Using the Gaussianized Coding Rate for Generative Modeling

Michael Psenka, Vishal Raman, Peter Tong

May 10, 2023

Abstract

Generative models all aim to model a target distribution by minimizing some "distributional distance" from the parametrized distribution to the target distribution. For the Linear Discriminative Representation (LDR) transcription framework, part of the novelty is the distributional distance used: a Gaussianized coding rate difference. Although this distributional distance hosts many benefits (efficiently computable, supports low-dimensional distributions), it is only a well-posed metric over Gaussian distributions. In this paper, we show twofold about the Gaussianized coding rate difference: (a) on a wider class of probability distributions, a class which a number of machine learning datasets fall into, the Gaussianized coding rate difference is still a well-posed metric, and (b) its simplicity allows for training stability that alternatives do not enjoy. We hope these theoretical insights guide future practitioners in how to design robust generative models using the Gaussianized coding rate distance.

1 Introduction

Distribution learning is a central task in machine learning: in its most loose definition, one aims to "learn" a probability distribution π using a finite set of samples $\{x_i\}_{i=1}^N$ from π. This is usually achieved in a parametric fashion: that is, we have a finite-dimensional parameter $\theta \in D$ and a "parametrization map" $\theta \to \pi_\theta$, and finally try to find $\theta^* \in \mathbb{R}^p$ such that $\pi_{\theta^*} \approx \pi$.

This task is usually approached through so-called *generative modeling*. Examples of generative modeling include *generative adversarial networks* (GANs) [6], *auto-encoders* (AEs) [9][7], and their various forms. A description of these methods can be found in [3] and its references.

A common concept used to both design and improve generative models is a "distributional distance": a metric $d(\pi, \hat{\pi}) : \mathcal{D}(\mathbb{R}^d) \times \mathcal{D}(\mathbb{R}^d) \to \mathbb{R}$ between two probability distributions $\pi, \hat{\pi} \in \mathcal{D}(\mathbb{R}^D)$ over $\mathbb{R}^d$. This allows us to mathematically determine how close π and $\hat{\pi}$ are, adding rigor to the statement $\pi_{\theta^*} \approx \pi$: we say $\pi_{\theta^*} \approx \pi$ if $d(\pi_{\theta^*}, \pi) \leq \epsilon$ for small enough $\epsilon > 0$. Distribution learning can then be thought of minimizing a certain "distributional distance," whose properties illuminate the properties of the respective generative model. A couple of examples:

- *KL-divergence* [11], the backbone of variational auto-encoders [8]. While this distance function is well-studied in information theory, the metric is only well-defined if both distributions $\pi, \hat{\pi} \in \mathcal{D}(\mathbb{R}^d)$ are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. This is broken if π or $\hat{\pi}$ have low-dimensional support, thus often in practice we need to artifically project $\pi, \hat{\pi}$ to a low-dimensional space of arbitrarily-chosen dimension.
- *Wasserstein distance* [17] (Earth-mover's distance), the core component for the popular *Wasserstein GAN* [1]. While the Wasserstein distance is well-defined on distributions with low-dimensional support, it is computationally intractible, even on some very simple distributions such as Gaussians. In practice, the Wasserstein distance is approximated, but these approximations are not well-understood and thus lead to hidden downsides.

Finally, we turn to the main subject of this paper: LDR transcription [3], a distribution learning model that is built off of the following distributional distance:

$$\begin{aligned}\Delta R(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) = &\frac{1}{2}\log\det\left(\boldsymbol{I} + \frac{\gamma}{2}(\boldsymbol{Z}\boldsymbol{Z}^\top + \hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^\top)\right) \\ &- \frac{1}{4}\log\det\left(\boldsymbol{I} + \gamma\boldsymbol{Z}\boldsymbol{Z}^\top\right) \\ &- \frac{1}{4}\log\det\left(\boldsymbol{I} + \gamma\hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^\top\right),\end{aligned} \tag{1}$$

where $\gamma := \frac{d}{n\epsilon^2}$ for some $\epsilon > 0$, and $\boldsymbol{Z}, \hat{\boldsymbol{Z}} \in \mathbb{R}^{d\times n}$ are matrices formed by collecting n samples from each of two distributions $\pi, \hat{\pi}$ supported on $\mathbb{R}^d$. This function comes from the coding rate of Gaussian distributions [14], where we measure the difference of coding the union of samples versus the sum of their coding rates. Eq. (1) enjoys the following nice properties:

- Eq. (1) is well defined for any distributions $\pi, \hat{\pi}$ that have finite variance, even if they are low-dimensional. While (1) is only defined on finite samples, we see by the central limit theorem that this metric does converge as we collect samples:

$$\begin{aligned}\lim_{n\to\infty} \Delta R(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) = \frac{1}{2}\log\det\left(\boldsymbol{I} + \frac{\alpha}{2}(\Sigma_\pi + \Sigma_{\hat{\pi}})\right) \\ - \frac{1}{4}\log\det\left(\boldsymbol{I} + \alpha\Sigma_\pi\right) \\ - \frac{1}{4}\log\det\left(\boldsymbol{I} + \alpha\Sigma_{\hat{\pi}}\right),\end{aligned} \tag{2}$$

  where $\alpha := \frac{d}{\epsilon^2}$, and $\Sigma_\pi = \mathrm{E}_\pi\, xx^\top := \int_{\mathbb{R}^d} xx^\top \pi(dx)$.
- Eq. (1) is interpretable, and thus practicioners can design algorithms using Eq. (1) in a principled way.
- Eq. (1), along with its gradients, are computable functions, as the function $\log\det(\cdot)$ has been studied extensively computationally.

Of course, as with any distributional distance, Eq. (1) has the following apparent (but empirically disproven) downsides:

- Eq. (1) is a "loose" distance function in the folllowing sense: $\Delta R(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) = 0$ if and only if $\boldsymbol{Z}\boldsymbol{Z}^\top = \hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^\top$ [18]. That is, two distributions have distance 0 only when their second moments are the same. This is well-posed between Gaussians, which are defined by their second moments, but two generic distributions may have the same second moments but wildly different behavior. This can be seen as a byproduct of the fact that $\Delta R(\cdot,\cdot)$ was derived assuming the input distributions are Gaussian.
- Difficulty to train. In the original LDR transcription paper, eq. (1) is optimized over a minimax objective, which as seen from many modern minimax distribution modelers, such as GANs, can lead to instability in training.

Of course, both of these downsides were empirically debunked in [3]: reconstruction accuracy was very high on image distributions known to be non-Gaussian, and the training process was very robust to hyperparameter tuning and initialization. However, these phenomena are not understood theoretically, thus making it difficult for outside practitioners to use the principles in [3] to make new generative models for their use case.

The purpose of this paper is to give clear, theoretical insight on why the Gaussianized coding rate difference in Eq. (1) is a good distributional distance to use for generative models, giving outside practitioners proper insight to develop their own models using Eq. (1). We show the two following ideal properties:

1. *The right amount of efficiency*: Eq. (1) is efficiently computable and interpretable because we assume the input distributions are Gaussians. However, if distributions are "close to" Gaussians in some way, then eq. (1) could be "close to" a proper metric. We show that:
   (a) Eq. (1) is a well-defined distance over distributions with *kurtosis* close to that of Gaussians, and
   (b) Experimental results on image datasets showing that they have kurtosis very similar to Gaussians.

   Thus, Eq. (1) works properly as a metric on exactly the kinds of distributions we see in machine learning.
2. *Training stability*: Eq. (1) is stable around its critical points. Unlike in GAN training [5], where a small perturbation around a critical point can result in a massive change in loss, small changes in parameter space at a critical point of $\Delta R(\boldsymbol{Z}, \hat{\boldsymbol{Z}})$ result in equivalently small changes in the loss, leading to more robust training.

2 The right amount of efficiency

As described in [14], Eq. (1) is derived by assuming that the distributions $\boldsymbol{Z}, \hat{\boldsymbol{Z}}$ are drawn from are Gaussian, yielding an efficient formula for computing the coding rate. However, distributions from realistic data are hardly Gaussian, so while Eq. (1) is computable on non-Gaussian distributions, it is not necessarily a well-defined metric outside of Gaussians: non-Gaussian distributions can be very different and still satisfy $\Delta R(Z, \hat{Z}) = 0$.

The purpose of this section is to show that this is not a fatal problem, as many realistic data distributions are "close to Gaussian" in the right way, making $\Delta R(Z, \hat{Z})$ remain a good distance of distributions while retaining is desirable computational properties. More concretely, we show this through the following:

1. We introduce the well-studied statistical quantity *kurtosis* [15] of a distribution, and describe how this measures the "closeness" of a distribution to a Gaussian distribution.
2. We show that if we consider $\Delta R(\boldsymbol{Z}, \hat{\boldsymbol{Z}})$ where $\boldsymbol{Z}, \hat{\boldsymbol{Z}}$ are sampled from distributions with close-to-Gaussian kurtosis, then $\Delta R(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) = 0 \implies \int_{\mathbb{R}^d} |\pi - \hat{\pi}|(dx) \leq \epsilon$ for some small ϵ. In other words, $\Delta R(\boldsymbol{Z}, \hat{\boldsymbol{Z}})$ is a proper metric on such distributions.
3. We give experimental data to show that a number of data distributions in modern machine learning have close-to-Gaussian kurtosis.
4. Finally, we show that simple neural network models preserve close-to-Gaussian kurtosis, thus training $\Delta R(f_\theta g_\eta f_\theta(\boldsymbol{X}), f_\theta(\boldsymbol{X}))$

2.1 Kurtosis

The *kurtosis* [15], denoted $k(\pi) \in \mathbb{R}$, of a distribution $\pi \in \mathcal{D}(\mathbb{R}^d)$ is a scalar measure that heuristically measures the "tailedness" of a distribution, and historically has been used to measure how close to Gaussian a distribution is: this is the property of interest for this paper.

First, we mathematically define the kurtosis. For a given dataset $X = \{x_1, x_2, x_3, ..., x_n\}, X \in \mathbb{R}^{d \times n}$, we define its kurtosis to be

$$k(X) = \frac{1}{n} \sum (x_i - \bar{x})^T \Sigma^{-1} (x_i - \bar{x}))^2 \tag{3}$$

where $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$, and $\Sigma := \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$.

Similarly, we define the kurtosis of a continuous distribution $\pi \in \mathcal{D}(\mathbb{R}^d)$ to be the following:

$$k(\pi) = \mathop{\mathrm{E}}_{\pi} \left((x - \mathrm{E}(x))^\top \Sigma^{-1} (x - \mathrm{E}(x))^\top \right)^2. \tag{4}$$

Here is a primary reason that kurtosis is used to measure how close to a Gaussian a distribution is: if π is any Gaussian over $\mathbb{R}^d$ with invertible covariance, then $k(\pi) = d(d+2)$ (proof in Appendix A.1). Note that is only defined if Σ is invertible; this is because the kurtosis measures a scaled version of "tailedness" and is ill-defined in any direction without variance. We can fix this by projecting to the subspace with positive variance using PCA on Σ, where we then get that for *any* Gaussian $\pi = \mathcal{N}(\mu, \Sigma)$, $k(\pi) = k(k+2)$, where $k = \text{rank}(\Sigma)$.

Thus, we use the adjusted kurtosis to measure a distribution's ratio of kurtosis compared to that of any full-rank Gaussian distribution:

$$k_{adjusted}(\pi) = \frac{k(\pi)}{d(d+2)} \tag{5}$$

The closer to 1 this number is, the closer to a Gaussian the kurtosis test indicates the distribution is [4].

2.2 ΔR on Gaussian-like distributions

When we say Gaussian-like, we mean distributions with low magnitude of *excess kurtosis*: $|k_{adjusted}(\pi) - 1| \leq \delta$. As we will show in the following section, many realistic data distributions fall into this category, even though they are non-Gaussian.

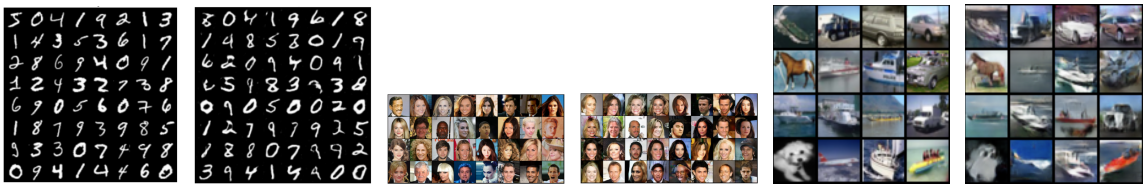
In this section of the main paper, we will give brief overviews with some intuiton of the following idea: for two distributions π_1, π_2 with low excess kurtosis, equating their second moments becomes a fairly stronger condition. Thus, if ΔR is minimized on distributions with low excess kurtosis, even if they are not necessarily Gaussians, this minimization will force the distributions to converge.

The first and most obvious note is the following, immediately following from the triangle inequality:

Lemma 2.1.
Denote $B_\delta \subset \mathcal{D}(\mathbb{R}^d)$ the set of distributions such that $|k_{adjusted}(\pi) - 1| \leq \delta$. It follows that $|k_{adjusted}(\pi_1) - k_{adjusted}(\pi_2)| \leq 2\delta$ for all $\pi_1, \pi_2 \in B_\delta$.

Thus, B_δ does act like a δ-ball and forces member distributions to be close together. The main question to ask now is if closeness in $k_{adjusted}$ means closeness in anything stronger or more interpretable. The answer here depends on how many more assumptions we can make on the distributions π_1, π_2:

1. Without any additional assumptions, we can assume the 4th moment of the distribution of 2-norm for each of π_1, π_2 differs by at most $2\delta C$, where C is a dimension-dependent constant. Using the theory of *cumulants* [2], we can conclude the following mouthful: the 4th order Taylor series of the log characteristic function $\log(\mathrm{E}_\pi\, e^{tX})$, where X is the 2-norm of the random variable distributed by π, differs by at most 2δ. Intuitively, the radial aggregate distribution of π_1, π_2 are similar up to a 4th order, rather than just the 2nd order similarity that equating second moments gives.



(a) MNIST X (b) MNIST $\hat{X}$ (c) Celeb-A X (d) Celeb-A $\hat{X}$ (e) CIFAR-10 X (f) CIFAR-10 $\hat{X}$

Figure 1: Visualizing the auto-encoding property of the learned ($\approx \hat{x} = g \circ f(x)$) on MNIST, Celeb-A and CIFAR-10 (zoom in for better visualization).

2. For further conclusions on how close π_1, π_2 are, we need further assumptions on their regularity. For example, if we assume the distributions of π_1, π_2 don't fluctuate much, then we can establish a scaled L^1 bound $\int_{\mathbb{R}^d} \|x\|_4^4 |\pi_1 - \pi_2|(dx) \leq C\delta$, where C is a dimension-dependent constant.

The end result for this section: if our distributions for $\boldsymbol{Z}, \hat{\boldsymbol{Z}}$ both have low excess kurtosis (which we measured is true for our image distributions), and are well-behaved in another intuitive manner, then forcing $\Delta(\boldsymbol{Z}, \hat{\boldsymbol{Z}}) = 0$ imposes π_1, π_2 to be close to each other beyond their second moments.

2.3 Simulation results

For this section, we conduct experiments trying to demonstrate that the distribution of the selected dataset have kurtosis very close to Gaussian. We selected MNIST[12], Celeb-A[13] and CIFAR-10[10] datasets. LDR has performed extremely well on both the first two datasets and slightly worse on CIFAR-10. From Dai's LDR Transcription paper, we see that both MNIST and Celeb-A have great reconstruction results as shown in figure 1 only using ΔR terms. The reconstructed image looks very "real" comparing to the input data. While for CIFAR-10, although we see sample wise alignment but deviates a little in places like color and less accurate. For the simulation, we select one class from each dataset because in the ΔR term, we calculate it based on class information. All tensors in the dataset are flattened to vectors and then calculate their kurtosis and adjusted kurtosis respectively.

Dataset	Kurtosis	Gaussian Kurtosis	Adjusted Kurtosis
Celeb-A [13]	9444769	9443320	1.001
MNIST[12]	133130	123200	1.080
CIFAR-10[10]	336591	251000	1.341

Table 1: Kurtosis Result for MNIST, Celeb-A and CIFAR-10 dataset

From table 1, we see that after adjusting the kurtosis to adjusted kurtosis, both Celeb-A and MNIST are very close to 1, which indicates that both datasets are very close to Gaussian and contains most of their information in the second order. This sets a very good foundation for the analysis in the next few sections, which shows how $\Delta(R)$ can be very good objective function for these datasets. However, for the CIFAR-10 Dataset, we see that it has a much larger adjusted Kurtosis. This statistically indicates that it deviates further from the Gaussian.

2.4 Simple neural networks preserve kurtosis

In this section, we will discuss how low excess kurtosis in the dateset(i.e X) gets transferred to lower-dimensional latent space(i.e Z). In most of the simple nerual networks include the one which LDR has performed the best - the Deep Convolutional Generative Adversarial Network(DCGAN) Network [16], network are consist of convolutional layers and batch norm layers, as shown in table 2.
Next we are going to prove that most parts in convolutional layers, which are fundamentally combination of linear transformation and batch norm layers preserve the kurtosis of the original data.

Theorem 2.2. *Suppose for a set of data X, with A as a set of invertible linear transformation. Let $k(X)$ denote the kurtosis of for any given set of data X. Then $k(X) = k(AX)$*

Proof. For convenience, we let $x = [x|1]$, $x' = [x'|1]$. In this case x = Ax+b can be written as x = Ax.

$$S = \sum (x_i - \hat{x})(x_i - \bar{x})^T \tag{6}$$

Gray image $\in \mathbb{R}^{32\times32\times1}$
4×4, stride=2, pad=1 conv 64 lReLU
4×4, stride=2, pad=1 conv. BN 128 lReLU
4×4, stride=2, pad=1 conv. BN 256 lReLU
4×4, stride=1, pad=0 conv 128

Table 2: Encoder for MNIST.

$$k = \frac{1}{n}\sum (x_i - \bar{x})^T S^{-1}(x_i - \bar{x}))^2 \tag{7}$$

since $x_i' = Ax$, $\hat{x'} = A\hat{x}$, Thus
$S' = \sum (x_i' - \hat{x_i'})(x_i' - \hat{x_i'})^T = \sum A(x_i - \hat{x_i})(x_i - \hat{x_i})^T A^T = A\sum (x_i - \hat{x_i})(x_i - \hat{x_i})^T A^T = ASA^T$
Thus, we have:

$$k' = \frac{1}{n}\sum (x_i' - \bar{x'})^T S'^{-1}(x_i' - \bar{x'}))^2 = \frac{1}{n}\sum (A(x_i - \bar{x_i}))^T S^{-1}(A(x_i - \bar{x_i})) \tag{8}$$

$$= \frac{1}{n}\sum ((x_i - \bar{x_i})^T A^T A^{-T} S^{-1} A^{-1} A(x_i - \bar{x_i}))^2 = \frac{1}{n}\sum ((x_i - \bar{x_i})^T S^{-1}(x_i - \bar{x_i}))^2 = k \tag{9}$$

□

Theorem 2.3. *Suppose for a set of data X, after a set of Batch Normalization, X'. Let $k(X)$ denote the kurtosis of for any given set of data X. Then $k(X) = k(X')$*

Proof. Once again, we define $\bar{x} = \frac{1}{n}\sum x_i$, the covariance of the Data X is:

$$S = \sum (x_i - \hat{x})(x_i - \bar{x})^T \tag{10}$$

and the kurtosis, followed by definition is:

$$k = \frac{1}{n}\sum (x_i - \bar{x})^T S^{-1}(x_i - \bar{x}))^2 \tag{11}$$

Since for each $x_i' = \frac{x_i - E[X]}{sd[X]}$ $S' = \sum (x_i' - \hat{x_i'})(x_i' - \hat{x_i'})^T = \sum x_i' x_i'^T = \sigma^2 S$ We have kurtosis being:

$$k' = \frac{1}{n}\sum (x_i' - \bar{x'})^T S'^{-1}(x_i' - \bar{x'}))^2 \tag{12}$$

$$= \frac{1}{n}\sum (x_i)^T \sigma^2 S^{-1}(x_i'))^2 \tag{13}$$

$$= \frac{1}{n}\sum (\frac{x_i - \bar{x_i}}{\sigma})^T \sigma^2 S^{-1} (\frac{x_i - \bar{x_i}}{\sigma})^2 \tag{14}$$

$$= \frac{1}{n}\sum (x_i - \bar{x_i})^T S^{-1}(x_i - \bar{x_i})^2 = k \tag{15}$$

□

3 Training Stability

A common problem for GAN training is stability [5], where small perturbations in the input parameters can lead to large deviations in the loss and thus lead to instability.

In this section, we show how the structure of $\Delta R(\boldsymbol{Z}, \hat{\boldsymbol{Z}})$ naturally leads to stability at critical points. In our analysis, we consider simple linear models: $\boldsymbol{Z} = \boldsymbol{FX}$, $\hat{\boldsymbol{Z}} = \boldsymbol{FGFX}$. While this is a simplification, we believe it is one that does more to highlight the true stability of ΔR. Increasing the complexity of f_θ, g_η will inevitably result in decreased stability of training ΔR: what we care about is how much additional instability training on ΔR adds on top of the instability of f_θ, g_η, which we now show is not much.

3.1 Linear Perturbation Bounds

Given $\boldsymbol{X} \in \mathbb{R}^D$, we define $\mathcal{T}^b_{\boldsymbol{X}}(\boldsymbol{F}, \boldsymbol{G}) := \Delta R(\boldsymbol{FX}, \boldsymbol{FGFX})$, the objective function we wish to train on. In this section, we give bounds on the perturbations $\mathcal{T}^b_{\boldsymbol{X}}(\boldsymbol{F}+\boldsymbol{M}, \boldsymbol{G})$ and $\mathcal{T}^b_{\boldsymbol{X}}(\boldsymbol{F}, \boldsymbol{G})$ where $(\boldsymbol{F}, \boldsymbol{G})$ are such that $\mathcal{T}^b_{\boldsymbol{X}}(\boldsymbol{F}, \boldsymbol{G}) = 0$ and M has a bounded norm $\|M\|_F \leq \delta$ for some fixed δ.

In order to do so, we will require the following facts:

Lemma 3.1. *For a positive definite square matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$,*

$$\operatorname{trace}(\boldsymbol{I}_d - \boldsymbol{X}^{-1}) \leq \log\det \boldsymbol{X} \leq \operatorname{trace}(\boldsymbol{X} - \boldsymbol{I}_d)$$

Lemma 3.2. *For a positive semidefinite matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$, $\log\det(\boldsymbol{I}_d + \boldsymbol{X}) \geq 0$.*

The proofs of the above lemmas are included in the Appendix.

Theorem 3.3. *Given $\boldsymbol{X} \in \mathbb{R}^D$, let $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ satisfy $\|\boldsymbol{M}\|_F \leq \delta < 1$ and $(\boldsymbol{F}, \boldsymbol{G})$ be so that $\boldsymbol{ZZ}^* = \hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^*$ where $\boldsymbol{Z} = \boldsymbol{FX}$, $\hat{\boldsymbol{Z}} = \boldsymbol{FGFX}$. Then,*

$$\mathcal{T}^b_{\boldsymbol{X}}(\boldsymbol{F}, \boldsymbol{G}+\boldsymbol{M}) \leq \frac{\gamma\|\boldsymbol{FX}\|_F^2}{2}\left(1 + \delta\|\boldsymbol{F}\|_F^2\left(1 + \frac{\|\boldsymbol{G}\|_F}{2}\right)\right).$$

Proof. First, note that by Lemma 3.2,

$$\begin{aligned}
\mathcal{T}^b_{\boldsymbol{X}}(\boldsymbol{F}, \boldsymbol{G}+\boldsymbol{M}) &= \Delta R(\boldsymbol{Z}, \boldsymbol{F}(\boldsymbol{G}+\boldsymbol{M})\boldsymbol{Z}) \\
&= \frac{1}{2}\log\det\left(\boldsymbol{I} + \frac{\gamma}{2}(\boldsymbol{ZZ}^* + \boldsymbol{F}(\boldsymbol{G}+\boldsymbol{M})\boldsymbol{Z}(\boldsymbol{F}(\boldsymbol{G}+\boldsymbol{M})\boldsymbol{Z})^*)\right) \\
&\quad - \frac{1}{4}\log\det\left(\boldsymbol{I} + \gamma\boldsymbol{ZZ}^*\right) \\
&\quad - \frac{1}{4}\log\det\left(\boldsymbol{I} + \gamma(\boldsymbol{F}(\boldsymbol{G}+\boldsymbol{M})\boldsymbol{Z}(\boldsymbol{F}(\boldsymbol{G}+\boldsymbol{M})\boldsymbol{Z})^*)\right) \\
&\leq \frac{1}{2}\log\det\left(\boldsymbol{I} + \frac{\gamma}{2}(\boldsymbol{ZZ}^* + \boldsymbol{F}(\boldsymbol{G}+\boldsymbol{M})\boldsymbol{Z}(\boldsymbol{F}(\boldsymbol{G}+\boldsymbol{M})\boldsymbol{Z})^*)\right)
\end{aligned}$$

Then, note that we can write

$$\begin{aligned}
\boldsymbol{F}(\boldsymbol{G}+\boldsymbol{M})\boldsymbol{Z}(\boldsymbol{F}(\boldsymbol{G}+\boldsymbol{M})\boldsymbol{Z})^* &= (\boldsymbol{FGZ} + \boldsymbol{FMZ})(\boldsymbol{FGZ} + \boldsymbol{FMZ})^* \\
&= (\hat{\boldsymbol{Z}} + \boldsymbol{FMZ})(\hat{\boldsymbol{Z}} + \boldsymbol{FMZ})^* \\
&= \hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^* + \hat{\boldsymbol{Z}}(\boldsymbol{FMZ})^* + \boldsymbol{FMZ}\hat{\boldsymbol{Z}}^* + \boldsymbol{FMZ}(\boldsymbol{FMZ})^* \\
&= \boldsymbol{ZZ}^* + \boldsymbol{FGZ}(\boldsymbol{FMZ})^* + \boldsymbol{FMZ}(\boldsymbol{FGZ})^* + \boldsymbol{FMZ}(\boldsymbol{FMZ})^*.
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathcal{T}^b_{\boldsymbol{X}}(\boldsymbol{F}, \boldsymbol{G}+\boldsymbol{M}) &\leq \frac{1}{2}\log\det\left(\boldsymbol{I} + \frac{\gamma}{2}(\boldsymbol{ZZ}^* + \boldsymbol{F}(\boldsymbol{G}+\boldsymbol{M})\boldsymbol{Z}(\boldsymbol{F}(\boldsymbol{G}+\boldsymbol{M})\boldsymbol{Z})^*)\right) \\
&= \frac{1}{2}\log\det\left(I + \gamma\boldsymbol{ZZ}^* + \frac{\gamma}{2}(\boldsymbol{FGZ}(\boldsymbol{FMZ})^* + \boldsymbol{FMZ}(\boldsymbol{FGZ})^* + \boldsymbol{FMZ}(\boldsymbol{FMZ}))^*\right).
\end{aligned}$$

By Lemma 3.1, the right-hand side is bounded by

$$\begin{aligned}
&\frac{1}{2}\operatorname{trace}\left(\gamma\boldsymbol{ZZ}^* + \frac{\gamma}{2}(\boldsymbol{FGZ}(\boldsymbol{FMZ})^* + \boldsymbol{FMZ}(\boldsymbol{FGZ})^* + \boldsymbol{FMZ}(\boldsymbol{FMZ})^*)\right) \\
&= \frac{\gamma}{2}\operatorname{trace}(\boldsymbol{ZZ}^*) + \frac{\gamma}{4}\operatorname{trace}(\boldsymbol{FGZ}(\boldsymbol{FMZ})^* + \boldsymbol{FMZ}(\boldsymbol{FGZ})^* + \boldsymbol{FMZ}(\boldsymbol{FMZ})^*) \\
&= \frac{\gamma}{2}\|\boldsymbol{Z}\|_F^2 + \frac{\gamma}{4}\operatorname{trace}(\boldsymbol{FGZ}(\boldsymbol{FMZ})^* + \boldsymbol{FMZ}(\boldsymbol{FGZ})^*) + \frac{\gamma}{4}\|\boldsymbol{FMZ}\|_F^2 \\
&= \frac{\gamma}{2}\|\boldsymbol{Z}\|_F^2 + \frac{\gamma}{4}(\langle\boldsymbol{FMZ}, \boldsymbol{FGZ}\rangle_F + \langle\boldsymbol{FGZ}, \boldsymbol{FMZ}\rangle_F) + \frac{\gamma}{4}\|\boldsymbol{FMZ}\|_F^2 \\
&\leq \frac{\gamma}{2}\|\boldsymbol{Z}\|_F^2 + \frac{\gamma}{2}\|\boldsymbol{FMZ}\|_F\|\boldsymbol{FGZ}\|_F + \frac{\gamma}{4}\|\boldsymbol{FMZ}\|_F^2 \\
&\leq \frac{\gamma}{2}\|\boldsymbol{Z}\|_F^2 + \frac{\gamma}{2}\|\boldsymbol{F}\|_F^2\|\boldsymbol{G}\|_F\|\boldsymbol{M}\|_F\|\boldsymbol{Z}\|_F^2 + \frac{\gamma}{4}\|\boldsymbol{F}\|_F^2\|\boldsymbol{M}\|^2\|\boldsymbol{Z}\|_F^2
\end{aligned}$$

where the last two lines follow from repeated applications of the Cauchy-Schwarz inequality.

Finally, since $\|\boldsymbol{M}\|_F < 1$, we have $\|\boldsymbol{M}\|_F^2 \leq \|\boldsymbol{M}\|_F \leq \delta$, it follows that

$$\begin{aligned}\frac{\gamma}{2}\|\boldsymbol{Z}\|_F^2 + \frac{\gamma}{2}\|\boldsymbol{F}\|_F^2\|\boldsymbol{G}\|_F\|\boldsymbol{M}\|_F\|\boldsymbol{Z}\|_F^2 + \frac{\gamma}{4}\|\boldsymbol{F}\|_F^2\|\boldsymbol{M}\|^2\|\boldsymbol{Z}\|_F^2 &= \frac{\gamma}{2}\|\boldsymbol{Z}\|_F^2\left(1 + \|\boldsymbol{F}\|_F^2\|\boldsymbol{G}\|_F\|\boldsymbol{M}\|_F + \frac{1}{2}\|\boldsymbol{F}\|_F^2\|\boldsymbol{M}\|^2\right) \\ &\leq \frac{\gamma}{2}\|\boldsymbol{Z}\|_F^2\left(1 + \delta\|\boldsymbol{F}\|_F^2\|\boldsymbol{G}\|_F + \frac{\delta}{2}\|\boldsymbol{F}\|_F^2\right) \\ &= \frac{\gamma\|\boldsymbol{F}\boldsymbol{X}\|_F^2}{2}\left(1 + \delta\|\boldsymbol{F}\|_F^2\left(1 + \frac{\|\boldsymbol{G}\|_F}{2}\right)\right)\end{aligned}$$

□

Following similar analysis in the $\boldsymbol{F}$ component, we obtain a corresponding theorem which is proved in the appendix:

Theorem 3.4. *Given $\boldsymbol{X} \in \mathbb{R}^D$, let $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ satisfy $\|\boldsymbol{M}\|_F \leq \delta < 1$ and $(\boldsymbol{F}, \boldsymbol{G})$ be so that $\boldsymbol{Z}\boldsymbol{Z}^* = \hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^*$ where $\boldsymbol{Z} = \boldsymbol{F}\boldsymbol{X}$, $\hat{\boldsymbol{Z}} = \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}$. Then,*

$$\mathcal{T}_{\boldsymbol{X}}^b(\boldsymbol{F} + \boldsymbol{M}, \boldsymbol{G}) \leq \frac{\gamma\|\boldsymbol{X}\|_F^2\|\boldsymbol{F}\|_F^2}{2}\left(1 + \frac{\delta\|\boldsymbol{G}\|_F^2\|\boldsymbol{F}\|_F}{2}\right).$$

Remark 3.5. By restricting to the case where $\|\boldsymbol{F}\|_F \leq 1$, $\|\boldsymbol{G}\|_F \leq 1$, $\|\boldsymbol{X}\|_F \leq 1$, we obtain the simpler bounds

$$\begin{aligned}\mathcal{T}_{\boldsymbol{X}}^b(\boldsymbol{F} + \boldsymbol{M}, \boldsymbol{G}) &\leq \frac{\gamma(2+\delta)}{4}, \\ \mathcal{T}_{\boldsymbol{X}}^b(\boldsymbol{F}, \boldsymbol{G} + \boldsymbol{M}) &\leq \frac{\gamma(1+3\delta)}{4}.\end{aligned}$$

4 Conclusions & future directions

In this project, we have provided theoretical insights into two ideal properties - efficiency and stability - by using eq. (1) as a measurement for the distance between distributions. We introduce kurtosis to measure how close a data is to the Gaussian Distribution. showed that simple neural networks mostly preserve the kurtosis in the dataset. Furthermore, by studying the $\mathcal{T}_X^b$ objective in the case of linear maps, we also provided an upper bound on the error when perturbing each input by a small amount. When normalizing our maps and data, we obtain bounds that are linear in the perturbation amount.

Of course, further researches and derivations are required for more complex neural networks such as the analysis on the Relu/Leaky-Relu layer's impact on kurtosis. While we chose kurtosis due to its historical significance with determining how Gaussian a distribution, there are likely better, less studied metrics out there to use in its place: our analysis highlights some possibilities for these. Additionally, further study is required to obtain tighter bounds on the linear perturbation errors.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[2] E. A. Cornish and R. A. Fisher. Moments and cumulants in the specification of distributions. *Revue de l'Institut international de Statistique*, pages 307–320, 1938.

[3] X. Dai, S. Tong, M. Li, Z. Wu, K. H. R. Chan, P. Zhai, Y. Yu, M. Psenka, X. Yuan, H. Y. Shum, and Y. Ma. Closed-loop data transcription to an ldr via minimaxing rate reduction, 2021.

[4] L. T. DeCarlo. On the meaning and use of kurtosis. *Psychological Methods*, 2:292–307, 1997.

[5] F. Farnia and A. Ozdaglar. Gans may have no nash equilibria. *arXiv preprint arXiv:2002.09124*, 2020.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[7] G. E. Hinton and R. S. Zemel. Autoencoders, minimum description length, and helmholtz free energy. *Advances in neural information processing systems*, 6:3–10, 1994.

[8] D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2014.

[9] M. A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991.

[10] A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research).

[11] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[12] Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[13] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[14] Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE transactions on pattern analysis and machine intelligence*, 29(9):1546–1562, 2007.

[15] K. V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 1970.

[16] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

[17] L. N. Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

[18] Y. Yu, K. H. R. Chan, C. You, C. Song, and Y. Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *arXiv preprint arXiv:2006.08558*, 2020.

A Appendix

The primary purpose of the appendix is to provide proofs for the mathematical statements made above.

A.1 Kurtosis proofs

We first address an elephant in the room: the requirement of kurtosis for the covariance to be full-rank, which is certainly not true in general. However, we can indeed generalized kurtosis to rank-deficient distributions:

Theorem A.1. *$k(\pi)$ is well-defined on distributinos with singular covariance matrix Σ*

Proof. Consider the truncated SVD $\Sigma = U_r S_r U_r^\top$, where we truncate off all 0 singular value terms. Then $x \to U_r^\top$ is an isometric transformation on the support of π, and the resulting covariance matrix $\Sigma' = U_r^\top \Sigma U_r$ is invertible, and we can measure the kurtosis $k(\pi)$ of the resulting distribution. □

Theorem A.2. *Denote $k(X)$ the estimated kurtosis as given in eq.* (3) *with a finite sample set $X \in \mathbb{R}^{d\times n}$. Then $k(X) = k(AX + b)$ for any invertible matrix $A \in GL(d)$ and any vector $b \in \mathbb{R}^d$.*

Proof. $k(X) = k(X + b)$ is clear from the fact we subtract the empirical mean from X to generate $\bar{X} = X - \frac{1}{n}\sum_{i=1}^n x_i = X + b - \frac{1}{n}\sum_{i=1}^n (x_i + b)$. Thus, we only need to show $k(X) = k(AX)$. Note the effect of applying an invertible transformation on the covariance matrix:

$$S_{AX} = \sum (Ax_i - A\bar{x})(Ax_i - A\bar{x})^\top = ASA^\top \tag{16}$$

, where $S := \sum (x_i - \hat{x})(x_i - \hat{x})^T$. Thus, we get the following for the kurtosis:

$$k(AX) = \frac{1}{n}\sum (Ax_i - A\bar{x})^T (ASA^\top)^{-1}(Ax_i - A\hat{x}))^2, \tag{17}$$

$$= \frac{1}{n}\sum ((x_i - \hat{x_i})A^\top A^{-\top} S^{-1} A^{-1} A(x_i - \hat{x_i}))^2, \tag{18}$$

$$= \frac{1}{n}\sum ((x_i - \hat{x_i})S^{-1}(x_i - \hat{x_i}))^2 = k(X). \tag{19}$$

□

We can prove the corresponding statement for the true kurtosis of continuous distributions:

Theorem A.3. *Denote $k(\pi)$ the kurtosis as given in eq.* (4) *with respect to the distribution $\pi \in \mathcal{D}(\mathbb{R}^d)$ such that $\Sigma_\pi := \mathrm{E}_\pi (x - \bar{x})(x - \bar{x})^\top$ is invertible. Then $k(\pi)$ is invariant under affine invertible linear transformations.*

Proof. We again reduce to the non-affine case since only the centered distribution is used, which is shift-invariant. For any invertible matrix A, the operation of multiplying a random variable sampled from π by A yields the distribution $\pi_A(dx) = \det(A)\pi(A^{-1}dx)$. Using the fact that $\int_{\mathbb{R}^d} f(x)\pi(A^{-1}x) = \frac{1}{\det(A)}\int_{\mathbb{R}^d} f(Ax)\pi(x)$ and $S_{\pi_A} = ASA^\top$, the proof follows exactly as in Theorem A.2. □

Theorem A.4. *Let $\pi = \mathcal{N}(\mu, \Sigma)$ be a Gaussian over $\mathbb{R}^d$, where Σ is invertible. Then the kurtosis $k(\pi) = d(d+2)$.*

Proof. Since we have shown that kurtosis is invariant to affine linear transformations in Theorem A.3, we can assume without loss of generality that $\pi = \mathcal{N}(0, I)$. First, we make the following simplifications:

$$k(\pi) = \mathrm{E}_\pi (x^\top x)^2 = \int_{\mathbb{R}^d} (x^\top x)^2 \pi(dx), \tag{20}$$

$$= \int_{\mathbb{R}^d} x^\top x x^\top x \pi(dx), \tag{21}$$

$$= \int_{\mathbb{R}^d} \mathrm{trace}(x^\top x x^\top x) \pi(dx), \tag{22}$$

$$= \int_{\mathbb{R}^d} \mathrm{trace}(x x^\top x x^\top) \pi(dx), \tag{23}$$

$$= \mathrm{trace}\left(\int_{\mathbb{R}^d} x x^\top x x^\top \pi(dx) \right), \tag{24}$$

$$= \mathrm{trace}\left(\int_{\mathbb{R}^d} x x^\top \|x\|_2^2 \pi(dx) \right). \tag{25}$$

We now show that $K := \int_{\mathbb{R}^d} x x^\top \|x\|_2^2 \pi(dx) = (d+2) I$, and thus $k(\pi) = \mathrm{trace}(K) = d(d+2)$. We start by showing the off-diagonal terms are 0:

$$K_{ij} = \int_{\mathbb{R}^d} x_i x_j \|x\|_2^2 \pi(dx), \tag{26}$$

$$= \int_{\mathbb{R}^d} x_i x_j (\sum_{k=1}^n x_k^2) \pi(dx), \tag{27}$$

$$= \int_{\mathbb{R}^d} x_i^3 x_j \pi(dx) + \int_{\mathbb{R}^d} x_i x_j^3 \pi(dx) + \int_{\mathbb{R}^d} x_i x_j (\sum_{k \neq i,j} x_k^2) \pi(dx), \tag{28}$$

$$= \int_{\mathbb{R}^d} x_i^3 \pi(dx) (\int_{\mathbb{R}^d} x_j \pi(dx)) + \int_{\mathbb{R}^d} x_i \pi(dx) (\int_{\mathbb{R}^d} x_j^3 \pi(dx)) + \int_{\mathbb{R}^d} x_i \pi(dx) (\int_{\mathbb{R}^d} x_j \pi(dx)) (\int_{\mathbb{R}^d} (\sum_{k \neq i,j} x_k^2 \pi(dx)), \tag{29}$$

$$= 0 + 0 + 0 = 0, \tag{30}$$

since $\mathrm{E}_\pi f(x) g(x) = \mathrm{E}_\pi f(x) \, \mathrm{E}_\pi g(x)$ if $f(x), g(x)$ independent, and $\mathrm{E}_\pi x = 0$. Now for the diagonal terms:

$$K_{ii} = \int_{\mathbb{R}^d} x_i^2 \|x\|_2^2 \pi(dx), \tag{31}$$

$$= \int_{\mathbb{R}^d} x_i^2 (\sum_{k=1}^n x_k^2) \pi(dx), \tag{32}$$

$$= \int_{\mathbb{R}^d} x_i^4 \pi(dx) + \int_{\mathbb{R}^d} x_i^2 (\sum_{k \neq i} \pi(dx) x_k^2), \tag{33}$$

$$= 3 + \int_{\mathbb{R}^d} x_i^2 \pi(dx) (\int_{\mathbb{R}^d} (\sum_{k \neq i} x_k^2) \pi(dx)), \tag{34}$$

$$= 3 + \int_{\mathbb{R}^d} (\sum_{k \neq i} x_k^2) \pi(dx), \tag{35}$$

$$= 3 + d - 1 = d + 2, \tag{36}$$

where (34) follows since we know the 4th moment of a scalar normal distribution is 3, (35) follows from unit variance of a scalar normal distribution, and (36) follows from the fact that for a $d-1$ dimensional normal distribution $\pi_{d-1} = \mathcal{N}(0, I) \in \mathcal{D}(\mathbb{R}^{d-1})$:

$$\int_{\mathbb{R}^{d-1}} \|x\|_2^2 \pi(dx) = \int_{\mathbb{R}^{d-1}} x^\top x \pi(dx), \tag{37}$$

$$= \int_{\mathbb{R}^{d-1}} \text{trace}(x^\top x)\pi(dx), \tag{38}$$

$$= \int_{\mathbb{R}^{d-1}} \text{trace}(xx^\top)\pi(dx), \tag{39}$$

$$= \text{trace}(\int_{\mathbb{R}^{d-1}} xx^\top \pi(dx)), \tag{40}$$

$$= \text{trace}(I) = d - 1. \tag{41}$$

Thus, $K = (d+2)I$ and $k(\pi) = \text{trace}(K) = d(d+2)$. □

A.2 Gaussian-like distributions proofs

Recall the definition of B_δ in Lemma 2.1. We show if the covariance matrices are normalized, then Gaussian-like distributions have similar 4th moments.

Without loss of generality, we assume all distributions are centered.

Theorem A.5. *Suppose $\pi_1, \pi_2 \in B_\delta$. Denote Σ_1, Σ_2 the covariance matrices of π_1, π_2. If the operator norms $\|\Sigma_1\|, \|\Sigma_2\| \leq 1$, then $|\text{E}_{\pi_1}\|x\|_2^4 - \text{E}_{\pi_1}\|x\|_2^4| \leq 2\delta(d)(d+1)$*

Proof. This follows from the following chain of inequalities:

$$|k_{adjusted}(\pi_1) - k_{adjusted}(\pi_2)| \leq 2\delta, \tag{42}$$

$$\implies |\int_{\mathbb{R}^d} (x^\top \Sigma_1^{-1} x)^2 \pi_1(dx) - \int_{\mathbb{R}^d} (x^\top \Sigma_2^{-1} x)^2 \pi_2(dx)| \leq 2\delta(d(d+2)), \tag{43}$$

$$\implies |\int_{\mathbb{R}^d} (x^\top x)^2 \pi_1(dx) - \int_{\mathbb{R}^d} (x^\top x)^2 \pi_2(dx)| \leq 2\delta(d(d+2)), \tag{44}$$

$$\implies |\int_{\mathbb{R}^d} \|x\|_2^4 \pi_1(dx) - \int_{\mathbb{R}^d} \|x\|_2^4 \pi_2(dx)| \leq 2\delta(d(d+2)), \tag{45}$$

which follows since $\|\Sigma x\| \leq 1 \forall x \implies \|\Sigma^{-1} x\| \geq 1 \forall x$, and further by monotonicity of $(\cdot)^2$. □

The above theorem gives the explicit connection of boudned excess kurtosis to similar 4th moments of the radial distribution. To see what this concludes about the characteristic functions, and namely similarity in 4th order Taylor polynomial expansions, we refer the reader to [2].

To make stronger assumptions, we need stronger assumptions on the distributions. One looser statement is the following:

Theorem A.6. *Suppose $\pi_1, \pi_2 \in B_\delta$ and both densities don't cross each other much: that is, either the Lebesgue measure of $\{x : \pi_1(x) \geq \pi_2(x)\}$ or $\{x : \pi_1(x) \leq \pi_2(x)\}$ is incredibly small. Then $\int_{\mathbb{R}^d} \|x\|_2^4 |\pi_1 - \pi_2|(dx) \lessapprox 2\delta(d)(d+1)$.*

Proof. From Theorem A.5, we know that $|\int_{\mathbb{R}^d} \|x\|_2^4 \pi_1(dx) - \int_{\mathbb{R}^d} \|x\|_2^4 \pi_2(dx)| \leq 2\delta(d(d+2))$. We can decompose into the following, defining the measurable set $A := \{x \mid \pi_1(x) \leq \pi_2(x)\}$ and assuming that the measure of A^c is incredibly small:

$$|\int_{\mathbb{R}^d} \|x\|_2^4 \pi_1(dx) - \int_{\mathbb{R}^d} \|x\|_2^4 \pi_2(dx)| = |\int_A \|x\|_2^4 (\pi_1 - \pi_2)(dx) + \int_{A^c} \|x\|_2^4 (\pi_1 - \pi_2)(dx)|, \tag{46}$$

$$\approx |\int_A \|x\|_2^4 (\pi_1 - \pi_2)(dx)|, \tag{47}$$

$$= \int_A \|x\|_2^4 |(\pi_1 - \pi_2)|(dx) \approx \int_{\mathbb{R}^d} \|x\|_2^4 |(\pi_1 - \pi_2)|(dx). \tag{48}$$

Thus, $\int_{\mathbb{R}^d} \|x\|_2^4 |\pi_1 - \pi_2|(dx) \lessapprox 2\delta(d)(d+1)$. □

Note that there was important hand-waving done here with the approximations, and we were not rigorous with the specific bounds. We aim in the future to improve these bounds; here, we simply want to show that with reasonable assumptions on the distributions π_1, π_2, we can make better bounds for how different π_1, π_2 can be overall.

A.3 Stability Proofs

Lemma A.7 (Lemma 3.1)**.** *For a positive definite square matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$,*

$$\operatorname{trace}(\boldsymbol{I}_d - \boldsymbol{X}^{-1}) \leq \log\det \boldsymbol{X} \leq \operatorname{trace}(\boldsymbol{X} - \boldsymbol{I}_d)$$

Proof. Let $\lambda_i(\boldsymbol{X})$ denote the eigenvalues of $\boldsymbol{X}$. We have

$$\begin{aligned}\log\det \boldsymbol{X} &= \log\left(\prod_{i=1}^{d} \lambda_i(\boldsymbol{X})\right)\\ &= \sum_{i=1}^{d} \log \lambda_i(\boldsymbol{X})\end{aligned}$$

Setting $0 < \lambda_i(X) = e^{\gamma_i}$, we have

$$\begin{aligned}\sum_{i=1}^{d} \log \lambda_i(\boldsymbol{X}) &= \sum_{i=1}^{d} \gamma_i\\ &\leq \sum_{i=1}^{d} e^{\gamma_i} - 1\\ &= \sum_{i=1}^{d} \lambda_i - 1\\ &= \operatorname{trace}(\boldsymbol{X} - \boldsymbol{I}_d)\end{aligned}$$

by the convexity of e^x. Similarly, we have

$$\begin{aligned}\sum_{i=1}^{d} \log \lambda_i(\boldsymbol{X}) &= \sum_{i=1}^{d} \gamma_i\\ &\geq \sum_{i=1}^{d} 1 - e^{-\gamma_i}\\ &= \sum_{i=1}^{d} 1 - \frac{1}{\lambda_i}\\ &= \operatorname{trace}(\boldsymbol{I}_d - \boldsymbol{X}^{-1})\end{aligned}$$

□

Lemma A.8 (Lemma 3.2)**.** *For a positive semidefinite matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$, $\log\det(\boldsymbol{I}_d + \boldsymbol{X}) \geq 0$.*

Proof. By Proposition 3.1, we have

$$\begin{aligned}\log\det(\boldsymbol{I}_d + \boldsymbol{X}) &\geq \operatorname{trace}(\boldsymbol{I}_d - (\boldsymbol{I}_d + \boldsymbol{X})^{-1})\\ &= d - \operatorname{trace}((I + X)^{-1})\\ &= d - \sum_{i=1}^{d} \frac{1}{1 + \lambda_i(X)}\\ &\geq d - \sum_{i=1}^{d} 1\\ &= 0.\end{aligned}$$

□

Theorem A.9 (Theorem 3.4)**.** *Given $\boldsymbol{X} \in \mathbb{R}^D$, let $\boldsymbol{M} \in \mathbb{R}^{D\times D}$ satisfy $\|\boldsymbol{M}\|_F \leq \delta < 1$ and $(\boldsymbol{F}, \boldsymbol{G})$ be so that $\boldsymbol{Z}\boldsymbol{Z}^* = \hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^*$ where $\boldsymbol{Z} = \boldsymbol{F}\boldsymbol{X}$, $\hat{\boldsymbol{Z}} = \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}$. Then,*

$$\mathcal{T}^b_{\boldsymbol{X}}(\boldsymbol{F} + \boldsymbol{M}, \boldsymbol{G}) \leq \frac{\gamma\|\boldsymbol{X}\|_F^2\|\boldsymbol{F}\|_F^2}{2}\left(1 + \frac{\delta\|\boldsymbol{G}\|_F^2\|\boldsymbol{F}\|_F}{2}\right).$$

Proof. Note that

$$\begin{aligned}(\boldsymbol{F}+\boldsymbol{M})\boldsymbol{G}(\boldsymbol{F}+\boldsymbol{M})\boldsymbol{X}\boldsymbol{X}^*(\boldsymbol{F}+\boldsymbol{M})^*\boldsymbol{G}^*(\boldsymbol{F}+\boldsymbol{M})^* &= \hat{\boldsymbol{Z}}\hat{\boldsymbol{Z}}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{M}^* \\ &+ \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{M}^* \\ &+ \boldsymbol{F}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{M}^* \\ &+ \boldsymbol{F}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{M}^* \\ &+ \boldsymbol{M}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{M}^* \\ &+ \boldsymbol{M}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{M}^* \\ &+ \boldsymbol{M}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{M}^* \\ &+ \boldsymbol{M}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{M}^*\end{aligned}$$

By Lemma 3.2, it suffices to consider the first term of $\mathcal{T}^b_{\boldsymbol{X}}(\boldsymbol{F}+\boldsymbol{M},)$. Hence, we have

$$\begin{aligned}\mathcal{T}^b_{\boldsymbol{X}}(\boldsymbol{F}+\boldsymbol{M},) &\le \frac{1}{2}\log\det(\boldsymbol{I}+\frac{\gamma}{2}(2\boldsymbol{Z}\boldsymbol{Z}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{M}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{M}^* \\ &\quad + \boldsymbol{F}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{M}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{M}^* \\ &\quad + \boldsymbol{M}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{M}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{M}^* \\ &\quad + \boldsymbol{M}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{M}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{M}^*) \\ &\le \frac{\gamma}{4}\operatorname{trace}(2\boldsymbol{Z}\boldsymbol{Z}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{M}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{M}^* \\ &\quad + \boldsymbol{F}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{M}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{F}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{M}^* \\ &\quad + \boldsymbol{M}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{M}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{F}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{M}^* \\ &\quad + \boldsymbol{M}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{F}^*\boldsymbol{G}^*\boldsymbol{M}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{F}^* + \boldsymbol{M}\boldsymbol{G}\boldsymbol{M}\boldsymbol{X}\boldsymbol{X}^*\boldsymbol{M}^*\boldsymbol{G}^*\boldsymbol{M}^*) \\ &\le \frac{\gamma}{2}\|\boldsymbol{Z}\|_F^2 + \frac{\gamma}{4}\|\boldsymbol{G}\|_F^2\|\boldsymbol{X}\|_F^2(4\|\boldsymbol{F}\|_F^3\|\boldsymbol{M}\|_F + 6\|\boldsymbol{F}\|_F^2\|\boldsymbol{M}\|_F^2 + 4\|\boldsymbol{F}\|_F\|\boldsymbol{M}\|_F^3 + \|\boldsymbol{M}\|_F^4) \\ &\le \frac{\gamma\|\boldsymbol{X}\|_F^2}{2}\left(\|\boldsymbol{F}\|_F^2 + \frac{1}{2}\left(\|\boldsymbol{G}\|_F^2(\|\boldsymbol{F}\|_F+\delta)^4 - \|\boldsymbol{F}\|_F^4\right)\right) \\ &\le \frac{\gamma\|\boldsymbol{X}\|_F^2}{2}\left(\|\boldsymbol{F}\|_F^2 + \frac{1}{2}\delta\|\boldsymbol{G}\|_F^2\|\boldsymbol{F}\|_F^3\right) \\ &= \frac{\gamma\|\boldsymbol{X}\|_F^2\|\boldsymbol{F}\|_F^2}{2}\left(1 + \frac{\delta\|\boldsymbol{G}\|_F^2\|\boldsymbol{F}\|_F}{2}\right).\end{aligned}$$

□