Certifiably Robust Variational Autoencoders

Ben Barrett[*,†] **Alexander Camuto**[†,‡] **Matthew Willetts**[¶,‡] **Tom Rainforth**[†]

Abstract

We derive bounds on the minimal size of an input perturbation required to change a variational autoencoder (VAE) reconstruction by more than an allowed amount, with these bounds depending on key parameters such as the Lipschitz constants of the encoder and decoder. Our bounds allow one to specify a desired level of robustness *upfront* and then train a VAE that is certified to achieve this robustness.

1 Introduction

Variational autoencoders (VAEs) are a powerful method for learning deep generative models [1, 2], yet like other methods [3] are susceptible to adversarial attacks. For example, VAEs can be induced to reconstruct images similar to an adversary's target through only moderate perturbation of the input image [4, 5, 6]. This is undesirable since VAEs have been used to improve the robustness of classifiers [7, 8], their encodings are commonly used in downstream tasks [9, 10], and their susceptibility to input perturbations challenges an original ambition that VAEs should capture "semantically meaningful [...] factors of variation in data" [11].

While previous work has already sought to obtain more robust VAEs empirically [12, 13, 14], this work lacks formal guarantees. This is a meaningful worry because in other model classes, robustification techniques showing promise empirically but lacking guarantees have later been circumvented by more sophisticated attacks [15, 16]. Further, though previous theoretical work [17] can ascertain robustness *post-training*, it cannot determine robustness *a priori*, before training.

Our work looks to alleviate these issues by providing VAEs whose robustness levels can be certified by design. In particular, we show how to construct *certifiably* robust VAEs by enforcing Lipschitz continuity in the encoder and decoder; we call the resulting models *Lipschitz-VAEs*.

We first derive a per-datapoint lower bound that guarantees a certain probability of a Lipschitz-VAE's reconstructions of distorted inputs being close to the reconstructions of undistorted inputs. Using this bound we can then obtain a margin that holds for all inputs. This second, *global* bound means that we can guarantee, for *any* input, that perturbations within the margin induce reconstructions that fall within a ball (of specified radius) of the original reconstruction with *at least* some specified probability. Since this margin does not depend on the value of the input data and can have its value specified *a priori* from a small number of network hyperparameters, it enables VAEs with chosen levels of robustness.

2 Background

VAEs Assume we want to learn a latent variable model with joint density $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})$, parameterized by θ, that captures observations $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \in \mathcal{X}$ generated according to an unknown process involving latent variables $\mathbf{z} \in \mathcal{Z}$. Since learning θ by maximum likelihood is typically intractable, variational inference introduces inference model $q_\phi(\mathbf{z}|\mathbf{x})$ [11], parametrized by ϕ, which yields a tractable lower bound on the marginal likelihood,

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})\right). \tag{1}$$

[*] Correspondence to `ben.neuber.barrett@gmail.com`. [†] University of Oxford [‡] Alan Turing Institute [¶] University College London

NeurIPS 2021 Workshop on Bayesian Deep Learning. Do not distribute.

In a VAE, θ and ϕ represent the parameters of the *decoder* and *encoder network* respectively. Having sampled $\mathbf{z}_i \sim q_\phi(\mathbf{z}|\mathbf{x}_i)$ on input $\mathbf{x}_i$, we will refer to $g_\theta(\mathbf{z}_i)$ as a *reconstruction* of $\mathbf{x}_i$, where $g_\theta(\cdot)$ denotes the *deterministic component of the decoder* [18].

Adversarial Attacks on VAEs Although much work has focused on classifiers, adversarial attacks have also been proposed for VAEs. Given original input $\mathbf{x}_o$ and the adversary's target output $\mathbf{x}_t$, the attacker seeks a perturbation $\boldsymbol{\delta} \in \mathcal{X}$ such that a VAE's reconstruction of the perturbed input $(\mathbf{x}_o + \boldsymbol{\delta})$ is similar to $\mathbf{x}_t$. The best performing attack in the current literature is a *latent space attack* [4, 5, 6], where for hyperparameter λ an adversary optimizes

$$\underset{\boldsymbol{\delta}:\ ||\boldsymbol{\delta}||_2}{\arg\min} \quad \mathrm{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}_o + \boldsymbol{\delta}) || q_\phi(\mathbf{z}|\mathbf{x}_t)\right) + \lambda ||\boldsymbol{\delta}||_2. \tag{2}$$

Defining Robustness, Lipschitz Continuity Since VAE reconstructions are typically continuous–valued and a VAE's encoder is usually a continuous distribution, any change to a VAE's input will almost surely result in a change in its reconstructions. This observation rules out robustness criteria that specify robustness using margins around inputs within which model outputs are constant [19, 20]. Further, VAEs are probabilistic: a VAE's outputs will vary even under the same input. To account for these considerations, we employ the robustness criterion of [17]:

Definition 2.1. ((r, ϵ)-robustness) For $r \in \mathbb{R}^+$ and $\epsilon \in [0, 1)$, a model f operating on a point $\mathbf{x}$ and outputting a continuous random variable is (r, ϵ)-robust to a perturbation $\boldsymbol{\delta}$ if and only if

$$\mathbb{P}\left[||f(\mathbf{x} + \boldsymbol{\delta}) - f(\mathbf{x})||_2 \leq r\right] > \epsilon.$$[2]

We term the probability above the *r-robustness probability*. The definition of (r, ϵ)-robustness naturally leads to the notion of an accompanying robustness margin [17]:

Definition 2.2. ((r, ϵ)-robustness margin) For $r \in \mathbb{R}^+$ and $\epsilon \in [0, 1)$, a model f has (r, ϵ)-robustness margin $R^{(r,\epsilon)}(\mathbf{x})$ about input $\mathbf{x}$ if $||\boldsymbol{\delta}||_2 < R^{(r,\epsilon)}(\mathbf{x}) \implies \mathbb{P}\left[||f(\mathbf{x} + \boldsymbol{\delta}) - f(\mathbf{x})||_2 \leq r\right] > \epsilon$.

A model with an (r, ϵ)-robustness margin on $\mathbf{x}$ can only be undermined by more than r by perturbations with norm less than $R^{(r,\epsilon)}(\mathbf{x})$ with probability less than $(1 - \epsilon)$.

Definition 2.3. (Lipschitz continuity) A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous if for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$, $||f(\mathbf{x}_1) - f(\mathbf{x}_2)||_2 \leq M||\mathbf{x}_1 - \mathbf{x}_2||_2$ for constant $M \in \mathbb{R}^+$. The least M for which this holds is called the *Lipschitz constant* of f, and if f has Lipschitz constant M, we say it is M-Lipschitz.

3 Certifiably Robust VAEs

Bounding the r-Robustness Probability We now introduce a VAE whose robustness levels can be certified by enforcing Lipschitz continuity in its encoder and decoder network. We can guarantee that this VAE's reconstructions will change only to a particular degree under distortions by bounding its r-robustness probability from below. Under the common choice of a diagonal-covariance multivariate Gaussian encoder with parameterization $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}; \mu_\phi(\mathbf{x}), \mathtt{diag}\left(\sigma^2_\phi(\mathbf{x})\right)\right)$, where $\mu_\phi : \mathcal{X} \to \mathbb{R}^{d_z}$ is the *encoder mean* and $\sigma_\phi : \mathcal{X} \to \mathbb{R}^{d_z}_{\geq 0}$ is the *encoder standard deviation*, the following result holds (proofs for all subsequent results provided in Appendix A).

Theorem 1 (Probability Bound). *Assume $q_\phi(\mathbf{z}|\mathbf{x})$ is as above and that the deterministic component of the Lipschitz-VAE decoder $g_\theta(\cdot)$ is a-Lipschitz, the encoder mean $\mu_\phi(\cdot)$ is b-Lipschitz, and the encoder standard deviation $\sigma_\phi(\cdot)$ is c-Lipschitz. Finally, let $\mathbf{z}_{\boldsymbol{\delta}} \sim q_\phi(\mathbf{z}|\mathbf{x} + \boldsymbol{\delta})$ and $\mathbf{z}_{\neg\boldsymbol{\delta}} \sim q_\phi(\mathbf{z}|\mathbf{x})$. Then for any $r \in \mathbb{R}^+$, any $\mathbf{x} \in \mathcal{X}$, and any input perturbation $\boldsymbol{\delta} \in \mathcal{X}$,*

$$\mathbb{P}\left[||g_\theta(\mathbf{z}_{\boldsymbol{\delta}}) - g_\theta(\mathbf{z}_{\neg\boldsymbol{\delta}})||_2 \leq r\right] \geq 1 - \min\left\{p_1(\mathbf{x}), p_2(\mathbf{x})\right\},$$

where

$$p_1(\mathbf{x}) := \min\left(1, \frac{a^2\left(b^2||\boldsymbol{\delta}||_2^2 + (c||\boldsymbol{\delta}||_2 + 2||\sigma_\phi(\mathbf{x})||_2)^2\right)}{r^2}\right)$$

$$p_2(\mathbf{x}) := \begin{cases} C(d_z)\frac{u(\mathbf{x})^{\frac{d_z}{2}}\exp\left\{-\frac{u(\mathbf{x})}{2}\right\}}{u(\mathbf{x}) - d_z + 2} & \left(\frac{r}{a} - b||\boldsymbol{\delta}||_2\right) \geq 0; d_z \geq 2; u(\mathbf{x}) > d_z - 2 \\ 1 & \textit{o.w.} \end{cases}$$

[2] We use the ℓ_2 norm but the following definitions could also be stated with respect to other norms.

for $u(\mathbf{x}) := \frac{\left(\frac{r}{a} - b||\boldsymbol{\delta}||_2\right)^2}{(c||\boldsymbol{\delta}||_2 + 2||\sigma_\phi(\mathbf{x})||_2)^2}$ *and constant* $C(d_z) := \frac{1}{\sqrt{\pi}} \exp\left\{\frac{1}{2}(d_z - (d_z - 1)\log d_z)\right\}$.

Theorem 1 tells us that a Lipschitz-VAE's r-robustness probability can be bounded in terms of r, the Lipschitz constants of the encoder and decoder, the norm of the encoder standard deviation, the dimension of the latent space, and the norm of the input perturbation.

Bounding the (r, ϵ)-Robustness Margin While Theorem 1 allows the r-robustness probability to be lower-bounded for a given input and input perturbation, we would like to guarantee a VAE's robustness at a given input to *all* input perturbations up to some magnitude (for a given ϵ). The following result lower-bounds the (r, ϵ)-robustness margin at a given input to provide such a guarantee.

Lemma 1.1 (Margin Bound)**.** *Given the assumptions of Theorem 1 and some* $\epsilon \in [0, 1)$*, the* (r, ϵ)*-robustness margin of this Lipschitz-VAE on input* $\mathbf{x}$*,*

$$R^{(r,\epsilon)}(\mathbf{x}) \geq \max\{m_1(\mathbf{x}), m_2(\mathbf{x})\} \quad \textit{for}$$

$$m_1(\mathbf{x}) := \frac{-4c||\sigma_\phi(\mathbf{x})||_2 + \sqrt{(4c||\sigma_\phi(\mathbf{x})||_2)^2 - 4\left(c^2 + b^2\right)\left(4||\sigma_\phi(\mathbf{x})||_2 - (1-\epsilon)\left(\frac{r}{a}\right)^2\right)}}{2\left(c^2 + b^2\right)}$$

and $m_2(\mathbf{x}) := \sup\left\{||\boldsymbol{\delta}||_2 : p_2(\boldsymbol{\delta}, \mathbf{x}) \leq (1-\epsilon)\right\}$, *where* $p_2(\boldsymbol{\delta}, \mathbf{x})$ *is as in Theorem 1.*[3]

A global margin can now be obtained by bounding $R^{(r,\epsilon)}(\mathbf{x})$ from below for all $\mathbf{x} \in \mathcal{X}$. The only input dependence is via $\sigma_\phi(\mathbf{x})$, which can be lifted by setting $\sigma_\phi(\mathbf{x}) = \boldsymbol{\sigma} \in \mathbb{R}^{d_z}_{\geq 0}$, a chosen hyperparameter. This can be done either during training (VAEs can be trained with a fixed encoder standard deviation without serious degradation in performance [21]), or afterwards, since all that matters to the bound is the value of $\boldsymbol{\sigma}$ at test time.

Theorem 2 (Global Margin Bound)**.** *Given the assumptions of Lemma 1.1, but with* $\sigma_\phi(\mathbf{x}) = \boldsymbol{\sigma} \in \mathbb{R}^{d_z}$*, the* (r, ϵ)*-robustness margin of this VAE for all inputs is*

$$R^{(r,\epsilon)} \geq \max\{m_1, m_2\} \quad \textit{for} \quad m_1 := \frac{\sqrt{-\left(4||\boldsymbol{\sigma}||_2^2 - (1-\epsilon)\left(\frac{r}{a}\right)^2\right)}}{b}$$

and $m_2 := \sup\{||\boldsymbol{\delta}||_2 : p_2(\boldsymbol{\delta}) \leq (1-\epsilon)\}$*, where* p_2 *is as in Theorem 1, but* $u := \frac{\left(\frac{r}{a} - b||\boldsymbol{\delta}||_2\right)^2}{4||\boldsymbol{\sigma}||_2^2}$.

Importantly, this result provides a robustness guarantee solely in terms of parameters we can choose ahead of training, most prominently the Lipschitz constants of the networks (as discussed in the next section) and $\boldsymbol{\sigma}$, the fixed encoder standard deviation. This distinguishes ours from previous work, which has only considered robustness in VAEs based on intractable model characteristics that must be empirically estimated after training [17].

Implementing Certifiably Robust VAEs Previously, we have taken the Lipschitz constants of the VAE's encoder and decoder as given. In practice, the Lipschitz constants of feed-forward and convolutional neural networks can be set using the approach of [22, 23]. Focusing on feed-forward networks, [22] shows that arbitrary Lipschitz constants can be chosen by composing 1-Lipschitz linear transformations, 1-Lipschitz activation functions, and appropriate scalings in each layer (we offer further exposition in Appendix B). This work's theoretical results might thus be utilized to provide practicable a priori robustness guarantees.

4 Conclusion

We have derived theoretical bounds on the degree of a VAE's robustness under input perturbation, with these bounds depending on parameters such as the Lipschitz constants of the VAE encoder and decoder networks. We have also seen that controlling these parameters, as existing methods permit, enables certification of a VAE's robustness ahead of training.

[3] We make explicit the dependence on $\boldsymbol{\delta}$.

References

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

[2] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[3] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[4] Pedro Tabacof, Julia Tavares, and Eduardo Valle. Adversarial images for variational autoencoders. *arXiv preprint arXiv:1612.00155*, 2016.

[5] George Gondim-Ribeiro, Pedro Tabacof, and Eduardo Valle. Adversarial attacks on variational autoencoders. *arXiv preprint arXiv:1806.04646*, 2018.

[6] Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. *2018 IEEE Security and Privacy Workshops (SPW)*, May 2018. doi: 10.1109/spw.2018.00014.

[7] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. *arXiv preprint arXiv:1805.09190*, 2018.

[8] Partha Ghosh, Arpan Losalka, and Michael J. Black. Resisting adversarial attacks using gaussian mixture variational autoencoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:541–548, Jul 2019. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.3301541.

[9] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, abs/1803.10122, 2018.

[10] Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. In *International Conference on Machine Learning*, pages 1480–1490. PMLR, 2017.

[11] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8245. doi: 10.1561/2200000056.

[12] Matthew Willetts, Alexander Camuto, Tom Rainforth, Stephen Roberts, and Chris Holmes. Improves vaes' robustness to adversarial attacks. *arXiv preprint arXiv:1906.00230*, 2019.

[13] Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dvijotham, and Pushmeet Kohli. Adversarially robust representations with smooth encoders. In *International Conference on Learning Representations*, 2020.

[14] Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dvijotham, Sven Gowal, and Pushmeet Kohli. The autoencoding variational autoencode. In *Advances in Neural Information Processing Systems*, 2020.

[15] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

[16] Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018.

[17] Alexander Camuto, Matthew Willetts, Stephen Roberts, Chris Holmes, and Tom Rainforth. Towards a theoretical understanding of the robustness of variational autoencoders. *arXiv preprint arXiv:2007.07365*, 2020.

[18] Abhishek Kumar and Ben Poole. On implicit regularization in β-vaes. *arXiv preprint arXiv:2002.00041*, 2020.

[19] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 2019.

[20] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, pages 11292–11303, 2019.

[21] Partha Ghosh, Mehdi S M Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From Variational to Deterministic Autoencoders. In *International Conference on Learning Representations*, 2020.

[22] Cem Anil, James Lucas, and Roger Grosse. Sorting out lipschitz function approximation. In *International Conference on Machine Learning*, pages 291–301, 2019.

[23] Qiyang Li, Saminul Haque, Cem Anil, James Lucas, Roger B Grosse, and Jörn-Henrik Jacobsen. Preventing gradient attenuation in lipschitz constrained convolutional networks. In *Advances in neural information processing systems*, pages 15390–15402, 2019.

[24] Jun Shao. Noncentral chi-squared, t- and f-distributions. Lecture, 2015. URL http://pages.stat.wisc.edu/~shao/stat609/stat609-13.pdf.

[25] Tadeusz Inglot. Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30(2):339–351, 2010.

[26] Todd Huster, Cho-Yu Jason Chiang, and Ritu Chadha. Limitations of the lipschitz constant as a defense against adversarial examples. *Lecture Notes in Computer Science*, page 16–29, 2019. ISSN 1611-3349. doi: 10.1007/978-3-030-13453-2_2.

[27] Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, pages 3835–3844, 2018.

A Proofs

Theorem 1 (Probability Bound)**.** *Assume $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}; \mu_\phi(\mathbf{x}), \mathtt{diag}\left(\sigma^2_\phi(\mathbf{x})\right)\right)$ and that the deterministic component of the Lipschitz-VAE decoder $g_\theta(\cdot)$ is a-Lipschitz, the encoder mean $\mu_\phi(\cdot)$ is b-Lipschitz, and the encoder standard deviation $\sigma_\phi(\cdot)$ is c-Lipschitz. Finally, let $\mathbf{z}_{\boldsymbol{\delta}} \sim q_\phi(\mathbf{z}|\mathbf{x}+\boldsymbol{\delta})$ and $\mathbf{z}_{\neg\boldsymbol{\delta}} \sim q_\phi(\mathbf{z}|\mathbf{x})$. Then for any $r \in \mathbb{R}^+$, any $\mathbf{x} \in \mathcal{X}$, and any input perturbation $\boldsymbol{\delta} \in \mathcal{X}$,*

$$\mathbb{P}\left[||g_\theta(\mathbf{z}_{\boldsymbol{\delta}}) - g_\theta(\mathbf{z}_{\neg\boldsymbol{\delta}})||_2 \leq r\right] \geq 1 - \min\left\{p_1(\mathbf{x}), p_2(\mathbf{x})\right\},$$

where

$$p_1(\mathbf{x}) := \min\left(1, \frac{a^2\left(b^2||\boldsymbol{\delta}||_2^2 + (c||\boldsymbol{\delta}||_2 + 2||\sigma_\phi(\mathbf{x})||_2)^2\right)}{r^2}\right)$$

and

$$p_2(\mathbf{x}) := \begin{cases} C(d_z)\frac{u(\mathbf{x})^{\frac{d_z}{2}}\exp\left\{-\frac{u(\mathbf{x})}{2}\right\}}{u(\mathbf{x})-d_z+2} & \left(\frac{r}{a} - b||\boldsymbol{\delta}||_2\right) \geq 0; d_z \geq 2; u(\mathbf{x}) > d_z - 2 \\ 1 & \textit{o.w.} \end{cases}$$

for $u(\mathbf{x}) := \frac{\left(\frac{r}{a} - b||\boldsymbol{\delta}||_2\right)^2}{(c||\boldsymbol{\delta}||_2 + 2||\sigma_\phi(\mathbf{x})||_2)^2}$ and constant $C(d_z) := \frac{1}{\sqrt{\pi}}\exp\left\{\frac{1}{2}(d_z - (d_z-1)\log d_z)\right\}$.

Proof. Since $g_\theta(\cdot)$ is a-Lipschitz,

$$||g_\theta(\mathbf{z}_1) - g_\theta(\mathbf{z}_2)||_2 \leq a||\mathbf{z}_1 - \mathbf{z}_2||_2 \tag{3}$$

for all $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}$.

Now assume $\mathbf{z}_1 \sim q_\phi(\mathbf{z}|\mathbf{x}_1)$ and $\mathbf{z}_2 \sim q_\phi(\mathbf{z}|\mathbf{x}_2)$ for some $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, such that $g_\theta(\mathbf{z}_1)$ and $g_\theta(\mathbf{z}_2)$ are random variables. Eq. (3) then implies

$$\left\{||g_\theta(\mathbf{z}_1) - g_\theta(\mathbf{z}_2)||_2 \leq r\right\} \supseteq \left\{a||\mathbf{z}_1 - \mathbf{z}_2||_2 \leq r\right\},$$

which in turn implies

$$\mathbb{P}\left[||g_\theta(\mathbf{z}_1) - g_\theta(\mathbf{z}_2)||_2 \leq r\right] \geq \mathbb{P}\left[a||\mathbf{z}_1 - \mathbf{z}_2||_2 \leq r\right]. \tag{4}$$

Letting $\mathbf{x}_1 = \mathbf{x} + \boldsymbol{\delta}$ and $\mathbf{x}_2 = \mathbf{x}$ such that $\mathbf{z}_1 = \mathbf{z}_{\boldsymbol{\delta}}$ and $\mathbf{z}_2 = \mathbf{z}_{\neg\boldsymbol{\delta}}$, $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}; \mu_\phi(\mathbf{x}), \mathtt{diag}\left(\sigma^2_\phi(\mathbf{x})\right)\right)$ means

$$\mathbf{z}_{\boldsymbol{\delta}} \sim q_\phi(\mathbf{z}|\mathbf{x}+\boldsymbol{\delta}) = \mathcal{N}\left(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}), \mathtt{diag}\left(\sigma^2_\phi(\mathbf{x}+\boldsymbol{\delta})\right)\right)$$

and

$$\mathbf{z}_{\neg\boldsymbol{\delta}} \sim q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mu_\phi(\mathbf{x}), \mathtt{diag}\left(\sigma^2_\phi(\mathbf{x})\right)\right).$$

Further, since samples from $q_\phi(\mathbf{z}|\cdot)$ are drawn independently in every VAE forward pass, we also know $\mathbf{z}_{\boldsymbol{\delta}}$ and $\mathbf{z}_{\neg\boldsymbol{\delta}}$ are independent, and thus, because the difference of independent multivariate Gaussian random variables is multivariate Gaussian,

$$\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}} \sim \mathcal{N}\left(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x}), \mathtt{diag}\left(\sigma^2_\phi(\mathbf{x}+\boldsymbol{\delta})\right) + \mathtt{diag}\left(\sigma^2_\phi(\mathbf{x})\right)\right).$$

Returning to (4), since $||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}||_2$ is a continuous random variable, we can write

$$\mathbb{P}\left[||g_\theta(\mathbf{z}_{\boldsymbol{\delta}}) - g_\theta(\mathbf{z}_{\neg\boldsymbol{\delta}})||_2 \leq r\right] \geq \mathbb{P}\left[||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}||_2 \leq \frac{r}{a}\right] = 1 - \mathbb{P}\left[||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}||_2 \geq \frac{r}{a}\right]. \tag{5}$$

The proof now diverges, yielding $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ respectively.

Obtaining $p_1(\mathbf{x})$: Recall $\mathcal{Z} = \mathbb{R}^{d_z}$, apply the definition of the ℓ_2 norm, and invoke Markov's Inequality to obtain

$$\mathbb{P}\left[||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}||_2 \geq \frac{r}{a}\right] = \mathbb{P}\left[\sum_{j=1}^{d_z}\left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}\right)_j^2 \geq \left(\frac{r}{a}\right)^2\right] \leq \frac{\mathbb{E}\left[\sum_{j=1}^{d_z}\left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}\right)_j^2\right]}{\left(\frac{r}{a}\right)^2}. \tag{6}$$

Now note that

$$\sum_{j=1}^{d_z} \left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}\right)_j^2 = \sum_{j=1}^{d_z} \left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j \frac{\left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}\right)_j^2}{\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j},$$

so that by the linearity of expectations,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{d_z} \left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}\right)_j^2\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{d_z} \left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j \frac{\left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}\right)_j^2}{\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j}\right] \\
&= \sum_{j=1}^{d_z} \left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j \mathbb{E}\left[\frac{\left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}\right)_j^2}{\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j}\right]. \qquad (7)
\end{aligned}$$

Because $\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}$ is diagonal-covariance multivariate Gaussian, the $\left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}\right)_j$ are jointly independent for all $j = 1, \ldots, d_z$, and so we recognize that

$$\frac{\left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}\right)_j^2}{\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j}$$

has a non-central χ^2 distribution with one degree of freedom and non-centrality parameter

$$\frac{\left(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})\right)_j^2}{\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j}.$$

Since for a non-central χ^2 random variable Y with n degrees of freedom and non-centrality parameter ϵ [24], $\mathbb{E}[Y] = n + \epsilon$, we have

$$\mathbb{E}\left[\frac{\left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}\right)_j^2}{\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j}\right] = 1 + \frac{\left(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})\right)_j^2}{\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j},$$

and so plugging into (7),

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{d_z} \left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}\right)_j^2\right] \\
&= \sum_{j=1}^{d_z} \left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j \left(1 + \frac{\left(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})\right)_j^2}{\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j}\right) \\
&= \sum_{j=1}^{d_z} \left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j + \sum_{j=1}^{d_z} \left(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})\right)_j^2.
\end{aligned}$$

Using

$$\sum_{j=1}^{d_z} \left(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})\right)_j^2 = ||\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})||_2^2$$

(the definition of the ℓ_2 norm), and

$$||\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})||_2 \leq b||\boldsymbol{\delta}||_2,$$

(since $\mu_\phi(\cdot)$ is b-Lipschitz), we obtain

$$\sum_{j=1}^{d_z} \left(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})\right)_j^2 = ||\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})||_2^2 \leq \left(b||\boldsymbol{\delta}||_2\right)^2 = b^2||\boldsymbol{\delta}||_2^2. \tag{8}$$

Similarly, using

$$\begin{aligned}
&\sum_{j=1}^{d_z} \left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j &(9)\\
&\leq \sum_{j=1}^{d_z} \sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta})_j + \sigma_\phi^2(\mathbf{x})_j + 2\sigma_\phi(\mathbf{x}+\boldsymbol{\delta})_j \sigma_\phi(\mathbf{x})_j &(10)\\
&= \sum_{j=1}^{d_z} \left(\sigma_\phi(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi(\mathbf{x})\right)_j^2 &(11)\\
&= \left(\sqrt{\sum_{j=1}^{d_z} \left(\sigma_\phi(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi(\mathbf{x})\right)_j^2}\right)^2 &(12)\\
&= ||\sigma_\phi(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi(\mathbf{x})||_2^2 &(13)
\end{aligned}$$

(where the above inequality follows from $\sigma_\phi : \mathcal{X} \to \mathbb{R}_{\geq 0}^{d_z}$, and the last equality follows from the definition of the ℓ_2 norm), and

$$\begin{aligned}
&||\sigma_\phi(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi(\mathbf{x})||_2\\
&= ||\sigma_\phi(\mathbf{x}+\boldsymbol{\delta}) - \sigma_\phi(\mathbf{x}) + 2\sigma_\phi(\mathbf{x})||_2\\
&\leq ||\sigma_\phi(\mathbf{x}+\boldsymbol{\delta}) - \sigma_\phi(\mathbf{x})||_2 + 2||\sigma_\phi(\mathbf{x})||_2\\
&\leq c||\boldsymbol{\delta}||_2 + 2||\sigma_\phi(\mathbf{x})||_2
\end{aligned}$$

(where the first inequality follows by the triangle inequality, and the second follows from the assumption that $\sigma_\phi(\cdot)$ is c-Lipschitz), we find

$$\sum_{j=1}^{d_z} \left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j \leq ||\sigma_\phi(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi(\mathbf{x})||_2^2 \leq \left(c||\boldsymbol{\delta}||_2 + 2||\sigma_\phi(\mathbf{x})||_2\right)^2. \tag{14}$$

Hence, returning to (6), we see

$$\begin{aligned}
&\frac{\mathbb{E}\left[\sum_{j=1}^{d_z} \left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}\right)_j^2\right]}{\left(\frac{r}{a}\right)^2}\\
&= \frac{\sum_{j=1}^{d_z} \left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j + \sum_{j=1}^{d_z} \left(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})\right)_j^2}{\left(\frac{r}{a}\right)^2}\\
&\leq \frac{b^2||\boldsymbol{\delta}||_2^2 + \left(c||\boldsymbol{\delta}||_2 + 2||\sigma_\phi(\mathbf{x})||_2\right)^2}{\left(\frac{r}{a}\right)^2}\\
&= \frac{a^2\left(b^2||\boldsymbol{\delta}||_2^2 + (c||\boldsymbol{\delta}||_2 + 2||\sigma_\phi(\mathbf{x})||_2)^2\right)}{r^2},
\end{aligned}$$

such that

$$\mathbb{P}\left[||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}||_2 \geq \frac{r}{a}\right] \leq \frac{\mathbb{E}\left[\sum_{j=1}^{d_z} \left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}\right)_j^2\right]}{\left(\frac{r}{a}\right)^2} \leq \frac{a^2\left(b^2||\boldsymbol{\delta}||_2^2 + (c||\boldsymbol{\delta}||_2 + 2||\sigma_\phi(\mathbf{x})||_2)^2\right)}{r^2}.$$

Noting that the right-most term is non-negative, and wanting to have a well-defined probability, we take

$$p_1(\mathbf{x}) := \min\left(1, \frac{a^2\left(b^2||\boldsymbol{\delta}||_2^2 + (c||\boldsymbol{\delta}||_2 + 2||\sigma_\phi(\mathbf{x})||_2)^2\right)}{r^2}\right),$$

such that

$$\mathbb{P}\left[||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}||_2 \geq \frac{r}{a}\right] \leq p_1(\mathbf{x}).$$

Obtaining $p_2(\mathbf{x})$**:** Return to Eq. (5). By the triangle inequality,

$$||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}||_2 \leq ||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}} - (\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x}))\,||_2 + ||\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})||_2,$$

and hence

$$\mathbb{P}\left[||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}||_2 \geq \frac{r}{a}\right] \tag{15}$$

$$\leq \mathbb{P}\left[(||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}} - (\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x}))\,||_2 + ||\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})||_2) \geq \frac{r}{a}\right] \tag{16}$$

$$= \mathbb{P}\left[||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}} - (\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x}))\,||_2 \geq \left(\frac{r}{a} - ||\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})||_2\right)\right]. \tag{17}$$

Then, again recalling $\mathcal{Z} = \mathbb{R}^{d_z}$,

$$\mathbb{P}\left[||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}} - (\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x}))\,||_2 \geq \left(\frac{r}{a} - ||\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})||_2\right)\right] \tag{18}$$

$$= \mathbb{P}\left[\sum_{j=1}^{d_z} \left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}} - (\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x}))\right)_j^2 \geq \left(\frac{r}{a} - ||\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})||_2\right)^2\right]$$

$$\leq \mathbb{P}\left[\sum_{j=1}^{d_z} \frac{\left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}} - (\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x}))\right)_j^2}{\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j} \geq \frac{\left(\frac{r}{a} - ||\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})||_2\right)^2}{\left(c||\boldsymbol{\delta}||_2 + 2||\sigma_\phi(\mathbf{x})||_2\right)^2}\right], \tag{19}$$

where the first equality uses the definition of the ℓ_2 norm, and the above inequality between probabilities uses the inequality from (14).

Now, since

$$\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}} \sim \mathcal{N}\left(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x}), \mathtt{diag}\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta})\right) + \mathtt{diag}\left(\sigma_\phi^2(\mathbf{x})\right)\right), \tag{20}$$

it follows that

$$\frac{\left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}} - (\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x}))\right)_j}{\sqrt{\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j}} \sim \mathcal{N}(0,1).$$

In particular, note that since $\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}$ is diagonal-covariance multivariate Gaussian, the

$$\frac{\left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}} - (\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x}))\right)_j}{\sqrt{\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j}}$$

are jointly independent for all $j = 1, \ldots, d_z$. Hence, because the sum of squares of d_z independent standard Gaussian random variables has a standard χ^2 distribution with d_z degrees of freedom,

$$\sum_{j=1}^{d_z} \frac{\left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}} - (\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x}))\right)_j^2}{\left(\sigma_\phi^2(\mathbf{x}+\boldsymbol{\delta}) + \sigma_\phi^2(\mathbf{x})\right)_j} =: Y \sim \chi_{d_z}^2.$$

Letting

$$u'(\mathbf{x}) := \frac{\left(\frac{r}{a} - ||\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})||_2\right)^2}{\left(c||\boldsymbol{\delta}||_2 + 2||\sigma_\phi(\mathbf{x})||_2\right)^2} \quad \text{and} \quad u(\mathbf{x}) := \frac{\left(\frac{r}{a} - b||\boldsymbol{\delta}||_2\right)^2}{\left(c||\boldsymbol{\delta}||_2 + 2||\sigma_\phi(\mathbf{x})||_2\right)^2},$$

we have $u'(\mathbf{x}) \geq u(\mathbf{x})$ by the assumption that $\mu_\phi(\cdot)$ is b-Lipschitz, since

$$||\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})||_2 \leq b||\boldsymbol{\delta}||_2,$$

and therefore

$$\left(\frac{r}{a} - ||\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})||_2\right) \geq \left(\frac{r}{a} - b||\boldsymbol{\delta}||_2\right)$$

(note also that $(c||\boldsymbol{\delta}||_2 + 2||\sigma_\phi(\mathbf{x})||_2)^2 \geq 0$). Then, using (19) with the requirement that

$$\left(\frac{r}{a} - ||\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})||_2\right) \geq \left(\frac{r}{a} - b||\boldsymbol{\delta}||_2\right) \geq 0$$

to ensure the inequality in (18) is meaningful,

$$\mathbb{P}\left[Y \geq u'(\mathbf{x})\right] \leq \mathbb{P}\left[Y \geq u(\mathbf{x})\right].$$

The tail bound for standard χ^2 random variables in (3.1) from [25] (which requires $u(\mathbf{x}) > d_z - 2$ and $d_z \geq 2$) then yields

$$\mathbb{P}\left[Y \geq u(\mathbf{x})\right] \leq C(d_z)\frac{u(\mathbf{x})^{\frac{d_z}{2}} \exp\left\{-\frac{u(\mathbf{x})}{2}\right\}}{u(\mathbf{x}) - d_z + 2}$$

for constant $C(d_z) := \frac{1}{\sqrt{\pi}} \exp\left\{\frac{1}{2}(d_z - (d_z - 1)\log d_z)\right\}$. Since the expression on the right-hand side is non-negative under the above conditions, we define

$$p_2(\mathbf{x}) := \begin{cases} C(d_z)\frac{u(\mathbf{x})^{\frac{d_z}{2}} \exp\left\{-\frac{u(\mathbf{x})}{2}\right\}}{u(\mathbf{x}) - d_z + 2} & \left(\frac{r}{a} - b||\boldsymbol{\delta}||_2\right) \geq 0; d_z \geq 2; u(\mathbf{x}) > d_z - 2 \\ 1 & \text{o.w.} \end{cases}$$

to ensure a well-defined probability. Then, by the inequalities starting from (15),

$$\mathbb{P}\left[||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}||_2 \geq \frac{r}{a}\right] \leq p_2(\mathbf{x}).$$

Obtaining the final bound: Choosing the least of $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ to obtain the tighter upper bound on $\mathbb{P}\left[||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}||_2 \geq \frac{r}{a}\right]$, we can plug in to (5), which gives

$$\begin{aligned} &\mathbb{P}\left[||g_\theta(\mathbf{z}_{\boldsymbol{\delta}}) - g_\theta(\mathbf{z}_{\neg\boldsymbol{\delta}})||_2 \leq r\right] \\ &\geq 1 - \mathbb{P}\left[||\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}||_2 \geq \frac{r}{a}\right] \\ &\geq 1 - \min\{p_1(\mathbf{x}), p_2(\mathbf{x})\}. \end{aligned}$$

■

Lemma 1.1 (Margin Bound)**.** *Given the assumptions of Theorem 1 and some $\epsilon \in [0, 1)$, the (r, ϵ)-robustness margin of this Lipschitz-VAE on input $\mathbf{x}$,*

$$R^{(r,\epsilon)}(\mathbf{x}) \geq \max\left\{m_1(\mathbf{x}), m_2(\mathbf{x})\right\}$$

for

$$m_1(\mathbf{x}) := \frac{-4c||\sigma_\phi(\mathbf{x})||_2 + \sqrt{\left(4c||\sigma_\phi(\mathbf{x})||_2\right)^2 - 4\left(c^2 + b^2\right)\left(4||\sigma_\phi(\mathbf{x})||_2 - (1-\epsilon)\left(\frac{r}{a}\right)^2\right)}}{2\left(c^2 + b^2\right)}$$

and $m_2(\mathbf{x}) := \sup\left\{||\boldsymbol{\delta}||_2 : p_2(\boldsymbol{\delta}, \mathbf{x}) \leq (1-\epsilon)\right\}$, where $p_2(\boldsymbol{\delta}, \mathbf{x})$ is as in Theorem 1.

Proof. By Theorem 1, for any input perturbation $\boldsymbol{\delta} \in \mathcal{X}$ and any input $\mathbf{x} \in \mathcal{X}$,

$$\mathbb{P}\left[||g_\theta(\mathbf{z}_{\boldsymbol{\delta}}) - g_\theta(\mathbf{z}_{\neg\boldsymbol{\delta}})||_2 \leq r\right] \geq 1 - \min\{p_1(\mathbf{x}), p_2(\mathbf{x})\}.$$

Hence, for our Lipschitz-VAE to be (r, ϵ)-robust to perturbation $\boldsymbol{\delta}$ on input $\mathbf{x}$ for threshold $\epsilon \in [0, 1)$, by Definition 2.1 it suffices that

$$1 - \min\{p_1(\mathbf{x}), p_2(\mathbf{x})\} > \epsilon.$$

Recalling Definition 2.2, since for a model f, $R^{(r,\epsilon)}(\mathbf{x})$ is defined by

$$||\boldsymbol{\delta}||_2 < R^{(r,\epsilon)}(\mathbf{x}) \implies \mathbb{P}\left[||f(\mathbf{x}+\boldsymbol{\delta}) - f(\mathbf{x})||_2 \leq r\right] > \epsilon,$$

for our Lipschitz-VAE $R^{(r,\epsilon)}(\mathbf{x})$ is at least the maximum perturbation norm such that

$$1 - \min\{p_1(\boldsymbol{\delta}, \mathbf{x}), p_2(\boldsymbol{\delta}, \mathbf{x})\} \geq \epsilon,$$

or equivalently,

$$\max\left\{\sup\left\{||\boldsymbol{\delta}||_2 : p_1(\boldsymbol{\delta}, \mathbf{x}) \leq (1-\epsilon)\right\},\ \sup\left\{||\boldsymbol{\delta}||_2 : p_2(\boldsymbol{\delta}, \mathbf{x}) \leq (1-\epsilon)\right\}\right\} \tag{21}$$

(where we make explicit the dependence on $\boldsymbol{\delta}$).

Denoting $m_1(\mathbf{x}) := \sup\left\{||\boldsymbol{\delta}||_2 : p_1(\boldsymbol{\delta}, \mathbf{x}) \leq (1-\epsilon)\right\}$ and rearranging, $m_1(\mathbf{x})$ becomes

$$\sup\left\{||\boldsymbol{\delta}||_2 : \left(c^2 + b^2\right)||\boldsymbol{\delta}||_2^2 + 4c||\sigma_\phi(\mathbf{x})||_2||\boldsymbol{\delta}||_2 + 4||\sigma_\phi(\mathbf{x})||_2^2 - (1-\epsilon)\left(\frac{r}{a}\right)^2 \leq 0\right\}.$$

Excluding the degenerate case of $c = 0$, that is assuming $c > 0$, this is attained at the maximum root of the quadratic equation

$$\left(c^2 + b^2\right)||\boldsymbol{\delta}||_2^2 + 4c||\sigma_\phi(\mathbf{x})||_2||\boldsymbol{\delta}||_2 + 4||\sigma_\phi(\mathbf{x})||_2^2 - (1-\epsilon)\left(\frac{r}{a}\right)^2 = 0,$$

provided a root exists, and so by the quadratic formula,

$$m_1(\mathbf{x}) = \frac{-4c||\sigma_\phi(\mathbf{x})||_2 + \sqrt{\left(4c||\sigma_\phi(\mathbf{x})||_2\right)^2 - 4\left(c^2+b^2\right)\left(4||\sigma_\phi(\mathbf{x})||_2 - (1-\epsilon)\left(\frac{r}{a}\right)^2\right)}}{2\left(c^2+b^2\right)}.$$

The second case does not admit a closed-form solution, so we will simply write

$$m_2(\mathbf{x}) := \sup\left\{||\boldsymbol{\delta}||_2 : p_2(\boldsymbol{\delta}, \mathbf{x}) \leq (1-\epsilon)\right\}.$$

Choosing the maximum of $m_1(\mathbf{x})$ and $m_2(\mathbf{x})$ then yields

$$R^{(r,\epsilon)}(\mathbf{x}) \geq \max\left\{m_1(\mathbf{x}), m_2(\mathbf{x})\right\}.$$

■

Theorem 2 (Global Margin Bound)**.** *Given the assumptions of Lemma 1.1, but with $\sigma_\phi(\mathbf{x}) = \boldsymbol{\sigma} \in \mathbb{R}^{d_z}$, the (r, ϵ)-robustness margin of this VAE for all inputs is*

$$R^{(r,\epsilon)} \geq \max\left\{m_1, m_2\right\},$$

where

$$m_1 := \frac{\sqrt{-\left(4||\boldsymbol{\sigma}||_2^2 - (1-\epsilon)\left(\frac{r}{a}\right)^2\right)}}{b}$$

and $m_2 := \sup\left\{||\boldsymbol{\delta}||_2 : p_2(\boldsymbol{\delta}) \leq (1-\epsilon)\right\}$, where p_2 is as in Theorem 1, but $u := \frac{\left(\frac{r}{a} - b||\boldsymbol{\delta}||_2\right)^2}{4||\boldsymbol{\sigma}||_2^2}$.

Proof. Given a fixed encoder standard deviation, that is substituting $\sigma_\phi(\mathbf{x}) = \boldsymbol{\sigma} \in \mathbb{R}^{d_z}$, we first have to derive a lower bound on the r-robustness probability to then bound the (r, ϵ)-robustness margin globally. We do this using the machinery of Theorem 1, which — lifting the now-redundant requirement that the encoder standard deviation be c-Lipschitz — can be invoked without loss of generality.

In the case of p_1 (recall the two bounds in the proof of Theorem 1), plugging in $\boldsymbol{\sigma}$ yields

$$\begin{aligned}
\mathbb{P}\left[||g_\theta(\mathbf{z}_{\boldsymbol{\delta}}) - g_\theta(\mathbf{z}_{\neg\boldsymbol{\delta}})||_2 \leq r\right] &\geq 1 - \frac{\mathbb{E}\left[\sum_{j=1}^{d_z}\left(\mathbf{z}_{\boldsymbol{\delta}} - \mathbf{z}_{\neg\boldsymbol{\delta}}\right)_j^2\right]}{\left(\frac{r}{a}\right)^2} \\
&= 1 - \frac{\sum_{j=1}^{d_z}\left(\boldsymbol{\sigma}^2 + \boldsymbol{\sigma}^2\right)_j + \sum_{j=1}^{d_z}\left(\mu_\phi(\mathbf{x}+\boldsymbol{\delta}) - \mu_\phi(\mathbf{x})\right)_j^2}{\left(\frac{r}{a}\right)^2} \\
&\geq 1 - \frac{b^2||\boldsymbol{\delta}||_2^2 + 4||\boldsymbol{\sigma}||_2^2}{\left(\frac{r}{a}\right)^2} \\
&= 1 - p_1
\end{aligned}$$

for $p_1 := \frac{a^2\left(b^2||\boldsymbol{\delta}||_2^2+4||\boldsymbol{\sigma}||_2^2\right)}{r^2}$, where the penultimate step follows by (8) and (14). In the case of p_2, we can directly substitute, obtaining

$$\mathbb{P}\left[||g_\theta(\mathbf{z}_{\boldsymbol{\delta}}) - g_\theta(\mathbf{z}_{\neg\boldsymbol{\delta}})||_2 \leq r\right] \geq 1 - p_2$$

for

$$p_2 := \begin{cases} C(d_z)\frac{u^{\frac{d_z}{2}}\exp\left\{-\frac{u}{2}\right\}}{u-d_z+2} & \left(\frac{r}{a} - b||\boldsymbol{\delta}||_2\right) \geq 0; d_z \geq 2; u > d_z - 2 \\ 1 & \text{o.w.} \end{cases}$$

and $u := \frac{\left(\frac{r}{a}-b||\boldsymbol{\delta}||_2\right)^2}{4||\boldsymbol{\sigma}||_2^2}$. Theorem 2 then follows by identical reasoning to Lemma 1.1. ■

B Implementing Lipschitz-VAEs

Previously, we have taken the Lipschitz constants of a VAE's encoder and decoder as given. In practice, ensuring the Lipschitz continuity of a deep learning architecture is non-trivial. Using [22] as a guide, we outline how to provably control the Lipschitz constants of an encoder and decoder network.[4]

We define a fully-connected network with L layers as the composition of linear transformations $\mathbf{W}_l$ and element-wise activation functions $\varphi_l(\cdot)$ for $l = 1, \dots, L$, where the output of the l-th layer

$$\mathbf{h}_l := \varphi_l(\mathbf{W}_l \mathbf{h}_{l-1}).$$

We let network input $\mathbf{x} =: \mathbf{h}_0$ and network output $\mathbf{y} := \mathbf{h}_L$.

Ensuring Lipschitz Continuity with Constant 1 We would like to ensure a fully-connected network is M-Lipschitz for arbitrary constant M. It has been shown that a natural way to achieve this is by first requiring Lipschitz continuity with constant 1 [22].

As 1-Lipschitz functions are closed under composition, if we can ensure that for every layer l, $\mathbf{W}_l$ and $\varphi_l(\cdot)$ are 1-Lipschitz, then the entire network will be 1-Lipschitz. Most commonly-used activation functions, such as the ReLU and Sigmoid, are already 1-Lipschitz [26, 27], and hence we need only ensure that $\mathbf{W}_l$ is also 1-Lipschitz.

This can be done by requiring $\mathbf{W}_l$ to be orthonormal, since $\mathbf{W}_l$ being 1-Lipschitz is equivalent to the condition

$$||\mathbf{W}_l||_2 := \sup_{||\mathbf{x}||_2 \leq 1} ||\mathbf{W}_l \mathbf{x}||_2 \leq 1, \tag{22}$$

where $||\mathbf{W}_l||_2$ equals the largest singular value of $\mathbf{W}_l$. The singular values of an orthonormal matrix all equal 1, and so the orthonormality of $\mathbf{W}_l$ implies (22) is satisfied.

In practice, $\mathbf{W}_l$ can be made orthonormal through an iterative algorithm called *Björck Orthonormalization*, which on input a matrix $\mathbf{A}$ finds the "nearest" orthonormal matrix to $\mathbf{A}$ [22]. Björck Orthonormalization is differentiable and so allows the encoder and decoder networks of a Lipschitz-VAE to be trained using gradient-based methods, just like a standard VAE.

Ensuring Lipschitz Continuity with Arbitrary Constants Now that we can train a 1-Lipschitz network, we would like to generalize this method to arbitrary Lipschitz constant M. To do so, note that if layer l has Lipschitz constant M_l, then the Lipschitz constant of the entire network is $M = \prod_{l=1}^{L} M_l$ [3].

Hence, for our L-layer fully-connected neural network to be M-Lipschitz, it suffices to ensure that each layer l has Lipschitz constant $M^{\frac{1}{L}}$. This is actually simple to achieve, because if we continue to assume $\varphi_l(\cdot)$ is 1-Lipschitz, Lipschitz constant $M^{\frac{1}{L}}$ in layer l follows from scaling the outputs of each layer's linear transformation by $M^{\frac{1}{L}}$.

Selecting Activation Functions While the above approach is sufficient to train networks with arbitrary Lipschitz constants, a result from [22] shows it is not sufficient to ensure the resulting networks are also expressive in the space of Lipschitz continuous functions. Informally, the result states that the expressivity of a Lipschitz-constrained network is limited when its activation functions are not gradient norm-preserving. Since activation functions such as the ReLU and the Sigmoid do not preserve the gradient norm, the expressivity of Lipschitz-constrained networks that use such activations will be limited.

To address this, [22] introduces a gradient norm-preserving activation function called *GroupSort*, which in each layer l groups the entries of matrix-vector product $\mathbf{W}_l \mathbf{h}_{l-1}$ into some number of groups, and then sorts the entries of each group by ascending order. It can be shown that when each group has size two,

$$(1 \quad 0)\,\text{GroupSort}\left(\begin{pmatrix} y \\ 0 \end{pmatrix}\right) = \text{ReLU}(y)$$

for any scalar y [22].

[4] For simplicity, we focus on fully-connected architectures, although the same ideas extend, for example, to convolutional architectures [23].