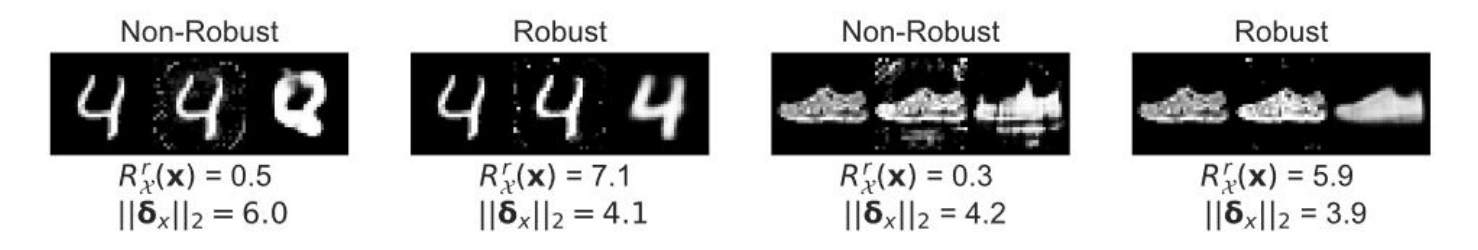
Robust VAEs

-Mayank Jain & Shikhar Agrawal
Prof. Suyash Awate

A preview of what’s coming

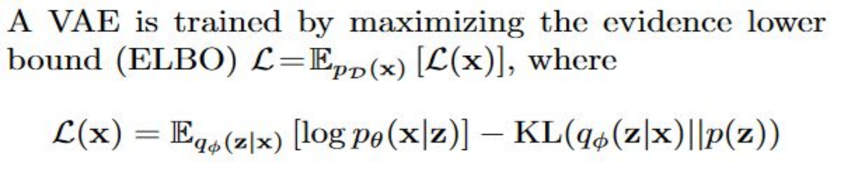
- We make inroads into understanding the robustness of Variational Autoencoders (VAEs) to adversarial attacks and other input perturbations.
- A hypothetical adversary can attack a VAE by applying small input perturbations to invoke meaningful changes in the encoding.
- There are currently no theoretical foundations for this robustness or even any frameworks or formalizations for exactly what it means for a VAE to be “robust.”
- As a first step to addressing these questions, we develop the first metric with which to evaluate the robustness of VAEs: r-robustness.
- 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 x within this margin.
- This in turn, allows us to provide a notion of a certifiably–robust reconstruction

Empirical view of ‘margin’


Non-Robust
$R^r_{\mathcal{X}}(\mathbf{x}) = 0.5$
$||\boldsymbol{\delta}_x||_2 = 6.0$
Robust
$R^r_{\mathcal{X}}(\mathbf{x}) = 7.1$
$||\boldsymbol{\delta}_x||_2 = 4.1$
Non-Robust
$R^r_{\mathcal{X}}(\mathbf{x}) = 0.3$
$||\boldsymbol{\delta}_x||_2 = 4.2$
Robust
$R^r_{\mathcal{X}}(\mathbf{x}) = 5.9$
$||\boldsymbol{\delta}_x||_2 = 3.9$

- Here we demonstrate that large $R^r_{\mathcal{X}}(\mathbf{x})$ are associated with model input pairs that are robust to adversarially generated input perturbations.
- We further derive a theoretical bound $R^r_{\mathcal{X}}(\mathbf{x})$ for as a function of the encoder variance and Jacobian. This provides insights into the characteristics of VAEs that contribute to robustness.
- We further demonstrate how these beneficial characteristics can be induced using methods introduced to learn disentangled representations.

VAE background

- A VAE simultaneously learns both a forward generative model, $p_\theta(\mathbf{x}|\mathbf{z})$, and an amortised $q_\phi(\mathbf{z}|\mathbf{x})$ approximate posterior distribution, q
- These are referred to as the decoder and encoder respectively, and a VAE can be thought of as a deep stochastic autoencoder.
- 

  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

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

The optimization is carried out using stochastic gradient descent with Monte Carlo samples.

- 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.
- 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.

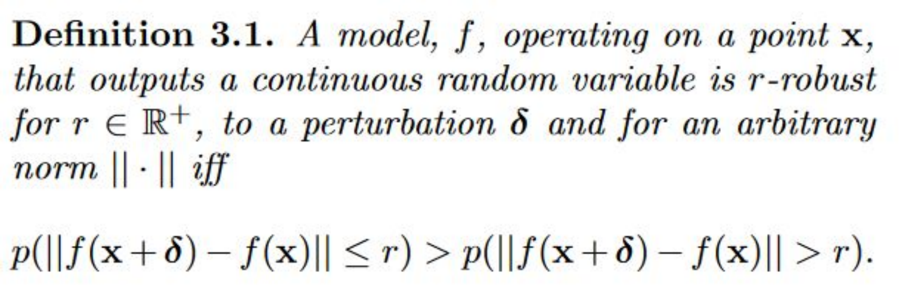
Disentangled VAEs

- Learning disentangled representations 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.
- One such method is the β-VAE which upweights the KL in the ELBO with a penalization factor:

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

Models trained under disentangling objectives have other beneficial properties, and have been shown to induce models that are more robust to adversarial attacks.

A probabilistic measure: r-robustness

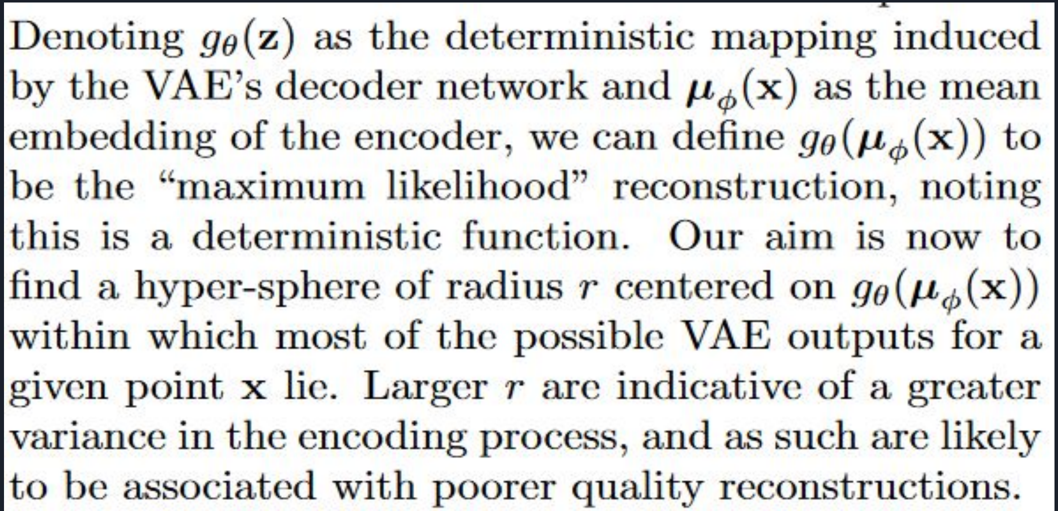
- Some of the most sophisticated deep learning classifiers can be broken by simply adding small perturbations to their inputs. A model's weakness to such perturbations is called its sensitivity.
- Defining such a margin for VAEs is conceptually more difficult as, in general, the reconstructions are continuous rather than discrete. To complicate matters further, a VAE's latent space is stochastic: the same input can result in different reconstructions.
- 

  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*

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

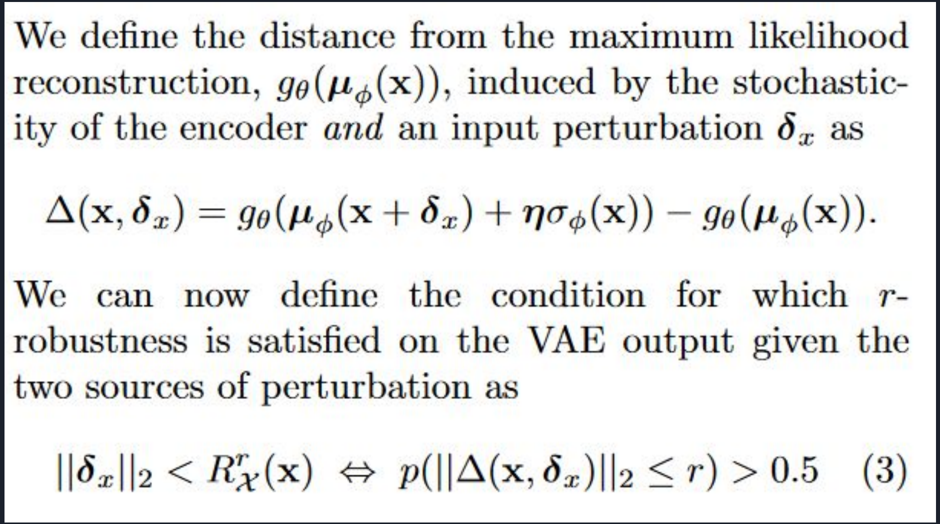
  We will assume from now on that the norm is taken to be the 2-norm || · ||

Redefining robustness for VAEs

- Given an input x, r-robustness dictates that we want to define some region in the reconstruction space within which most of the decoded samples from the latent embedding z will fall.
- 

  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.

Redefining robustness for VAEs (II)

- 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.
- 

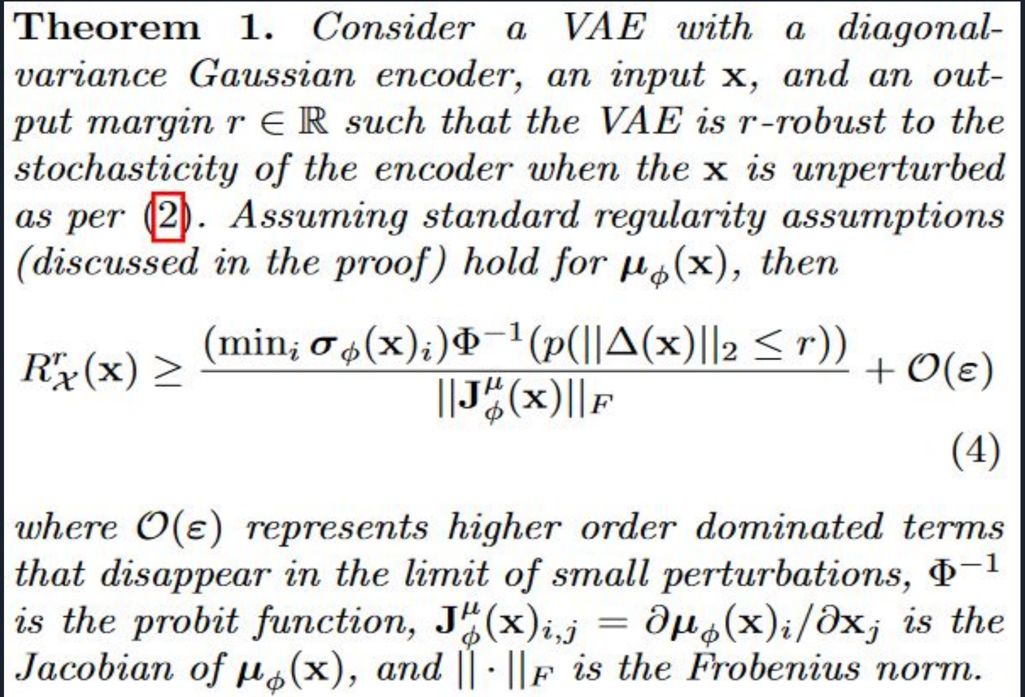
We define the distance from the maximum likelihood reconstruction, $g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))$, induced by the stochasticity of the encoder *and* an input perturbation $\boldsymbol{\delta}_x$ as

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

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

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

Characterizing the margin

- We would like to understand what characteristics of the VAE are likely to 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 the margin.
- 

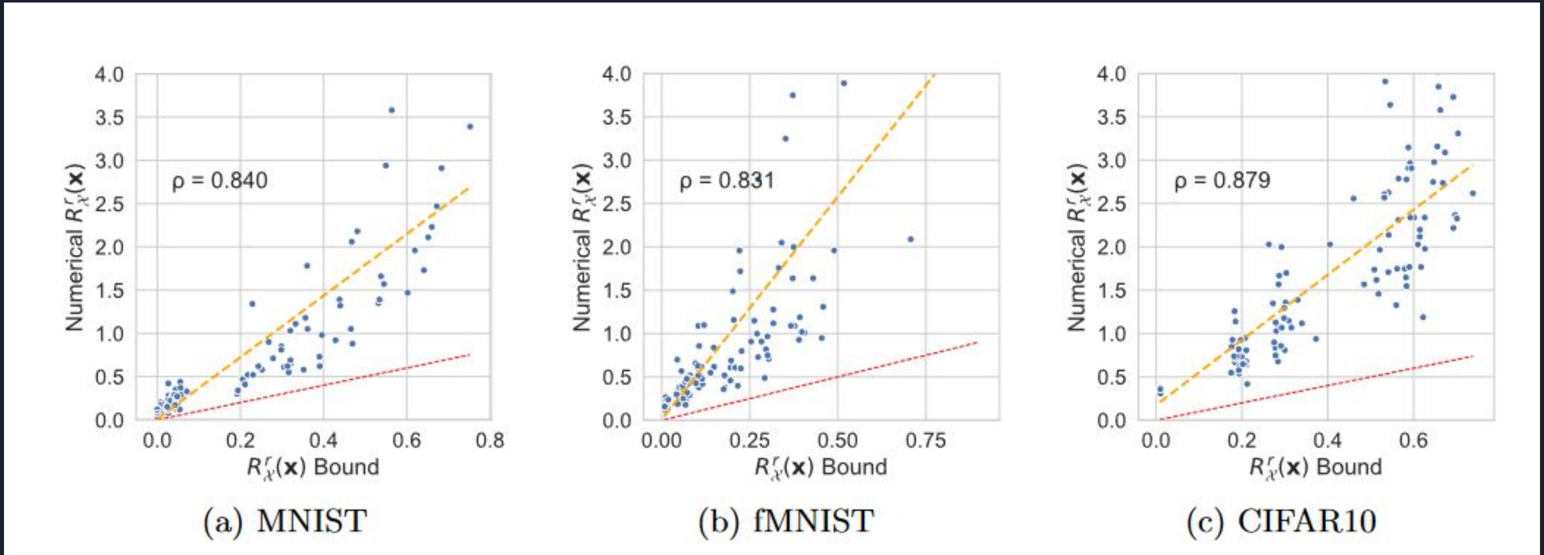
Theorem 1. *Consider a VAE with a diagonal-variance Gaussian encoder, an input* $\mathbf{x}$*, and an output margin* $r \in \mathbb{R}$ *such that the VAE is* r*-robust to the stochasticity of the encoder when the* $\mathbf{x}$ *is unperturbed as per (2). Assuming standard regularity assumptions (discussed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

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

where $\mathcal{O}(\varepsilon)$ *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.*

Accuracy of the Theorem

- This bound is based on a first order approximation of μ around the original input x. As such, the result is particularly applicable to networks 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 as long as the margin is within a linear part. This gives us margins for which VAEs are certifiably robust, up to first order expansions, to adversarial perturbations on their inputs.

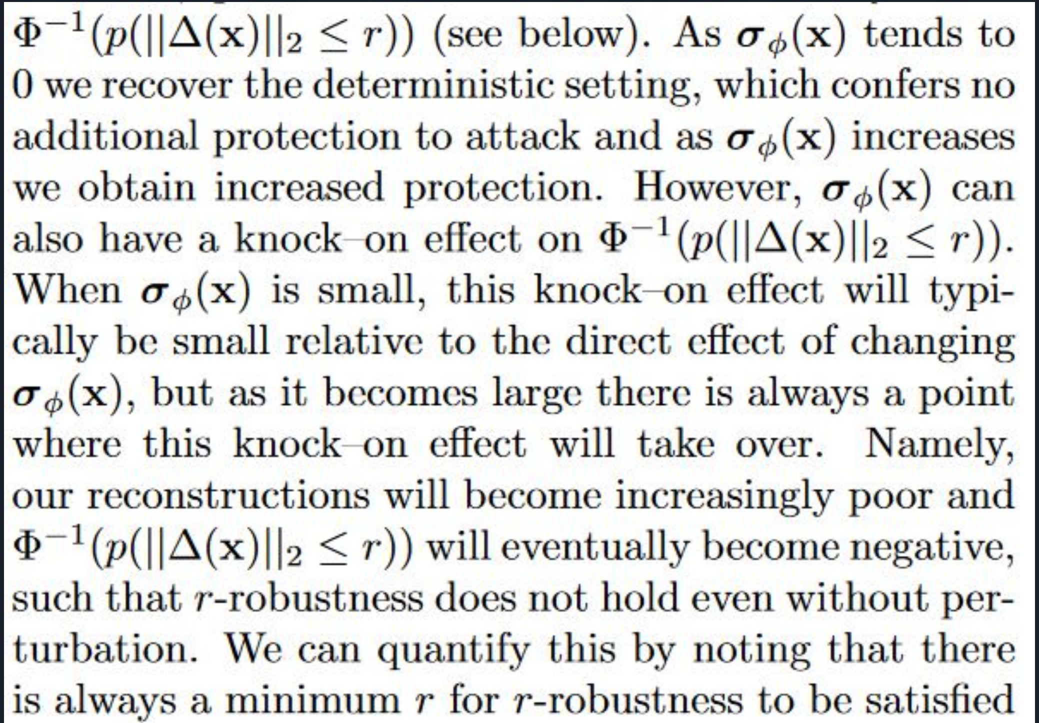


(a) MNIST

(b) fMNIST

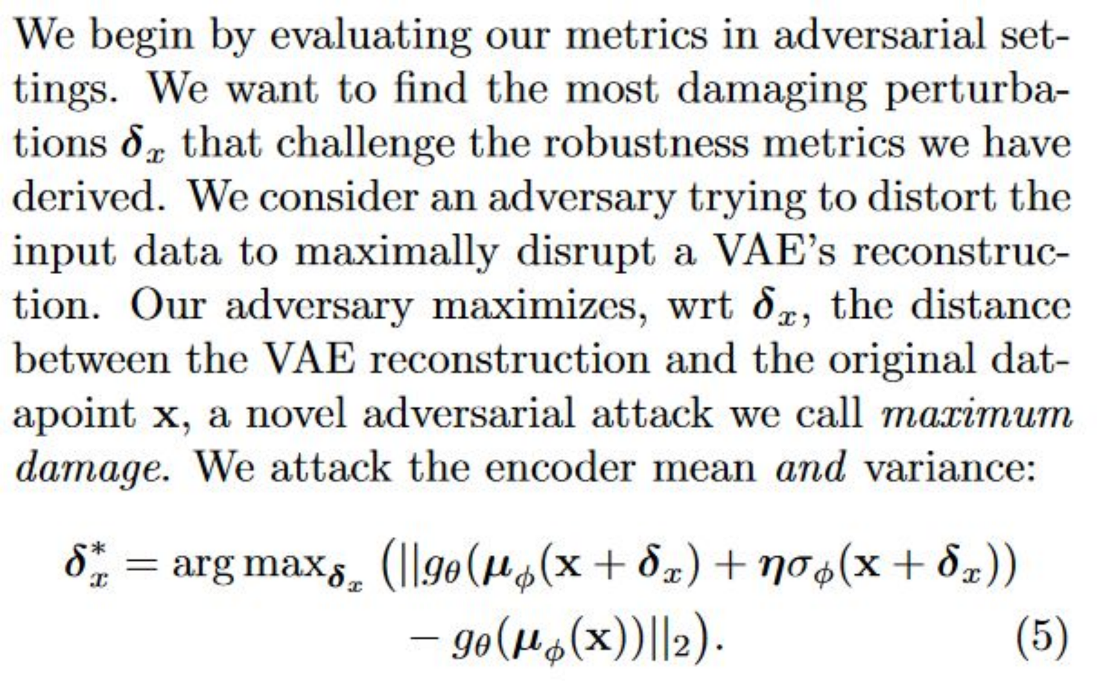
(c) CIFAR10

Accuracy of the Theorem (II)



$\Phi^{-1}(p(||\Delta(\mathbf{x})||_2 \leq r))$ (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

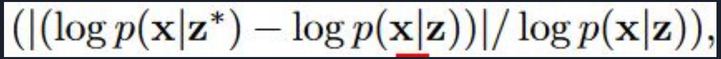
Empirical investigations



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:

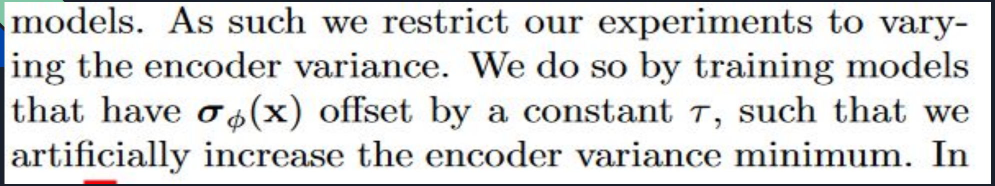
$$\boldsymbol{\delta}_x^* = \arg\max_{\boldsymbol{\delta}_x} \big(||g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}+\boldsymbol{\delta}_x) + \boldsymbol{\eta}\sigma_\phi(\mathbf{x}+\boldsymbol{\delta}_x)) - g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))||_2\big). \qquad (5)$$

quantify the degradation in model performance as the relative log likelihood degradation

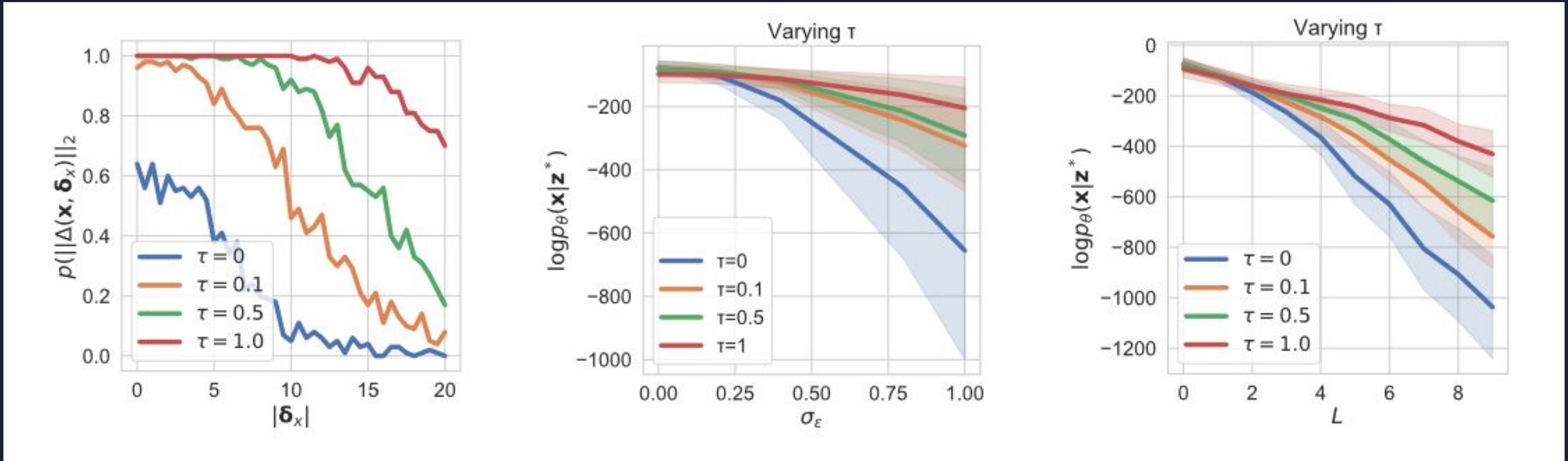


$$(|(\log p(\mathbf{x}|\mathbf{z}^*) - \log p(\mathbf{x}|\mathbf{z}))|/\log p(\mathbf{x}|\mathbf{z})),$$

Empirical Investigations (II)

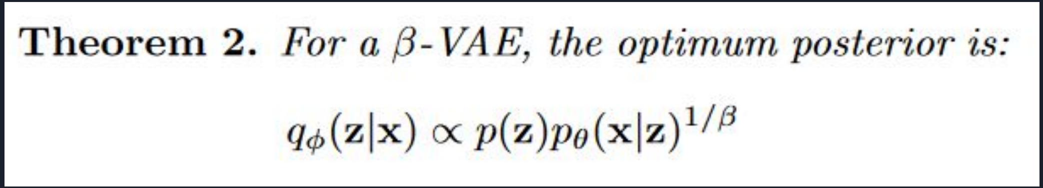


models. As such we restrict our experiments to varying the encoder variance. We do so by training models that have $\boldsymbol{\sigma}_\phi(\mathbf{x})$ offset by a constant τ, such that we artificially increase the encoder variance minimum. In



Robustness of Disentangled VAEs

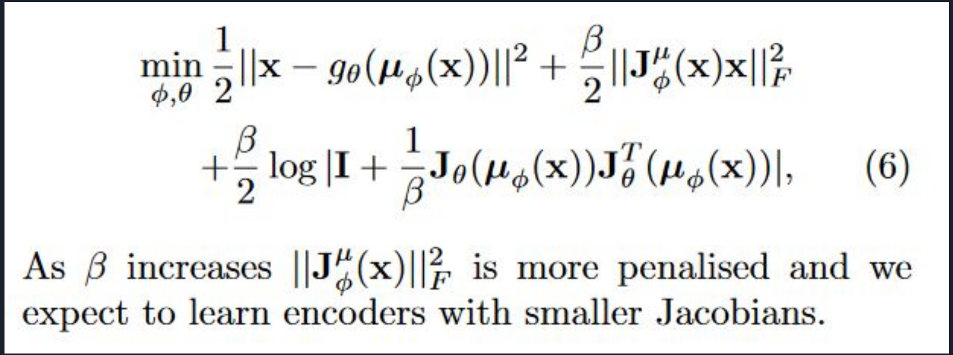
- Disentangling increases encoder variance:



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

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

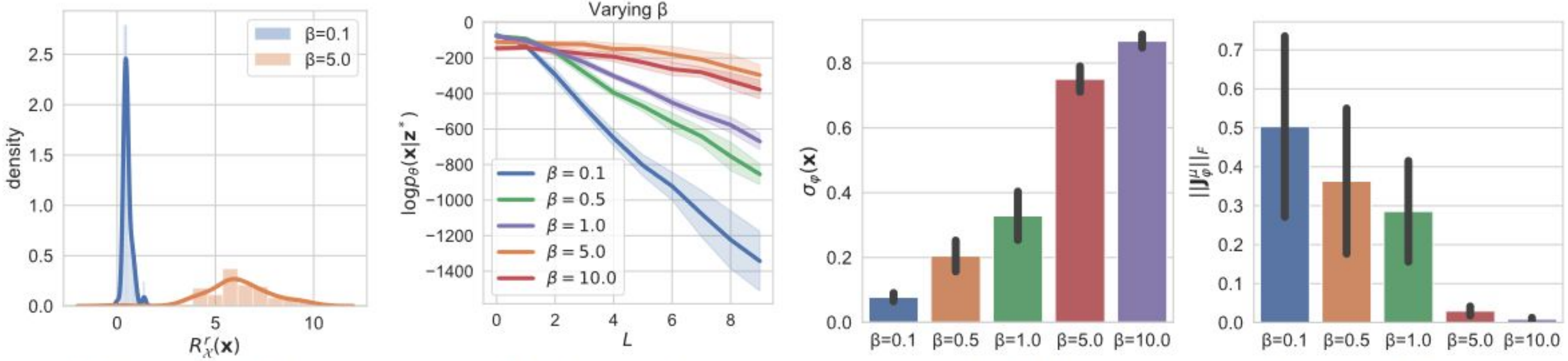
Disentangling penalizes Jacobian norm:



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

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

Empirical proof of robustness of disentangled VAEs



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