On the Geometry of the Latent Space of Deep Generative Models

Naren Manoj
Department of Computer Science
The University of Texas at Austin
Austin, TX 78705, USA
manoj.narens@utexas.edu

August 22, 2018

Abstract

In this paper, we identify a geometric property of the latent space of Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). We call this property the *class radius* and show that the class radius can be upper-bounded by various diversity measures of the samples from the generator. Therefore, a method to compute the class radius results in a one-sided test to determine the diversity of the generative model.

1 Introduction

Deep generative models are powerful methods used to represent a distribution over natural images. Two popular instances of deep generative models are Variational Autoencoders (VAEs) (Kingma & Welling (2013)) and Generative Adversarial Networks (GANs) (Goodfellow et al. (2014)). However, theoretical understanding of these models is still lacking. In particular, it is unclear how some properties of VAEs and GANs contribute to or detract from diversity and sample quality.

One important quality of deep generative models is that they generate their samples as follows. A vector of some fixed dimensionality, known as a latent vector, is sampled from a space known as a latent space according to some probability distribution over the latent space. This vector is then pushed through a neural network, and an image is outputted. The goal of generative models is to model the data distribution - that is, match the distribution of output images to some desired distribution over natural images. Since there is both significant structure induced by the parameterization of the neural network generators and geometry induced by this sampling procedure, there are interesting properties relating to each that may lend insights into how well deep generative models might be modeling their target distributions.

In this paper, we claim that one geometric property of the latent space of a deep generative model, which we term the *class radius*, can be used to bound a function of the diversity of the generated samples. In particular, **knowing the class radius yields an upper bound on the diversity of the samples as measured by collision probability**. Our analysis uses results obtained by and closely parallels the methods used by Fawzi et al. (2018). We then suggest various ways to exploit this theoretical observation practically. Specifically, a method to compute the class radius would result in a one-sided test for the diversity of a generative model as measured by collision probability.

The rest of this paper is organized as follows. In Section 2, we define important terminology. In Section 3, we state and discuss our theoretical claims relating the class radius to the

diversity of a generative model. Finally, in Section 4, we contextualize our work with respect to other efforts in related areas and suggest problems of further interest.

2 Notation and Terms

In this section, we formally define and explain the terms we use throughout the remainder of this paper.

Definition 2.1 (Deep Generative Model)**.** *A* deep generative model *is a (possibly randomized) function, denoted g, that accepts some vector $z \in \mathbb{R}^d$ for some d as input and outputs some image $I \in \mathcal{I}$, where $\mathcal{I}$ represents the space of images.*

Definition 2.2 (Encoder)**.** *An* encoder *is a (possibly randomized) function, denoted f, that accepts some input image $I \in \mathcal{I}$ (where, again $\mathcal{I}$ is the space of images) and outputs some vector $z \in \mathbb{R}^d$.*

State of the art implementations of deep generative models and encoders include Variational Autoencoders[1], Generative Adversarial Networks (GANs)[2][3], and their respective variants. Throughout the remainder of this paper, we use the terms *latent vector* and *coding vector* interchangeably with the input vector for a model g. We call the input space of g and the output space of f the *latent space* or the *coding space*.

Definition 2.3 (Class)**.** *Suppose we are given a classifier into K classes $c : \mathcal{I} \to [K]$ that accepts as input an image and outputs a class label which is an integer index in $[K] \coloneqq \{i\}_{i=1}^{K}$. Then, the* class *of an image I under c is given by:*

$$K_{c(I)} \coloneqq \{z' \mid c\left(g\left(z'\right)\right) = c(I)\}$$

In words, K_i is the subset of the latent space that is classified into label i when a point $z \in K_i$ is passed through g and subsequently classified by c.

Observe that our definition of class is very general. Notably, a class could be the set of latent vectors mapping to a particular classification of an image under some classifier (for instance, K_i for some i could be all the latent vectors mapping to images of dogs). Similarly, a class could be the set of latent vectors mapping to every image at most ϵ away from some reference image in ℓ_p distance.

Furthermore, notice that K_i need not be connected even for simple functions g and classifiers c.

Example 2.1. *Suppose z is distributed according to a standard univariate Gaussian, $g(z) = z^2$, and $c(x) = 1 + \mathbb{1}_{x \geq 1}$. Thus, we have:*

$$K_1 = (-1, 1)$$
$$K_2 = (-\infty, -1] \cup [1, \infty)$$

Note that K_2 is not connected.

In Example 2.2, we discuss a slightly more interesting scenario to help clarify our terminology.

Example 2.2. *Consider a GAN $g : \mathbb{R}^2 \to [-5, 5]^2$. The GAN intends to represent an instance of the* 2D-Grid *distribution, which is a mixture of 25 2D spherical Gaussians with means $(2i, 2j), -2 \leq i, j \leq 2$ and variances 0.0025.*

Next, define the classification function c below:

$$c(p) = \underset{-2 \leq i,j \leq 2}{\arg\min} \|(2i, 2j) - p\|_2$$

Thus, $c(p)$ is sending p to its nearest lattice point with only even coordinates.

[1] Kingma & Welling (2013)

[2] Goodfellow et al. (2014)

[3] Though a GAN does not come with an encoder by default, numerous GAN flavors involve training an encoder.

Now, see Figure 2.1. The left image depicts classes. Each distinctly colored region represents points in the same class. The right image shows the GAN's representation of the target distribution. The GAN's distribution is shown with the orange dots, while the target distribution is shown by the blue dots.

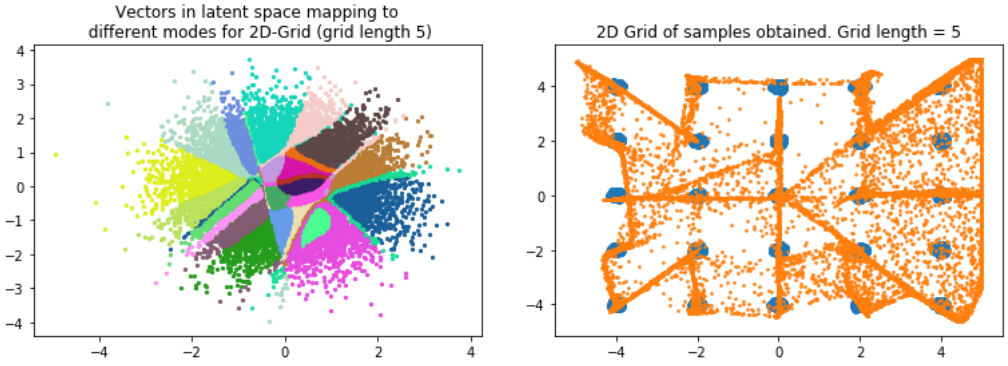

Figure 2.1: Left image: Points sharing the same color are in the same class under the classification function c and GAN g defined in Example 2.2. Right image: Mapping of points in the space depicted in the left image under the same GAN g. Each blue dot represents a distinct class, and each orange dot represents a sample obtained by applying g to a vector sampled from the left image. Thus, the classification function c sends each orange dot to its closest blue dot.

The central focus of this paper is addressing the *class radius* of a generative model with respect to some classifier. Roughly speaking, the class radius answers the following question - on average, for a randomly selected point z in the latent space, how far away is the nearest point in a different class?

To formalize this idea, we introduce the following definition.

Definition 2.4 (Class Radius)**.** *Consider the following distribution Q over $\mathbb{R}^d$ with probability density function q where $\nu(A)$ denotes the Gaussian measure of set A:*

$$q(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{x^T x/2} \cdot \frac{1}{\sum \nu(K_i)} & x \in \bigcup_{i=1}^{K} K_i \\ 0 & \textit{otherwise} \end{cases}$$

Thus, Q is the Gaussian distribution over $\mathbb{R}^d$ conditioned on x belonging to a class.

Then, the class radius *is the expected distance from a random point $z \sim Q$ in some class to the closest point in a different class. Specifically, define $r(z)$ below, and let $c(g(z)) = i$:*

$$r(z) := \inf_{z' \in \bigcup_{j \neq i} K_j} \|z - z'\|$$

Then the class radius is:

$$\mathop{\mathbb{E}}_{z \sim Q}[r(z)]$$

Our central claim is that the class radius is upper bounded by a decreasing function of the diversity of the generative model. Thus, if we know the class radius, we can then claim that the diversity cannot exceed a certain measurement for various quantifications of diversity. As a result, a method to compute the class radius results in a one-sided test that lower bounds the probability that two samples from the model fall into the same class.

3 Theory: A High Class Radius is Bad for Diversity

In this section, we state a bound directly relating the *collision entropy* of the generated distribution to the class radius. Thus, **the class radius can be bounded by a function of**

the diversity of the generative model. Under a reasonable probability decay assumption (in particular, the maximum probability of any class is a decreasing function in K, the number of classes), we show that our upper bound is a decreasing function in the number of classes K. Thus, by a simple squeeze argument, we have that the class radius decays to 0 as the number of classes grows arbitrarily large. Additionally, we show that our bound is smallest when the generator generates a uniform distribution across its K modes. This may imply that a generative model has learned a potentially more useful distribution; we provide some basic experiments to back up this idea.

The rest of this section is organized as follows. We first formally state and define the problem setting under which we work. Next, we state our central claims. Finally, we discuss some definite and potential practical implications of our results.

Notation and Problem Setting Now, suppose that we are given access to a *deterministic* generative model g that accepts as input some vector $z \sim \mathcal{N}\left(0, \text{diag}\left(\vec{1}_d\right)\right)$ and outputs some image $I \in \mathcal{I}$, where d is the dimension of the latent space and $\mathcal{I}$ is the space of images. Furthermore, suppose we have a classification function c with K possible outputs. Let the probability that an image of class K_i is generated be p_i, and let the vector of these probabilities be p. Let $\varphi(x)$ denote the standard univariate Gaussian probability density function (PDF), $\Phi(x)$ denote the standard univariate Gaussian cumulative distribution function (CDF), and $\Phi^{-1}(x)$ denote the inverse of $\Phi(x)$.

Assumptions For the purpose of our proofs, we enforce the following reasonable assumption on the generated distribution:

$$\|p\|_\infty \leq \min\left(\frac{1}{5}, h(K)\right) \text{ such that } h(K) = \Omega\left(\frac{1}{K}\right) \text{ and } h(K) = o(1)$$

Notice that this just means that the maximum probability of any class does not remain constant; instead, the maximum probability decays as the number of classes K increases. One way to view this constraint is to observe that we are enforcing some kind of balance on the classes. By limiting the maximum probability of any class, we are ensuring that no class becomes too asymptotically dominant.

Furthermore, let d be the dimension of the latent space. Then, we assume:

$$\bigcup_{i=1}^{K} K_i = \mathbb{R}^d$$

This assumption essentially means that every vector in the latent space gets mapped to some image in the true distribution.

Now, we state our central theoretical result. Informally, we can upper bound the class radius based on the distribution induced by our generative model.

Theorem 3.1. *If $0 < p_i \leq 1/5, \forall i \in [K]$, then the following holds regarding the expected class radius of each image in the latent space of our generative model:*

$$\mathbb{E}\left[r(z)\right] \leq \frac{\log\left(4\pi \log\left(1/\left\|p\right\|_2^2\right)\right)}{\sqrt{2\log\left(1/\left\|p\right\|_2^2\right)}}$$

Proof. We defer a full proof to the appendices; see the proof of Theorem B.1. □

Observe that $\|p\|_2^2$ is exactly the following:

$$\mathop{\mathbb{P}}_{z_1, z_2 \sim \mathcal{N}\left(0, \text{diag}\left(\vec{1}_d\right)\right)}\left[f\left(g\left(z_1\right)\right) = f\left(g\left(z_2\right)\right)\right]$$

Furthermore, recall that $-\log\left(\|p\|_2^2\right)$ is the second-order Rényi entropy, which is also known as the collision entropy[4]. Thus, our upper bound on the class radius can also be viewed as a measure of the diversity of the generated distribution, which supports the use of the class radius as a measure of diversity. We formalize this in the below corollary.

Corollary 3.2. *If $H_\alpha(p)$ denotes the α-order Rényi entropy of the distribution specified by probabilities p, then for all $\alpha \geq 2$, we have:*

$$\mathbb{E}\left[r(z)\right] \leq \frac{\log\left(4\pi H_\alpha(p)\right)}{\sqrt{2H_\alpha(p)}}$$

Proof. Use the fact that for all $\alpha_1 \leq \alpha_2$, $H_{\alpha_1}(p) \geq H_{\alpha_2}(p)$. Next, observe that for $\alpha = 2$, the inequality in the corollary statement becomes exactly that stated in Theorem 3.1. Finally, use the fact that $\log\left(4\pi x\right)/\sqrt{2x}$ is decreasing in x. □

Furthermore, notice that inverting the bound in Theorem 3.1 yields an upper bound on $-\log\left(\|p\|_2^2\right)$. As a result, having access to $\mathbb{E}\left[r(z)\right]$ for some generative model g and classification function f translates into an upper bound on the diversity of g with respect to f.

We can also show that to minimize the bound in Theorem 3.1, we must take $p_i = 1/K$. We state this in the following theorem.

Theorem 3.3. *The upper bound stated in Theorem 3.1 is minimized when $p_i = 1/K$. Informally, if the generative model induces the uniform distribution over possible images, then the expected class radius is smallest. The bound in Theorem 3.1 then becomes:*

$$\mathbb{E}\left[r(z)\right] \leq \frac{\log\left(4\pi \log\left(K\right)\right)}{\sqrt{2\log\left(K\right)}}$$

Proof. We defer a full proof to the appendices; see the proof of Theorem B.2. □

We now show that the bound in Theorem 3.1 is decreasing in K. Thus, as K grows, we must have that $\mathbb{E}\left[r(z)\right]$ approaches 0. We formalize and prove this claim below.

Theorem 3.4. *Suppose that we impose the following constraint on p_i:*

$$p_i \in \left[0, \min\left(\frac{1}{5}, h(K)\right)\right]$$

where $h(K)$ is a function of K such that $h(K) = o(1)$ and $h(K) = \Omega(1/K)$. Then, we have:

$$\lim_{K\to\infty} \mathbb{E}\left[r(z)\right] = 0$$

Proof. We defer a full proof to the appendices; see the proof of Theorem B.3. □

3.1 Towards a Test to Evaluate Model Diversity

We now propose several ways to leverage these observations in practice. One interesting problem is to determine a test statistic that can be computed given a generator g and a classifier f. This test statistic can then be used to make claims regarding the diversity of the generative models. In particular, if we could easily compute $\mathbb{E}\left[r(z)\right]$, then we could compare two generative models. The theorems above suggest that a higher value of $\mathbb{E}\left[r(z)\right]$ might imply that g is less diverse with respect to f (recall that our upper bounds decrease as the generated distribution approaches uniformity and as the number of outputs K grows arbitrarily large). Unfortunately, since our bound is one-sided, we cannot conclusively state

[4]Recall that the α-order Rényi entropy is $\alpha\left(1-\alpha\right)^{-1}\log\left(\|p\|_\alpha\right)$. Taking the limit as α approaches 1 yields the Shannon entropy, which is simply $-\sum p_i \log\left(p_i\right)$) (Bosyk et al. (2012)). We obtain the collision entropy by setting $\alpha = 2$.

whether one model is more diverse than the other; instead, this test would only provide evidence towards such a conclusion.

As a first step towards demonstrating the viability of such a test, we approximate $\mathbb{E}[r(z)]$ for a toy example. We train multiple GANs on variants of the 2D-Grid synthetic dataset[5] (Lin et al. (2017); Srivastava et al. (2017)) and observe various degrees of mode collapse on each. We use this simple synthetic dataset because measuring mode collapse is easy when the modes are known and because the results of this task are easy to visualize. The results are shown in Figure 3.1. It is clear from this figure that the GANs capturing more modes enjoy a lower class radius value, which indicates that obtaining an easily computable approximation to $\mathbb{E}[r(z)]$ would likely be useful in comparing the diversity of two generative models.

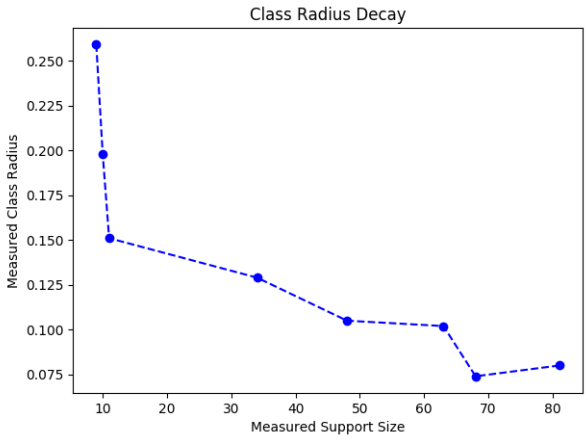

Figure 3.1: In the above figure, we plot the (discretized) support size of a GAN trained on the 2D-Grid task against the approximation of the class radius we compute.

Another problem of interest is to use $\mathbb{E}[r(z)]$ or some proxy to $\mathbb{E}[r(z)]$ as a regularizer for training a generative model. Such a method should penalize a lack of diversity and should therefore encourage the model to learn a better distribution.

3.2 Connection to Adversarial Robustness

In this section, we point out a tangential implication of our theoretical results. Our proofs of the theorems in the preceding sections are heavily inspired by those by Fawzi et al. (2018), as their problem setting is equivalent to ours. Fawzi et al. (2018) consider the problem of determining robustness against *adversarial examples*. Adversarial examples are perturbed inputs to a classifier with the intention of changing the classifier outputs. The adversarial input is usually nearly identical to the original input. Due to the equivalence of the problem settings, we can apply our results to the adversarial examples domain.

Perhaps the biggest implication of our results to adversarial examples is that the expected in-distribution robustness is smallest when each class is equiprobable, a result of Theorem 3.3. Furthermore, as a result of Theorem 3.4, we have that the expected in-distribution robustness decays to 0 as long as the probability of the most probable class decays with the number of classes K. This provides some theoretical basis for the results obtained by Jordan et al. (2018) in which the authors observe that far smaller perturbations are required to adversarially attack an ImageNet classifier than are required to attack a CIFAR-

[5]The 2D-Grid task consists of a target distribution that is a mixture of $(2n+1)^2$ spherical two-dimensional Gaussians with variances 0.0025 and means $(2i, 2j), -n \leq i, j \leq n$. The classification function c we use sends each generated point to its nearest mean, as done in the special case in Example 2.2.

10 classifier. Since ImageNet contains far more classes than CIFAR-10 and no class in either dataset is vastly overrepresented, the conditions required for our theorems to apply are met.

4 Conclusion

In this paper, we identified a geometric property of the latent space of generative models called the *class radius*. We proved that a high class radius implies an upper bound on the diversity of the generative model. We then linked this insight to various measurements of diversity of the generated probability distribution as well as to implications regarding the robustness of classifiers to adversarial perturbations. Finally, we proposed a few directions which leverage our theoretical insights for practical purposes.

Future Work As previously mentioned, our work leaves open the problem of leveraging our theoretical observations to improve the training of generative models and to potentially devise additional techniques to evaluate existing generative models. A more broad issue of interest is developing a more theoretical understanding of generative models, which we hope our work contributes towards and sparks further research in.

Related Work On the theoretical side, the paper by Fawzi et al. (2018) is most similar to ours. Our bounds are based on results the authors obtain for the adversarial examples domain and our analysis is heavily inspired by theirs. In this paper, the authors prove that under some assumptions, every classifier is vulnerable to small adversarial perturbations. One key assumption they make is that the data distribution on which the classifier operates comes from a generative model whose latent vectors are distributed according to a spherical Gaussian with identity covariance. This assumption is commonly reflected in generative models such as VAEs and GANs and their respective variants.

On the empirical side, there are various papers that make similar contributions to ours. Arora et al. (2018) attack the problem of determining diversity of GANs by leveraging the birthday paradox. Their so-called birthday paradox test hinges on the fact that high collision probabilities are evidence of low diversity. This provides another angle from which to view our theoretical result relating collision probabilities to the class radius. Additionally, this work provides one solution to a problem we pose - namely, producing a test that can be used to evaluate the diversity of a generative model. Borji (2018) describes a zoo of existing methods to evaluate generative models, and Khrulkov & Oseledets (2018) and Santurkar et al. (2017) present more recent methods to evaluate generative models. However, except for Khrulkov & Oseledets (2018), no authors of recent generative model evaluation strategies leverage geometric insights.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do GANs learn the distribution? some theory and empirics. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=BJehNfW0-`.

Ali Borji. Pros and cons of gan evaluation measures. *arXiv preprint arXiv:1802.03446*, 2018.

GM Bosyk, M Portesi, and A Plastino. Collision entropy and optimal uncertainty. *Physical Review A*, 85(1):012108, 2012.

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. *arXiv preprint arXiv:1802.08686*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

He Huang, Phillip Yu, and Changhu Wang. An introduction to image synthesis with generative adversarial nets. *arXiv preprint arXiv:1803.04469*, 2018.

Matt Jordan, Naren Manoj, Surbhi Goel, and Alex Dimakis. Combined adversarial attacks. *NIPS 2018 (submitted)*, 2018.

Valentin Khrulkov and Ivan Oseledets. Geometry score: A method for comparing generative adversarial networks. *arXiv preprint arXiv:1802.02664*, 2018.

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

Fei-Fei Li, Justin Johnson, and Serena Yeung. Lecture notes for convolutional neural networks for visual recognition, May 2018.

Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. Pacgan: The power of two samples in generative adversarial networks. *arXiv preprint arXiv:1712.04086*, 2017.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Shibani Santurkar, Ludwig Schmidt, and Aleksander Madry. A classification-based perspective on gan distributions. *arXiv preprint arXiv:1711.00970*, 2017.

Akash Srivastava, Lazar Valkoz, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems*, pp. 3308–3318, 2017.

A Deep Generative Models: A Whirlwind Tour

In this section, we give a brief background on the deep generative models we consider in this paper.

A *generative model* solves the following problem (Li et al. (2018)).

Given training data, generate samples from the same distribution as the training data.

A.1 Variational Autoencoders

In this section, we provide a brief overview of Variational Autoencoders. For a more thorough survey, see the work by Doersch (2016). Most of the material in this subsection is borrowed from this survey.

A Variational Autoencoder (VAE) consists of two components. The first is an encoder, which is a potentially nondeterministic function sending inputs to latent codes. The second is a decoder, which is also a potentially nondeterministic function sending latent codes to outputs. The vanilla VAE objective minimizes two terms. The first term is a discrepancy measure between the distribution of encoder outputs and some prespecified distribution over latent codes. The second term is a reconstruction error measuring, essentially, how lossy the compression is that the encoder performs.

We can break down the flow of information through a VAE down via the following steps.

1. Suppose X is a random variable denoting an image randomly sampled from the data distribution $\mathcal{X}$. Suppose x is one instance of X.
2. Encode X by sampling from an encoding distribution P. Thus, the output Z, a random variable denoting the obtained latent code, is distributed according to $P(Z|X = x)$. Notice that if our encoder is deterministic, then $P(Z|X = x)$ has all its mass at one point x.
3. Suppose z is one instance of Z. We now decode Z by sampling from a decoding distribution Q. Here, the output $\hat{X}$, a random variable denoting the output, is distributed according to $Q\left(\hat{X}|Z = z\right)$. Again, if our decoder is deterministic, then $Q\left(\hat{X}|Z = z\right)$ has all its mass at one point z.

We now have two objectives.

- The distribution over latent codes, $P(Z|X)$ where $X \sim \mathcal{X}$, should be close to our pre-imposed target distribution over latent codes. In our work, we would like the latent codes to follow a Gaussian as closely as possible. To penalize distance from a Gaussian, we use a Kullback-Leibler Divergence term as a loss. Thus, we add $KL\left(P(Z|X), \mathcal{N}\left(0, \text{diag}\left(\vec{1}_d\right)\right)\right)$ to our total loss.
- The reconstruction $\hat{X}$ should be as close to X as possible on average. In this paper, we penalize the squared ℓ_2 norm between the two images. We thus add $\mathbb{E}\left[\left\|\hat{X} - X\right\|_2^2\right]$ to our total loss.

A.2 Generative Adversarial Networks

In this section, we provide a brief overview of Generative Adversarial Networks. For a more thorough survey, see the work by Huang et al. (2018). Most of the material in this subsection is borrowed from this survey.

A Generative Adversarial Network (GAN) is a deep learning algorithm consisting of two neural networks - a *generator*, which generates images as a function of latent vectors sampled from some distribution $\mathcal{Z}$, and a *discriminator*, which classifies input images as coming from

the data distribution $\mathcal{X}$ or from the distribution specified by the generator (given by $G(\mathcal{Z})$, the pushforward of $\mathcal{Z}$ through G).

Algorithm A.1 describes how a GAN is trained. At each step, the generator tries to fool the discriminator by generating realistic images. Thus, the objective of the generator is to maximize the misclassifications that the discriminator makes, and the objective of the discriminator is to maximize its accuracy. This therefore yields the following objective:

$$\min_G \max_D V(D,G) \coloneqq \mathop{\mathbb{E}}_{X\sim\mathcal{X}} [\log(D(X))] + \mathop{\mathbb{E}}_{z\sim\mathcal{Z}} [\log(1 - D(G(z)))]$$

Algorithm A.1 GAN training

Input: Data distribution to model $\mathcal{X}$
Initialize generator G and discriminator D
while G not converged **do**
    Obtain latent vector z from some distribution
    Obtain $X_f = G(z)$ (generated outputs)
    Obtain X_r (real training data)
    Train discriminator by minimizing $\log(D(x_r)) + \log(1 - D(x_f))$
    Train generator by maximizing $\log(D(x_f))$
end while
Output: G, D

GANs have the benefit of generating very sharp images. However, GANs suffer from a problem known as *mode collapse*, in which the generator effectively fails to produce samples from numerous modes in $\mathcal{X}$. Several training modifications have been proposed to mitigate mode collapse (Lin et al. (2017); Salimans et al. (2016); Arjovsky et al. (2017)). On the other hand, researchers have devised methods to detect mode collapse (Arora et al. (2018)), and, more broadly, evaluate the extent to which the GAN has learned the original data manifold (Khrulkov & Oseledets (2018)).

B Proofs

In this section, we restate and prove our main results. We begin by showing that our problem setting is equivalent to the one considered by Fawzi et al. (2018). Thus, the bounds the authors obtain are applicable to the problem we consider. We then state and prove intermediate lemmas we employ in our proofs of our main claims. Next, we state and prove our main theorems. Finally, we reproduce the proofs presented by Fawzi et al. (2018) of the lemmas we use for our results.

Equivalence to Fawzi et al. (2018) We show that the results obtained by Fawzi et al. (2018) apply to our problem setting. First, notice that the definition of *in-distribution robustness* is equivalent to Definition 2.4. Next, observe that the assumptions on our classification function match the assumptions on the discriminator detailed by Fawzi et al. (2018). Furthermore, note that the distribution on the latent vectors in both settings is the same. These summarize the assumptions and definitions we make and are equal to those made by Fawzi et al. (2018). Thus, the results the authors obtain are applicable in our setting as well, and we may proceed.

B.1 Lemmas

We first state bounds without proof derived in Fawzi et al. (2018). These bounds form the foundation of our results.

Lemma B.1. *Define:*

$$a_{\neq i} \coloneqq \Phi^{-1}\left(\mathbb{P}\left[\bigcup_{j\neq i} K_j\right]\right)$$

where K_j represents the set of points in latent space mapping to class K_j. Then, the below inequality follows:

$$\mathbb{E}\left[r(z)\right] \leq \sum_{i=1}^{K} -a_{\neq i}\Phi(-a_{\neq i}) + \frac{e^{-a_{\neq i}^2/2}}{\sqrt{2\pi}}$$

Proof. We defer a full proof to Section B.3; see Lemma B.12 for details. □

Lemma B.2. *If $K \geq 5$, then:*

$$\Phi^{-1}\left(1 - \frac{1}{K}\right) \geq \sqrt{\log\left(\frac{K^2}{4\pi \log\left(K\right)}\right)}$$

Proof. We defer a full proof to Section B.3; see Lemma B.13 for details. □

We also state the following three important equalities regarding the standard Gaussian PDF, CDF, and inverse CDF, respectively.

Lemma B.3. *If we define:*

$$\varphi(x) \coloneqq \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{B.1}$$

$$\Phi(x) \coloneqq \mathop{\mathbb{P}}_{t \sim \mathcal{N}(0,1)}\left[t \leq x\right] = \int_{-\infty}^{x} \varphi(t)\,\mathrm{d}t \tag{B.2}$$

Then the following hold:

$$\varphi(x) = \varphi(-x) \tag{B.3}$$

$$\Phi(-x) = 1 - \Phi(x) \tag{B.4}$$

$$\Phi^{-1}(1-x) = -\Phi^{-1}(x) \tag{B.5}$$

Proof. As a result of Equation B.1, observe that Equation B.3 immediately follows. To see why Equation B.2 holds, note the following for positive x:

$$\mathbb{P}\left[t \geq x\right] = \mathbb{P}\left[t \leq -x\right] = \Phi(-x)$$

Furthermore, observe the following:

$$\mathbb{P}\left[t \leq x\right] = 1 - \mathbb{P}\left[t \geq x\right] = \Phi(x)$$

Substituting yields:

$$\Phi(x) + \Phi(-x) = 1$$

as desired. For Equation B.5, notice the following manipulations:

$$\begin{aligned} x &=: \Phi(y) \\ 1 - x &= 1 - \Phi(y) \\ &= \Phi(-y) \\ \Phi^{-1}(1-x) &= -y \\ &= -\Phi^{-1}(x) \end{aligned}$$

as desired. □

Using these, we derive the following lemma, whose notation is largely borrowed from Fawzi et al. (2018).

Lemma B.4. *We can rewrite the upper bound derived from Lemma B.1 as follows:*

$$\sum_{i=1}^{K} -a_{\neq i}\Phi(-a_{\neq i}) + \frac{e^{-a_{\neq i}^2/2}}{\sqrt{2\pi}} = \sum_{i=1}^{K} \Phi^{-1}(p_i) \cdot p_i + \frac{e^{-\Phi^{-1}(p_i)^2/2}}{\sqrt{2\pi}}$$

Proof. We first simplify the following a little:

$$\begin{aligned}\Phi(-a_{\neq i}) &= 1 - \Phi(a_{\neq i}) \\ &= 1 - \Phi\left(\Phi^{-1}\left(\mathbb{P}\left[\bigcup_{j\neq i} K_j\right]\right)\right) \\ &= 1 - \mathbb{P}\left[\bigcup_{j\neq i} K_j\right] \\ &= \mathbb{P}\left[K_i\right] \\ &=: p_i\end{aligned}$$

Now, observe the following equality:

$$a_{\neq i} = \Phi^{-1}\left(1 - p_i\right) = -\Phi^{-1}(p_i)$$

We use this to put our target expression exclusively in terms of our probability distribution:

$$f(p) \coloneqq \sum_{i=1}^{K} \Phi^{-1}(p_i) \cdot p_i + \frac{e^{-\Phi^{-1}(p_i)^2/2}}{\sqrt{2\pi}}$$

as desired. □

We now revisit the assumption we impose on the vector of generated probabilities p. Under this assumption, we have the following.

Lemma B.5. *Consider the following constraints on p:*

$$\begin{aligned}\|p\|_\infty \leq v = \min\left(\frac{1}{5}, h(K)\right) \text{ such that } h(K) = \Omega\left(\frac{1}{K}\right) \text{ and } h(K) = o(1) \\ \|p\|_1 = 1\end{aligned}$$

Then, we have:

$$\frac{1}{K} \leq \|p\|_2^2 \leq \min\left(\frac{1}{5}, 3h(K)\right)$$

Proof. We begin with the lower bound. Notice that the following must hold for any vector $p \in \mathbb{R}^K$:

$$\|p\|_1^2 \cdot \frac{1}{K} \leq \|p\|_2^2$$

Setting $\|p\|_1 = 1$ immediately gives the desired.

We now proceed with the upper bound. Consider the following instantiation of p:

$$\begin{aligned}p_i = v, \forall i \in \left\{1, \ldots, \left\lfloor\frac{1}{v}\right\rfloor\right\} \\ p_{\lfloor 1/v\rfloor + 1} = 1 - v\left\lfloor\frac{1}{v}\right\rfloor \\ p_i = 0 \text{ otherwise}\end{aligned}$$

Observe that this instantiation of p satisfies both constraints we imposed earlier. Now, consider the following manipulations:

$$\begin{aligned}
\|p\|_2^2 &= \sum p_i^2 \\
&= \left\lfloor \frac{1}{v} \right\rfloor v^2 + \left(1 - v \left\lfloor \frac{1}{v} \right\rfloor\right)^2 \\
&= \left\lfloor \frac{1}{v} \right\rfloor v^2 + \left(1 - 2v \left\lfloor \frac{1}{v} \right\rfloor + v^2 \left\lfloor \frac{1}{v} \right\rfloor^2\right) \\
&\leq \frac{1}{v} \cdot v^2 + 1 - 2v \left\lfloor \frac{1}{v} \right\rfloor + v^2 \cdot \frac{1}{v^2} \\
&= v + 2 - 2v \left\lfloor \frac{1}{v} \right\rfloor \\
&\leq v + 2 - 2v \left(\frac{1}{v} - 1\right) \\
&= 3v
\end{aligned}$$

We now show that this is an optimal assignment. Observe that this allocation of mass follows a greedy strategy - we allocate as much mass as we can into each bucket until we run out. To see that the greedy strategy works, note that at each step, we can either allocate as much mass as we can to a bucket, or spread that same mass over several other buckets. To show that allocating as much mass as we can into one bucket alone and leaving the rest empty is optimal, it suffices to show the following for any vector x:

$$\|x\|_1^2 = \left(\sum x_i\right)^2 \geq \|x\|_2^2 = \sum x_i^2$$

This is clearly true, and we may conclude. □

We now state various intermediate results we use to prove our main claims.

Lemma B.6. *The following function is concave if $x \in (0, 1)$:*

$$g(x) \coloneqq \frac{\log\left(4\pi \log\left(1/x\right)\right)}{\sqrt{2 \log\left(1/x\right)}}$$

Proof. One sufficient condition for g to be concave is that its second derivative, written below, is always nonpositive on the same domain:

$$g''(x) = \frac{-2\log\left(1/x\right)\left(\log\left(4\pi\log\left(1/x\right)\right) - 2\right) + 3\log\left(4\pi\log\left(1/x\right)\right) - 8}{4\sqrt{2}x^2\left(\log\left(1/x\right)\right)^{5/2}}$$

Note that the denominator is always positive on this domain, so it suffices to show that the numerator is always nonpositive on the domain. To clean up notation, make the following substitution:

$$y = \log\left(\frac{1}{x}\right)$$

Then, note that $y \geq 0$. We now wish to show that the below expression is always nonpositive for all nonnegative y:

$$\begin{aligned}
&-2y\left(\log\left(4\pi y\right) - 2\right) + 3\log\left(4\pi y\right) - 8 \\
&= -2y\left(\log\left(\frac{4\pi}{e^2}\right) + \log\left(y\right)\right) + \log\left(\frac{64\pi^3}{e^8}\right) + 3\log\left(y\right)
\end{aligned}$$

Notice that as a result of convexity, the following two inequalities hold:

$$\begin{aligned}
-x\log\left(x\right) &\leq -x + 1 \\
\log\left(x\right) &\leq x - 1
\end{aligned}$$

We use these to our advantage:

$$\begin{aligned}
&-2y\left(\log\left(\frac{4\pi}{e^2}\right)+\log(y)\right)+\log\left(\frac{64\pi^3}{e^8}\right)+3\log(y)\\
&\leq -2y\left(\log\left(\frac{4\pi}{e^2}\right)+1\right)+\log\left(\frac{64\pi^3}{e^6}\right)+3\log(y)\\
&\leq -2y\left(\log\left(\frac{4\pi}{e^2}\right)+1\right)+\log\left(\frac{64\pi^3}{e^6}\right)+3(y-1)\\
&=\log\left(\frac{e^5}{16\pi^2}\right)y+\log\left(\frac{64\pi^3}{e^9}\right)
\end{aligned}$$

This is simply a linear equation; thus, to check that it is nonpositive over $y \geq 0$, it is enough to check that the slope and vertical intercepts are nonpositive. This is clearly true here, and we may therefore conclude. □

Lemma B.7. *The following function is increasing when $x \in (0, 1/2]$:*

$$g(x) \coloneqq \frac{\log(4\pi\log(1/x))}{\sqrt{2\log(1/x)}}$$

Proof. One sufficient condition for g to be increasing on this domain is that its first derivative, written below, is always positive on the same domain:

$$g'(x) = \frac{\log(4\pi\log(1/x)) - 2}{2\sqrt{2}x\left(\log(1/x)\right)^{3/2}}$$

Note that the denominator is always positive on this domain, so it suffices to show that the numerator is always positive on the domain. Specifically, we wish to show that for $x \in (0, 1/2]$:

$$\log\left(4\pi\log\left(\frac{1}{x}\right)\right) > 2$$

Since the LHS is always increasing in $1/x$, it follows that to minimize the LHS, we must minimize $1/x$. This occurs by setting $x = 1/2$. The desired result immediately follows. □

B.2 Theorems

Theorem B.1 (Restatement of Theorem 3.1)**.** *If $0 < p_i \leq 1/5, \forall i \in [K]$, then the following holds regarding the expected class radius of each image in the latent space of our generative model:*

$$\mathbb{E}[r(z)] \leq \frac{\log\left(4\pi\log\left(1/\left\|p\right\|_2^2\right)\right)}{\sqrt{2\log\left(1/\left\|p\right\|_2^2\right)}}$$

Proof. If $p_i \leq 1/5$, then by Lemma B.2, we have the following:

$$\begin{aligned}
\Phi^{-1}(p_i) &\leq -\sqrt{\log\left(\frac{1/p_i^2}{4\pi\log(1/p_i)}\right)}\\
-\Phi^{-1}(p_i)^2 &\leq -\log\left(\frac{1/p_i^2}{4\pi\log(1/p_i)}\right)\\
e^{-\Phi^{-1}(p_i)^2} &\leq \frac{4\pi\log(1/p_i)}{1/p_i^2}\\
e^{-\Phi^{-1}(p_i)^2/2} &\leq \sqrt{\frac{4\pi\log(1/p_i)}{1/p_i^2}}\\
&= p_i\sqrt{4\pi\log(1/p_i)}
\end{aligned}$$

We therefore have:

$$\begin{aligned}
f(p) &\leq \sum_{i=1}^{K} -p_i \sqrt{\log\left(\frac{1/p_i^2}{4\pi \log(1/p_i)}\right)} + p_i \sqrt{2\log(1/p_i)} \\
&= \sum_{i=1}^{K} p_i \left(\sqrt{2\log(1/p_i)} - \sqrt{\log\left(\frac{1/p_i^2}{4\pi \log(1/p_i)}\right)}\right) \\
&= \sum_{i=1}^{K} p_i \left(\sqrt{2\log(1/p_i)} - \sqrt{2\log(1/p_i) - \log(4\pi \log(1/p_i))}\right) \\
&= \sum_{i=1}^{K} p_i \left(\frac{\log(4\pi \log(1/p_i))}{\sqrt{2\log(1/p_i)} + \sqrt{2\log(1/p_i) - \log(4\pi \log(1/p_i))}}\right) \\
&\leq \sum_{i=1}^{K} p_i \left(\frac{\log(4\pi \log(1/p_i))}{\sqrt{2\log(1/p_i)}}\right)
\end{aligned}$$

By Lemma B.6, using weights p_i, we can apply Jensen's Inequality:

$$\begin{aligned}
f(p) &\leq g\left(\|p\|_2^2\right) \\
&= \frac{\log\left(4\pi \log\left(1/\|p\|_2^2\right)\right)}{\sqrt{2\log\left(1/\|p\|_2^2\right)}}
\end{aligned}$$

as desired. □

Theorem B.2 (Restatement of Theorem 3.3)**.** *The upper bound stated in Theorem 3.1 is minimized when $p_i = 1/K$. Informally, if the generative model induces the uniform distribution over possible images, then the expected class radius is smallest. The bound in Theorem 3.1 then becomes:*

$$\mathbb{E}[r(z)] \leq \frac{\log(4\pi \log(K))}{\sqrt{2\log(K)}}$$

Proof. By Lemma B.7, if we wish to minimize g, it suffices to minimize $\|p\|_2^2$ as long as $\|p\|_2^2 < 1/2$. By Lemma B.5, this must be satisfied. Now, observe that for a vector $p \in \mathbb{R}^K$, we must have $\|p\|_1 \leq \sqrt{K}\|p\|_2$. Since $\|p\|_1 = 1$, we have $1/\|p\|_2^2 \geq K$. We then use this in the bound from Theorem 3.1 to obtain the desired result.

We also present an information-theoretic proof. It is well-known that Rényi entropies achieve their maximum of $\log(K)$ when the distribution p is uniform (intuitively, this means that the distribution maximizing uncertainty is the uniform distribution). We then use the fact that the following function is clearly decreasing:

$$\frac{\log(4\pi x)}{\sqrt{2x}}$$

and we arrive at the same conclusion.

To obtain the exact form of the bound, use the fact that setting $p_i = 1/K$ yields $\|p\|_2^2 = 1/K$. □

Theorem B.3 (Restatement of Theorem 3.4)**.** *Suppose that we impose the following constraint on p_i:*

$$p_i \in \left[0, \min\left(\frac{1}{5}, h(K)\right)\right]$$

where $h(K)$ is a function of K such that $h(K) = o(1)$ and $h(K) = \Omega(1/K)$. Then, we have:

$$\lim_{K \to \infty} \mathbb{E}[r(z)] = 0$$

Proof. By Theorem 3.1, we have:

$$\mathbb{E}\left[r(z)\right] \leq \frac{\log\left(4\pi \log\left(1/\left\|p\right\|_2^2\right)\right)}{\sqrt{2\log\left(1/\left\|p\right\|_2^2\right)}}$$

Then, using the results from Lemmas B.5 and B.7, we have:

$$\mathbb{E}\left[r(z)\right] \leq \frac{\log\left(4\pi \log\left(\min\left(\frac{1}{5}, 3h(K)\right)\right)\right)}{\sqrt{2\log\left(1/\min\left(\frac{1}{5}, 3h(K)\right)\right)}}$$

Using Lemma B.7 along with the fact that $h(K) = o(1)$ yields the following:

$$\lim_{K\to\infty} \frac{\log\left(4\pi \log\left(\min\left(\frac{1}{5}, 3h(K)\right)\right)\right)}{\sqrt{2\log\left(1/\min\left(\frac{1}{5}, 3h(K)\right)\right)}} = 0$$

Since $\mathbb{E}\left[r(z)\right] \geq 0$, the desired result must follow by the squeeze theorem. □

B.3 Proofs of Prior Results

With the intent of making this paper more self-contained, we reproduce here the proofs originally obtained by Fawzi et al. (2018). We begin with several more foundational lemmas; we then proceed to our main proofs.

B.3.1 Intermediate Results

Lemma B.8 (Gaussian Isoperimetric Inequality)**.** *Let ν_d be the Gaussian measure on $\mathbb{R}^d$. Let $A \subseteq \mathbb{R}^d$ and let $A_\eta = \{z \mid \exists z' \in A \text{ such that } \left\|z' - z\right\|_2 \leq \eta\}$. If $\Phi(x)$ denotes the cumulative distribution function (CDF) as previously defined, then:*

$$\nu_d\left(A_\eta\right) \geq \Phi\left(\Phi^{-1}\left(\nu_d\left(A\right)\right) + \eta\right)$$

Lemma B.9 (Bounds on the Normal CDF)**.** *If $x \geq 0$, $\varphi(x)$ denotes the standard Guassian probability density function (PDF), and denotes the $\Phi(x)$ denotes the standard Gaussian cumulative distribution function (CDF) as previously defined, then:*

$$1 - \varphi(x) \cdot \frac{2}{x + \sqrt{x^2 + 8/\pi}} \leq \Phi(x) \leq 1 - \varphi(x) \cdot \frac{2}{x + \sqrt{x^2 + 4}}$$

Lemma B.10. *Define:*

$$a_{\neq i} \coloneqq \Phi^{-1}\left(\mathbb{P}\left[\bigcup_{j\neq i} K_j\right]\right)$$

where K_j represents the set of points in latent space mapping to class K_j. Then, the below inequality follows:

$$\mathbb{P}\left[r(z) \leq \eta\right] \geq \sum_{i=1}^{K} \left(\Phi\left(a_{\neq i} + \eta\right) - \Phi\left(a_{\neq i}\right)\right)$$

Proof. Define the following sets:

$$K_{\neq i} \coloneqq \bigcup_{j\neq i} K_j$$
$$K_i^\eta \coloneqq \{z \in K_i \mid \exists z' \in K_{\neq i} \text{ such that } \left\|z - z'\right\|_2 \leq \eta\}$$

Thus, $K_{\neq i}$ is the set of all points belonging to some class but not in class i, and K_i^η is the set of all points in K_i that are η away from a different class. Observe that we wish to find the probability mass of K_i^η for every i; summing this will yield our final answer.

Next, notice that if we set $A = K_{\neq i}$, then in the notation of Lemma B.8, $A_\eta = K_{\neq i} \cup K_i^\eta$. With this setting of A, we can therefore use Lemma B.8:

$$\begin{aligned}\nu_d\left(A_\eta\right) &\geq \Phi\left(\Phi^{-1}\left(\nu_d\left(A\right)\right) + \eta\right) \\ \nu_d\left(K_{\neq i} \cup K_i^\eta\right) &\geq \Phi\left(\Phi^{-1}\left(\nu_d\left(K_{\neq i}\right)\right) + \eta\right)\end{aligned}$$

Since $K_{\neq i}$ and K_i^η are disjoint, we can split the term in the LHS:

$$\begin{aligned}\nu_d\left(K_{\neq i} \cup K_i^\eta\right) &\geq \Phi\left(\Phi^{-1}\left(\nu_d\left(K_{\neq i}\right)\right) + \eta\right) \\ \nu_d\left(K_{\neq i}\right) + \nu_d\left(K_i^\eta\right) &\geq \Phi\left(\Phi^{-1}\left(\nu_d\left(K_{\neq i}\right)\right) + \eta\right) \\ \nu_d\left(K_i^\eta\right) &\geq \Phi\left(\Phi^{-1}\left(\nu_d\left(K_{\neq i}\right)\right) + \eta\right) - \nu_d\left(K_{\neq i}\right)\end{aligned}$$

Summing from $i = 1$ up to $i = K$ and swapping out notation yields the desired result. □

Lemma B.11. *The following holds:*

$$\int_0^\infty \Phi\left(-a - \eta\right) \mathrm{d}\eta = -a\Phi(-a) + \frac{e^{-a^2/2}}{\sqrt{2\pi}}$$

Proof. Use u-substitution. Let $u = a + \eta$. Thus, $\mathrm{d}u = \mathrm{d}\eta$ and if $\eta = 0$, then $u = a$. This allows us to rewrite our target statement:

$$\int_a^\infty \Phi\left(-u\right) \mathrm{d}u = -a\Phi(-a) + \frac{e^{-a^2/2}}{\sqrt{2\pi}}$$

To compute the LHS, observe the following manipulations:

$$\begin{aligned}\int \Phi(x) \,\mathrm{d}x &= \int x\varphi(x) + \Phi(x) - x\varphi(x) \,\mathrm{d}x \\ &= \int x\varphi(x) + \Phi(x) \,\mathrm{d}x - \int x\varphi(x) \,\mathrm{d}x \\ &= x\Phi(x) + \varphi(x)\end{aligned}$$

Thus:

$$\begin{aligned}\int \Phi(-u) \,\mathrm{d}u &= \int \left(1 - \Phi(u)\right) \mathrm{d}u \\ &= u - u\Phi(u) - \varphi(u) \\ &= u\left(1 - \Phi(u)\right) - \varphi(u) \\ &= u\Phi(-u) - \varphi(u)\end{aligned}$$

Taking the final evaluation gives:

$$\begin{aligned}\int_a^\infty \Phi\left(-u\right) \mathrm{d}u &= \lim_{u\to\infty} \left(u\Phi(-u) - \varphi(u)\right) - \left(a\Phi(-a) - \varphi(a)\right) \\ &= -a\Phi(-a) + \varphi(a)\end{aligned}$$

as desired, where the final line follows from Lemma B.9 and the squeeze theorem. □

B.3.2 Main Results

Lemma B.12 (Restatement of Lemma B.1)**.** *Define:*

$$a_{\neq i} \coloneqq \Phi^{-1}\left(\mathbb{P}\left[\bigcup_{j\neq i} K_j\right]\right)$$

where K_j represents the set of points in latent space mapping to class K_j. Then, the below inequality follows:

$$\mathbb{E}\left[r(z)\right] \leq \sum_{i=1}^K -a_{\neq i}\Phi(-a_{\neq i}) + \frac{e^{-a_{\neq i}^2/2}}{\sqrt{2\pi}}$$

Proof. Observe that the following holds:

$$\mathbb{E}\left[r(z)\right] = \int_0^\infty \mathbb{P}\left[r(z) \geq \eta\right] \mathrm{d}\eta$$

By Lemma B.10, we have:

$$\mathbb{P}\left[r(z) \geq \eta\right] \leq 1 - \sum_{i=1}^{K} \left(\Phi\left(a_{\neq i} + \eta\right) - \Phi\left(a_{\neq i}\right)\right)$$

Therefore:

$$\begin{aligned}
\mathbb{E}\left[r(z)\right] &= \int_0^\infty \mathbb{P}\left[r(z) \geq \eta\right] \mathrm{d}\eta \\
&\leq \int_0^\infty \left(1 - \sum_{i=1}^{K} \left(\Phi\left(a_{\neq i} + \eta\right) - \Phi\left(a_{\neq i}\right)\right)\right) \mathrm{d}\eta
\end{aligned}$$

Observe that:

$$1 = \sum_{i=1}^{K} \left(1 - \Phi\left(a_{\neq i}\right)\right)$$

Substitute this into the above:

$$\begin{aligned}
\mathbb{E}\left[r(z)\right] &\leq \int_0^\infty \left(\sum_{i=1}^{K} \left(1 - \Phi\left(a_{\neq i}\right)\right) - \sum_{i=1}^{K} \left(\Phi\left(a_{\neq i} + \eta\right) - \Phi\left(a_{\neq i}\right)\right)\right) \mathrm{d}\eta \\
&= \int_0^\infty \sum_{i=1}^{K} \left(1 - \Phi\left(a_{\neq i} + \eta\right)\right) \mathrm{d}\eta \\
&= \int_0^\infty \sum_{i=1}^{K} \Phi\left(-a_{\neq i} - \eta\right) \mathrm{d}\eta \\
&= \sum_{i=1}^{K} \int_0^\infty \Phi\left(-a_{\neq i} - \eta\right) \mathrm{d}\eta
\end{aligned}$$

By Lemma B.11, we can rewrite the above:

$$\mathbb{E}\left[r(z)\right] \leq \sum_{i=1}^{K} -a_{\neq i}\Phi(-a_{\neq i}) + \frac{e^{-a_{\neq i}^2/2}}{\sqrt{2\pi}}$$

which is exactly the desired. □

Lemma B.13 (Restatement of Lemma B.2)**.** *If $K \geq 5$, then:*

$$\Phi^{-1}\left(1 - \frac{1}{K}\right) \geq \sqrt{\log\left(\frac{K^2}{4\pi \log\left(K\right)}\right)}$$

Proof. Since $\Phi(x)$ is monotonically increasing, we can rewrite the target statement:

$$1 - \frac{1}{K} \geq \Phi\left(\sqrt{\log\left(\frac{K^2}{4\pi \log\left(K\right)}\right)}\right)$$

By Lemma B.9, we have:

$$\Phi(x) \leq 1 - \varphi(x) \cdot \frac{2}{x + \sqrt{x^2 + 4}}$$

It therefore suffices to show that for $x = \sqrt{\log\left(K^2 / \left(4\pi \log\left(K\right)\right)\right)}$:

$$\varphi(x) \cdot \frac{2}{x + \sqrt{x^2 + 4}} \geq \frac{1}{K}$$

Now, notice the following manipulations:

$$\begin{aligned}\varphi(x) \cdot \frac{2}{x+\sqrt{x^2+4}} &= \varphi(x) \cdot \frac{\frac{1}{\sqrt{2}} \cdot 2}{\frac{1}{\sqrt{2}} \cdot \left(x+\sqrt{2}\sqrt{x^2/2+2}\right)} \\ &= \frac{e^{-x^2/2}}{\sqrt{2}\sqrt{\pi}} \cdot \frac{\frac{1}{\sqrt{2}} \cdot 2}{\frac{1}{\sqrt{2}} \cdot \left(x+\sqrt{2}\sqrt{x^2/2+2}\right)} \\ &= \frac{e^{-x^2/2}}{\sqrt{\pi}} \cdot \frac{1}{x/\sqrt{2}+\sqrt{x^2/2+2}}\end{aligned}$$

Set $y = x/\sqrt{2}$ and continue, using the fact that $y + \sqrt{y^2+2} \leq 2\sqrt{y^2+1}$:

$$\begin{aligned}\varphi(x) \cdot \frac{2}{x+\sqrt{x^2+4}} &= \frac{e^{-x^2/2}}{\sqrt{\pi}} \cdot \frac{1}{x/\sqrt{2}+\sqrt{x^2/2+2}} \\ &= \frac{e^{-y^2}}{\sqrt{\pi}} \cdot \frac{1}{y+\sqrt{y^2+2}} \\ &\geq \frac{e^{-y^2}}{2\sqrt{\pi}} \cdot \frac{1}{\sqrt{y^2+1}}\end{aligned}$$

Set $z = y^2$. We now wish to show:

$$\frac{e^{-z}}{2\sqrt{\pi}} \cdot \frac{1}{\sqrt{z+1}} \geq \frac{1}{K}$$

Or, equivalently:

$$\log(K) \geq z + \log\left(\sqrt{4\pi z+4\pi}\right) = \left(\log(K) - \frac{1}{2}\log(4\pi\log(K))\right) + \log\left(\sqrt{4\pi z+4\pi}\right)$$

This amounts to showing:

$$\log\left(\sqrt{4\pi\log(K)}\right) \geq \log\left(\sqrt{4\pi z+4\pi}\right)$$

Which is the same as:

$$\log(K) \geq z+1$$

Substituting in z gives another equivalent expression to prove:

$$\log(4\pi\log(K)) \geq 2$$

The LHS is clearly increasing, and it is easy to check that the inequality holds at $K = 5$. This is the desired result, and we may conclude. $\square$