Reforming Generative Autoencoders via Goodness-of-Fit Hypothesis Testing

Aaron Palmer
Computer Science Dept.
University of Connecticut
Storrs, CT 06269

Dipak K. Dey
Statistics Dept.
University of Connecticut
Storrs, CT 06269

Jinbo Bi
Computer Science Dept.
University of Connecticut
Storrs, CT 06269

Abstract

Generative models, while not new, have taken the deep learning field by storm. However, the widely used training methods have not exploited the substantial statistical literature concerning parametric distributional testing. Having sound theoretical foundations, these goodness-of-fit tests enable parts of the black box to be stripped away. In this paper we use the Shapiro-Wilk and propose a new multivariate generalization of Shapiro-Wilk to respectively test for univariate and multivariate normality of the code layer of a generative autoencoder. By replacing the discriminator in traditional deep models with the hypothesis tests, we gain several advantages: objectively evaluate whether the encoder is actually embedding data onto a normal manifold, accurately define when convergence happens, explicitly balance between reconstruction and encoding training. Not only does our method produce competitive results, but it does so in a fraction of the time. We highlight the fact that the hypothesis tests used in our model asymptotically lead to the same solution of the L_2-Wasserstein distance metrics used by several generative models today.

1 INTRODUCTION

Recently a large variety of generative models have been proposed such as generative adversarial networks and generative autoencoders. A widely-used way to construct such a network requires training of a generator and a discriminator. There are great needs to understand the statistical foundation of these generative models. On the other hand, there exists substantial statistical literature concerning parametric distributional hypothesis testing with a solid theoretical base. One particular group of deep generative models which, we show in this study, can benefit from hypothesis testing is the generative autoencoder (GAE). The objective of these models is to reconstruct the input as accurately as possible, while constraining the code layer to a specified distribution, usually normal. Often times once training has ended, this code layer distribution does not in fact match the required distribution. The spirit of these models is to embed data into a distribution that matches the prior to enable sampling, and thus it is of utmost importance we have ways to assess the quality of the fit. In other words, if the embedded distribution does not match the prior that is used to sample and generate instances, the method does not work in theory.

In this paper we propose to use goodness-of-fit hypothesis tests of normality on the code layer of an autoencoder as a new type of critic in both the univariate and multivariate case. Doing so leads to an adversary-free optimization problem. These hypothesis tests provide a more direct way to measure if the data representation, the latent code layer, matches a pre-specified distribution. More specifically, we test for normality using a composite test:

$$H_0 : \mathbf{X} \in \mathcal{G} \qquad \text{vs} \qquad H_1 : \mathbf{X} \notin \mathcal{G} \tag{1}$$

where $\mathcal{G} = \{\pi : \pi = \mathcal{N}(\mu, \mathbf{\Sigma}), -\infty < \mu < \infty, \mathbf{\Sigma}$ is positive semi-definite (p.s.d)$\}$. Many tests for comparing two distributions can be used in our model[1]. We specifically focus on the well studied univariate Shapiro-Wilk test (Shapiro and Wilk, 1965) and propose a new multivariate generalization of the Shapiro-Wilk test to demonstrate the effectiveness of the new method. We further highlight a link between these methods and those based on the Wasserstein distance by drawing attention to the fact that the Shapiro-Wilk test and the L_2-Wasserstein distance lead to the same asymptotic solution.

The remainder of the paper follows as such. Section 2 covers existing work that is most closely related to our method. In section 3 we present the basics of hypothesis testing followed by a recap of the Shapiro-Wilk test and propose its multivariate generalization. Section 4 describes the new method in detail, followed by theoretical analysis in Section 5 where we explore the linkage between the hypothesis tests and several distance-based methods of training. Section 6 discusses the empirical results. Section 7 presents a discussion of the new method, followed by a conclusion section 8.

Notation. Boldface capital letters, e.g., $\mathbf{Y}$, denote matrices while boldface lower case, e.g., $\mathbf{y}$, denote vectors. Scalar values are denoted by lower case letters and no font change, such as y_i. Upper case without font change denote test statistics. Calligraphic capital letters, e.g., $\mathcal{Y}$, denote sets. Probability density functions (PDF) are represented as $p(\mathbf{z})$, while cumulative distribution functions are the upper case version $P(\mathbf{z})$. Any modifications to this are explained in their respective context.

2 RELATED WORK

The original generative adversarial network (GAN) (Goodfellow et al., 2014) sparked a surge in generative models, parameterized by neural networks, that has yet to abate. As this paper is focused on autoencoders, GAN can be thought of as just the decoder part of a regular autoencoder. This decoder, $G(\cdot)$, endeavors to learn a mapping from the sampled prior, $p(\mathbf{z})$, to the data distribution $p(\mathbf{x})$. An auxillary network called the discriminator, $D(\cdot)$, serves to discern how close the generated data distribution, $p_g(\mathbf{z})$, is to the true data distribution $p(\mathbf{x})$. Training a GAN amounts to the widely known two-player minimax game $\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_d(\mathbf{x})}[logD(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[log(1-D(G(\mathbf{z})))]$.

Using an autoencoder styled network to train a generative model is not new; the main goal being to understand the code layer distribution given data, $q(\mathbf{z}|\mathbf{x})$, while minimizing a reconstruction error. These GAEs tend to fall into several classes dictated by their training and generating mechanisms, and include adversarial methods, variational methods, MCMC based procedures, and the most closely related to our work, statistical hypothesis tests.

The adversarial autoencoder (AAE) (Makhzani et al., 2015) is a modification of the original GAN in which an encoder network is included, and where the discriminator is shifted from the decoder network to the latent code space. The encoder creates the encoding distribution $q(\mathbf{z}|\mathbf{x})$ which defines an aggregated posterior distribution $q(\mathbf{z}) = \int_{\mathbf{x}} q(\mathbf{z}|\mathbf{x}) p_{data}(\mathbf{x}) d\mathbf{x}$ where $p_{data}(\mathbf{x})$ is the input data distribution. As in GAN, training the AAE, in part, amounts to a minimax game between a generator network, $G(\cdot)$, and a discriminator network, $D(\cdot)$ where the objective is to have $q(\mathbf{z})$ match $p(\mathbf{z})$, the specified prior distribution defined over the latent space $\mathcal{Z}$. The decoder maps back to data space $\mathcal{X}$ giving $p(\mathbf{x}|\mathbf{z})$. A reconstruction loss is minimizes as usual.

However, aside from possible mode collapse issues, there are questions regarding how the generator and discriminator should be balanced during training, the issue of when to stop still has not been satisfactorily addressed. Within the adversarially trained autoencoders it is possible to vary the loss function used in the discriminator. One possible change is to use a Wasserstein loss (Arjovsky et al., 2017) (WGAN) which alieviates the vanishing gradient problem. Improvements to the WGAN include the addition of a gradient penalty (Gulrajani et al., 2017). In (Tolstikhin et al., 2017) the authors modify the AAE to use a Wasserstein distance between the target distribution and the model distribution.

Variational autoencoder (VAE) methods include the work of (Kingma and Welling, 2013; Rezende et al., 2014; Mnih and Gregor, 2014). Despite being some of the most successful methods for generation, they have been found to produce unrealistic or blurry samples (Dosovitskiy and Brox, 2016). The VAE model makes use of a random decoder mapping $p(\mathbf{x}|\mathbf{z})$. Moreover, there is no auxillery network needed for discriminaion. A third line of thought comes from modification of the traditional autoencoder paradigm so as to recover the density using MCMC. These include (Rifai et al., 2012; Bengio et al., 2013b, 2014) and attempt to use contraction operators, or denoising criteria in order to generate a Markov chain by repeated perturbations during the encoding phase. However, it has been a challenge to ensure adequate mixing in that process (Bengio et al., 2013a).

To the best of our knowledge there is only one method, aside from our work, that falls into the class of statistical hypothesis tests for training generative networks. It is based on the maximum mean discrepancy (MMD) (Gretton et al., 2007, 2012). Two works utilizing the MMD for training came out simultaneously (Li et al., 2015) and (Dziugaite et al., 2015), each taking a different approach. Li et al. used the MMD on features learned from the autoencoder to shape the distribution of the output layer of the network to create a generative moment matching network (GMMN). On the other hand, Dziugaite et al. applied the MMD to directly compare the generated against true data. This latter method is the closest to our work. When using the MMD, the bandwidth parameter in the kernel plays a crucial role in determining the statistical efficiency of MMD, and it is still an open problem how to find its optimal value. Moreover, using kernels in MMD requires that the computation of the objective function scales quadratically with the amount of data. This is due to the requirement of a linear increase in sample size as dimensionality increases, and is necessary to ensure the power, covered next, goes to 1 as $n \to \infty$. In (Li et al., 2017) the authors propose to mitigate the bandwidth problem by using adversarial kernel learning to replace the fixed Gaussian kernel in the GMMN, while in (Sutherland et al., 2016) the authors propose to maximize the power of the statistical test based on the MMD.

3 STATISTICAL HYPOTHESIS TESTS

Distinguishing between two distributions is often carried out in the form of a hypothesis test. Suppose θ is a quantity of interest, the format of a hypothesis test between the null, H_0, and the alternative, H_1, is: $H_0 : \theta \in \mathbf{\Theta_0}$ vs $H_1 : \theta \in \mathbf{\Theta_0^c}$. To understand a hypothesis test the concept of statistical power is required. Two types of errors associated with hypothesis testing exist: type I, and type II. A type I error occurs when the null is rejected when it is true; the rate of this is called α. A type II error occurs when the null is not rejected when the alternative is true, its rate defined as β. Power is defined to be $1 - \beta$. It is not possible to control both type I and type II errors; therefore it is necessary to pre-specify the α one is willing to tolerate. The more powerful the test, the better. By utilizing information about the null distribution, a test statistic can be computed such that, for a given α, if its value is unlikely to be observed, then H_0 is rejected in favor of the alternative hypothesis' conclusion. This discerning threshold is called the critical value, and comes from the null distribution of the test statistic. When this distribution is known, a p-value, which is the observed significance level of the test, can be calculated. The p-value is bounded between 0 and 1, and can be interpretted as the probability the test statistic being at least as extreme as the test statistic calculated on the sample under H_0. The p-value affords the ability for different users to judge whether to reject or fail to reject H_0.

Testing for goodness-of-fit takes up the task of testing whether the underlying data distribution belongs to some given family of distribution functions. One such example is determining if a sample $x_1, ..., x_n$ is normally distributed or not, for which the hypothesis test is denoted in Eq.(1). For a more thorough discussion of hypothesis testing and goodness-of-fit see (Casella and Berger, 2002; Lehmann and Romano, 2006).

Hypothesis testing techniques fall into several sub-categoriess. (Seier, 2002) describes these sub-categories as those tests belonging to: skewness and kurtosis tests, empirical distribution function tests, regression and correlation tests, and others. Hypothesis tests can also be split into parametric vs non-parametric. Parametric hypothesis tests make assumptions about the underlying distribution while non-parametric hypothesis tests (also called distribution-free tests) do not. In this paper we focus on a test containing a parametric null hypothesis.

3.1 UNIVARIATE: SHAPIRO-WILK TEST

A goodness-of-fit test for normality is the Shapiro-Wilk (SW) test. In a Monte Carlo simulation by (Razali et al., 2011) the authors compared the power of the SW test to several non-parametric tests on various alternative distributions concluding it was the most powerful. The SW test is a composite parametric test to determine if a univariate data sample comes from a normal distribution. The original SW test was limited to a sample size between 3 and 50, but (Royston, 1982) extended the approach to use up to 2000 samples. The SW original test statistic, W, is calculated as

$$W = \frac{(\sum_{i=i}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{2}$$

In the Eq.(2) $x_{(i)}$ represents the i^{th} ordered statistic of the sample. The constants a_i are given by $(a_1, a_2, ..., a_n) = \frac{\mathbf{m}^T\mathbf{V}^{-1}}{(\mathbf{m}^T\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m})^{1/2}}$, where $\mathbf{m}$ is a vector consisting of the n expected values of the order statistics of independent and identically distributed random variables samples from the standard normal distribution. $\mathbf{V}$ is the covariance matrix of those order statistics. The most extreme order statistics are weighted the largest, and decrease when approaching the median. This property of the SW test will be important in later sections. Calculation of the constants a_i can be computationally demanding, prompting Royston in (Royston, 1992) to approximate these coefficients. He found that for $12 \leq n \leq 2000$ a two-parameter log-normal distribution fitted the upper half of the empirical distribution of $1 - W$. The associated p-value for W is referred to the upper tail of $\mathcal{N}(0, 1)$. Using the hypothesis test defined in Eq.(1), the SW test fails to reject H_0 if $W > W_\alpha$, or the p-value is larger than α, where W_α is the critical value based on the chosen confidence level. Three analytical properties that will become useful in the later sections, originally presented as lemmas in (Shapiro and Wilk, 1965), are cited as follows:

Lemma 3.1. *W is scale and origin invariant.*

Lemma 3.2. *W has a maximum value of 1*

Lemma 3.3. *The minimum value of W is $\frac{na_1^2}{(n-1)}$*

As our approach makes use of this test as a new loss function, one must be cognizant of its computational complexity. Enjoying the benefits of a strong test at the cost of long run time may not be appealing. However, the a_i are calculated a single time prior to training and are stored. The actual time complexity of the SW test during training is $O(nlog(n))$.

3.2 A NEW MULTIVARIATE GENERALIZATION OF SHAPIRO-WILK

Definition 3.1. (Multivariate Normal (Rao et al., 1973)) A d-dimensional random variable $\mathbf{u}$, that is, a random variable $\mathbf{u}$ taking values in E_d (Euclidean space of d-dimensions) is said to have a d-variate normal distribution $\mathcal{N}_d$ if and only if every linear function of $\mathbf{u}$ has a univariate normal distribution.

In a review by (Mecklin and Mundfrom, 2005) the authors noted more than 50 methods for testing multivariate normality. However, finding a multivariate test that is both powerful, and has low time complexity proved challenging. This necessitated the creation of a new multivariate hypothesis test that was able to make use of the strengths of SW, and relies on a well-known characterization of the multivariate normal (MVN) distribution.

Proposition 3.4. $\mathbf{x} \sim \mathcal{N}_d(\mu, \mathbf{\Sigma})$ *if and only if* $\mathbf{z} = \mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \mu) \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I})$.

Letting $\bar{\mathbf{x}}$ and $\mathbf{S}$ be respectively the sample mean and covariance matrix, define $\mathbf{S}^{-\frac{1}{2}}$ as the symmetric positive definite square root of the inverse of $\mathbf{S}$. Therefore, when $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \sim \mathcal{N}_d(\mu, \mathbf{\Sigma})$, then $\mathbf{z}_i = \mathbf{S}^{-\frac{1}{2}}(\mathbf{x}_i - \bar{\mathbf{x}}) \quad \forall i = 1, ..., n$ should be approximately $\mathcal{N}_d(\mathbf{0}, \mathbf{I})$. Under the assumption that observations are independent, and writing $\mathbf{z}_i = (z_{i1}, z_{i2}, ..., z_{id})^T$, then $z_{ij} \sim \mathcal{N}(0, 1)$ approximately for each $j = 1, ..., d$ and $i = 1, ..., n$.

To test the null hypothesis that the sample $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ is from $\mathcal{N}_d(\mu, \mathbf{\Sigma})$ where μ and $\mathbf{\Sigma}$ are unknown we propose to vectorize the entire $\mathbf{Z}$ matrix as $\mathbf{z}_{vec} = \mathbf{vec}(\mathbf{z}) = (z_{11}, z_{12}, ..., z_{nd})^T$, and then use the SW test statistic of Eq.(2) on $\mathbf{z}_{vec}$. Under H_0, W is expected to be close to 1. This multivariate generalization of the Shapiro-Wilk test (MSW) does not require any corrections for multiple testing, nor any simulation to calculate new critical values. Furthermore, it inherits the same good power properties while keeping the test complexity at $O(nd\text{log}(nd))$.

3.3 SHAPIRO-WILK ASYMPTOTICS

For years after the original (Shapiro and Wilk, 1965) paper, the distribution of W remained unknown. While variations of the original W statistic were proposed, it was the modification by (De Wet et al., 1972) that produced the first correlation normality test with known asymptotic distribution. The de Wet-Venter statistic $W^\star$ is defined as

$$W^\star = \sum_i \left(\frac{x_{(i)} - \bar{x}}{s_n} - \Phi^{-1}\left[\frac{i}{(n+1)}\right]\right)^2, \qquad (3)$$

where $\Phi^{-1}(\cdot)$ is the inverse normal cumulative density function. In their paper it was shown that $W^\star$ converges in distribution to

$$2n(1 - {W^\star}^{\frac{1}{2}}) - a_n \xrightarrow{\mathcal{D}} \xi, \qquad (4)$$

where $\xi = \sum_3^\infty \frac{y_i^2 - 1}{i}$, $\{y_i, i \geq 1\}$ is a sequence of independent and identically distributed $\mathcal{N}(0, 1)$ random variables, and with $a_n = \frac{1}{n+1}\left\{ \sum_{i=1}^n \frac{j(1-j)}{(\phi\{\Phi^{-1}(j)\})^2} - \frac{3}{2} \right\}$, where $j = \frac{i}{n+1}$, and $\phi(\cdot)$ is the standard normal density. (Verrill and Johnson, 1983; Fotopolous et al., 1984) showed that the Shapiro-Francia (Shapiro and Francia, 1972) statistic $W^\dagger$ and the de Wet-Venter statistic $W^\star$ were asymptotically equivalent via convergence in probability $n({W^\star}^{\frac{1}{2}} - {W^\dagger}^{\frac{1}{2}}) \xrightarrow{\mathcal{P}} 0$. (Leslie et al., 1986) produced the final result connecting Shapiro-Wilk to Shapiro-Francia showing that $n(W^{\frac{1}{2}} - {W^\dagger}^{\frac{1}{2}}) \xrightarrow{\mathcal{P}} 0$.

4 PROPOSED GENERATIVE MODEL

We propose to replace the discriminator neural network with a goodness-of-fit hypothesis test; specifically the Shapiro-Wilk hypothesis test, and its multivariate generalization. As the main idea here, maximizing the associated test statistic forces the encoder to encode data to a distribution (from which the decoder learns to generate data) so that the null hypothesis is not rejected, hence allowing $q(\mathbf{z})$ to be indistinguishable from the true distribution $p(\mathbf{z})$. Lemma 3.2 gives a target value for maximization, and from lemma 3.1 it can be seen that maximizing W to $W \geq W_\alpha$ results in the encoding distribution $q(\mathbf{z}|\mathbf{x})$ indistinguishable from the family $\mathcal{G} = \{\pi : \pi = \mathcal{N}(\mu, \mathbf{\Sigma})\}$.

Constraining the network to map $q(\mathbf{z}|\mathbf{x})$ to a specific distribution in the Gaussian class, for instance $\mathcal{N}(\mathbf{0}, \mathbf{I})$, as is often done, is not necessary. It may present challenges when generating data if the decoder is not robust to deviations from the prior, $p(\mathbf{z})$. Our new model allows the network to find the right $q(\mathbf{z}) = \pi^\star \in \mathcal{G} = \{\pi : \pi = \mathcal{N}(\mu, \mathbf{\Sigma})\}$ that minimizes the reconstruction loss without requiring a specific prior as long as it is in the class. In the univariate case, SW is directly applied to the encoded data, and the decoder works off of this code layer. When training is complete, $(\hat{\mu}, \hat{\mathbf{\Sigma}})$ are estimated using all data and the decoder generates data from $\mathcal{N}(\hat{\mu}, \hat{\mathbf{\Sigma}})$. In the multivariate case, however, we include a whitening step in the code layer, which is necessary in order to use the proposed multivariate SW. In other words, let $\mathbf{y}$ be the encoded data of k samples, and $(\bar{\mathbf{y}}, \mathbf{S})$ be the respective sample mean and covariance, we whiten the encoded data as $\mathbf{S}^{-\frac{1}{2}}(\mathbf{y} - \bar{\mathbf{y}})$. Then, Prop.3.4 allows the decoder to work off samples coming from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. In our empirical evaluation, we found that this whitening step also helped when other hypothesis tests were used in the proposed approach (see the supplement).

Now the presence of normality for $q(\mathbf{z})$ can be directly tested, i.e., cannot be rejected if W passes a critical value W_α. The overall optimization problem that our neural network solves is formulated as

$$\min ||\mathbf{X} - F_\psi(G_\theta(\mathbf{X}))||_2^2 \text{ s.t. } W(G_\theta(\mathbf{X})) > W_\alpha, \tag{5}$$

where $G_\theta(\cdot)$ is the encoder, and $F_\psi(\cdot)$ is the decoder, respectively parameterized by θ and ψ. Using the mathematically equivalent multi-objective loss, we can also find the solution $G_\theta^\star, F_\psi^\star = \arg\min_{G_\theta, F_\psi} ||\mathbf{X} - F_\psi(G_\theta(\mathbf{X}))||_2^2 - \lambda W(G_\theta(\mathbf{X}))$ for some proper value of $\lambda > 0$ although we propose an algorithm that directly solves Eq.(5).

4.1 OPTIMIZATION OF W

Eq.(5) is commonly optimized using a flavor of gradient descent with mini-batches, e.g., Adam (Kingma and Ba, 2014) which is used in Alg.1. The following proposition characterizes how to compute the gradient of W.

Proposition 4.1. *Let k be the size of the mini-batch. For any layer ℓ, denote $\mathbf{Y}^\ell = \mathbf{\Theta}^\ell \mathbf{Y}^{(\ell-1)}$, where $\mathbf{Y}^{(\ell-1)}$ is an $(n \times k)$ matrix of arbitrary activation from layer $(\ell - 1)$, $\mathbf{Y}^\ell$ is an $(m \times k)$ matrix of linear activation for layer ℓ, and $\mathbf{\Theta}^\ell$ is the $(m \times n)$ matrix of parameters connecting the layers. Let $\mathbf{y}$ be the $(mk \times 1)$ ascendingly sorted vectorization of $\mathbf{Y}^\ell$. Then, $\mathbf{y}$ can be computed by $\mathbf{A}\theta$ where $\mathbf{A}$ and θ are the re-organized $\mathbf{Y}^{\ell-1}$ and the vectorization of $\mathbf{\Theta}^\ell$. Specifically, $\mathbf{A}$ is an $(mk \times mn)$ matrix with each row containing the relevant $\mathbf{Y}^{(\ell-1)}$ data for a particular node's activation. The gradient of W can be computed by*

$$\nabla_\theta W = \frac{2\mathbf{a}^T\mathbf{A}\theta}{\theta^T\mathbf{Z}\theta}\mathbf{a}^T\mathbf{A}\left[\mathbf{I} - \frac{\theta\theta^T\mathbf{Z}}{\theta^T\mathbf{Z}\theta}\right], \tag{6}$$

where $\mathbf{Z} = \mathbf{A}(\mathbf{I} - \frac{\mathbf{J}}{mk})\mathbf{A}^T$, $\mathbf{I}$ is mk-dimensional identity matrix, and $\mathbf{J}$ is a $(mk \times mk)$ matrix of ones.

Unlike the adversarial framework the hypothesis testing model is straightforward to train. As shown in the pseudocode for this method in Alg.1, only the encoder part of the GAE updates when H_0 is rejected (by re-optimizing G_θ to reach $W > W_\alpha$ or a p-value if available). When whitening is used, and if $W > W_\alpha$, the decoder can generate new data by sampling with respect to $\mathcal{N}(\mathbf{0}, \mathbf{I})$. It must be reiterated that failure to reject does *not* imply normality, however in practice this procedure works well.

4.2 INNER LOOP TERMINATION

Alg.1 follows a conditional alternating optimization procedure, or can also be referred to as a feasible direction method. The outer loop seeks to minimize the reconstruction loss, whereas the inner loop evaluates the hypothesis testing and identifies the updates of θ that satisfy: $W > W_\alpha$, a constraint that implies failure to reject H_0. Note that the inner loop is activated only when the condition is *not* already met. A fundamental question is whether this "while" loop will terminate given we solve a highly non-convex optimization problem (ultimately, whether we can find a θ such that the $q(\mathbf{z})$ stays within the Gaussian class). In (Bottou, 1991a) results were given for a general non-convex setting and show that under specific conditions the computation will converge. The following theorem is cited from that paper for which the proof can be found in (Bottou, 1991b).

Algorithm 1 Hypothesis Testing Autoencoder

```
Require: $\mathbf{X}$=training data, $N$=number of iterations,
    $W_\alpha$=critical value, $\lambda$=regularization coefficient,
    $m$=size of mini-batch > $p$=dimension of latent code
    layer to be tested
Initialize: $\theta, \psi$
for $i = 0$ to $N$ do
    Sample next mini-batch from training data $\mathbf{X}_m$
    $(\theta, \psi) \leftarrow$ Adam$(\nabla_{(\theta,\psi)} ||\mathbf{X}_m - F_\psi(G_\theta(\mathbf{X}_m))||_2^2)$
    Compute $W(G_\theta(\mathbf{X}_m))$
    while $W(G_\theta(\mathbf{X}_m)) \leq W_\alpha$ do
        Sample next mini-batch from training data $\mathbf{X}_m$
        $\theta \leftarrow$ Adam$(\nabla_{(\theta)} \lambda ||W_\alpha - W(G_\theta(\mathbf{X}_m))||_2^2)$
        Compute $W(G_\theta(\mathbf{X}_m))$
    end while
    Estimate $(\hat{\mu}, \hat{\mathbf{\Sigma}})$        (Not necessary if whitened)
end for
```

Theorem 4.2. *For any measure $dP(\mathbf{z})$, if the cost $C(\theta) = \mathbb{E}(J(\mathbf{z}, \theta))$ is differentiable up to the third derivatives where J is an objective function to be optimized, with bounded second and third derivatives, and if the following assertions are true,*

(i) $\forall\theta, E(H(\theta)) = \nabla_\theta C(\theta)$

(ii) $\sum_{t=1}^\infty \epsilon_t = \infty, \quad \sum_{t=1}^\infty \epsilon_t^2 < \infty$

(iii) $\exists A, B, \quad \forall\theta, \quad \mathbb{E}(H(\theta)^2) < A + BC(\theta)$

(iv) $\exists C_{min}, \quad \forall\theta, \quad C_{min} < C(\theta)$

then $C(\theta_t)$ converges with probability 1 and $\nabla_\theta C(\theta_t)$ converges to 0 with probability 1.

In our case, $W(\theta)$ is C, thus $H(\theta)) \equiv \nabla_\theta W(\theta)$, and ϵ_t is the learning rate. The inner loop terminates according to Thm.4.2 if its conditions are all satisfied (the detailed proof is given in a supplement). Using the same argument of (Bottou, 1991b) regarding the similarity of simulated annealing, denoting $q_t(\theta)$ the density of probability that θ_t follows, by Thm.4.2 the support of $q_t(\theta)$ converges to the set of extrema of $W(\theta)$, i.e., $\theta_t \rightarrow \{\theta | W(\theta) = 1\}$

thus $W(\theta_t) \to 1$. In fact, it is not necessary to train until $W(\theta) = 1$, so the procedure exits once $W > W_\alpha$.

5 THEORETICAL EQUIVALENCY

There exists a link between a distance based method for comparing the goodness-of-fit of two distributions and hypothesis testing discovered in (del Barrio et al., 1999). The general class of Wasserstein distances is studied in (Villani, 2008). We recite the definition here.

Definition 5.1. (Wasserstein Distances) Let (χ, d) be a Polish metric space, and let $p \in [1, \infty)$. For any two probability measures μ, ν on χ, the Wasserstein distance of order p between μ and ν is defined by

$$W_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_\chi d(x,y)^p d\pi(x,y) \right)^{\frac{1}{p}}$$

$$= \inf \left\{ \left[\mathbb{E} d(X,Y)^p \right]^{\frac{1}{p}}, \ \text{law}(X) = \mu, \ \text{law}(Y) = \nu \right\},$$

where $\Pi(\mu, \nu)$ is the set of all joint probability measures. When the Polish metric space under consideration is the one-dimensional Euclidean space, $W(\mu, \nu) = W_2(\mu, \nu)$.

Of primary interest is the L_2-Wasserstein distance. It is possible to consider the distance between distributions P_1 and P_2, defined by $\mathcal{W}(P_1, P_2) = \Big[\int_0^1 (F_1^{-1}(t) - F_2^{-1}(t))^2 dt \Big]^{\frac{1}{2}}$, where F_1^{-1} and F_2^{-1} are the quantile functions of P_1 and P_2 defined to be $F_i^{-1}(t) = \inf\{s : F_i(s) \geq t\}$ for $i = 1, 2$. The distance between a distribution with cumulative distribution function (CDF) F, mean μ_0 and standard deviation σ_0, and the class of *all* normal distributions can be written as $\mathcal{W}^2(F, \mathcal{G}') = \inf\{\mathcal{W}^2(F, \pi), \pi \in \mathcal{G}'\}$, where $\mathcal{G}' = \{\pi : \pi = \Phi\Big(\frac{x-\mu}{\sigma}\Big), -\infty < \mu < \infty, \sigma > 0\}$.

By expressing a normal random variable with mean μ and variance σ^2 as $F^{-1}(p) = \mu + \sigma\Phi^{-1}(p)$, it can be shown that $\frac{\mathcal{W}^2(F,\mathcal{G}')}{\sigma_0^2} = 1 - \frac{\Big(\int_0^1 (F^{-1}(t))\Phi^{-1}(t)dt\Big)^2}{\sigma_0^2}$. With a random sample of data, $x_1, x_2, ..., x_n$, with underlying CDF F, define $\mathcal{R}_n = \frac{\mathcal{W}^2(F_n,\mathcal{G}')}{S_n^2} = 1 - \frac{\hat{\sigma}^2}{S_n^2}$, where $\hat{\sigma}_n = \int_0^1 F_n^{-1}(t)\Phi^{-1}(t)dt$ and S_n^2 is the sample variance. $\mathcal{R}_n$ can be utilized as a test statistic for testing the composite null hypothesis that the data are normally distributed, and it belongs to the class of minimum distance tests.

5.1 L$_2$-WASSERSTEIN ASYMPTOTICS

To study the null asymptotics of $\mathcal{R}_n$, assuming normality, (del Barrio et al., 1999) used approximations of quantile processes by Brownian bridges, $B(t)$. Under normality del Barrio et al. show that there exist constants a_n such that $n\mathcal{R}_n - a_n \xrightarrow{\mathcal{D}} \int_0^1 \hat{B}^2(t) - E\hat{B}^2(t)dt$ where $\hat{B} = \frac{(B - \langle B,1\rangle 1 - \langle B, \Phi^{-1}\rangle \Phi^{-1}}{\phi(\Phi^{-1})}$. By applying principle component decomposition, (del Barrio et al., 1999) obtains the final result and is repeated here for clarity.

Theorem 5.1. *Let $\{X_n\}_n$ be a sequence of i.i.d normal random variables. Then*

$$\mathcal{R}_n - a_n \xrightarrow{\mathcal{D}} -\frac{3}{2} + \sum_{j=3}^{\infty} \frac{Y_j^2 - 1}{j},$$

where $\{Y_n\}_n$ is a sequence of i.i.d $\mathcal{N}(0,1)$ random variables with $a_n = \int_{\frac{1}{(n+1)}}^{\frac{n}{(n+1)}} \frac{t(1-t)}{(\phi(\Phi^{-1}(t)))^2} dt$

Cross referencing the asymptotics in Section 3.3, we find the L_2-Wasserstein normality test to be equivalent to the Shapiro-Wilk test, thus attaining similar power properties.

6 EMPIRICAL EVALUATION

We evaluated our method, the hypothesis testing based autoencoder using univariate Shapiro-Wilk (HTAE-SW), and its multivariate generalization (HTAE-MSW) on the standard MNIST digits dataset (LeCun et al., 1998), comparing it against the models believed to be the most closely related: the adversarial autoencoder (Makhzani et al., 2015) (AAE), the adversarial autoencoder with Wasserstein discriminator loss (WAAE), and autoencoder with maximum mean discrepancy loss for the critic (MMDAE as the model is not adversarial). We note that MMD constitutes a true hypothesis test where an α-level test may be performed by way of permutation or approximation tests. Using it as such would lead to the HTAE-MMD model, a variant of our model. However, we used it as others did by simply optimizing it for comparison against the HTAE. To further evaluate the proposed approach, four additional goodness-of-fit tests were used to replace the (M)SW test. A supplementary material provided more results and discussion on Royston's H (Royston, 1983) (HTAE-R), Mardia's Skewness (Mardia, 1970) (HTAE-M), Malkovich-Afifi (Malkovich and Afifi, 1973) (HTAE-MA), and Henze-Zirkler (Henze and Zirkler, 1990) (HTAE-HZ). All models used $||x - \hat{x}||_2^2$ for the reconstruction loss.

The network architecture was a fully connected class conditional autoencoder with conditioning done at the code layer. Two hidden layers are used between the input and code layer, each consisting of 784 nodes. The decoder contained the same structure. For AAE and WAAE, discriminators contained 2 layers each of 784 nodes with their respective losses. Models requiring sampling from $p(\mathbf{z})$ used Gaussian priors of appropriate dimension $p(\mathbf{z}) = \mathcal{N}_d(0, I)$. Dropout (Srivastava et al., 2014) was used with a "keep" probability equal to 0.9. The Adam optimizer was used with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and $\eta = 0.001$. The hypothesis test was computed at the mini-batch level consisting of 100 samples. The hypothesis tests require an α to be set. While techinically a hyperparameter, it is the level of the test with a clear meaning that training proceeds with respect to this level of confidence. A more stringent α-level tends to increase inner loop iterations. We found the commonly used $\alpha = 0.05$ to work well. Two experiments were conducted to guage the efficacy of the hypothesis testing method: a univariate and multivariate test.

Several criteria were used to assess each model: reconstruction loss, generative quality, normality constraints, prior matching (where applicable), and run-time. We generated images of hand-written digits, and monitored the reconstruction loss during training. By considering the hypothesis test statistic as an objective measure for rejecting normality, the test statistic was monitored for each iteration, and its corresponding p-value was plotted, during training for all models. In particular, the null hypothesis was rejected when the p-value was less than 0.05; larger p-values were preferred. Q-Q plots were also provided. By plotting the theoretical quantiles of the normal distribution against the empirical quantiles of the data in a Q-Q plot, any departure from the straight line provides evidence against normality. This can be used as a diagnostic measure after training has completed. The run times are included in Table 1.

6.1 UNIVARIATE: SHAPIRO-WILK

There was only a single node in the latent code layer in the univariate case. Results from the univariate case can be seen in Fig.1. Along with the plots mentioned above, the univariate case presented an opportunity to monitor the trajectory of $(\hat{\mu}, \hat{\sigma})$. Initial $(\hat{\mu}, \hat{\sigma})$ were calculated using the initialized network weights. By monitering the p-values, it appeared that the $q(\mathbf{z})$ distributions from many other methods were not in fact normal. Of the Q-Q plots, only HTAE-SW maintained close enough proximity to the straight line. Based on the final batch, WAAE and MMDAE were able to match the prior distribution (μ, σ) parameters fairly closely, however neither maintained normality. AAE could neither match parameters nor maintain normality. As HTAE-SW was not restricted to a specific normal, it had the opportunity to explore the normal class for an optimal distribution for the given data.

6.2 MULTIVARIATE GENERALIZATION OF SHAPIRO-WILK

For the multivariate methods a latent code dimension of 8 was used. By employing the new multivariate generalization of the SW test (MSW) it was possible to use Q-Q plots to lend visual support to the p-value outcome, however trajectory plots were no longer an option. Each model was run for 100,000 iterations with plots shown in Fig.2, and run time shown in Table 1 when the code layer had 8 nodes. The generated images seemed to get visually better the higher the simple moving average of the p-values was. Q-Q plots for the last mini-batch show improper tail behaviors, for normality, in all models but HTAE-MSW. As before, the p-value should be greater than $\alpha = 0.05$ to fail to reject the null; again the higher the p-value the better. For models that failed to reject H_0 but had poor generative quality, this suggested several possibilities: training time needed to be increased, more focus should be given to reconstruction, or the network size should be increased. As can be seen in Table 1 HTAE methods were substantially faster in all cases.

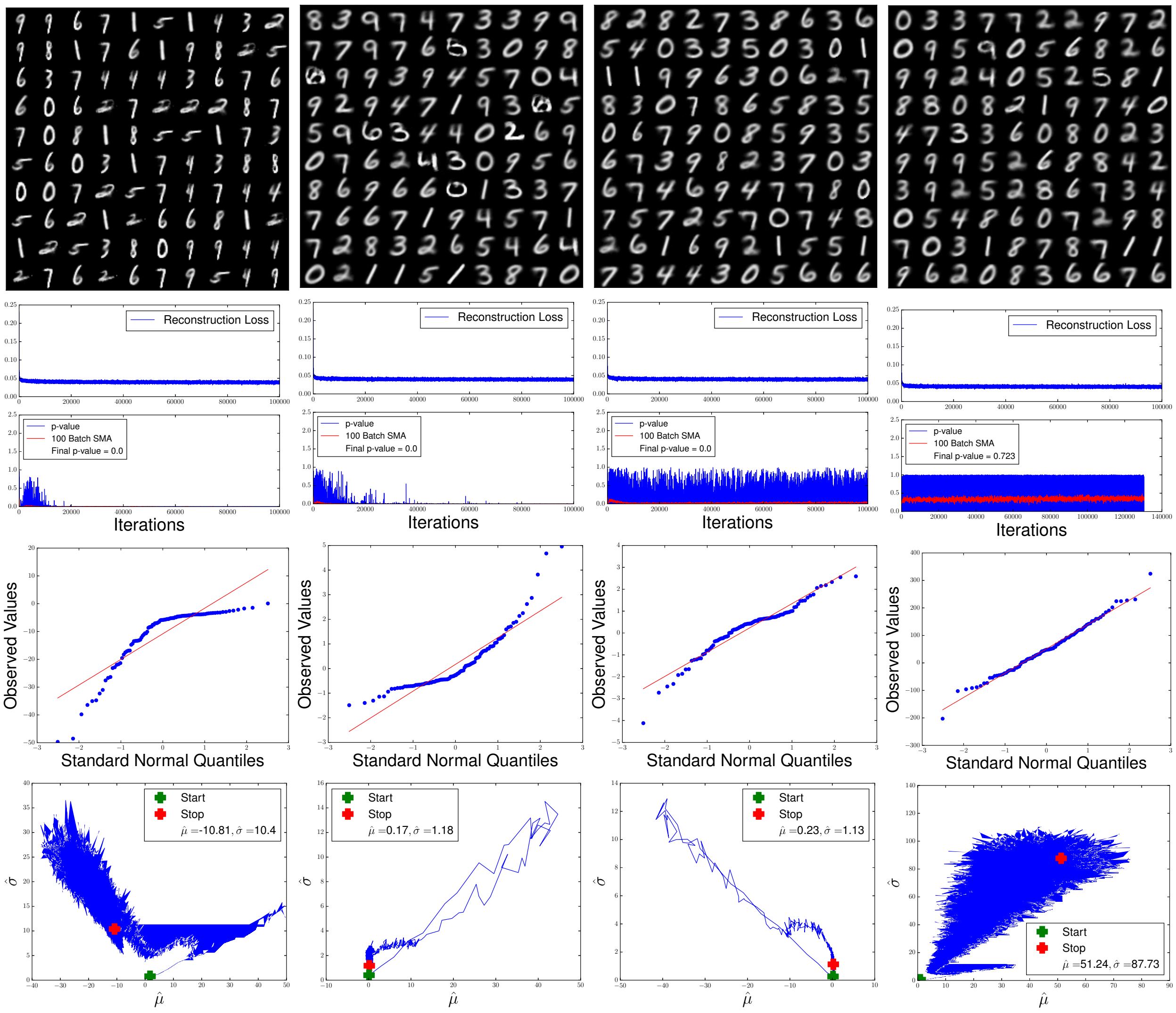
Table 1: Run time in seconds for 10^5 iterations in the 1-D and 8-D cases using an NVIDIA GTX 1080Ti GPU.

Method	1-D	8-D
AAE	765.39	982.61
WAAE	904.95	1092.19
MMDAE	597.67	756.69
HTAE-(M)SW	346.16	548.08

7 DISCUSSION

Our empirical results suggest that substituting a hypothesis test, notably W and its the new multivariate generalization, which do not require pre-specifying a mean and covariance, may be a competitive alternative to other members of the GAE class. Allowing $q(\mathbf{z})$ to deviate in the feasible space, $\mathcal{G}$, during training means sampling is done with respect to $q(\mathbf{z}) = \hat{\pi} = \mathcal{N}(\hat{\mu}, \hat{\mathbf{\Sigma}}) \in \mathcal{G}$ where $(\hat{\mu}, \hat{\mathbf{\Sigma}})$ are estimated (or when whitening is used in the multivariate case, with respect to $\mathcal{N}(\mathbf{0}, \mathbf{I})$). Consequently, there is less need to worry about discrepancy between the distribution we *want* to sample from, and the distribution we *are* sampling from, as we tacitly interpret failure to reject as within the class of normals. This need *not* be the case in the other models. However, as HTAE-MSW does make use of whitening, ensuring enough samples to adequately estimate $\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}$ and $\hat{\mu}$ is a necessity.

The AAE and WAAE both require training of a discriminator network. This network, with size on the order of the encoder or decoder, increases training time. Moreover, training of the discriminator needs to be scheduled in balance with that of the generator, and how exactly this should be done is still an open problem. On the other hand, using the hypothesis test abolishes this problem



(a) AAE (b) WAAE (c) MMDAE (d) HTAE-SW

Figure 1: Row one illustrates sample digits generated by each model. Row two shows the reconstruction loss, and the simple moving average (SMA) of the p-value for a batch size of 100, along with the final mini-batch terminating p-value. Row three contains the Q-Q plots, and row four plots the trajectory of $(\hat{\mu}, \hat{\sigma})$ over the course of training.

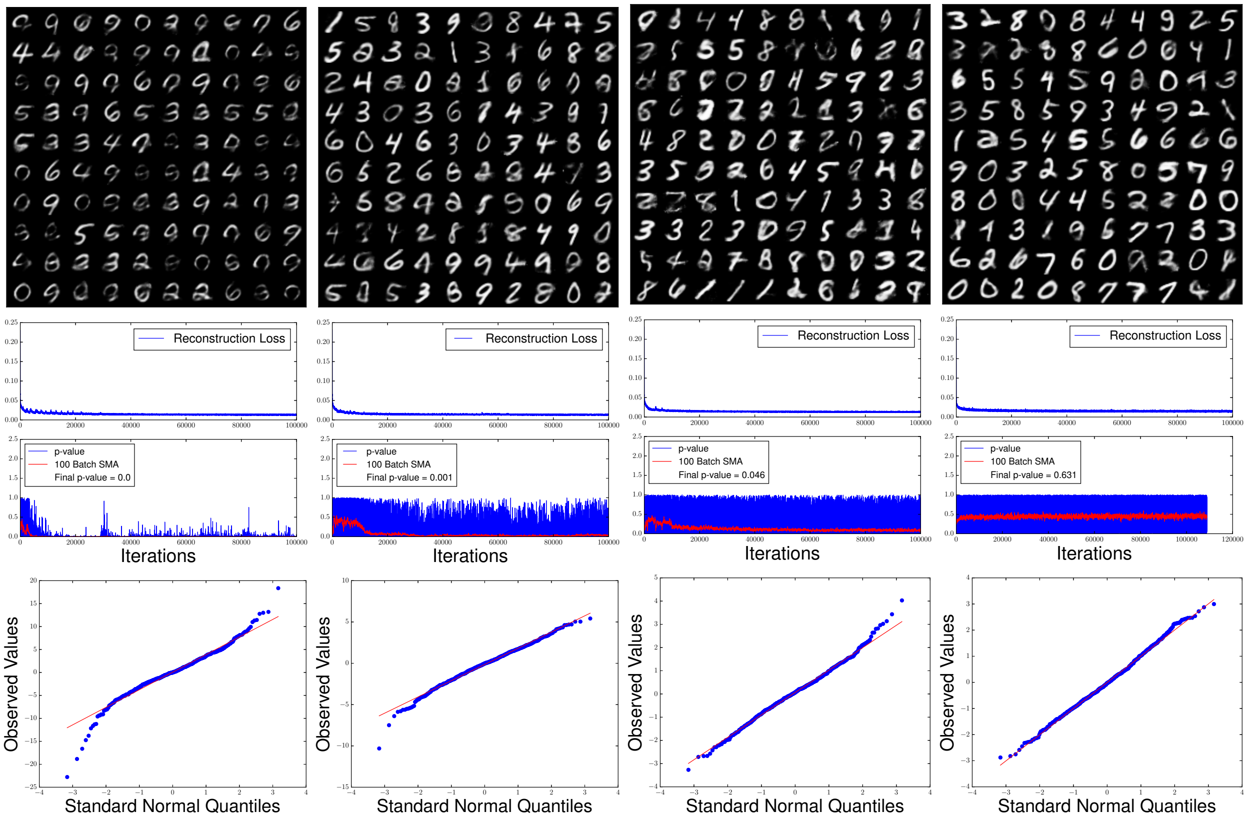
completely. By utilizing the critical values (or p-values) for the test statistic it is now possible to know *precisely* when to alternate between enforcing prior constraints, and minimizing the reconstruction loss. In our empirical evaluation, we also observed that using a parametric hypothesis test could improve gradient updates by utilizing information about the null distribution. A full exploration into this mechanism is left for future investigation.

A concern may be raised that Thm.4.2 merely guarantees the "while" loop will terminate, yet provides no indication of when it terminates. In all experiments, the speed never proved to be an issue. The inner loop was able to ensure $q(\mathbf{z}) \in \mathcal{G}$ in a very small number of iterations as can be seen in Fig.3 in supplementary materials. We attempt to understand why this is the case in the near future.

We experimented with four other goodness-of-fit tests (see the supplement please), but none provided the benefits that the Shapiro-Wilk and its generalization did. Issues included longer run times, and weaker power. The search for greater power motivates the following conjecture.

Conjecture 1. The more powerful the hypothesis test, the more precise the null distribution information contained within the test statistic that can be transmitted to the encoder to update θ via the gradient.

A trade-off between the power of the test, time complex-



(a) AAE

(b) WAAE

(c) MMDAE

(d) HTAE-MSW

Figure 2: Similar to Fig.1, the top row plots random samples generated from each model. Row two contains the p-values with a 100 batch SMA. Row three are Q-Q plots associated with MSW.

ity and reconstruction quality likely exists. What is the cheapest computational complexity of a hypothesis test available for a given power? We expect that future research will reveal more on these questions.

8 CONCLUSION AND FUTURE WORK

In this paper we have proposed a new method for training generative autoencoders by explicitly testing the distribution of the code layer output via univariate and multivariate parametric hypothesis tests. We have shown a number of benefits to using such an approach including: objectively verifying if training has indeed pushed $q(\mathbf{z}) \in \mathcal{G}$, the ability to utilize the critical value of a hypothesis test as a threshold for determining when to switch between reconstruction and encoding iterations. Our method produces competitive results while showing computational efficiency. Moreover, we explored the link between the Shapiro-Wilk hypothesis test and the L_2-Wasserstein distance between two distributions.

Given the large numbers of univariate and multivariate parametric hypothesis tests available, it remains to be seen how others compare when used in a generative autoencoder. In fact, any distribution for which a hypothesis test can be derived can be used for training a latent code layer. Furthermore, the proposed method of training can be applied to any of the models that takes a generative autoencoder style network. This initial work brings up additional interesting problems, so our future work will dive into the questions raised. It is also worth asking how other hypothesis testing methods can be included into neural network training.

Acknowledgments

We thank Vladamir Pozdnyakov for insightful discussions, especially regarding the multivariate test. This work was supported by National Science Foundation (NSF) grants: IIS-1320586, CCF-1514357, and IIS-1718738. J. Bi was also supported by National Institutes of Health (NIH) grants: R01DA037349 and K02DA043063 as well as NSF grants: DBI-1356655 and IIS-1447711.

References

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Y. Bengio, G. Mesnil, Y. Dauphin, and S. Rifai. Better mixing via deep representations. In *International Conference on Machine Learning*, pages 552–560, 2013a.

Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, pages 899–907, 2013b.

Y. Bengio, E. Laufer, G. Alain, and J. Yosinski. Deep generative stochastic networks trainable by backprop. In *International Conference on Machine Learning*, pages 226–234, 2014.

L. Bottou. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8):0, 1991a.

L. Bottou. *Une Approche Theorique de L'Apprentissage Connexionniste et Applications A La Reconnaissance de la Parole*. PhD thesis, 1991b.

G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

J. De Wet et al. Asymptotic distributions of certain test criteria of normality. *South African Statistical Journal*, 6 (2):135–149, 1972.

E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, and J. M. Rodríguez-Rodríguez. Tests of goodness of fit based on the l2-wasserstein distance. *Annals of Statistics*, pages 1230–1239, 1999.

A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*, pages 658–666, 2016.

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

S. Fotopolous, J. Leslie, and M. Stephens. Errors in approximations for expected normal order statistics with an application to goodness-of-fit. Technical report, Technical Report, 1984.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5769–5779, 2017.

N. Henze and B. Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods*, 19(10):3595–3617, 1990.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

J. Leslie, M. A. Stephens, and S. Fotopoulos. Asymptotic distribution of the shapiro-wilk w for testing for normality. *The Annals of Statistics*, pages 1497–1506, 1986.

C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2200–2210, 2017.

Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727, 2015.

A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

J. F. Malkovich and A. Afifi. On tests for multivariate normality. *Journal of the american statistical association*, 68(341):176–179, 1973.

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

C. J. Mecklin and D. J. Mundfrom. A monte carlo comparison of the type i and type ii error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation*, 75(2):93–107, 2005.

A. Mnih and K. Gregor. Neural variational inference and learning in belief networks. *arXiv preprint arXiv:1402.0030*, 2014.

C. R. Rao, C. R. Rao, M. Statistiker, C. R. Rao, and C. R. Rao. *Linear statistical inference and its applications*, volume 2. Wiley New York, 1973.

N. M. Razali, Y. B. Wah, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33, 2011.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

S. Rifai, Y. Bengio, Y. Dauphin, and P. Vincent. A generative process for sampling contractive auto-encoders. *arXiv preprint arXiv:1206.6434*, 2012.

J. Royston. An extension of shapiro and wilk's w test for normality to large samples. *Applied statistics*, pages 115–124, 1982.

J. Royston. Some techniques for assessing multivarate normality based on the shapiro-wilk w. *Applied Statistics*, pages 121–133, 1983.

P. Royston. Approximating the shapiro-wilk w-test for non-normality. *Statistics and Computing*, 2(3):117–119, 1992.

E. Seier. Comparison of tests for univariate normality. *InterStat Statistical Journal*, 1:1–17, 2002.

S. S. Shapiro and R. Francia. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67(337):215–216, 1972.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4): 591–611, 1965.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*, 2016.

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

S. P. Verrill and R. A. Johnson. *The asymptotic distributions of censored data versions of the Shapiro-Wilk test of normality statistic*. University of Wisconsin, Department of Statistics, 1983.

C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.