Appendix for Towards a Theoretical Understanding of the Robustness of Variational Autoencoders

Alexander Camuto, Matthew Willetts, Stephen Roberts, Chris Holmes, Tom Rainforth

A Choosing r for r-robustness

Proposition 1. *For any input and perturbation, a necessary requirement for a VAE with a Gaussian encoder to satisfy r-robustness is that*

$$r > \sqrt{2\text{Tr}(\boldsymbol{\Sigma}(\mathbf{x}))} + \mathcal{O}(\boldsymbol{\varepsilon}) \tag{1}$$

where $\boldsymbol{\Sigma}(\mathbf{x}) = \mathbf{J}_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))\boldsymbol{\sigma}^2_\phi(\mathbf{x})\mathbf{J}^T_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))$, $(\mathbf{J}_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))_{i,j} = \partial g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))_i/\partial(\boldsymbol{\mu}_\phi(\mathbf{x}))_j$, *and* $\mathcal{O}(\boldsymbol{\varepsilon})$ *represents higher order terms that tend to zero in the limit* $\boldsymbol{\sigma}_\phi(\mathbf{x}) \to \mathbf{0}$.

We provide empirical confirmations in Appendix D.1.1 that show that the r for r-robustness scales with the encoder variance.

Proof. Let $g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))$ be the result of mapping to the encoder mean ($\boldsymbol{\mu}_\phi$) and then decoding to the likelihood mean (g_θ), and $\Delta(\mathbf{x}) = g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\eta} \circ \boldsymbol{\sigma}_\phi(\mathbf{x})) - g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))$ we want to find a bound for r for which:

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

where as before $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Here we can invoke Taylor's theorem on $g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\eta} \circ \boldsymbol{\sigma}_\phi(\mathbf{x}))$ around the deterministic mapping $\boldsymbol{\mu}_\phi(\mathbf{x})$. Namely, if we assume that all terms in Hessian of $g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x})$ are finite (i.e. $|\partial^2 g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x})_i/\partial \boldsymbol{\mu}_\phi(\mathbf{x})_j \boldsymbol{\mu}_\phi(\mathbf{x})_k| < \infty \, \forall i, j, k$), then we have:

$$g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\epsilon}) = g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x})) + \mathbf{J}_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))(\boldsymbol{\eta} \circ \boldsymbol{\sigma}_\phi(\mathbf{x})) + \mathcal{O}(\boldsymbol{\varepsilon}) \tag{3}$$

where $\mathcal{O}(\varepsilon)$ represents asymptotically dominated higher order terms that go to zero in the limit of small $\boldsymbol{\sigma}_\phi(\mathbf{x}))$ and $\mathbf{J}_\theta$ is defined element-wise as:

$$\mathbf{J}_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))_{i,j} = \frac{\partial g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))_i}{\partial(\boldsymbol{\mu}_\phi(\mathbf{x}))_j} \tag{4}$$

Note that $\mathbf{J}_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))(\boldsymbol{\eta} \circ \boldsymbol{\sigma}_\phi(\mathbf{x}))$ is distributed according to the multivariate Gaussian

$$\mathcal{N}(0, \mathbf{J}_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))\boldsymbol{\sigma}^2_\phi(\mathbf{x})\mathbf{J}^T_\theta(\boldsymbol{\mu}_\phi(\mathbf{x})))$$

Given these definitions

$$p(||\Delta(\mathbf{x})||_2 \leq r) > p(||\Delta(\mathbf{x})||_2 > r) \tag{5}$$
$$\Leftrightarrow p(||\Delta(\mathbf{x})||_2 \leq r) > 0.5 \tag{6}$$
$$\Leftrightarrow p(||\mathbf{J}_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))(\boldsymbol{\eta} \circ \boldsymbol{\sigma}_\phi(\mathbf{x})) + \mathcal{O}(\boldsymbol{\varepsilon})||_2 < r) > 0.5 \tag{7}$$
$$\Leftrightarrow p(||\mathbf{J}_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))(\boldsymbol{\eta} \circ \boldsymbol{\sigma}_\phi(\mathbf{x})) + \mathcal{O}(\boldsymbol{\varepsilon})||^2_2 < r^2) > 0.5 \tag{8}$$

We must now consider the distribution of the square norm of $\boldsymbol{\mathcal{E}}(\mathbf{x}) = \mathbf{J}_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))(\boldsymbol{\eta} \circ \boldsymbol{\sigma}_\phi(\mathbf{x}))$. Let

$$Q(\boldsymbol{\mathcal{E}}(\mathbf{x})) = ||\boldsymbol{\mathcal{E}}(\mathbf{x})||^2_2 = \boldsymbol{\mathcal{E}}(\mathbf{x})^T \boldsymbol{\mathcal{E}}(\mathbf{x}) \tag{9}$$
$$\boldsymbol{\Sigma}(\mathbf{x}) = \mathbf{J}_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))\boldsymbol{\sigma}^2_\phi(\mathbf{x})\mathbf{J}^T_\theta(\boldsymbol{\mu}_\phi(\mathbf{x})) \tag{10}$$
$$\mathbf{Y}(\mathbf{x}) = \boldsymbol{\Sigma}(\mathbf{x})^{-\frac{1}{2}} \boldsymbol{\mathcal{E}}(\mathbf{x}) \tag{11}$$

Given that we restrict ourselves to positive activation functions, $\mathbf{J}_\theta$ is positive and $\mathbf{\Sigma}(\mathbf{x})$ will be positive semi definite and is invertible. As such we have $Q(\boldsymbol{\mathcal{E}}) = \mathbf{Y}^T(\mathbf{x})\mathbf{\Sigma}(\mathbf{x})\mathbf{Y}(\mathbf{x})$.

Using the spectral decomposition theorem we can write that $\mathbf{\Sigma}(\mathbf{x}) = \mathbf{P}^T(\mathbf{x})\mathbf{\Lambda}(\mathbf{x})\mathbf{P}(\mathbf{x})$ where $\mathbf{P}^T(\mathbf{x})\mathbf{P}(\mathbf{x}) = \mathbf{I}$ and $\mathbf{\Lambda}(\mathbf{x})$ is the diagonal matrix of the eigenvalues of $\mathbf{\Sigma}(\mathbf{x})$, $\lambda_1, \ldots, \lambda_{d_\mathcal{X}}$, where $d_\mathcal{X}$ is the dimensionality of the data-space. Given that $\mathbf{\Sigma}(\mathbf{x})$ is positive semi definite $\mathbf{\Lambda}(\mathbf{x})$ will only have positive values.

Let $\mathbf{U}(\mathbf{x}) = \mathbf{P}(\mathbf{x})\mathbf{Y}(\mathbf{x}) = \mathbf{P}(\mathbf{x})\mathbf{\Sigma}(\mathbf{x})^{-\frac{1}{2}}\boldsymbol{\mathcal{E}}(\mathbf{x})$, which is multivariate Gaussian with identity matrix and zero mean. We have that:

$$
\begin{aligned}
Q(\boldsymbol{\mathcal{E}}) &= \mathbf{Y}^T(\mathbf{x})\mathbf{\Sigma}(\mathbf{x})\mathbf{Y}(\mathbf{x}) && (12)\\
&= \mathbf{Y}^T(\mathbf{x})\mathbf{P}^T(\mathbf{x})\mathbf{\Lambda}(\mathbf{x})\mathbf{P}(\mathbf{x})\mathbf{Y}(\mathbf{x}) && (13)\\
&= \mathbf{U}^T(\mathbf{x})\mathbf{\Lambda}(\mathbf{x})\mathbf{U}(\mathbf{x}) && (14)
\end{aligned}
$$

As such:

$$
\sum_{i=1}^{d_\mathcal{X}} (\boldsymbol{\mathcal{E}}_i)^2 = \mathbf{U}^T(\mathbf{x})\mathbf{\Lambda}(\mathbf{x})\mathbf{U}(\mathbf{x}) = \sum_{i=1}^{d_\mathcal{X}} \lambda_i (\mathbf{U}_i(\mathbf{x}))^2, \quad \lambda_i(\mathbf{U}_i(\mathbf{x}))^2 \sim \Gamma\left(\frac{1}{2}, 2\lambda_i\right) \tag{15}
$$

This comes from the fact that for $\lambda_i\mathbf{X}, \mathbf{X} \sim \Gamma\left(\frac{1}{2}, 2\right)$ we have that $\lambda_i\mathbf{X} \sim \Gamma\left(\frac{1}{2}, 2\lambda_i\right)$.

To establish a lower bound on r, we use Markov's inequality which states that:

$$
p(||\boldsymbol{\mathcal{E}}(\mathbf{x}) + \mathcal{O}(\boldsymbol{\varepsilon})||_2^2 > r^2) < \frac{\mathbb{E}\,||\boldsymbol{\mathcal{E}}(\mathbf{x}) + \mathcal{O}(\boldsymbol{\varepsilon})||_2^2}{r^2} \tag{16}
$$

Here $\mathbb{E}\,||\boldsymbol{\mathcal{E}}(\mathbf{x})||_2^2 = \mathbb{E}\sum_{i=1}^{d_\mathcal{X}} (\boldsymbol{\mathcal{E}}_i(\mathbf{x}))^2 = \mathbb{E}\sum_{i=1}^{d_\mathcal{X}} (\lambda_i(\mathbf{U}_i)^2(\mathbf{x}))^2$, which is simply $\sum_{i=1}^{d_\mathcal{X}} \lambda_i$.

Recall that we want: $p(||\boldsymbol{\mathcal{E}}(\mathbf{x}) + \mathcal{O}(\boldsymbol{\varepsilon})||_2^2 > r^2) < 0.5$. As such

$$
r > \sqrt{2\sum_{i=1}^{d_\mathcal{X}} \lambda_i + \mathcal{O}(\boldsymbol{\varepsilon})} = \sqrt{2\mathrm{Tr}(\mathbf{\Sigma}(\mathbf{x}))} + \mathcal{O}(\boldsymbol{\varepsilon}) \tag{17}
$$

□

B Margin for r-robustness in $\mathcal{X}$

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*

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

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

Proof. Suppose we have an r for which r-robustness is satisfied before any perturbation is added to the VAE input. First we want to establish a margin in the latent space $\mathcal{Z}$ for which our model is robust given a perturbation in the latent space.

To do this, we first define

$$
\Delta_e(\mathbf{y}) = g_\theta(\mathbf{y}) - g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x})), \quad \mathbf{y} \in \mathcal{Z} \tag{18}
$$

where g_θ is the decoder network and $\mathbf{y}$ is an arbitrary realization of the latents. Note here that there is an implicit dependency on $\mathbf{x}$, but as this input is fixed we will ignore this dependency throughout. Let $\mathbf{z} = \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\eta} \circ \boldsymbol{\sigma}_\phi(\mathbf{x})$ be the random variable produced by the embedding, i.e. the latent sampled by the encoder. We want to find a bound R_e^r for which:

$$\|\boldsymbol{\delta}_z\|_2 \leq R_e^r \quad \Leftrightarrow \quad p(||\Delta_e(\mathbf{z}+\boldsymbol{\delta}_z)||_2 \leq r) > p(||\Delta_e(\mathbf{z}+\boldsymbol{\delta}_z)||_2 > r) \tag{19}$$

such that r-robustness is satisfied on the decoder output when we apply a deterministic a perturbation $\boldsymbol{\delta}_z$ of maximum size R_e^r to the random variable $\mathbf{z}$. Note that all the stochasticity is contained in $\boldsymbol{\eta}$.

Let A^r denote the set of $\boldsymbol{\delta}_z$ for which (19) holds and conversely let B^r be the set of $\boldsymbol{\delta}_z$ for which it does not. By assumption in the Theorem, then $\mathbf{0} \in A^r$ as the unperturbed input satisfies r-robustness. Moreover, we also have that this unperturbed input $\boldsymbol{\delta}_z$ has a probability $p_\Delta(\mathbf{0}) := p(||\Delta_e(\mathbf{z})||_2 \leq r) = p(||\Delta(\mathbf{x})||_2 \leq r) > 0.5$ of returning a reconstruction with r of $g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))$.

Now we know that $\mathbf{z}$ is a Gaussian random variable and so regardless of form of the decoder, $p_\Delta(\boldsymbol{\delta}_z) := p(||\Delta_e(\mathbf{z}+\boldsymbol{\delta}_z)||_2 \leq r)$ must vary smoothly as we change $\boldsymbol{\delta}_z$. In essence, as we increase the size of the perturbation $\boldsymbol{\delta}_z$ slowly from zero, the distribution of $\mathbf{z}+\boldsymbol{\delta}_z$ will still have most of its mass of the same region $\mathbf{z}$. When coupled with the fact that we have some "excess probability" $p_\Delta(\mathbf{0}) - 0.5$ beyond what it is needed for r-robustness, there must be at certain degree to which we can increase $\boldsymbol{\delta}_z$ before all this excess probability is "used up". We can then use this to construct a bound for R_e^r by considering the minimum $\boldsymbol{\delta}_z$ to break r-robustness in the "worst-case" setting for the boundary between A^r and B^r.

Intuitively as shown in Figure B.1, and also more formally using the Neyman-Pearson lemma (Neyman & Pearson, 1933) by analogy to the approach of Cohen et al. (2019), this worst case setting will occur when the boundary between A^r and B^r is a straight line perpendicular to the direction of lowest variance for $\mathbf{z}$ (remembering that this is Gaussian distributed) and $\boldsymbol{\delta}_z$ is increased in this direction of lowest variance. In essence, this is the setup where our excess probability is used up most quickly for a given $\|\boldsymbol{\delta}_z\|_2$. By assumption in the theorem statement, we are using a diagonal covariance encoder and so this direction of lowest variance is the latent variable corresponding to $\arg\min_i \boldsymbol{\sigma}_\phi(\mathbf{x})_i$. Further, by noting that we need only consider the marginal distribution in this dimension, it is straightforward to see that the bound is reached when

$$\|\boldsymbol{\delta}_z\|_2 = \left(\min_i \boldsymbol{\sigma}_\phi(\mathbf{x})_i\right) \Phi^{-1}\left(p_\Delta(\mathbf{0})\right) = \left(\min_i \boldsymbol{\sigma}_\phi(\mathbf{x})_i\right) \Phi^{-1}\left(p(||\Delta(\mathbf{x})||_2 \leq r)\right) \tag{20}$$

where Φ^{-1} is the inverse cumulative distribution function for a unit Gaussian, i.e. the probit function. Note that this yields $\|\boldsymbol{\delta}_z\|_2 = 0$ if $p(||\Delta(\mathbf{x})||_2 \leq r) = 0$, such we get the excepted result that our margin is zero is r-robustness only just holds without an input perturbation.

Next we need to relate $\|\boldsymbol{\delta}_z\|_2$ to $\|\boldsymbol{\delta}_x\|_2$. Here we can straightforwardly invoke Taylor's theorem on $\boldsymbol{\mu}_\phi(\mathbf{x}+\boldsymbol{\delta}_x)$ around the original input $\mathbf{x}$. Namely, if we assume that all terms in Hessian of $\mu_\phi(\mathbf{x})$ are finite (i.e. $|\partial^2 \boldsymbol{\mu}_\phi(\mathbf{x})_i / \partial \mathbf{x}_j \mathbf{x}_k| < \infty \, \forall i,j,k$), then we have

$$\boldsymbol{\delta}_z = \boldsymbol{\mu}_\phi(\mathbf{x}+\boldsymbol{\delta}_x) - \boldsymbol{\mu}_\phi(\mathbf{x}) = \mathbf{J}_\phi^\mu(\mathbf{x})\boldsymbol{\delta}_x + \mathcal{O}(\varepsilon) \tag{21}$$

where $\mathcal{O}(\varepsilon)$ represents asymptotically dominated higher order terms that go to zero in the limit of small $\boldsymbol{\delta}_x$. We thus have

$$\|\boldsymbol{\delta}_x\|_2 \leq \frac{\|\boldsymbol{\delta}_z\|_2}{\|\mathbf{J}_\phi^\mu(\mathbf{x})\|_F} + \mathcal{O}(\varepsilon) \tag{22}$$

where $\mathcal{O}(\varepsilon)$ again represents asymptotically dominated higher order terms (note though these are not the same terms as in (21)). To complete the proof we now simply combine this with (20) to give the $\|\boldsymbol{\delta}_x\|_2$ at which the bound is reached and thus the $R_{\mathcal{X}}^r(\mathbf{x})$ quoted in the theorem, namely

$$R_{\mathcal{X}}^r(\mathbf{x}) \geq \frac{(\min_i \boldsymbol{\sigma}_\phi(\mathbf{x})_i)\Phi^{-1}(p(||\Delta(\mathbf{x})||_2 \leq r))}{||\mathbf{J}_\phi^\mu(\mathbf{x})||_F} + \mathcal{O}(\varepsilon) \tag{23}$$

where the inequality comes from the fact that the $\boldsymbol{\delta}_z$ we derived was the worst possible case (i.e. smallest $\boldsymbol{\delta}_z$ which might reach the bound).

□

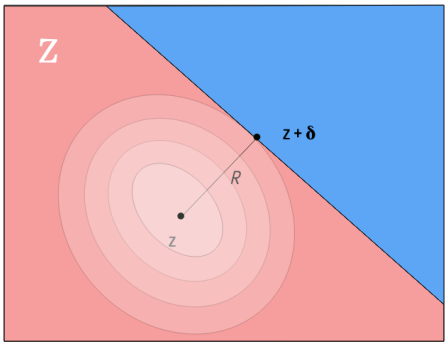


Figure B.1: Illustration of the boundary R we are measuring in $\mathcal{Z}$. Red represents spaces where A^r is satisfied. Blue represent spaces where B^r is satisfied. The concentric ellipsoids centered on $\mathbf{z}$ are the contours of $\mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi(\mathbf{x}))$. R is the minimum distance δ for which A^r is satisfied. The line dividing the two spaces represent the Neyman-Pearson "worst-case" model and is along the direction of minimum *variance*, $\min_i \boldsymbol{\sigma}^2_\phi(\mathbf{x})_i$.

C β-VAE Optimal Posterior

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

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

Proof. Here we use calculus of variations to obtain optimal posteriors for β-VAEs. The objectives we are optimising are over the whole dataset $\mathcal{D} = \{\mathbf{x}_i\}, i = 1, .., N$, with empirical data density $\rho(\mathbf{x}) = \frac{1}{N}\sum_i^N \delta(\mathbf{x} - \mathbf{x}_i)$.

The evidence lower bound for a β-VAE is

$$\mathcal{L}_\beta(\mathcal{D}; \theta, \phi) = \mathbb{E}_{\rho(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta \,\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))\right]. \tag{24}$$

This is easier to work with written explicitly as integrals. Note that as we are going to be finding the optimal $q_\phi(\mathbf{z}|\mathbf{x})$ we must add a constraint so that it integrates to 1.

$$\mathcal{L}_\beta(\mathcal{D}; \theta, \phi) = \int \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z}\, \rho(\mathbf{x})\big[q_\phi(\mathbf{z}|\mathbf{x})[\log p_\theta(\mathbf{x}|\mathbf{z}) - \beta \log q_\phi(\mathbf{z}|\mathbf{x}) + \beta \log p(\mathbf{z})] + \lambda(\mathbf{x})(q_\phi(\mathbf{z}|\mathbf{x}) - 1)\big] \tag{25}$$

For brevity, going forward $p_\theta(\mathbf{x}|\mathbf{z}) = p$, $p(\mathbf{z}) = \pi$, $q_\phi(\mathbf{z}|\mathbf{x}) = q$. We also view $\mathcal{L}$ as depending on q, p directly.

To proceed with calculus of variations, we substitute $q \to q + \epsilon$, where ϵ is a small function that goes to zero appropriately fast for large $\mathbf{x}, \mathbf{z}$. Thus we expand $\mathcal{L}$ to first order in q to find $\frac{\delta \mathcal{L}}{\delta q}$. The form of q for which this gradient is zero gives us the optimum q for this functional.

$$\mathcal{L}_\beta(q + \epsilon) = \int \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z}\, \rho(\mathbf{x})\big[(q + \epsilon)[\log p - \beta \log(q + \epsilon) + \beta \log \pi] + \lambda(\mathbf{x})(q + \epsilon - 1)\big] \tag{26}$$

Recall that $\log(1 + x) \approx x$ to first order. Thus $\log(q + \epsilon) \approx \log q + \frac{\epsilon}{q}$ to first order. So,

$$\mathcal{L}_\beta(q + \epsilon) = \int \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z}\, \rho(\mathbf{x}) q\big[\log p - \beta \log q + \beta \log \pi - \beta \frac{\epsilon}{q}\big] \tag{27}$$

$$+ \int \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z}\, \rho(\mathbf{x}) \epsilon\big[\log p - \beta \log q + \beta \log \pi - \beta \frac{\epsilon}{q}\big] \tag{28}$$

$$+ \int \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z}\, \rho(\mathbf{x}) \lambda(\mathbf{x})(q - 1) + \int \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z}\, \rho(\mathbf{x}) \lambda(\mathbf{x}) \epsilon + \int \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z}\, \rho(\mathbf{x}) O(\epsilon^2). \tag{29}$$

Rearranging we find

$$\mathcal{L}_\beta(q+\epsilon) = \mathcal{L}_\beta(q) + \int \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{z}\,\rho(\mathbf{x})\epsilon\big[\log p - \beta\log q + \beta\log\pi - \beta + \lambda(\mathbf{x})\big] + \int \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{z}\,\rho(\mathbf{x})O(\epsilon^2) \tag{30}$$

$$= \mathcal{L}_\beta(q) + \int \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{z}\,\frac{\delta\,\mathcal{L}_\beta}{\delta q}\epsilon + \int \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{z}\,\rho(\mathbf{x})O(\epsilon^2) \tag{31}$$

At the optimum value of q the functional will have vanishing functional derivative $\frac{\delta\,\mathcal{L}_\beta}{\delta q}$, so

$$\log p - \beta\log q + \beta\log\pi - \beta + \lambda(\mathbf{x}) = 0, \tag{32}$$

$$\log q = \frac{1}{\beta}\log p + \log\pi + C(\mathbf{x}). \tag{33}$$

Exponentiating we find the optimal q to be

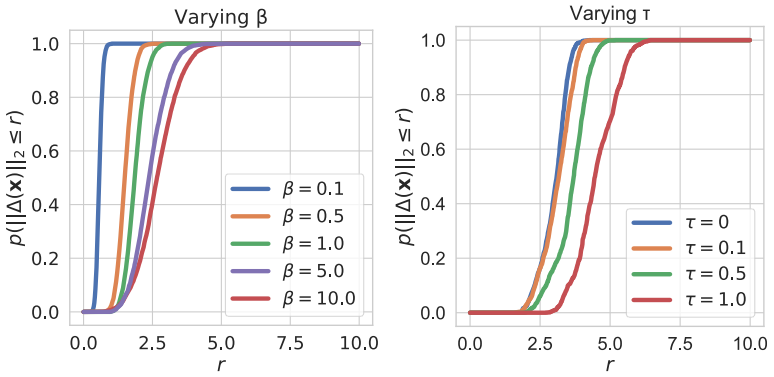
$$q_\phi(\mathbf{z}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}p_\theta(\mathbf{x}|\mathbf{z})^{\frac{1}{\beta}}p(\mathbf{z}), \tag{34}$$

where Z is an appropriate normalising constant. This completes the proof. □

D Empirical Calculation of the Bounds

D.1 Estimating the minimum r

D.1.1 Results



(a) (b)

Figure D.2: Here we show that the minimum r for which $p(||\Delta(\mathbf{x})||_2 \leq r) = 0.5$ increases with β and τ, where β is the penalty applied to the KL in β-VAEs and τ is an offset added to the encoder standard deviation $\boldsymbol{\sigma}_\phi(\mathbf{x})$. This probability, estimated as detailed below in Appendix D.1.2, increases with r, but increases more slowly for large β (a) and large τ (b). In such models the encoding process has higher variance resulting in a greater spread of reconstructions, confirming Proposition 1 in Appendix A that the minimum r for r-robustness increases with the encoder variance.

D.1.2 Algorithm

Algorithm 1: Estimating r

Result: r such that $p(||\Delta(\mathbf{x})||_2 \leq r) > 0.5$
m, *step*, *samples*, $\mathbf{x}$, $r \leftarrow 0$, $p(||\Delta(\mathbf{x})||_2 \leq r) \leftarrow 0$;
while $p(||\Delta(\mathbf{x})||_2 \leq r) < m$ **do**
  $d \leftarrow \{\}$;
  for $i \leftarrow 1$ **to** *samples* **by** 1 **do**
    $\mathbf{s} \sim \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}_\phi(\mathbf{x}))$;
    $s_d \leftarrow ||g_\theta(\mathbf{s}) - g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))||_2$;
    d.insert(s_d);
  end
  $r \leftarrow r + step$;
  $p(||\Delta(\mathbf{x})||_2 \leq r) \leftarrow \frac{\text{Sum}(d<r)}{nsamples}$;
end

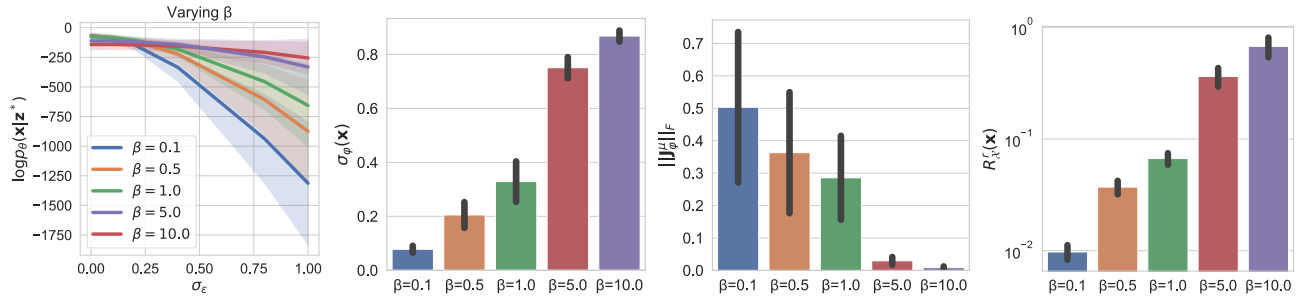
D.2 Estimating $R^r_{\mathcal{X}}(\mathbf{x})$

Algorithm 2: Estimating $R^r_{\mathcal{X}}(\mathbf{x})$

Result: $R^r_{\mathcal{X}}(\mathbf{x})$ such that $p(||\Delta(\mathbf{x}, \boldsymbol{\delta}_x)||_2 \leq r) > 0.5$
step, *samples*, $\mathbf{x}$, r, $p(||\Delta(\mathbf{x}, \boldsymbol{\delta}_x)||_2 \leq r) \leftarrow 0$, $R^r_{\mathcal{X}}(\mathbf{x}) \leftarrow 10$, *restarts* $\leftarrow 5$;
while $p(||\Delta(\mathbf{x}, \boldsymbol{\delta}_x)||_2 \leq r) < 0.5$ **do**
  for $j \leftarrow 1$ **to** *restarts* **by** 1 **do**
    $d \leftarrow \{\}$;
    for $i \leftarrow 1$ **to** *samples* **by** 1 **do**
      $\boldsymbol{\delta}_x \leftarrow$ max damage attack constrained to the norm $R^r_{\mathcal{X}}(\mathbf{x})$;
      $\mathbf{s} \sim \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x} + \boldsymbol{\delta}_x), \boldsymbol{\sigma}_\phi(\mathbf{x} + \boldsymbol{\delta}_x))$
      $s_d \leftarrow ||g_\theta(\mathbf{s}) - g_\theta(\boldsymbol{\mu}_\phi(\mathbf{x}))||_2$;
      d.insert(s_d);
    end
    $R^r_{\mathcal{X}}(\mathbf{x}) \leftarrow R^r_{\mathcal{X}}(\mathbf{x}) - step$;
    $p(||\Delta(\mathbf{x}, \boldsymbol{\delta}_x)||_2 \leq r) \leftarrow \frac{\text{Sum}(d<r)}{nsamples}$;
  end
end

E β-VAE Sensitivity Experiments

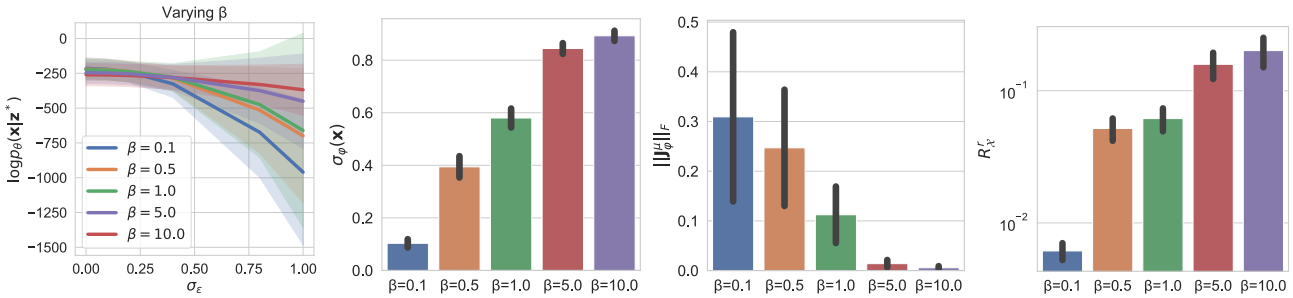
(a) $\beta = 0.1$ (b) $\beta = 0.5$ (c) $\beta = 1$ (d) $\beta = 5$ (e) $\beta = 10$



(f) VAE Sensitivity (g) Encoder Variance (h) Encoder Jacobian (i) $R^r_{\mathcal{X}}(\mathbf{x})$ bound

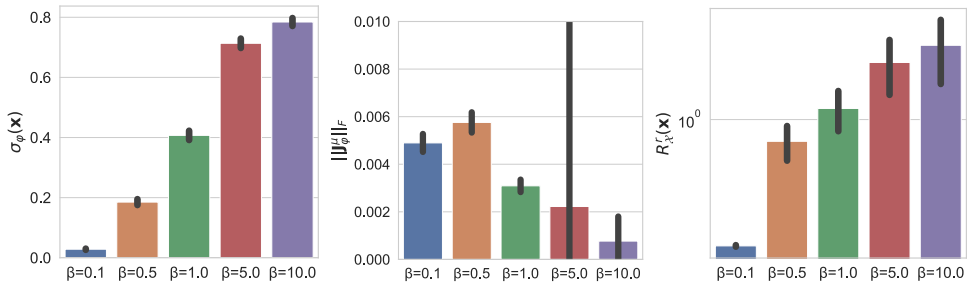
Figure E.3: Here we illustrate that β-VAEs, trained MNIST, with higher β penalties generalise better and are less sensitive to input perturbations. The first row (a)-(e) shows samples drawn from the latent space prior that are then fed through the VAE decoder. It is clear that as β increases, so too does the quality of generated samples. (f) shows the sensitivity of the VAE to input perturbations. We add zero-mean Gaussian noise of variance σ^2_ϵ to the VAE input to form a noisy input $\mathbf{x}^*$ and embedding $\mathbf{z}*$. We then measure the likelihood of the original point $\mathbf{x}$ under this noisy embedding. σ^2_ϵ is thus an approximation of the margin of robustness of the VAE, if the VAE's likelihood does not change even for high variance noise, it must have a large margin of robustness ($R^r_{\mathcal{X}}(\mathbf{x})$). The likelihood of $\mathbf{x}$ is quasi constant, under increasing noise variance, for high values of β. This supports our analysis that such models have higher $R^r_{\mathcal{X}}(\mathbf{x})$. Figures (g) and (h) show that the encoder variance and that the norm of the encoder Jacobian ($||\mathbf{J}^\mu_\phi(\mathbf{x})||_F$) increase as β increases, supporting our analysis that the changes in these values underpin the robustness observed. In (i) we calculate the bound for $R^r_{\mathcal{X}}(\mathbf{x})$ from Theorem 1 where we ignore higher order terms. We select r such that $p_{A^r}(\mathbf{x}) = 0.9$, which is a relatively strict metric for robustness. In (f-i) confidence intervals correspond to the standard deviations of values over the entire MNIST dataset. Taken as a whole these experiments support our analysis that the margin $R^r_{\mathcal{X}}(\mathbf{x})$ increases with β as in Theorem 2, in conjuction with the norm of the encoder Jacobian and the encoder variance, supporting Theorem 1.

(a) $\beta = 0.1$ (b) $\beta = 0.5$ (c) $\beta = 1$ (d) $\beta = 5$ (e) $\beta = 10$



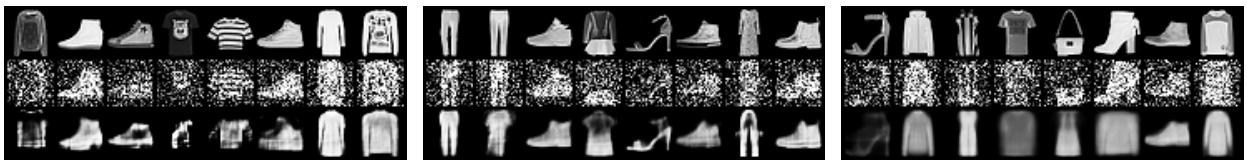
(f) VAE Sensitivity (g) Encoder Variance (h) Encoder Jacobian (i) $R^r_\mathcal{X}(\mathbf{x})$ bound

Figure E.4: Here we illustrate that β-VAEs, trained on fashion-MNIST, with higher β penalties generalise better and are less sensitive to input perturbations. The first row (a)-(e) shows samples drawn from the latent space prior that are then fed through the VAE decoder. It is clear that as β increases, so too does the quality of generated samples. (f) shows the sensitivity of the VAE to input perturbations. We add zero-mean Gaussian noise of variance σ_ϵ^2 to the VAE input to form a noisy input $\mathbf{x}^*$ and embedding $\mathbf{z}*$. We then measure the likelihood of the original point $\mathbf{x}$ under this noisy embedding. σ_ϵ^2 is thus an approximation of the margin of robustness of the VAE, if the VAE's likelihood does not change even for high variance noise, it must have a large margin of robustness ($R^r_\mathcal{X}(\mathbf{x})$). The likelihood of $\mathbf{x}$ is quasi constant, under increasing noise variance, for high values of β. This supports our analysis that such models have higher $R^r_\mathcal{X}(\mathbf{x})$. Figures (g) and (h) show that the encoder variance and that the encoder Jacobian norm ($||\mathbf{J}^\mu_\phi(\mathbf{x})||_F$) increase as β increases, supporting our analysis that the changes in these values underpin the robustness observed. In (i) we calculate the bound for $R^r_\mathcal{X}(\mathbf{x})$ from Theorem 1 where we ignore higher order terms. We select r such that $p_{A^r}(\mathbf{x}) = 0.9$, which is a relatively strict metric for robustness. In (f-i) confidence intervals correspond to the standard deviations of values over the entire fashion-MNIST dataset. Taken as a whole these experiments support our analysis that the margin $R^r_\mathcal{X}(\mathbf{x})$ increases with β as in Theorem 2, in conjuction with the norm of the encoder Jacobian and the encoder variance, supporting Theorem 1.



(a) Encoder Variance (b) Encoder Jacobian (c) $R^r_{\mathcal{X}}(\mathbf{x})$ bound

Figure E.5: Here we illustrate that β-VAEs, trained on CIFAR10, with higher β have larger margins of robustness. Figures (a) and (b) show that the encoder variance and that the encoder Jacobian norm ($||\mathbf{J}^\mu_\phi(\mathbf{x})||_F$) increase as β increases, supporting our analysis that the changes in these values underpin the robustness observed. In (i) we calculate the bound for $R^r_{\mathcal{X}}(\mathbf{x})$ from Theorem 1 where we ignore higher order terms. We select r such that $p_{A^r}(\mathbf{x}) = 0.9$, which is a relatively strict metric for robustness. In (a-c) confidence intervals correspond to the standard deviations of values over the entire dataset. Taken as a whole these experiments support our analysis that the margin $R^r_{\mathcal{X}}(\mathbf{x})$ increases with β as in Theorem 2, in conjuction with the norm of the encoder Jacobian and the encoder variance, supporting Theorem 1.



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

Figure E.6: We show reconstructions of noisy data for VAEs trained with $\beta \in \{0.1, 1, 10\}$ on Fashion-MNIST. The first row corresponds to the original image, the second to noised a image $\mathbf{x} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, (0.5^2)\mathbf{I})$. Clearly larger β models are less sensitive to noise, supporting our analysis that increasing β increases the margin of robustness to perturbations.

F Network Hyperparameters

All networks used the same hyperparameters. Namely networks were trained for 100 epochs with the Adam optimizer, with a learning rate of 0.001 and a batch size of 512.

For MNIST and fashion-MNIST networks for the encoder variance and encoder mean were two hidden layer multi-layer perceptrons (MLPs) with 400 units per layer, which shared their first layer. Similarly the decoder was a two layer MLP with 400 units per layer. For these datasets we used a latent space size of 20.

For CIFAR10 we used 4-layer MLPs with 400 units per layer for the encoder and decoder networks and used a 64-dimensional latent space.