustness to adversarial attacks using jacobian regularization. *Lecture Notes in Computer Science*, (pp. 525–541).

Kingma, D. P. & Welling, M. (2014). Auto-encoding variational bayes. In *ICLR*.

Kos, J., Fischer, I., & Song, D. (2018a). Adversarial examples for generative models. *IEEE Symposium on Security and Privacy Workshops*, (pp. 36–42).

Kos, J., Fischer, I., & Song, D. (2018b). Adversarial Examples for Generative Models. In *IEEE Security and Privacy Workshops* (pp. 36–42).

Kumar, A. & Poole, B. (2020). On Implicit Regularization in β-VAEs.

Locatello, F., Bauer, S., Lucie, M., Rätsch, G., Gelly, S., Schölkopf, B., & Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*.

Mathieu, E., Rainforth, T., Siddharth, N., & Teh, Y. W. (2019). Disentangling disentanglement in variational autoencoders. In *ICML*.

Miller, J. W. & Dunson, D. B. (2019). Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527), 1113–1125.

Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *IEEE CVPR*.

Neyman, J. & Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. 231, 73–108.

Nguyen, A. T. & Raff, E. (2019). Adversarial Attacks, Regression, and Numerical Stability Regularization. In *The AAAI Workshop on Engineering Dependable and Secure Machine Learning Systems*.

Papernot, N., Mcdaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. *IEEE European Symposium on Security and Privacy*, (pp. 372–387).

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014).

Stochastic backpropagation and approximate inference in deep generative models. In *ICML*.

Rolinek, M., Zietlow, D., & Martius, G. (2019). Variational autoencoders pursue PCA directions (by accident. In *IEEE CVPR*.

Schott, L., Rauber, J., Bethge, M., & Brendel, W. (2019). Towards the first adversarially robust neural network model on MNIST. In *ICLR*.

Shamir, A., Safran, I., Ronen, E., & Dunkelman, O. (2019). A Simple Explanation for the Existence of Adversarial Examples with Small Hamming Distance.

Sokolić, J., Giryes, R., Sapiro, G., & Rodrigues, M. R. (2017). Robust Large Margin Deep Neural Networks. *IEEE Transactions on Signal Processing*, 65(16), 4265–4280.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In *ICLR*.

Tabacof, P., Tavares, J., & Valle, E. (2016). Adversarial Images for Variational Autoencoders.

Uesato, J., O'Donoghue, B., Oord, A. v. d., & Kohli, P. (2018). Adversarial risk and the dangers of evaluating against weak attacks. *arXiv*.

Wenzel, F., Roth, K., Veeling, B. S., Świątkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., & Nowozin, S. (2020). How Good is the Bayes Posterior in Deep Neural Networks Really? *arXiv*.

Willetts, M., Camuto, A., Rainforth, T., Roberts, S., & Holmes, C. (2019). Improving VAEs' Robustness to Adversarial Attack.