AUTO-ENCODING GOODNESS OF FIT

Anonymous authors
Paper under double-blind review

ABSTRACT

For generative autoencoders to learn a meaningful latent representation for data generation, a careful balance must be achieved between reconstruction error and how close the distribution in the latent space is to the prior. However, this balance is challenging to achieve due to a lack of criteria that work both at the mini-batch (local) and aggregated posterior (global) level. Goodness of fit (GoF) hypothesis tests provide a measure of statistical indistinguishability between the latent distribution and a target *distribution class*. In this work, we develop the Goodness of Fit Autoencoder (GoFAE), which incorporates hypothesis tests at two levels. At the mini-batch level, it uses GoF test statistics as regularization objectives. At a more global level, it selects a regularization coefficient based on higher criticism, i.e., a test on the uniformity of the local GoF p-values. We justify the use of GoF tests by providing a relaxed L_2-Wasserstein bound on the distance between the latent distribution and target prior. We propose to use GoF tests and prove that optimization based on these tests can be done with stochastic gradient (SGD) descent on a compact Riemannian manifold. Empirically, we show that our higher criticism parameter selection procedure balances reconstruction and generation using mutual information and uniformity of p-values respectively. Finally, we show that GoFAE achieves comparable FID scores and mean squared errors with competing deep generative models while retaining statistical indistinguishability from Gaussian in the latent space based on a variety of hypothesis tests.

1 INTRODUCTION

Training generative autoencoders (GAEs) to generate quality samples requires balancing low reconstruction error with a regularization loss that encourages the latent representation to be meaningful for data generation (Hinton & Salakhutdinov, 2006; Ruthotto & Haber, 2021). Variational autoencoders (VAEs) achieve this balance by maximizing a lower bound on the data log-likelihood, which consists of a reconstruction term and a Kullback-Leibler (KL) divergence term between the posterior and prior distributions (Kingma & Welling, 2013; Rezende et al., 2014). The development of new GAE *meta-priors* (Bengio et al., 2013) generalizes assumptions about the data distribution in the latent space and have enabled the learning of representations that are hierarchically organized (Sønderby et al., 2016; Vahdat & Kautz, 2020), semi-supervised (Kingma et al., 2014), clustered (Makhzani et al., 2015; Lechat et al., 2021), or disentangled (Kumar et al., 2018; Horan et al., 2021). Meta-priors are typically enforced with an additive term in the GAE loss function that minimizes a statistical distance, e.g., KL divergence, between the posterior and a specific prior (Tschannen et al., 2018).

Another important class of statistical distances used in GAE meta-priors are based on optimal transport (OT) and Wasserstein distance, including the Wasserstein (WAE) (Tolstikhin et al., 2017), Sinkhorn (Patrini et al., 2018) and sliced-Wasserstein autoencoders (SWAE) (Kolouri et al., 2018). Importantly, OT-based GAEs encourage the *aggregated* posterior density to match an assumed prior distribution (compared to, e.g., VAEs that typically minimize the per-sample discrepancy) and have been shown to avoid the disadvantages of VAEs and adversarial methods (Tolstikhin et al., 2017). However, in all cases, a balance between reconstruction and generation is required (Alemi et al., 2018) to prevent decoupling between the posterior and the prior (Hoffman & Johnson, 2016) or vanishing KL between the two (Bowman et al., 2016). Regardless of the selected meta-prior, it is difficult to decide when the posterior is *close enough* to the prior due to: (a) the absence of tight constraints on the statistical distances; (b) their distributions across mini-batches; and (c) the difference in scale between reconstruction and regularization objectives. If the posterior deviates from the prior, GAEs may generate samples of poor quality or with incorrect proportions (Rubenstein et al., 2018).

In contrast to statistical distances, goodness-of-fit (GoF) tests are statistical hypothesis tests that measure the indistinguishability between the posterior and a prior *distribution class* (Stephens, 2017), typically the class of Gaussians. While both parametric (Ridgeway & Mozer, 2018; Palmer et al., 2018) and nonparametric (Ding et al., 2019) hypothesis tests have been used in GAEs, prior work exhibits both theoretical and practical issues. In GAEs, GoF test statistics are optimized *locally* in mini-batches and no mechanism exists to determine if the *aggregated* posterior is indistinguishable from the prior. If GoF test p-values are small (i.e., mini-batches are distinguishable from the prior), then sampling quality is poor; conversely, an abundance of large GoF p-values may result in poor reconstruction as the posterior matches too closely to the prior at the mini-batch level (Lucas et al., 2019). Careful selection of regularization coefficients is required to ensure that the statistical distance is *close enough* but *not too close* to balance reconstruction and generation. Further, test statistic optimization in GAEs suffers from identifiability issues, unbounded domains, and gradient singularities and there exists no principled algorithm to optimize GoF using SGD.

Our Contributions: We develop and justify a framework for parametric test statistic optimization, resulting in a novel GAE that optimizes GoF tests for normality, and an algorithm for regularization coefficient selection based on p-value higher criticism (Donoho et al., 2004). Note that the GoF tests are not only for Gaussians with nonsingular covariance matrices, but *all* Gaussians, so they can handle situations where the data distribution is concentrated on or closely around a manifold with dimension smaller than the ambient space. The framework uses Gaussian priors as it is a standard option [(Doersch, 2016; Bousquet et al., 2017; Kingma et al., 2019)], and also because normality tests are much better understood than GoF tests for other distributions, with many more tests and more accurate calculation of p-values available. The framework can be modified to use other priors in a straightforward way provided that the same level of understanding can be gained on GoF tests for those priors as for normality tests. All proofs are deferred to the appendix. Specifically, our contributions are summarized as follows.

- We propose a framework (Sec. 2) for bounding the statistical distance between the posterior and the prior in GAEs, which forms the theoretical foundation for a deterministic GAE - Goodness of Fit Autoencoder (GoFAE), that directly optimizes GoF hypothesis test statistics.
- We examine four GoF tests of normality based on correlation and empirical distribution functions (Sec. 3). Each GoF test focuses on a different aspect of Gaussianity, e.g., moments or quantiles.
- A model selection method using higher criticism of p-values is proposed (Sec. 3), which enables global normality testing of the aggregated posterior and is test-based instead of performance-based. This method helps determine the range of the regularization coefficient that well balances reconstruction and generation using uniformity of p-values respectively (Fig. 2b).
- We show that gradient based optimization of test statistics for normality can be complicated by identifiability issues, unbounded domains, and gradient singularities; we propose a SGD that optimizes over a Riemannian manifold (Stiefel manifold in our case) that effectively solves our GAE formulation with convergence analysis (Sec. 4).
- We show that GoFAE achieves comparable FID scores and mean squared error on three datasets while retaining statistical indistinguishability from Gaussian in the latent space (Sec. 5).

2 A Generative Auto-Encoder Framework

Let $(\mathcal{X}, \mathbb{P}_X)$ and $(\mathcal{Z}, \mathbb{P}_Z)$ be two probability spaces. In our setup, $\mathbb{P}_X$ is the true, but unknown, data distribution while $\mathbb{P}_Z$ is a pre-specified, or prior distribution on the latent space $\mathcal{Z}$.[1] An implicit generative model aims to find a possibly stochastic mapping $G : \mathcal{Z} \to \mathcal{X}$, known as the decoder, such that the distribution of $G(Z)$, i.e., $\mathbb{P}_{G(Z)}$ matches $\mathbb{P}_X$, where $Z \sim \mathbb{P}_Z$. One approach is based on optimal transport (OT) (Villani, 2008). Recall that if μ and ν are two distributions on a space $\mathcal{E}$ and $c(u, v)$ is a cost function on $\mathcal{E}$, then the OT cost to transfer μ to ν is $\mathcal{T}_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(U,V)\sim\pi}[c(U, V)]$, where $\Pi(\mu, \nu)$ is the set of all joint probability distributions with marginals μ and ν. If $\mathcal{E}$ is endowed with a metric d and $c(u, v) = d(u, v)^p$ with $p > 0$, then $d_{W_p}(\mu, \nu) = \mathcal{T}_c(\mu, \nu)^{1/p}$ is known as the L_p-Wasserstein distance. In principle, the decoder G can be learned by solving $\min_G \mathcal{T}_c(\mathbb{P}_X, \mathbb{P}_{G(Z)})$. Unfortunately, the minimization is intractable. Instead, the idea of the autoencoder approach is to find a possibly stochastic mapping $F : \mathcal{X} \to \mathcal{Z}$, known as the encoder, such that the distribution of $Y = F(X)$, i.e., $\mathbb{P}_Y$ matches the prior $\mathbb{P}_Z$. This idea combined with the generative modeling leads to GAEs, which learn the encoder and decoder by

[1] This should be distinguished from the priors in Bayesian statistics.

solving the following optimization problem

$$\min_{F,G}[\mathbb{E}_{X\sim\mathbb{P}_X, Y=F(X)}c(X, G(Y)) + \lambda D(\mathbb{P}_Y, \mathbb{P}_Z)], \tag{1}$$

where $\lambda > 0$ is a regularization coefficient and D is a penalty on statistical discrepancy. In the GAE framework, $Y = F(X)$ is the latent variable or code with $\mathbb{P}_Y$ referred to as the aggregated posterior, $G(Y) = G(F(X))$ the reconstruction of X, and $G(Z)$ the generated data. When $\mathcal{X}$ and $\mathcal{Z}$ are Euclidean spaces, a common choice for c is the squared Euclidean distance, leading to the L_2-Wasserstein distance d_{W_2} and the following bounds.

Proposition 1. *If $\mathcal{X}$ and $\mathcal{Z}$ are Euclidean spaces and the decoder G is differentiable, then (a) $d_{W_2}(\mathbb{P}_{G(Y)}, \mathbb{P}_{G(Z)}) \leq \|\nabla G\|_\infty d_{W_2}(\mathbb{P}_Y, \mathbb{P}_Z)$, (b) $d_{W_2}(\mathbb{P}_X, \mathbb{P}_{G(Z)}) \leq [\mathbb{E}\|X - G(Y)\|^2]^{1/2} + \|\nabla G\|_\infty d_{W_2}(\mathbb{P}_Y, \mathbb{P}_Z)$.*

Combined with the triangle inequality for d_{W_2}, (a) and (b) imply that when the reconstruction error $[\mathbb{E}\|X - G(Y)\|^2]^{1/2}$ is small, proximity between the latent distribution $\mathbb{P}_Y$ and the prior $\mathbb{P}_Z$ is sufficient to ensure proximity of the generated data distribution $\mathbb{P}_{G(Z)}$ and the data distribution $\mathbb{P}_X$. Both (a) and (b) are similar to (Patrini et al., 2018), though here a less tight bound – the square-error – is used in (b) as it is an easier objective to optimize. In practice the encoder and decoder are networks trained using mini-batches of data. Henceforth, we denote them by $F_{\boldsymbol{\theta}}$ and $G_{\boldsymbol{\phi}}$ respectively, with $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ being the parameters of the corresponding networks. Given a mini-batch of size m, $\{X_i\}_{i=1}^m$, the L_2-Wasserstein distance between the empirical distribution of $\{Y_i = F_{\boldsymbol{\theta}}(X_i)\}_{i=1}^m$ and $\mathbb{P}_Z$ is a natural approximation of $d_{W_2}(\mathbb{P}_Y, \mathbb{P}_Z)$ because it is a consistent estimator of the latter as $m \to \infty$, although in general is biased (Xu et al., 2020).

Proposition 2. *Let $\hat{\mathbb{P}}_{X,m}$ be the empirical distribution of samples $\{X_i\}_{i=1}^m$, and $\hat{\mathbb{P}}_{Y,m}$ be the empirical distribution of $\{Y_i = F_{\boldsymbol{\theta}}(X_i)\}_{i=1}^m$. Assume that $F_{\boldsymbol{\theta}}(X)$ is differentiable with respect to X with bounded gradients $\nabla F_{\boldsymbol{\theta}}(X)$. Then, (a) $d_{W_2}(\hat{\mathbb{P}}_{Y,m}, \mathbb{P}_Y) \leq \|\nabla F_{\boldsymbol{\theta}}\|_\infty d_{W_2}(\hat{\mathbb{P}}_{X,m}, \mathbb{P}_X)$, (b) $d_{W_2}(\mathbb{P}_Y, \mathbb{P}_Z) \leq d_{W_2}(\hat{\mathbb{P}}_{Y,m}, \mathbb{P}_Z) + \|\nabla F_{\boldsymbol{\theta}}\|_\infty d_{W_2}(\hat{\mathbb{P}}_{X,m}, \mathbb{P}_X)$.*

For reasonably large m, $d_{W_2}(\hat{\mathbb{P}}_{X,m}, \mathbb{P}_X)$ is small, so from (b), the proximity of the latent distribution and the prior is mainly controlled by the proximity between the empirical distribution $\hat{\mathbb{P}}_{Y,m}$ and $\mathbb{P}_Z$. Together with Proposition 1, this shows that in order to achieve close proximity between $\mathbb{P}_{G(Z)}$ and $\mathbb{P}_X$, we need to have a strong control on the proximity between $\hat{\mathbb{P}}_{Y,m}$ and $\mathbb{P}_Z$ in addition to a good reconstruction. To this end, we propose to use hypothesis tests (HTs) associated with the Wasserstein distance. Note that HTs associated with other statistical distances can also be used in conjunction. Many standard statistical distances are not metric. We will consider several well-known tests, including the Shapiro-Wilk (SW) test (Shapiro & Wilk, 1965). While the SW aims to control the Wasserstein distance, the other tests aim to control statistical distances dominated by the Wasserstein distance. Although at the *population* level, the Wasserstein distance provides stronger separation between different distributions, at the *sample* level, noisy data can exhibit various properties that some tests are sensitive to while others are not. Therefore, it is useful to exploit different tests.

3 The Need for a Higher Criticism: From Local to Global Testing

GoF tests assess whether a sample of data, $\{Y_i\}_{i=1}^m$, is drawn from a distribution class $\mathcal{G}$. The most general opposing hypotheses are $\mathcal{H}_0 : \mathbb{P}_Y \in \mathcal{G}$ vs $\mathcal{H}_1 : \mathbb{P}_Y \notin \mathcal{G}$. A test is specified by a test statistic T and a significance level $\alpha \in (0, 1)$. The observed value of the statistic is $T(\{Y_i\}_{i=1}^m)$ and its (lower tail) p-value is $P(T \leq T(\{Y_i\}_{i=1}^m) \,|\, \mathcal{H}_0)$. $\mathcal{H}_0$ is rejected in favor of $\mathcal{H}_1$ at level α if the p-value is less than α, or equivalently, if $T(\{Y_i\}_{i=1}^m \leq T_\alpha$, where T_α is the α-th quantile of T under $\mathcal{H}_0$. The probability integral transform (PIT) states that if T has a continuous distribution under $\mathcal{H}_0$, the p-values are uniformly distributed on $(0, 1)$ (Murdoch et al., 2008). Multivariate normality (MVN) has received the most attention in the study on multivariate GoF tests. A random variable $Y \in \mathbb{R}^d$ has a normal distribution if and only if for every $\mathbf{u} \in \mathcal{S}$, $Y^T\mathbf{u}$, the projection of Y on $\mathbf{u}$, follows a UVN distribution (Rao et al., 1973), where $\mathcal{S} = \{\mathbf{u} \in \mathbb{R}^d \,|\, \|\mathbf{u}\| = 1\}$ is the unit sphere in $\mathbb{R}^d$. Furthermore, if Y has a non-normal distribution, the set of $\mathbf{u} \in \mathcal{S}$ with $Y^T\mathbf{u}$ being normal has Lebesque measure 0 (Shao & Zhou, 2010). Thus to test on MVN, one way is to apply GoF tests on univariate normality (UVN) to a set of (random) projections of Y and then make a decision based on the resulting collection of test statistics. Many univariate tests exist (Looney, 1995; Mecklin &

Mundfrom, 2005), each focusing on different distributional characteristics. Recently, GoF tests based on reproducing kernel Hilbert space theory have been developed (Liu et al., 2016; Chwialkowski et al., 2016). In this work, we only use traditional GoF tests. We next briefly describe correlation based (CB) tests because they are directly related to the Wasserstein distance. For other test types, see appendix §B. In essence, CB tests on UVN (Dufour et al., 1998) are based on assessment of the correlation between empirical and UVN quantiles. Two common CB tests are Shapiro-Wilk (SW) and Shapiro-Francia (SF) (Shapiro & Francia, 1972). Their test statistics will be denoted by T_{SW} and T_{SF} respectively. Both are closely related to d_{W_2}. For example, for $m \gg 1$, T_{SW} is a strictly decreasing function of $d_{W_2}(\hat{\mathbb{P}}_{Y,m}, \mathcal{N}(0,1))$ (del Barrio et al., 1999), i.e., large T_{SW} values correspond to small d_{W_2}, justifying the rule that $T_{SW} > T_\alpha$ in order *not* to reject $\mathcal{H}_0$.

The Benefits of GoF Testing in Autoencoders. GoF tests on MVN inspect if $\mathbb{P}_Y$ is equal to some $\mathbb{P}_Z \in \mathcal{G}$, where $\mathcal{G}$ is the class of all Gaussians, i.e., MVN distributions. The primary justification for preferring $\mathbb{P}_Z$ as a Gaussian to the more commonly specified isotropic Gaussian (Makhzani et al., 2015; Kingma & Welling, 2013; Higgins et al., 2017; Dai & Wipf, 2019) is that its covariance $\mathbf{\Sigma}$ is allowed to implicitly adapt during training. For example, $\mathbf{\Sigma}$ can be of full rank, allowing for correlations between the latent variables, or singular (i.e., a degenerate MVN, Fig. S1). The ability to adapt to degenerate MVN is of particular benefit. If $F_{\boldsymbol{\theta}}$ is differentiable and the dimension of the latent space is greater than the intrinsic dimension of the data distribution, then the latent variable $Y = F_{\boldsymbol{\theta}}(X)$ only takes values in a manifold whose dimension is less than that of the latent space (Rubenstein et al., 2018). If $\mathbb{P}_Z$ is only allowed to be an isotropic Gaussian, or more generally, a Gaussian with a covariance matrix of full rank, the regularizer in Eq. 1 will promote $F_{\boldsymbol{\theta}}$ that can fill the latent space with $F_{\boldsymbol{\theta}}(X)$, leading to poor sample quality and wrong proportions of generated images (Rubenstein et al., 2018). Since GoF tests use the class $\mathcal{G}$ of all Gaussians, they may help avoid such a situation. Note that this precludes the use of whitening, i.e. the transformation $\mathbf{\Sigma}^{-1/2}[\mathbf{X} - \mathbb{E}(\mathbf{X})]$ to get to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ as the covariance matrix $\mathbf{\Sigma}$ may be singular. In other words, whitening confines available Gaussians to a subset much smaller than $\mathcal{G}$; see the cartoon Fig.(1). Notably, a similar benefit was exemplified in the 2-Stage VAE, where decoupling manifold learning from learning the probability measure stabilized FID scores and sharpened generated samples when the intrinsic dimension is less than the dimension of the latent space (Dai & Wipf, 2019).

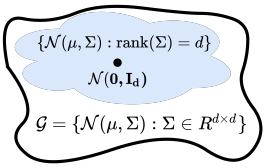


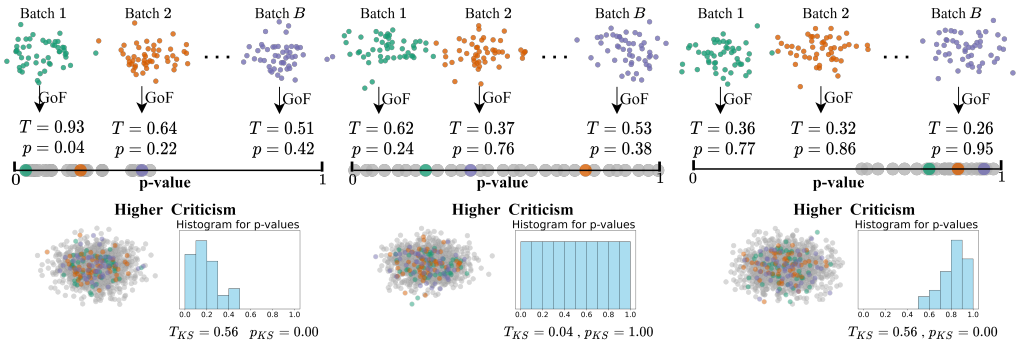
Figure 1: Illustration of distribution family $\mathcal{G}$.

There are several additonal benefits of GoF testing. First, it allows higher criticism (HC) to be used, a topic to be covered shortly. As mentioned in the introduction, GoF tests act at the local level (mini-batch), while HC applies to the aggregated level (training set). Second, many GoF tests produce closed-form test statistics, and thus do not require tuning and provide easily understood output. Lastly, testing for MVN via projections is unaffected by rank deficiency. This is *not* the case with the majority of multivariate GoF tests, e.g., those that act directly on the matrix of observations like the HZ test (Henze & Zirkler, 1990). In fact, any affine invariant test statistic for MVN must be a function of Mahalanobis distance, consequently requiring non-singularity (Henze, 2002).

3.1 A Local Perspective: Goodness of Fit for Normality

In light of Section 2, the aim is to ensure the proximity of the data distribution $\mathbb{P}_X$ and the distribution $\mathbb{P}_{G(Z)}$ by finding $F_{\boldsymbol{\theta}}, G_{\boldsymbol{\phi}}$ such that the reconstruction error is small and the distribution $\mathbb{P}_Y$ of the latent variable $Y = F_{\boldsymbol{\theta}}(X)$ matches the prior $\mathbb{P}_Z$. If the prior is taken to be in $\mathcal{G}$, then given a reconstruction loss function d, we can formulate the problem as $\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}[d(X, G_{\boldsymbol{\phi}}(Y))]$ s.t. $Y = F_{\boldsymbol{\theta}}(X)$ is Gaussian. To enforce the constraint in the minization, a HT can be conducted to decide whether or not to reject the null $\mathcal{H}_0 : \mathbb{P}_{F_{\boldsymbol{\theta}}(X)} \in \mathcal{G}$. Without loss of generality, suppose that we use a test statistic T such that $\mathcal{H}_0$ is rejected when T is *small*. It is natural to include the rejection criterion as a constraint to the optimization problem $\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}[d(X, G_{\boldsymbol{\phi}}(Y))]$, s.t. $\mathbb{E}[T(F_{\boldsymbol{\theta}}(X))] > T_\alpha$. Rewriting with regularization coefficient λ leads to the Lagrangian $\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}[d(X, G_{\boldsymbol{\phi}}(Y))] + \lambda(T_\alpha - \mathbb{E}[T(F_{\boldsymbol{\theta}}(X))])$. Since $\lambda \geq 0$, the final objective can be simplified to $\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}[d(X, G_{\boldsymbol{\phi}}(Y)) - \lambda T(F_{\boldsymbol{\theta}}(X))]$.

When the network is trained using a *single* mini-batch of size m, making Y *less* statistically distinguishable from $\mathcal{G}$ amounts to increasing $\mathbb{E}[T(\{F_{\boldsymbol{\theta}}(X_i)\}_{i=1}^m)]$, where X_i are i.i.d. $\sim \mathbb{P}_X$. However, if the training results in none of the mini-batches yielding small T values to reject $\mathcal{H}_0$ at a given α level, it also indicates a mismatch between the latent and prior distributions. This type of mismatch cannot be detected at the mini-batch level; a more global view is detailed in the next section.



(a) Under (b) Correct (c) Over

Figure 2: Demonstration of model selection. When the regularization coefficient λ is too small (a) or too big (c), GAE is under (or over) regulated, resulting in large values of higher criticism (shown here an example of Kolmogorov–Smirnov (KS) test T_{KS} of the uniformity of GoF test p-values over all mini-batches. (b) shows the case when an appropriate λ is chosen with a small T_{KS} value.

3.2 A Global Perspective: Goodness of Fit for Uniformity - Higher Criticism

Neural networks are powerful universal function approximators (Hornik et al., 1989; Cybenko, 1989). With sufficient capacity it may be possible to train $F_{\boldsymbol{\theta}}$ to *overfit*, that is, to produce *too* many mini-batches with large p-values. Under $\mathcal{H}_0$, the PIT posits p-value uniformity. Therefore, it is expected that after observing many mini-batches, approximately a fraction α of them will have p-values that are less than α. This idea of a more encompassing test is known as Tukey's higher criticism (HC) (Donoho et al., 2004). While *each* mini-batch GoF test is concerned with optimizing for indistinguishability from the prior distribution class $\mathcal{G}$, the HC test is concerned with testing whether the *collection* of p-values is uniformly distributed on $(0, 1)$, which may be accomplished through the Kolmogorov–Smirnov uniformity test. See Algorithm 1 for a pseudo-code and Fig. 2 for an illustration of the HC process. HC and GoF form a symbiotic relationship; HC cannot exist by itself, and GoF by itself may over or under fit, producing p-values in incorrect proportions.

Algorithm 1 Evaluating Higher Criticism

Require: Trained encoder $F_{\boldsymbol{\theta}}$, $\{\mathbf{x}_i\}_{i=1}^N$, GoF test T

for $i = 1 : \lfloor N/m \rfloor$ **do**
  Randomly sample mini-batch $\mathbf{X}$ of size m
  $\mathbf{Y} = G_{\boldsymbol{\theta}}(\mathbf{X})$
  if projection required **then**
    $T^\star = T(\mathbf{Y}\mathbf{u})$, where $\mathbf{u} \in \mathcal{S}$
  else
    $T^\star = T(\mathbf{Y})$
  Calculate p-value of $T^\star$ and store it
Use KS to evaluate uniformity of p-value set

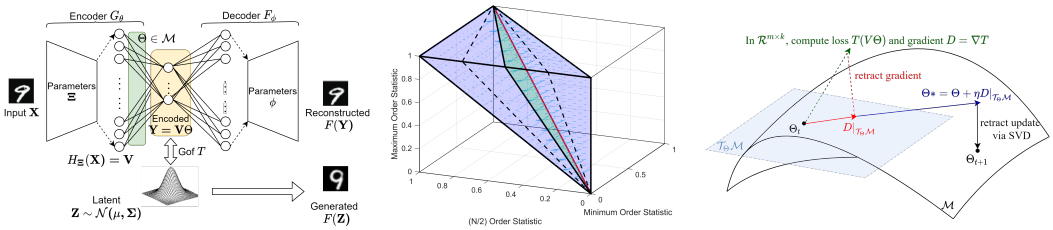
4 Riemannian SGD for Optimizing a Goodness of Fit Autoencoder

A GoF test statistic T is not merely an evaluation mechanism. Its gradient will impact how the model learns what characteristics of a sample indicate strong deviation from normality, carrying over to what the aggregated posterior $\mathbb{P}_Y$ becomes. The following desiderata may help when selecting T. In the following, we denote a collection of sample observations by bold $\mathbf{X}$, $\mathbf{V}$, and $\mathbf{Y}$.

- **HT-Trainable:** $T(F_{\boldsymbol{\theta}}(\mathbf{X}))$ is almost everywhere continuously differentiable in feasible region Ω.
- **HT-Consistent:** There exists a $\boldsymbol{\theta}^\star$ in Ω such that $F_{\boldsymbol{\theta}^\star}(\mathbf{X})$ becomes indistinguishable from the target distribution in the sense that $\mathcal{H}_0$ cannot be rejected.

HT-Trainable is needed if gradient based methods are used to optimize the network parameters. Consider a common encoder architecture $F_{\boldsymbol{\theta}}$ that consists of multiple feature extraction layers (forming a mapping $H_{\boldsymbol{\Xi}}(\mathbf{X})$) followed by a fully connected layer with parameter $\boldsymbol{\Theta}$. Thus, $F_{\boldsymbol{\theta}}(\mathbf{X}) = H_{\boldsymbol{\Xi}}(\mathbf{X})\boldsymbol{\Theta} = \mathbf{V}\boldsymbol{\Theta} = \mathbf{Y}$, and $\boldsymbol{\theta} = \{\boldsymbol{\Xi}, \boldsymbol{\Theta}\}$. The last layer is linear with no activation function as shown in Fig. 3a. With this design, normality can be optimized with respect to the last layer $\boldsymbol{\Theta}$ as discussed below. Given an HT-Consistent statistic T, it is natural to seek a solution, $\boldsymbol{\theta}^\star$.

Theorem 3. *Suppose $\mathbf{V} \in \mathbb{R}^{m\times k}$ is of full row rank. For $\boldsymbol{\Theta} \in \mathbb{R}^{k\times d}$, define $\mathbf{Y} = \mathbf{V}\boldsymbol{\Theta}$. Denote $T_{SW} = T_{SW}(\mathbf{Y}\mathbf{u})$ where $\mathbf{u} \in \mathbb{R}^d$ is a unit vector. Then, T_{SW} is differentiable with respect to $\boldsymbol{\Theta}$ almost everywhere, and $\nabla_{\boldsymbol{\Theta}} T_{SW} = \mathbf{0}$ if and only if $T_{SW} = 1$.*



(a) Full architecture for GoFAE (b) T_{SW} polytope (c) Optimization on $\mathcal{M}$

Figure 3: (a) GoFAE Architecture. (b) Example of singularity. Blue region is a polytope created as $\mathcal{U} = \{\mathbf{x} = [x_1, x_2, x_3] \in \mathbb{R}^3 : 0 \leq x_1 \leq x_2 \leq x_3 \leq 1\}$. For $\forall \mathbf{x} \in \mathcal{U}$, the coordinates of $\mathbf{x}$ are also its order statistics, the min, median and max coordinates corresponds to the x, y and z-axis. The green simplex is where $T_{SW}(\{x_1, x_2, x_3\}) = 1$, and the region built with the dotted lines and the red line creates an acceptance region, outside of which is the rejection region. The light blue arrows are the derivatives at corresponding points and the red boundary line corresponds to the singularity where the gradient blows up. (c) Visualization for GoF optimization on a Stiefel manifold $\mathcal{M}$.

This theorem justifies the use of T_{SW} as an objective function according to HT-Conditions. The largest possible value of T_{SW} is 1, corresponding to an inability to detect deviation from normality no matter what level α is specified. See the appendix for other HT choices.

Identifiability, Singularities, and Stiefel Manifold. If $\mathbf{Y} = \mathbf{V}\mathbf{\Theta}$ is Gaussian for some $\mathbf{\Theta}$, then $\mathbf{V}\mathbf{\Theta}\mathbf{M}$ is also Gaussian for any nonsingular matrix $\mathbf{M}$. Thus, any matrix of the form $\mathbf{\Theta}\mathbf{M}$ is a solution. This leads to several problems when optimizing an objective function containing $\mathbf{\Theta}$ as a variable and the test statistic as part of the objective. First, there is an identifiability issue since, without any restrictions, the set of solutions is infinite. Here, all that matters is the direction of the projection, so restricting the space to the unit sphere is reasonable. Second, almost all GoF tests for normality in the current literature are affine invariant. However, the gradient and Hessian of the test statistics will not be bounded. Fig. 3b illustrates an example of such a singularity issue.

Theorem 4. *Let $T(\{\mathbf{y}_i\}_{i=1}^m)$, $m \geq 3$ be any affine invariant test statistic that is non-constant and differentiable wherever $\mathbf{y}_i$ are not all equal. Then, for any $\mathbf{b}$, as $(\mathbf{y}_1, \ldots, \mathbf{y}_m) \to (\mathbf{b}, \ldots, \mathbf{b})$, $\sup \|\nabla T(\{\mathbf{y}_i\}_{i=1}^m)\| \to \infty$, where $\|\cdot\|$ is the Frobenius norm.*

If $\mathbf{\Theta}$ is searched without being kept away from zero or the diagonal line, traditional SGD results will not be applicable to yield convergence of $\mathbf{\Theta}$. A common strategy might be to lower bound the smallest singular value of $\mathbf{\Theta}^T\mathbf{\Theta}$. However, it does not solve the issue that any re-scaling of $\mathbf{\Theta}$ leads to another solution, so $\mathbf{\Theta}$ must be also upper bounded. It is thus desirable to restrict $\mathbf{\Theta}$ in such a way that avoids getting close to singular by both upper and lower bounding it in terms of its singular values. We propose to restrict $\mathbf{\Theta}$ to the compact Riemannian manifold of orthonormal matrices, i.e. $\mathbf{\Theta} \in \mathcal{M} = \{\mathbf{\Theta} \in \mathbb{R}^{k\times d} : \mathbf{\Theta}^T\mathbf{\Theta} = \mathbf{I}_d\}$ which is also known as the Stiefel manifold. This imposes a feasible region for $\boldsymbol{\theta} \in \Omega = \{\{\mathbf{\Xi}, \mathbf{\Theta}\} : \mathbf{\Theta}^T\mathbf{\Theta} = \mathbf{I}_d\}$. Optimization over a Stiefel manifold has been studied previously (Nishimori & Akaho, 2005; Absil et al., 2009; Cho & Lee, 2017; Bécigneul & Ganea, 2018; Huang et al., 2018; 2020; Li et al., 2020).

4.1 Optimizing Goodness of Fit Test Statistics

A Riemannian metric provides a way to measure lengths and angles of tangent vectors of a smooth manifold. For the Stiefel manifold $\mathcal{M}$, in the Euclidean metric, we have $\langle \mathbf{\Gamma}_1, \mathbf{\Gamma}_2 \rangle = \text{trace}(\mathbf{\Gamma}_1^T\mathbf{\Gamma}_2)$ for any $\mathbf{\Gamma}_1, \mathbf{\Gamma}_2$ in the tangent space of $\mathcal{M}$ at $\mathbf{\Theta}$, $\mathcal{T}_\mathbf{\Theta}\mathcal{M}$. It is known that $\mathcal{T}_\mathbf{\Theta}\mathcal{M} = \{\mathbf{\Gamma} : \mathbf{\Gamma}^T\mathbf{\Theta} + \mathbf{\Theta}^T\mathbf{\Gamma} = \mathbf{0}, \mathbf{\Theta} \in \mathcal{M}\}$ (Li et al., 2020). For arbitrary matrix $\mathbf{\Theta} \in \mathbb{R}^{k\times d}$, the retraction back to $\mathcal{M}$, denoted $\mathbf{\Theta}\big|_\mathcal{M}$, can be performed via a singular value decomposition (SVD), $\mathbf{\Theta}\big|_\mathcal{M} = \mathbf{R}\mathbf{S}^T$, where $\mathbf{\Theta} = \mathbf{R}\mathbf{\Lambda}\mathbf{S}^T$ and $\mathbf{\Lambda}$ is the $d \times d$ diagonal matrix of singular values, $\mathbf{R} \in \mathbb{R}^{k\times d}$ has orthonormal columns, i.e. $\mathbf{R} \in \mathcal{M}$, and $\mathbf{S} \in \mathbb{R}^{d\times d}$ is an orthogonal matrix. The retraction of the gradient $\mathbf{D} = \nabla_\mathbf{\Theta} T_\star$ to $\mathcal{T}_\mathbf{\Theta}\mathcal{M}$, denoted by $\mathbf{D}\big|_{\mathcal{T}_\mathbf{\Theta}\mathcal{M}}$, can be accomplished by $\mathbf{D}\big|_{\mathcal{T}_\mathbf{\Theta}\mathcal{M}} = \mathbf{D} - \mathbf{\Theta}(\mathbf{\Theta}^T\mathbf{D} + \mathbf{D}^T\mathbf{\Theta})/2$ (Fig. 3c). The process of mapping Euclidean gradients to $\mathcal{T}_\mathbf{\Theta}\mathcal{M}$, updating $\mathbf{\Theta}$, and retracting back to $\mathcal{M}$ is well

known (Nishimori & Akaho, 2005). The next result states that the Riemannian gradient for T_{SW} has a finite second moment, which is needed for our convergence theorem in §4.2.

Theorem 5. *Let $T = T_{SW}$ be as in Theorem 3. Denote by $\nabla_{\boldsymbol{\theta}}$, ∇_{Θ}, the Riemannian gradient w.r.t. $\boldsymbol{\theta}$, $\boldsymbol{\Theta}$, respectively, and $\|\cdot\|$ the Frobenius norm. Let $\mathbf{B} = (\mathbf{I} - \mathbf{J}/m)\mathbf{V}$, i.e., $\mathbf{B}$ is obtained by subtracting from each row of $\mathbf{V}$ the mean of the rows. Suppose (a) $\sup_{\boldsymbol{\Xi}}(\mathbb{E}[\|\mathbf{B}\|^4] + \mathbf{E}[\|\nabla_{\boldsymbol{\Xi}}\mathbf{B}\|^4]) < \infty$, and (b) $\sup_{\|\mathbf{x}\|=1,\boldsymbol{\Xi}} \mathbb{E}[\|\mathbf{B}\mathbf{x}\|^{-4}] < \infty$. Then, $\sup_{\boldsymbol{\theta}} \mathbb{E}[\|\nabla_{\boldsymbol{\theta}} T\|^2] < \infty$.*

4.2 Optimizing the Goodness of Fit Autoencoder

Given a mini-batch of size m, $\{\mathbf{x_i}\}_{i=1}^m$, the GoFAE loss function is written as[2] :

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \{\mathbf{x}_i\}_{i=1}^m) = \frac{1}{m}\sum_{i=1}^{m} d(\mathbf{x}_i, G_{\boldsymbol{\phi}}(F_{\boldsymbol{\theta}}(\mathbf{x}_i))) \pm \lambda T(\{F_{\boldsymbol{\theta}}(\mathbf{x}_i)\}_{i=1}^m), \tag{2}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\Xi}, \boldsymbol{\Theta}\}$. Since $\boldsymbol{\Theta} \in \mathcal{M}$, model training requires Riemannian SGD. The proof of Bonnabel (2013) for the convergence of SGD on a Riemannian manifold, which extended from the Euclidean case (Bottou, 1998), requires conditions on step size, differentiability, and a uniform bound on the stochastic gradient.We show that this result holds under a much weaker condition, eliminating the need for a uniform bound. While we apply the resulting theorem to our specific problem with a Stiefel manifold, we emphasize that Theorem 6 is applicable to other suitable Riemannian manifolds.

Theorem 6. *Let $\mathcal{M}$ be a connected Riemannian manifold with injectivity radius uniformly bounded from below by $I > 0$. Let $C \in C^{(3)}(\mathcal{M})$ and R_w be a twice continuously differentiable retraction. Let $z_0, z_1, \ldots$ be i.i.d. $\sim \zeta$ taking values in $\mathcal{Z}$. Let $H : \mathcal{Z} \times \mathcal{M} \to \mathcal{TM}$ be a measurable function such that $\mathbb{E}[H(\zeta, w)] = \nabla C(w)$ for all $w \in \mathcal{M}$, where $\mathcal{TM}$ is the tangent bundle of $\mathcal{M}$. Consider the SGD update $w_{t+1} = R_{w_t}(-\gamma_t H(z_t, w_t))$ with stepsize $\gamma_t > 0$ satisfying $\sum \gamma_t^2 < +\infty$ and $\sum \gamma_t = +\infty$. Suppose there exists a compact set K such that all $w_t \in K$ and $\sup_{w \in K} \mathbb{E}[\|H(\zeta, w)\|^2] \leq A^2$ for some $A > 0$. Then, $C(w_t)$ converges almost surely and $\nabla C(w_t) \to 0$ almost surely.*

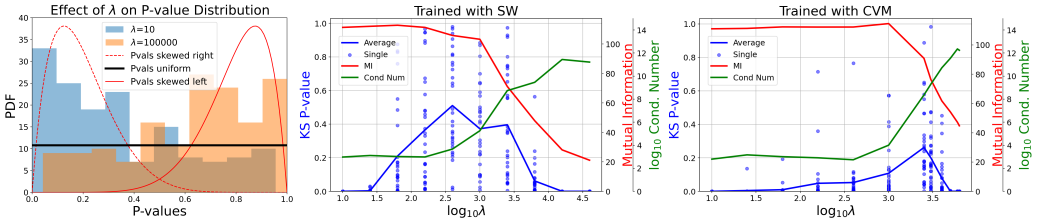
In our context, $C(\cdot)$ corresponds to Eq. (2). The parameters of GoFAE are $\{\boldsymbol{\theta}, \boldsymbol{\phi}\} = \{\boldsymbol{\Xi}, \boldsymbol{\Theta}, \boldsymbol{\phi}\}$ where $\boldsymbol{\Xi}, \boldsymbol{\phi}$ are defined in Euclidean space, a Riemannian manifold endowed with the Euclidean metric, and $\boldsymbol{\Theta}$ on the Stiefel manifold. Therefore, $\{\boldsymbol{\Xi}, \boldsymbol{\Theta}, \boldsymbol{\phi}\}$ is a product manifold that is also Riemannian and $\{\boldsymbol{\Xi}, \boldsymbol{\Theta}\ \boldsymbol{\phi}\}$ are updated simultaneously. Convergence of GoFAE holds from Proposition 5 and Theorem 6 provided that $\boldsymbol{\Xi}, \boldsymbol{\phi}$ stay within a compact set. Algorithm 2 depicts the GoFAE pseudo-code.

5 Experiments

We evaluate generation and reconstruction performance, normality, and informativeness of GoFAE using several GoF statistics on the MNIST (LeCun et al., 1998), CelebA (Liu et al., 2015) and CIFAR10 (Krizhevsky et al., 2009) datasets and compare to several other GAE models. We emphasize that our goal is not to merely to produce competitive evaluation metrics, but to provide a principled way to balance the reconstruction versus prior matching trade-off. For MNIST and CelebA the architectures are based on Tolstikhin et al. (2017), while CIFAR10 is from Lippe (2022). The aim is to keep the architecture consistent across models with the exception of method specific components. Please see the appendix for complete architectures (§C), training details (§D), and additional results (§E). The following results are from experiments on CelebA.

Effect of $\boldsymbol{\lambda}$ and Mutual Information (MI). We investigated the balance between generation and reconstruction in the GoFAE by considering mini-batch (local) GoF test p-value distributions, aggregated posterior (global) normality, and MI as a function of λ (Fig. 4). Global normality as assessed through the higher criticism (HC) principle for p-value uniformity denoted by KS_{unif} and computed using Algorithm 1. We trained GoFAE on CelebA using T_{SW} as the test statistic for $\lambda = 10$ (Fig. 4a, blue) and $\lambda = 100,000$ (Fig. 4a, yellow). For small λ, the model emphasizes reconstruction and the p-value distribution will be skewed right since the penalty for deviating from normality is small (Fig. 4a, red dashed line). As λ increases, more emphasis is placed on normality and, for some λ, p-values are expected to be uniformly distributed (Fig. 4a, black line). If λ is too large, the p-value distribution will be left-skewed (Fig. 4a, red solid line), which corresponds to overfitting the prior.

[2]The sign of λ is determined by the test statistic. Since T_{SW} is our exemplar it would be negative.



(a) p-value distribution (b) Shapiro-Wilk (c) Cramer von Mises

Figure 4: Effects of λ on p-value distribution and mutual information for GoFAE models.

We assessed the normality of GoFAE mini-batch encodings, using 30 repetitions of Algorithm 1, for different λ and KS_{unif} (Figs. 4b-4c). The blue points and blue solid line represent the KS_{unif} test p-values and mean, respectively. HC rejects global normality (KS_{unif}<0.05) when the test statistic p-values are right-skewed (too many mini-batches deviate from normality) or left-skewed (mini-batches are *too* normal); this GoFAE property discourages the posterior from over-fitting the prior. The range of λ values for which the distribution of GoF test p-values is indistinguishable from uniform is where the mean $KS_{unif} \geq 0.05$ (Figs. 4b-4c).

Algorithm 2 GoFAE Optimization

Require: test T, learning rates: η_1, η_2, max iter. J
Initialize: $\boldsymbol{\Xi}, \boldsymbol{\Theta}, \boldsymbol{\phi}$
while $j < J$ **do**
  Randomly sample mini-batch of size m, $\mathbf{X}$
  $\mathbf{V} = F^{(1)}_{\boldsymbol{\Xi}_j}(\mathbf{X})$
  $\mathbf{Y} = F^{(2)}_{\boldsymbol{\Theta}_j}(\mathbf{V}) = \mathbf{V}\boldsymbol{\Theta}_j$
  if test requires projection **then**
    $T^\star = T(\mathbf{Y}\mathbf{u})$, where $\mathbf{u} \in \mathcal{S}$
  else
    $T^\star = T(\mathbf{Y})$
  $\mathcal{L} = d(\mathbf{X}, G_{\boldsymbol{\phi}}(\mathbf{Y})) \pm \lambda T^\star$
  $\boldsymbol{\Xi}_{j+1} = \boldsymbol{\Xi}_j - \eta_1 \nabla_{\boldsymbol{\Xi}} \mathcal{L}$ or other optim
  $\boldsymbol{\phi}_{j+1} = \boldsymbol{\phi}_j - \eta_1 \nabla_{\boldsymbol{\phi}} \mathcal{L}$ or other optim
  $\mathbf{D} = \nabla_{\boldsymbol{\Theta}} T^\star$
  $\boldsymbol{\Gamma} = \mathbf{D}\big|_{\mathcal{T}_{\boldsymbol{\Theta}}\mathcal{M}} = \mathbf{D} - \boldsymbol{\Theta}_j(\boldsymbol{\Theta}_j^T\mathbf{D} + \mathbf{D}^T\boldsymbol{\Theta}_j)/2$
  $\boldsymbol{\Theta}'_{j+1} = \boldsymbol{\Theta}_j + \eta_2 \boldsymbol{\Gamma}$
  Compute $\mathbf{R}\boldsymbol{\Lambda}\mathbf{S}^T = \text{SVD}(\boldsymbol{\Theta}'_{j+1})$
  $\boldsymbol{\Theta}_{j+1} = \boldsymbol{\Theta}'_{j+1}\big|_{\mathcal{M}} = \mathbf{R}\mathbf{S}^T$

Finally, we used the estimated $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ for each model to draw $N = 10^4$ samples $\{\mathbf{z}_i\}_{i=1}^N$ and generate images $\{G_{\boldsymbol{\phi}}(\mathbf{z}_i)\}_{i=1}^N$. These images are then encoded, giving the set $\{\tilde{\mathbf{z}}_i\}_{i=1}^N$ with $\tilde{\mathbf{z}}_i = F_{\boldsymbol{\theta}}(G_{\boldsymbol{\phi}}(\mathbf{z}_i))$, resulting in a Markov chain $\{\mathbf{z}_i\}_{i=1}^N \to \{G_{\boldsymbol{\phi}}(\mathbf{z}_i)\}_{i=1}^N \to \{\tilde{\mathbf{z}}_i\}_{i=1}^N$. Assuming $\tilde{\mathbf{z}}_i$ are also normal, the data processing inequality gives a lower bound on the MI (Figs. 4b-4c, red line). As λ increases the MI diminishes, suggesting the posterior overfits the prior as the encoding becomes independent of the data. The unimodality of the average KS_{unif} and monotonicity of MI suggests λ can be selected solely based on KS_{unif} without reference to a performance criterion.

Gaussian Degeneracy. Due to noise in the data, numerically the aggregate posterior can never be singular. Nevertheless, the experiments indicate that the GoF test can push the aggregated posterior to become "more and more singular" in the sense that the condition number of its covariance matrix, $\kappa(\hat{\boldsymbol{\Sigma}})$, becomes increasingly large. As λ continues to increase, $\kappa(\hat{\boldsymbol{\Sigma}})$ also increases (Figs. 4b-4c, green line), implying a Gaussian increasingly concentrated around a lower-dimensional linear manifold. The same trend is shown in Fig.5, which depicts the spectrum of singular values (SV) of $\hat{\boldsymbol{\Sigma}}$ for each λ after training with T_{SW}. From the figure, it is seen that while the spectrum does not exhibit extremely large SVs, it has many small SVs. Notably in Figs. 4b-4c even when λ is not too large, $\kappa(\hat{\boldsymbol{\Sigma}})$ is relatively large, indicating a shift to lower-dimension. However, MI remains relatively large and the KS_{unif} test suggests the aggregated posterior is still indistinguishable from Gaussian. Together, these results are evidence that the GoFAE can adapt as needed to a reduced-dimension representation while maintaining the representation's informativeness.

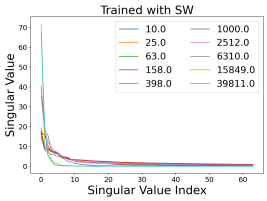


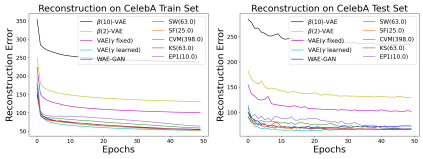
Figure 5: SV of $\hat{\boldsymbol{\Sigma}}$

Reconstruction, Generation, and Normality. We assessed the quality of the generated and test set reconstructed images using Fréchet Inception Distance (FID) (Heusel et al., 2017) and mean-square error (MSE) based on 10^4 samples (Table 1). We compared GoFAE with AE (Bengio et al., 2013), VAE (Kingma & Welling, 2013), VAE with learned γ (Dai & Wipf, 2019), 2-Stage VAE (Dai & Wipf, 2019), β-VAE (Higgins et al., 2017) and WAE-GAN (Tolstikhin et al., 2017); convergence was assessed by tracking the test set reconstruction error over training epochs (Tables S2-S8 and Fig. S9). Figs. 6a, 6b are for models presented in Table 1. We selected the smallest λ whose mean p-value of the HC

Table 1: Evaluation of CelebA by MSE, FID scores, and samples with p-values from higher criticism.

Algorithm	MSE ↓	FID Score ↓		Kolmogorov-Smirnov Uniformity Test				
		Recon.	Gen.	SW	SF	CVM	KS	EP
AE-Baseline	69.32	40.28	126.43	0.0(0.0)	0.0(0.0)	0.0(0.0)	0.0(0.0)	0.0(0.0)
VAE (fixed γ)	101.78	49.29	56.78	0.01(0.05)	0.0(0.01)	0.10(0.19)	0.30(0.30)	0.23(0.27)
VAE (learned γ)	68.09	34.54	58.15	0.0(0.0)	0.0(0.0)	0.0(0.0)	0.0(0.0)	0.0(0.0)
2-Stage-VAE	68.09	34.54	49.55	0.51(0.26)	0.49(0.27)	0.49(0.29)	0.52(0.28)	0.49(0.32)
$\beta(10)$-VAE	235.37	99.59	101.30	0.56(0.31)	0.46(0.35)	0.47(0.28)	0.38(0.25)	0.44(0.27)
$\beta(2)$-VAE	129.72	60.72	65.65	0.19(0.21)	0.06(0.11)	0.42(0.29)	0.52(0.30)	0.44(0.27)
WAE-GAN	72.70	38.23	48.39	0.01(0.04)	0.0(0.0)	0.11(0.18)	0.14(0.19)	0.13(0.15)
GoFAE-SW(63)	**66.11**	34.69	43.76	0.14(0.23)	0.28(0.26)	0.01(0.03)	0.08(0.17)	0.08(0.17)
GoFAE-SF(25)	66.35	**33.89**	**43.3**	0.04(0.09)	0.12(0.20)	0.0(0.01)	0.02(0.04)	0.01(0.03)
GoFAE-CVM(398)	67.43	37.50	46.74	0.19(0.22)	0.34(0.29)	0.05(0.08)	0.16(0.27)	0.19(0.28)
GoFAE-KS(63)	66.62	35.14	43.93	0.0(0.00)	0.0(0.00)	0.0(0.00)	0.00(0.00)	0.00(0.00)
GoFAE-EP(10)	67.03	39.71	48.53	0.03(0.08)	0.06(0.10)	0.00(0.00)	0.00(0.00)	0.02(0.04)

uniformity test, KS_{unif}, was greater than 0.05 for GoFAE models. The CB GoF tests have the most competitive performance on FID and test set MSE. We assessed the normality of mini-batch encodings across each method using several GoF tests for normality combined with KS_{unif}. We ran 30 repetitions of Algorithm 1 for each method and reported the mean (std) of the KS_{unif} p-values in Table 1. In addition to superior MSE and FID scores, the GoFAE models obtained uniform p-values under varying GoF tests. Variability across GoF tests highlights the fact that different tests are sensitive to different distributional characteristics. Qualitative and quantitative assessments (§E.3) and convergence plots (§E.4) are given in the appendix for MNIST, CIFAR-10, and CelebA. Finally, an ablation study provided empirical justification for Riemannian SGD in the GoFAE (§E.5).



(a) Train (b) Test

Figure 6: Convergence of reconstruction error on the CelebA training (a) and testing (b) sets.

6 Summary and Conclusion

We presented the GoFAE, a deterministic GAE based on optimal transport and Wasserstein distances that optimizes GoF hypothesis test statistics. We showed that gradient based optimization of GoFAE induces identifiability issues, unbounded domains, and gradient singularities, which we resolve using Riemannian SGD. By using GoF statistics to measure deviation from the prior *class* of Gaussians, our model is capable of implicitly adapting its covariance matrix during training from full-rank to singular, which we demonstrate empirically. We developed a performance agnostic model selection algorithm based on higher criticism of p-values for global normality testing of the aggregated posterior. Collectively, empirical results show that GoFAE achieves comparable reconstruction and generation performance while retaining statistical indistinguishability from Gaussian in the latent space.

Several limitations in our methods provide opportunities for future work. First, a most powerful GoF test for normality does not exist. Each test is sensitive to certain characteristics of the null distribution. In GAEs, the induced distribution is unknown; hence we can't know which test is best for the task. Solutions might include a weighted combination of tests or randomly selecting one for each iteration. Second, we solved difficulties in high-dimensional scaling of UVN GoF tests by projection, which can be slow for univariate tests because many samples from the unit sphere are required. Third, the increased interpretability of test statistic regularization comes with the danger of p-hacking by permuting the GoF test and regularization coefficient until significance is achieved (Head et al., 2015). This issue applies more broadly, can lead to biased and misleading findings, and possibly be used deceptively by wrongfully claiming statistical significance (Brodeur et al., 2020); however, p-hacking can be detected with binomial tests (Gerber & Malhotra, 2008; Simmons et al., 2011), ameliorated by correcting for multiple comparisons (Streiner, 2015), or avoided with Bayesian posterior probabilities (Gelman & Loken, 2013). Lastly, the theoretical principles concerning overfitting while using a test statistic as a regularizer are not well understood and requires further investigation.

REFERENCES

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pp. 159–168. PMLR, 2018.

Theodore W Anderson, Donald A Darling, et al. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics*, 23(2):193–212, 1952.

Ludwig Baringhaus and Norbert Henze. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35(1):339–348, 1988.

Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. *arXiv preprint arXiv:1810.00760*, 2018.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Léon Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schoelkopf. From optimal transport to generative modeling: the vegan cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1002. URL `https://aclanthology.org/K16-1002`.

Abel Brodeur, Nikolai Cook, and Anthony Heyes. Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110(11):3634–60, 2020.

Minhyung Cho and Jaehyung Lee. Riemannian approach to batch normalization. In *Advances in Neural Information Processing Systems*, pp. 5225–5235, 2017.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International conference on machine learning*, pp. 2606–2615. PMLR, 2016.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

RalphB D'Agostino. *Goodness-of-fit-techniques*. Routledge, 2017.

Bin Dai and David Wipf. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019.

Eustasio del Barrio, Juan A Cuesta-Albertos, Carlos Matrán, and Jesús M Rodríguez-Rodríguez. Tests of goodness of fit based on the l2-wasserstein distance. *Annals of Statistics*, pp. 1230–1239, 1999.

Lizhong Ding, Mengyang Yu, Li Liu, Fan Zhu, Yong Liu, Yu Li, and Ling Shao. Two generator game: Learning to sample via linear goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pp. 11260–11271, 2019.

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

David Donoho, Jiashun Jin, et al. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

Jean-Marie Dufour, Abdeljelil Farhat, Lucien Gardiol, and Lynda Khalaf. Simulation-based finite sample normality tests in linear regressions. *The Econometrics Journal*, 1(1):154–173, 1998.

Thomas W Epps and Lawrence B Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983.

Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 348, 2013.

Alan S Gerber and Neil Malhotra. Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research*, 37(1):3–30, 2008.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Megan L Head, Luke Holman, Rob Lanfear, Andrew T Kahn, and Michael D Jennions. The extent and consequences of p-hacking in science. *PLoS Biol*, 13(3):e1002106, 2015.

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.

Norbert Henze. Invariant tests for multivariate normality: a critical review. *Statistical papers*, 43(4): 467–506, 2002.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, 2016.

Daniella Horan, Eitan Richardson, and Yair Weiss. When is unsupervised disentanglement possible? *Advances in Neural Information Processing Systems*, 34, 2021.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Lei Huang, Xianglong Liu, Bo Lang, Adams Yu, Yongliang Wang, and Bo Li. Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Lei Huang, Xianglong Liu, Jie Qin, Fan Zhu, Li Liu, and Ling Shao. Projection based weight normalization: Efficient method for optimization on oblique manifold in dnns. *Pattern Recognition*, 105:107317, 2020.

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

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

Diederik P Kingma, Danilo J Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, pp. 3581–3589, 2014.

Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Yuta Koike. High-dimensional central limit theorems for homogeneous sums. *arXiv preprint arXiv:1902.03809*, 2019.

Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. Sliced-wasserstein autoencoder: An embarrassingly simple generative model. *arXiv preprint arXiv:1804.01947*, 2018.

Selcuk Korkmaz, Dincer Goksuluk, and Gokmen Zararsiz. Mvn: An r package for assessing multivariate normality. *The R Journal*, 6(2):151–162, 2014.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

Alexis Lechat, Stéphane Herbin, and Frédéric Jurie. Semi-supervised class incremental learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 10383–10389. IEEE, 2021.

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

Jun Li, Li Fuxin, and Sinisa Todorovic. Efficient riemannian optimization on the stiefel manifold via the cayley transform. *arXiv preprint arXiv:2002.01113*, 2020.

Phillip Lippe. UvA Deep Learning Tutorials. `https://uvadlc-notebooks.readthedocs.io/en/latest/`, 2022.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pp. 276–284. PMLR, 2016.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Stephen W Looney. How to use tests for univariate normality to assess multivariate normality. *The American Statistician*, 49(1):64–70, 1995.

James Lucas, George Tucker, Roger B Grosse, and Mohammad Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. *Advances in Neural Information Processing Systems*, 32, 2019.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Christopher J Mecklin and Daniel 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.

Duncan J Murdoch, Yu-Ling Tsai, and James Adcock. P-values are random variables. *The American Statistician*, 62(3):242–245, 2008.

Yasunori Nishimori and Shotaro Akaho. Learning algorithms utilizing quasi-geodesic flows on the stiefel manifold. *Neurocomputing*, 67:106–135, 2005.

Aaron Palmer, Dipak Dey, and Jinbo Bi. Reforming generative autoencoders via goodness-of-fit hypothesis testing. In *UAI*, pp. 1009–1019, 2018.

Giorgio Patrini, Marcello Carioni, Patrick Forre, Samarth Bhargav, Max Welling, Rianne van den Berg, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. *arXiv preprint arXiv:1810.01118*, 2018.

Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Mathematischer Statistiker, Calyampudi Radhakrishna Rao, and Calyampudi Radhakrishna Rao. *Linear statistical inference and its applications*, volume 2. Wiley New York, 1973.

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

Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. *Advances in Neural Information Processing Systems*, 31:185–194, 2018.

Nathan Ross. Fundamentals of stein's method. *Probability Surveys*, 8:210–293, 2011.

Paul K Rubenstein, Bernhard Schoelkopf, and Ilya Tolstikhin. On the latent space of wasserstein auto-encoders. *arXiv preprint arXiv:1802.03761*, 2018.

Lars Ruthotto and Eldad Haber. An introduction to deep generative modeling. *arXiv preprint arXiv:2103.05180*, 2021.

Yongzhao Shao and Ming Zhou. A characterization of multivariate normality through univariate projections. *Journal of multivariate analysis*, 101(10):2637–2640, 2010.

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

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

Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3745–3753, 2016.

Michael A Stephens. *Goodness-of-fit techniques*. Routledge, 2017.

David L Streiner. Best (but oft-forgotten) practices: the multiple problems of multiplicity—whether and how to correct for many statistical tests. *The American journal of clinical nutrition*, 102(4): 721–728, 2015.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:2007.03898*, 2020.

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

Tianlin Xu, Li Kevin Wenliang, Michael Munn, and Beatrice Acciaio. Cot-gan: Generating sequential data via causal optimal transport. *Advances in Neural Information Processing Systems*, 33:8798–8809, 2020.

A PROOFS

A.1 PROOF OF PROPOSITION 1.

Recall that in the setup of the result, X is a random variable taking values in $\mathbb{R}^m$, Y and Z are random variables taking values in $\mathbb{R}^d$, $F : \mathbb{R}^m \to \mathbb{R}^d$ and $G : \mathbb{R}^d \to \mathbb{R}^m$ are two mappings, and X and Y are linked by $Y = F(X)$.

Proposition 1. *If $\mathcal{X}$ and $\mathcal{Z}$ are Euclidean spaces and the decoder G is differentiable, then (a) $d_{W_2}(\mathbb{P}_{G(Y)}, \mathbb{P}_{G(Z)}) \leq \|\nabla G\|_\infty d_{W_2}(\mathbb{P}_Y, \mathbb{P}_Z)$, (b) $d_{W_2}(\mathbb{P}_X, \mathbb{P}_{G(Z)}) \leq [\mathbb{E}\|X - G(Y)\|^2]^{1/2} + \|\nabla G\|_\infty d_{W_2}(\mathbb{P}_Y, \mathbb{P}_Z)$.*

Proof. (a) Recall that for any two probability measures μ and ν on the same Euclidean space,

$$d^2_{W_2}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int \|y - z\|^2 \pi(\mathrm{d}z, \mathrm{d}z).$$

Let $f(x, y) = (G(x), G(y))$. For any $Q \in \Pi(\mathbb{P}_Y, \mathbb{P}_Z)$, the measure $Q'(\cdot) = Q \circ f^{-1}(\cdot) \in \Pi(\mathbb{P}_{G(Y)}, \mathbb{P}_{G(Z)})$, so

$$\begin{aligned} d^2_W(\mathbb{P}_{G(Y)}, \mathbb{P}_{G(Z)}) \leq \int \|y' - z'\|^2 Q'(\mathrm{d}y', \mathrm{d}z') &= \int \|G(y) - G(z)\|^2 Q(\mathrm{d}y, \mathrm{d}z) \\ &\leq \|\nabla G\|^2_\infty \int \|y - z\|^2 Q(\mathrm{d}y, \mathrm{d}z). \end{aligned}$$

Taking infimum of the right hand side over all $Q \in \Pi(\mathbb{P}_Y, \mathbb{P}_Z)$ gives the result.

(b) By triangular inequality for d_{W_2} and part (a),

$$\begin{aligned} d_{W_2}(\mathbb{P}_X, \mathbb{P}_{G(Z)}) &\leq d_{W_2}(\mathbb{P}_X, \mathbb{P}_{G(Y)}) + d_{W_2}(\mathbb{P}_{G(y)}, \mathbb{P}_{G(Z)}) \\ &\leq [\mathbb{E}\|X - G(Y)\|^2]^{1/2} + \|\nabla G\|_\infty d_{W_2}(\mathbb{P}_Y, \mathbb{P}_Z). \end{aligned}$$

□

A.2 PROOF OF PROPOSITION 2.

Proposition 2. *Let $\hat{\mathbb{P}}_{X,m}$ be the empirical distribution of samples $\{X_i\}_{i=1}^m$, and $\hat{\mathbb{P}}_{Y,m}$ be the empirical distribution of $\{Y_i = F_{\boldsymbol{\theta}}(X_i)\}_{i=1}^m$. Assume that $F_{\boldsymbol{\theta}}(X)$ is differentiable with respect to X with bounded gradients $\nabla F_{\boldsymbol{\theta}}(X)$. Then, (a) $d_{W_2}(\hat{\mathbb{P}}_{Y,m}, \mathbb{P}_Y) \leq \|\nabla F_{\boldsymbol{\theta}}\|_\infty d_{W_2}(\hat{\mathbb{P}}_{X,m}, \mathbb{P}_X)$, (b) $d_{W_2}(\mathbb{P}_Y, \mathbb{P}_Z) \leq d_{W_2}(\hat{\mathbb{P}}_{Y,m}, \mathbb{P}_Z) + \|\nabla F_{\boldsymbol{\theta}}\|_\infty d_{W_2}(\hat{\mathbb{P}}_{X,m}, \mathbb{P}_X)$.*

Proof. (a) is a direct consequence of Proposition 1 (a). (b) follows by combining (a) with $d_{W_2}(\mathbb{P}_Y, \mathbb{P}_Z) \leq d_{W_2}(\hat{\mathbb{P}}_{Y,m}, \mathbb{P}_Z) + d_{W_2}(\hat{\mathbb{P}}_{Y,m}, \mathbb{P}_Y)$. □

A.3 PROOF OF REMARK IN SECTION 3 LINE 168.

Remark. *$d_{W_2}(\hat{\mathbb{P}}_{Y,m}, \mathbb{P}_Z) \leq k_{\alpha,m}$ strongly suggests small $d_{W_2}(\mathbb{P}_Y, \mathbb{P}_Z)$.*

Proof. First, it is reasonable to assume that, as long as the family of encoders $F_{\boldsymbol{\theta}}(\cdot)$ is regular enough, $\|\nabla F_{\boldsymbol{\theta}}(\cdot)\|_\infty$ is bounded for all $\boldsymbol{\theta}$ by, say $M > 0$. Second, one can expect that $\lim_{m\to\infty} d_{W_2}(\hat{\mathbb{P}}_{X,m}, \mathbb{P}_X) \xrightarrow{\text{Prob.}} 0$, so for m reasonably large, the probability for $d_{W_2}(\hat{\mathbb{P}}_{X,m}, \mathbb{P}_X)$ to be large is small. Also, for m large enough, $k_{\alpha,m}$ is small. Therefore, if $d_{W_2}(\mathbb{P}_Y, \mathbb{P}_Z)$ is large (note that it is a constant), then by Theorem 2(b), the event that $d_{W_2}(\hat{\mathbb{P}}_{Y,m}, \mathbb{P}_Z) \leq k_{\alpha,m}$ is a sub-event of $d_{W_2}(\hat{\mathbb{P}}_{X,m}, \mathbb{P}_X) \geq M^{-1}(d_{W_2}(\mathbb{P}_Y, \mathbb{P}_Z) - k_{\alpha,m})$. The latter is an event of small probability. □

A.4 PROOF OF THEOREM 4.

As described in Section B.1, the Shapiro–Wilk statistic of a sample $X_1, \ldots, X_m$ is

$$T_{SW} = \frac{(\sum_{i=1}^m a_i X_{(i)})^2}{\sum_{i=1}^m (X_i - \bar{X})^2},$$

where $X_{(1)} \leq X_{(2)} \cdots \leq X_{(m)}$ are the order statistics and a_i are certain constants that are all different. Let $\mathbf{a} = (a_1, \ldots, a_m)^T$. We only need the fact that $\|\mathbf{a}\| = 1$ and $\mathbf{1}^T\mathbf{a} = 0$, where $\mathbf{1}$ is the vector of m 1's.

Theorem 4. *Suppose $\mathbf{V} \in \mathbb{R}^{m\times k}$ is of full row rank. For $\mathbf{\Theta} \in \mathbb{R}^{k\times d}$, define $\mathbf{Y} = \mathbf{V\Theta}$. Denote $T_{SW} = T_{SW}(\mathbf{Yu})$ where $\mathbf{u} \in \mathbb{R}^d$ is a unit vector. Then, T_{SW} is differentiable with respect to $\mathbf{\Theta}$ almost everywhere, and $\nabla_{\mathbf{\Theta}} T_{SW} = \mathbf{0}$ if and only if $T_{SW} = 1$.*

Proof. Let $\mathbf{V}_*$ be a row permutation of $\mathbf{V}$ such that the coordinates of $\mathbf{V}_*\mathbf{\Theta u}$ are in increasing order. Denote by $\mathbf{I}$ the $m \times m$ identity matrix and $\mathbf{J} = \mathbf{1}\mathbf{1}^T$. Put $\mathbf{B} = (\mathbf{I} - \mathbf{J}/m)\mathbf{V}_*$ and note that $\mathbf{a}^T\mathbf{V}_* = \mathbf{a}^T\mathbf{B}$ as $\mathbf{a}^T\mathbf{J} = \mathbf{a}^T\mathbf{1}\mathbf{1}^T = (\mathbf{1}^T\mathbf{a})^T\mathbf{1} = \mathbf{0}$. Then

$$T_{SW} = \frac{(\mathbf{a}^T\mathbf{V}_*\mathbf{\Theta u})^2}{\|(\mathbf{I} - \mathbf{J}/m)\mathbf{V}_*\mathbf{\Theta u}\|^2} = \frac{(\mathbf{a}^T\mathbf{B\Theta u})^2}{\|\mathbf{B\Theta u}\|^2}. \tag{S3}$$

It is easy to see that T_{SW} is differentiable at $\mathbf{\Theta}$ if all $\mathbf{v}_i^T\mathbf{\Theta u}$ are different, where $\mathbf{v}_1^T, \ldots, \mathbf{v}_m^T$ are the rows of $\mathbf{V}$. Also, in this case, $\mathbf{V}_*$ is unique and $\mathbf{B\Theta u} \neq 0$. On the other hand, for $i \neq j$, since $\mathbf{u} \neq \mathbf{0}$ and $\mathbf{v}_i \neq \mathbf{v}_j$ as $\mathbf{V}$ is of full row rank, the set of $\mathbf{\Theta}$ with $(\mathbf{v}_i - \mathbf{v}_j)^T\mathbf{\Theta u} = 0$ is a strict linear subspace of $\mathbb{R}^{k\times d}$, so has Lebesgue measure 0. Then T_{SW} is differentiable in $\mathbf{\Theta}$ almost everywhere.

Fix $\mathbf{\Theta}$ with all the coordinates of $\mathbf{V\Theta u}$ being different. Then

$$\nabla_{\mathbf{\Theta}} T_{SW} = \frac{2(\mathbf{a}^T\mathbf{B\Theta u})}{\|\mathbf{B\Theta u}\|^2}\mathbf{B}^T\left[\mathbf{a} - \frac{(\mathbf{a}^T\mathbf{B\Theta u})(\mathbf{B\Theta u})}{\|\mathbf{B\Theta u}\|^2}\right]\mathbf{u}^T. \tag{S4}$$

Note that $\mathbf{a}^T\mathbf{B\Theta u}$ is a scalar. Also, $\mathbf{a}^T\mathbf{B\Theta u} \neq 0$, for otherwise from (S3) $T_{SW} = 0$, contradicting Lemma 3 (Shapiro & Wilk, 1965).

Suppose $\nabla_{\mathbf{\Theta}} T_{SW} = \mathbf{0}$. Then from (S4),

$$\mathbf{B}^T\left[\mathbf{a} - \frac{(\mathbf{a}^T\mathbf{B\Theta u})(\mathbf{B\Theta u})}{\|\mathbf{B\Theta u}\|^2}\right]\mathbf{u}^T = \mathbf{0}.$$

Let $c = \frac{\|\mathbf{B\Theta u}\|^2}{\mathbf{a}^T\mathbf{B\Theta u}}$ and $\mathbf{h} = \mathbf{B\Theta u} - c\mathbf{a}$. Then the above equation can be written as

$$\mathbf{B}^T\left[\mathbf{a} - \frac{\mathbf{B\Theta u}}{c}\right]\mathbf{u}^T = -c^{-1}\mathbf{B}^T\mathbf{h}\mathbf{u}^T = \mathbf{0},$$

giving $\mathbf{B}^T\mathbf{h}\mathbf{u}^T\mathbf{u} = \|\mathbf{u}\|^2\mathbf{B}^T\mathbf{h} = \mathbf{0}$. From $\mathbf{u} \neq \mathbf{0}$, $\mathbf{0} = \mathbf{h}^T\mathbf{B} = \mathbf{h}^T(\mathbf{I} - \mathbf{J}/m)\mathbf{V}_* = (\mathbf{h} - k\mathbf{1})^T\mathbf{V}_*$ with $k = (\mathbf{h}^T\mathbf{1})/m$. Because $\mathbf{V}_*$ is of full row rank, $\mathbf{h} = k\mathbf{1}$. However, $\mathbf{1}^T\mathbf{h} = \mathbf{1}^T\mathbf{B\Theta u} - c\mathbf{1}^T\mathbf{a}$. Since $\mathbf{1}^T\mathbf{B} = \mathbf{1}^T(\mathbf{I} - \mathbf{J}/m)\mathbf{V} = \mathbf{0}$ and as pointed earlier, $\mathbf{1}^T\mathbf{a} = 0$, then $\mathbf{1}^T\mathbf{h} = 0$. As a result $k = 0$, giving $\mathbf{h} = 0$, i.e. $\mathbf{B\Theta u} = c\mathbf{a}$. Plugging this into (S4) and noting that $\|\mathbf{a}\| = 1$, $T_{SW} = 1$.

On the other hand, by Lemma 2 of (Shapiro & Wilk, 1965), $T_{SW} = 1$ if and only if $\mathbf{B\Theta u} = s\mathbf{a}$ for some scalar $s \neq 0$. Clearly in this case all the coordinates of $\mathbf{B\Theta u}$ are different, so T_{SW} is differentiable at $\mathbf{\Theta}$. Since 1 is the maximum value of T_{SW}, then $\nabla_{\mathbf{\Theta}} T_{SW} = \mathbf{0}$. □

A.5 Proof of Theorem 5.

Theorem 5. *Let $T(\{\mathbf{y}_i\}_{i=1}^m)$, $m \geq 3$ be any affine invariant test statistic that is non-constant and differentiable wherever $\mathbf{y}_i$ are not all equal. Then, for any $\mathbf{b}$, as $(\mathbf{y}_1, \ldots, \mathbf{y}_m) \to (\mathbf{b}, \ldots, \mathbf{b})$, $\limsup \|\nabla T(\{\mathbf{y}_i\}_{i=1}^m)\| \to \infty$, where $\|\cdot\|$ is the Frobenius norm.*

Proof. First, $\forall s > 0$, $T(s\mathbf{y}_1 + \mathbf{b}, \ldots, s\mathbf{y}_m + \mathbf{b}) = T(\mathbf{y}_1, \ldots, \mathbf{y}_m)$ then $\nabla T(s\mathbf{y}_1 + \mathbf{b}, \ldots, s\mathbf{y}_m + \mathbf{b}) = \frac{1}{s}\nabla T(\mathbf{y}_1, \ldots, \mathbf{y}_m)$. By assumption $\nabla T(\mathbf{y}_1, \ldots, \mathbf{y}_m) \neq 0$ for some $(\mathbf{y}, \ldots, \mathbf{y}_m)$. Then let $s \to 0$, $\|\nabla T(s\mathbf{y}_1 + \mathbf{b}, \ldots, s\mathbf{y}_m + \mathbf{b})\| = \frac{1}{s}\|\nabla T(\mathbf{y}_1, \ldots, \mathbf{y}_m)\| \to \infty$. □

A.6 Proof of Theorem 6.

Recall $\mathbf{X} = (X_1, \ldots, X_m)$ and $\mathbf{V} = H_{\mathbf{\Xi}}(\mathbf{X})$, where the X_i's are i.i.d. sample observations. Note that $\mathbf{V}$ is a function of $\mathbf{\Xi}$ and $\mathbf{X}$ but not a function of $\mathbf{\Theta}$.

Theorem 6. Let $T = T_{SW}$ be as in Theorem 4. Denote by $\nabla_{\boldsymbol{\theta}}$, ∇_{Θ}, the Riemannian gradient w.r.t. $\boldsymbol{\theta}$, $\boldsymbol{\Theta}$, etc., respectively, and $\|\cdot\|$ the Frobenius norm. Let $\mathbf{B} = (\mathbf{I} - \mathbf{J}/m)\mathbf{V}$, i.e., $\mathbf{B}$ is obtained by subtracting from each row of $\mathbf{V}$ the mean of the rows. Suppose

(a) $\sup_{\boldsymbol{\Xi}}(\mathbb{E}[\|\mathbf{B}\|^4] + \mathbf{E}[\|\nabla_{\boldsymbol{\Xi}}\mathbf{B}\|^4]) < \infty$,

(b) $\sup_{\|\mathbf{x}\|=1,\boldsymbol{\Xi}} \mathbb{E}[\|\mathbf{B}\mathbf{x}\|^{-4}] < \infty$.

Then, $\sup_{\boldsymbol{\theta}} \mathbb{E}[\|\nabla_{\boldsymbol{\theta}} T\|^2] < \infty$.

Proof. Since all the assumptions and assertion of Theorem 6 are invariant under any permutation of the rows of $\mathbf{V}$, for ease of notation and without loss of generality, suppose the coordinates of $\mathbf{V}\boldsymbol{\Theta}\mathbf{u}$ are in increasing order. By the construction, the coordinates of $\mathbf{B}\boldsymbol{\Theta}\mathbf{u}$ are also increasing and have mean $\mathbf{0}$. Then

$$T = \frac{(\mathbf{a}^T\mathbf{B}\boldsymbol{\Theta}\mathbf{u})^2}{\|\mathbf{B}\boldsymbol{\Theta}\mathbf{u}\|^2}.$$

By $\nabla_{\boldsymbol{\theta}} T = (\nabla_{\boldsymbol{\Theta}} T, \nabla_{\boldsymbol{\Xi}} T)$, it suffices to show $\sup_{\boldsymbol{\theta}} \mathbb{E}(\|\nabla_{\boldsymbol{\Theta}} T\|^2) < \infty$ and $\sup_{\boldsymbol{\theta}} \mathbb{E}(\|\nabla_{\boldsymbol{\Xi}} T\|^2) < \infty$.

Regard the Stiefel manifold $\mathcal{M}$ as a subset of $\mathbb{R}^{k\times d}$ equipped with the inner product $\langle \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2\rangle = \mathrm{tr}(\boldsymbol{\Gamma}_1^T\boldsymbol{\Gamma}_2)$, which is simply the Euclidean inner product of the vectorized $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$. The Riemannian gradient $\nabla_{\boldsymbol{\Theta}}$ is obtained under this inner product and is equal to the orthogonal projection of $\nabla_{\boldsymbol{\Theta}}$ onto $\mathcal{T}_{\boldsymbol{\Theta}}\mathcal{M}$, where $\nabla_{\boldsymbol{\Theta}}$ denotes the Euclidean gradient. Then $\|\nabla_{\boldsymbol{\Theta}} T\| \le \|\nabla_{\boldsymbol{\Theta}} T\|$, so it is enough to show $\sup_{\boldsymbol{\theta}} \mathbb{E}(\|\nabla_{\boldsymbol{\Theta}} T\|^2) < \infty$. From (S4),

$$\nabla_{\boldsymbol{\Theta}} T = \frac{2(\mathbf{a}^T\mathbf{B}\boldsymbol{\Theta}\mathbf{u})}{\|\mathbf{B}\boldsymbol{\Theta}\mathbf{u}\|^2}\mathbf{B}^T\left[\mathbf{a} - \frac{(\mathbf{a}^T\mathbf{B}\boldsymbol{\Theta}\mathbf{u})(\mathbf{B}\boldsymbol{\Theta}\mathbf{u})}{\|\mathbf{B}\boldsymbol{\Theta}\mathbf{u}\|^2}\right]\mathbf{u}^T.$$

Then by $\|\mathbf{u}\| = 1$,

$$\|\nabla_{\boldsymbol{\Theta}} T\| \le \frac{4\|\mathbf{a}\|^2\|\mathbf{B}\|}{\|\mathbf{B}\boldsymbol{\Theta}\mathbf{u}\|},$$

so by Cauchy-Schwarz inequality

$$\begin{aligned}\mathbb{E}(\|\nabla_{\boldsymbol{\Theta}} T\|^2) &\le 16\|\mathbf{a}\|^4\left[\mathbb{E}(\|\mathbf{B}\|^4)\mathbb{E}(\|\mathbf{B}\boldsymbol{\Theta}\mathbf{u}\|^{-4})\right]^{1/2} \\ &\le 16\|\mathbf{a}\|^4\left[\mathbb{E}(\|\mathbf{B}\|^4)\sup_{\|\mathbf{x}\|=1,\boldsymbol{\Xi}}\mathbb{E}(\|\mathbf{B}\mathbf{x}\|^{-4})\right]^{1/2},\end{aligned}$$

where the second inequality is due to the independence of $\mathbf{u}$ and $\mathbf{B}$. Then by assumptions (b) and (c), $\sup_{\boldsymbol{\theta}} \mathbb{E}(\|\nabla_{\boldsymbol{\Theta}} T\|^2) < \infty$.

Next, since $\boldsymbol{\Xi}$ lives in a Euclidean space, the Riemannian gradient of T w.r.t. $\boldsymbol{\Xi}$ consists of the partial derivatives of T w.r.t. of its coordinates. For each coordinate ξ of $\boldsymbol{\Xi}$,

$$\partial_\xi T := \frac{\partial T}{\partial \xi} = 2\left[\frac{\mathbf{a}^T(\partial_\xi\mathbf{B})\boldsymbol{\Theta}\mathbf{u}(\mathbf{a}^T\mathbf{B}\boldsymbol{\Theta}\mathbf{u})}{\|\mathbf{B}\boldsymbol{\Theta}\mathbf{u}\|^2} - \frac{(\mathbf{a}^T\mathbf{B}\boldsymbol{\Theta}\mathbf{u})^2(\mathbf{B}\boldsymbol{\Theta}\mathbf{u})^T(\partial_\xi\mathbf{B})\boldsymbol{\Theta}\mathbf{u}}{\|\mathbf{B}\boldsymbol{\Theta}\mathbf{u}\|^4}\right].$$

Then by $\|\boldsymbol{\Theta}\mathbf{u}\| = 1$,

$$|\partial_\xi T| \le \frac{4\|\mathbf{a}\|^2\|\partial_\xi\mathbf{B}\|}{\|\mathbf{B}\boldsymbol{\Theta}\mathbf{u}\|}.$$

Taking the sum of $(\partial_\xi T)^2$ over all the coordinates of $\boldsymbol{\Xi}$ then yields

$$\|\nabla_{\boldsymbol{\Xi}} T\|^2 \le \frac{16\|\mathbf{a}\|^4\|\nabla_{\boldsymbol{\Xi}}\mathbf{B}\|^2}{\|\mathbf{B}\boldsymbol{\Theta}\mathbf{u}\|^2}.$$

So by Cauchy-Schwarz inequality

$$\begin{aligned}\mathbb{E}(\nabla_{\boldsymbol{\Xi}} T)^2 &\le 16\|\mathbf{a}\|^4\left[\mathbb{E}(\|\nabla_{\boldsymbol{\Xi}}\mathbf{B}\|^4)\mathbb{E}(\|\mathbf{B}\boldsymbol{\Theta}\mathbf{u}\|^{-4})\right]^{1/2} \\ &\le 16\|\mathbf{a}\|^4\left[\mathbb{E}(\|\nabla_{\boldsymbol{\Xi}}\mathbf{B}\|^4)\sup_{\|\mathbf{x}\|=1,\boldsymbol{\Xi}}\mathbb{E}(\|\mathbf{B}\mathbf{x}\|^{-4})\right]^{1/2},\end{aligned}$$

where the second inequality is again due to the independence of $\mathbf{u}$ and $\mathbf{B}$. Then by assumptions (b) and (c), $\sup_{\boldsymbol{\theta}} \mathbb{E}(\|\nabla_{\boldsymbol{\Xi}} T\|^2) < \infty$. □

A.7 PROOF OF THEOREM 7.

The following result is a relaxation of Theorems 1 and 2 in (Bonnabel, 2013). Let $(\gamma_t)_{t\geq 0} = (\gamma_0, \gamma_1, \gamma_2, \ldots)$ be a sequence of step sizes. Suppose $C(\cdot)$ is a three times continuously differentiable cost function on a smooth connected Riemannian manifold $\mathcal{M}$, $(z_t)_{t\geq 0}$ is a sequence of i.i.d. random variables taking values in a measurable space $\mathcal{Z}$, and $H(\cdot,\cdot)$ is a measurable function on $\mathcal{Z} \times \mathcal{TM}$ such that $\mathbb{E}_z H(z,w) = \nabla C(w)$, where $\mathcal{TM}$ is the tangent bundle of $\mathcal{M}$ and $z \sim z_t$. Also suppose $w_0 \in \mathcal{M}$ is independent of $(z_t)_{t\geq 0}$.

Theorem 7. *Let $\mathcal{M}$ be a connected Riemannian manifold with injectivity radius uniformly bounded from below by $I > 0$. Let R_w be a twice continuously differentiable retraction, and consider the update $w_{t+1} = R_{w_t}(-\gamma_t H(z_t, w_t))$ with step size γ_t. Assume $(\gamma_t)_{t\geq 0}$ satisfy $\sum \gamma_t^2 < +\infty$, and $\sum \gamma_t = +\infty$. Suppose there exists a compact set K such that $w_t \in K\ \forall t \geq 0$. We also suppose that the gradient is bounded in K, i.e. for all $w \in K$ we have $\mathbb{E}_z[\|H(z,w)\|^2] \leq A^2$ for some $A > 0$. Then, $C(w_t)$ converges almost surely and $\nabla C(w_t) \to 0$ almost surely.*

As in (Bonnabel, 2013), the proof of Theorem 7 starts with the the following result which is of interest in its own right.

Proposition A.1. *Let $(\gamma_t)_{t\geq 0}$ and $\mathcal{M}$ be as in Theorem 7. Consider the update $w_{t+1} = \exp_{w_t}(-\gamma_t H(z_t, w_t))$, where $\exp_w$ is the exponential map at w. Suppose there exists a compact set K such that $w_t \in K$ for all $t \geq 0$. We also suppose for some $A > 0$, $\mathbb{E}_z(\|H(z,w)\|^2) \leq A^2$ for all $w \in K$. Then, $C(w_t)$ converges a.s. and $\nabla C(w_t) \to 0$ a.s.*

Proof of Proposition A.1. The proof builds upon the one for Theorem 1 (Bonnabel, 2013). Let $\mathcal{F}_t = \sigma(w_0, z_0, \ldots, z_{t-1})$. Then for $t \geq 1$, w_t is $\mathcal{F}_t$ measurable. If $\gamma_t \|H(z_t, w_t\| < I$, then $\exp_{w_t}\{sH(z_t, w_t)\}_{0\leq s\leq \gamma_t}$ is the geodesic linking w_{t+1} and w_t, so as in the proof of Theorem 1 in (Bonnabel, 2013), the Taylor formula yields

$$C(w_{t+1}) - C(w_t) \leq -\gamma_t \langle H(z_t, w_t), \nabla C(w_t)\rangle + \gamma_t^2 \|H(z_t, w_t)\|^2 k_2,$$

where $\langle\, ,\, \rangle$ is the Riemannian inner product at $\mathcal{T}_{w_t}\mathcal{M}$. The inequality has the same form as equation (5) in (Bonnabel, 2013) but with $k_2 = \sup_{w\in K_0} \|\nabla^2 C\|$, i.e., $\|(\nabla^2 C(w))v\| \leq k_2 \|v\|$ for all $w \in K_0$ and $v \in \mathcal{T}_w\mathcal{M}$, where K_0 is the compact set of all points with distance at most I from K. Define events

$$E_{-1} = \emptyset, \quad E_t = \{\|H(z_t, w_t)\| < I/\gamma_t\}, \quad t \geq 0. \tag{S5}$$

Denote by $\mathbf{1}_E$ the indicator function of an event E. Then by $C(w) \geq 0$,

$$C(w_{t+1})\mathbf{1}_{E_t} \leq C(w_t) - \gamma_t \langle H(z_t, w_t), \nabla C(w_t)\rangle \mathbf{1}_{E_t} + \gamma_t^2 \|H(z_t, w_t)\|^2 k_2. \tag{S6}$$

Taking expectations conditional on $\mathcal{F}_t$ on both sides of the equality,

$$\begin{aligned}\mathbb{E}[C(w_{t+1})\mathbf{1}_{E_t} \,|\, \mathcal{F}_t] &\leq C(w_t) - \gamma_t \mathbb{E}(\langle H(z_t, w_t), \nabla C(w_t)\rangle \,|\, \mathcal{F}_t) \\ &\quad + \gamma_t \mathbb{E}(\langle H(z_t, w_t), \nabla C(w_t)\rangle \mathbf{1}_{E_t^c} \,|\, \mathcal{F}_t) + \gamma_t^2 \mathbb{E}(\|H(z_t, w_t)\|^2 \,|\, \mathcal{F}_t) k_2.\end{aligned} \tag{S7}$$

Since z_t is independent from $\mathcal{F}_t$ while w_t is $\mathcal{F}_t$-measurable,

$$\begin{aligned}\mathbb{E}(\langle H(z_t, w_t), \nabla C(w_t)\rangle \,|\, \mathcal{F}_t) &= \mathbb{E}_z(\langle H(z, w_t), \nabla C(w_t)\rangle) = \|\nabla C(w_t)\|^2, \\ \mathbb{E}(\|H(z_t, w_t)\|^2 \,|\, \mathcal{F}_t) &= \mathbb{E}_z(\|H(z, w_t)\|^2) \leq A^2,\end{aligned} \tag{S8}$$

where in $\mathbb{E}_z(\cdot)$, w_t is treated as a fixed value. On the other hand,

$$\begin{aligned}\mathbb{E}(\langle H(z_t, w_t), \nabla C(w_t)\rangle \mathbf{1}_{E_t^c} \,|\, \mathcal{F}_t) &= \langle \mathbb{E}[H(z_t, w_t)\mathbf{1}_{E_t^c} \,|\, \mathcal{F}_t], \nabla C(w_t)\rangle \\ &\leq \mathbb{E}[\|H(z_t, w_t)\| \mathbf{1}_{E_t^c} \,|\, \mathcal{F}_t] k_1,\end{aligned}$$

where $k_1 = \sup_K \|\nabla C\|$. Since $E_t^c = \{\|H(z_t, w_t)\| \geq I/\gamma_t\}$, by Markov inequality,

$$\mathbb{E}[\|H(z_t, w_t)\| \mathbf{1}_{E_t^c} \,|\, \mathcal{F}_t] \leq \mathbb{E}[\|H(z_t, w_t)\|^2/(I/\gamma_t) \,|\, \mathcal{F}_t] \leq \gamma_t A^2/I. \tag{S9}$$

Then from (S7),

$$\mathbb{E}[C(w_{t+1})\mathbf{1}_{E_t} \,|\, \mathcal{F}_t] \leq C(w_t) + \gamma_t^2 A^2 k - \gamma_t \|\nabla C(w_t)\|^2 \tag{S10}$$

with $k = k_2 + k_1/I$. Let

$$N_t = C(w_t)\mathbf{1}_{E_{t-1}} + A^2 k \sum_{s \geq t} \gamma_s^2 - \sum_{s<t} C(w_s)\mathbf{1}_{E_{s-1}^c}.$$

Since $\sum_{s<t+1} C(w_s)\mathbf{1}_{E_{s-1}^c}$ is $\mathcal{F}_t$-measurable, then from (S10),

$$\mathbb{E}(N_{t+1} \,|\, \mathcal{F}_t) \leq N_t - \gamma_t \|\nabla C(w_t)\|^2. \tag{S11}$$

Therefore, N_t is a supermartingale. Let

$$\xi = \sum_{s \geq 0} C(w_s)\mathbf{1}_{E_{s-1}^c}.$$

Let $k_0 = \sup_{w \in K} C(w)$. Then $|N_t| \leq k_0 + A^2 k \sum_{s \geq 0} \gamma_s^2 + \xi$. On the other hand, by Fubini's theorem followed by Markov inequality,

$$\mathbb{E}\xi \leq k_0 \sum_{s \geq 0} \mathbb{P}\{\|H(z_s, w_s)\| \geq I/\gamma_t\} \leq k_0 \sum_{s \geq 0} \frac{\gamma_s^2}{I^2} \sup_{w \in K} \mathbb{E}_z \|H(z, w)\|^2 < \infty.$$

Then N_t is uniformly integrable, i.e., $\sup_t \mathbb{E}(|N_t|\mathbf{1}_{|N_t \geq c|}) \to 0$ as $c \to \infty$. Then by martingale convergence theorem, N_t converges a.s. Moreover, from the display and Borel-Cantelli lemma, all but a finite number of $\mathbf{1}_{E_t^c}$ are 0, a.s. As a result, it then follows that $C(w_t)$ converges a.s.

To show that $\nabla C(w_t) \to 0$ a.s., by Doob's decomposition, $N_t = M_t - Z_t$, where M_t is a martingale and Z_t is increasing and $\mathcal{F}_{t-1}$-measurable with $Z_0 = 0$, and both M_t and Z_t are uniformly integrable, giving $Z_t \uparrow Z \geq 0$ with $\mathbb{E}Z < \infty$. Put $p = \|\nabla C\|^2$. From (S11),

$$\sum_t \gamma_t p(w_t) \leq \sum_t (Z_{t+1} - Z_t) = Z < \infty \quad \text{a.s.}$$

Then, by $\sum \gamma_t = \infty$, to show $\nabla C(w_t) \to 0$, it suffices to show that $p(w_t)$ converges a.s.

Similar to (S6),

$$p(w_{t+1})\mathbf{1}_{E_t} - p(w_t) \leq -2\gamma_t \langle \nabla C(w_t), (\nabla^2 C(w_t))H(z_t, w_t)\rangle \mathbf{1}_{E_t} + \gamma_t^2 \|H(z_t, w_t)\|^2 k_4,$$

where k_4 is an upper bound on $\|\nabla^2 p\|$ on K_0. For ease of notation, write $\langle \nabla C, (\nabla^2 C)H\rangle_t$ for $\langle \nabla C(w_t), (\nabla^2 C(w_t))H(z_t, w_t)\rangle$, $\|H\|_t$ for $\|H(z_t, w_t)\|$, and so on. Then

$$\begin{aligned}
\mathbb{E}(p(w_{t+1})\mathbf{1}_{E_t} - p(w_t) \,|\, \mathcal{F}_t) &\leq -2\gamma_t \mathbb{E}(\langle \nabla C, (\nabla^2 C)H\rangle_t \mathbf{1}_{E_t} \,|\, \mathcal{F}_t) + \gamma_t^2 k_4 \mathbb{E}(\|H\|_t^2 \,|\, \mathcal{F}_t) \\
&\leq -2\gamma_t \mathbb{E}(\langle \nabla C, (\nabla^2 C)H\rangle_t \mathbf{1}_{E_t} \,|\, \mathcal{F}_t) + \gamma_t^2 k_4 A^2.
\end{aligned}$$

On the other hand, recalling that $k_1 = \sup_K \|\nabla C\|$ and $k_2 = \sup_{K_0} \|\nabla^2 C\|$,

$$\begin{aligned}
|\mathbb{E}(\langle \nabla C, (\nabla^2 C)H\rangle_t \mathbf{1}_{E_t} \,|\, \mathcal{F}_t)| &\leq |\mathbb{E}(\langle \nabla C, (\nabla^2 C)H\rangle_t \,|\, \mathcal{F}_t)| + |\mathbb{E}(\langle \nabla C, (\nabla^2 C)H\rangle_t \mathbf{1}_{E_t^c} \,|\, \mathcal{F}_t)| \\
&\leq |\langle \nabla C, (\nabla^2 C)\nabla C\rangle_t| + \|\nabla C\|_t \cdot \mathbb{E}(\|(\nabla^2 C)H\|_t \mathbf{1}_{E_t^c} \,|\, \mathcal{F}_t) \\
&\leq k_2 \|\nabla C\|_t^2 + k_2 \|\nabla C\|_t \cdot \mathbb{E}(\|H\|_t \mathbf{1}_{E_t^c} \,|\, \mathcal{F}) \\
&\leq k_2 \|\nabla C\|_t^2 + k_3 \|\nabla C\|_t \cdot \mathbb{E}(\|H\|_t \mathbf{1}_{E_t^c} \,|\, \mathcal{F}) \\
&\leq k_2 p(w_t) + (k_1 k_2/I) A^2 \gamma_t.
\end{aligned}$$

where the last line follows from (S9). Combining the above two displays,

$$\mathbb{E}(p(w_{t+1})\mathbf{1}_{E_t} - p(w_t)\mathbf{1}_{E_{t-1}} \,|\, \mathcal{F}_t) \leq q(w_t) := p(w_t)\mathbf{1}_{E_{t-1}^c} + 2k_2 \gamma_t p(w_t) + k_* \gamma_t^2 A^2, \tag{S12}$$

where $k_* = k_4 + 2k_1 k_2/I$. From the above proof, $\sum \mathbb{E}q(w_t) < \infty$. Then by a similar argument for the convergence of $C(w_t)$ based on submartingale convergence, $p(w_t)$ converges a.s. □

Proof of Theorem 7. The proof builds upon the one for Theorem 2 (Bonnabel, 2013). There are constants $\varrho > 0$ and $0 < I_0 \leq I$, such that $d(R_w(v), \exp_w(v)) \leq \varrho \|v\|^2$ for all $w \in K$ and $v \in \mathcal{T}_w \mathcal{M}$ with $\|v\| \leq I_0$. Without loss of generality, suppose $\varrho > 1$ and $I_0 = I < 1/\varrho$, for otherwise we can decrease I.

Define the same events E_t as in (S5) and let constants $k_0, \ldots, k_4$ be defined as in the proof of Proposition A.1. Let $w_t^* = \exp_{w_t}(-\gamma_t H(z_t, w_t))$. Then

$$d(w_{t+1}^*, w_{t+1})\mathbf{1}_{E_t} \leq \gamma_t^2 \varrho \|H(z_t, w_t)\|^2.$$

By assumption, $w_{t+1} \in K$. Then on the event E_t, $d(w_{t+1}^*, w_{t+1}) \leq \varrho I^2 < I$, so $w_{t+1} \in K_0$, where K_0 is defined in the proof of Proposition A.1. Then

$$\begin{aligned} C(w_{t+1})\mathbf{1}_{E_t} - C(w_t) &\leq C(w_{t+1}^*)\mathbf{1}_{E_t} - C(w_t) + |C(w_{t+1}) - C(w_{t+1}^*)|\mathbf{1}_{E_t} \\ &\leq C(w_{t+1}^*)\mathbf{1}_{E_t} - C(w_t) + d(w_{t+1}, w_{t+1}^*)k_1\mathbf{1}_{E_1} \\ &\leq C(w_{t+1}^*)\mathbf{1}_{E_t} - C(w_t) + \gamma_t^2\|H(z_t, w_t)\|^2\varrho k_1. \end{aligned}$$

Then by (S8) and (S10),

$$\mathbb{E}(C(w_{t+1})\mathbf{1}_{E_t} - C(w_t) \,|\, \mathcal{F}_t) \leq \gamma_t^2 A^2 k - \gamma_t \|\nabla C(w_t)\|^2,$$

where $k = k_2 + k_1/I + \varrho k_1$. Then, following the same argument as in the proof of Proposition A.1, $C(w_t)$ converges a.s. and $\sum_{t\geq 0} \gamma_t \mathbb{E}(\|\nabla C(w_t)\|^2) < \infty$, and to show $p(w_t) := \|\nabla C(w_t)\|^2 \to 0$, it suffices to show $p(w_t)$ converges a.s. We have

$$\begin{aligned} p(w_{t+1})\mathbf{1}_{E_t} - p(w_t) &\leq |p(w_{t+1}) - p(w_{t+1}^*)|\mathbf{1}_{E_t} + p(w_{t+1}^*)\mathbf{1}_{E_t} - p(w_t) \\ &\leq k_4\gamma_t^2\varrho\|H(z_t, w_t)\|^2 + p(w_{t+1}^*)\mathbf{1}_{E_t} - p(w_t). \end{aligned}$$

Then by (S8) and (S12),

$$\mathbb{E}(p(w_{t+1})\mathbf{1}_{E_t} - p(w_t)\mathbf{1}_{E_{t-1}} \,|\, \mathcal{F}_t) \leq p(w_t)\mathbf{1}_{E_{t-1}^c} + 2k_2\gamma_t p(w_t) + (2k_4 + 2k_1k_2/I)\gamma_t^2 A^2.$$

Then, with the same argument following (S12), $p(w_t)$ converges a.s. □

B GOODNESS-OF-FIT TESTS

In this section, we present several commonly used GoF tests. They are grouped into three classes: tests based on correlation (CB), tests based on empirical distribution function (EDF), and tests based on empirical characteristic function (ECF). The CB GoF tests were first covered in Section 3. The latter two are covered here.

Empirical Distribution Function (EDF) Tests: UVN EDF tests are based on the discrepancy between the empirical and hypothesized distribution functions (D'Agostino, 2017), encompassing two broad classes: supremum tests, and quadratic tests. Kolmogorov–Smirnov (KS) is a supremum test, measuring the largest absolute distance. Two popular quadratic tests are Cramér–von Mises (CVM), which measures the integrated quadratic deviation weighted by a function Ψ, and Anderson–Darling (AD) (Anderson et al., 1952), which gives higher weight to distribution tails.

Empirical Characteristic Function (ECF) Tests: ECF tests are based on the weighted integral of the difference between the ECF and its pointwise limit, including Epps–Pulley (EP) for UVN (Epps & Pulley, 1983) and MVN (Baringhaus & Henze, 1988) and the Henze–Zirkler (HZ) test, a generalization of EP (Henze & Zirkler, 1990).

For consistency, T_* and d_* represent respectively the test statistic and corresponding statistical distance for each test. For a random sample $X_1, X_2, \ldots X_m$ i.i.d. $\sim F_X$, denote its sample mean, sample variance and empirical distribution function by $\bar{X}$, S_m, and $\hat{F}_{X,m}$, respectively. If the X_i's are univariate, we further arrange them in order, to get the order statistics $X_{(1)} \leq X_{(2)} \leq \cdots X_{(m)}$.

B.1 CB CLASS: SHAPIRO–WILK, SHAPIRO–FRANCIA

Let $X_1, \ldots, X_m$ be an i.i.d. univariate sample from $\sim F_X$. To test the null hypothesis that F_X is a normal distribution, the Shapiro–Wilk test statistic is defined as

$$T_{SW} = \frac{\left(\sum_{i=1}^m a_i X_{(i)}\right)^2}{\sum_{i=1}^m \left(X_i - \bar{X}\right)^2},$$

where $\mathbf{a} = (a_1, a_2, \ldots, a_m)^T$ is obtained via

$$\mathbf{a} = \frac{\mathbf{M}^{-1}\mathbf{c}}{\|\mathbf{M}^{-1}\mathbf{c}\|}$$

with $\mathbf{c} = (c_1, \ldots, c_m)^T$ and $\mathbf{M}$ being the mean vector and covariance matrix, respectively, of the order statistics of m independent $N(0,1)$ random variables. A sample version of the corresponding L_2-Wasserstein distance can be formulated as

$$d_{W_2}^2 = S_m^2 - \left(\int_0^1 \hat{F}_{X,m}^{-1}(t)\Phi^{-1}(t)\,\mathrm{d}t\right)^2$$

with $\Phi(\cdot)$ being the distribution function of standard normal (del Barrio et al., 1999). Following the same notations for SW test, the Shapiro–Francis test on normality (Shapiro & Francia, 1972) is defined as

$$T_{SF} = \frac{\left(\sum_{i=1}^m b_i X_{(i)}\right)^2}{\sum_{i=1}^m \left(X_i - \bar{X}\right)^2},$$

where $\mathbf{b} = (b_1, b_2, \ldots, b_m)^T$ is obtained via

$$\mathbf{b} = \frac{\mathbf{c}}{\|\mathbf{c}\|}.$$

B.2 EDF CLASS: CRAMER–VON MISES, KOLMOGOROV–SMIRNOV

Let $X_1, \ldots, X_m$ be an i.i.d. univariate sample $\sim F_X$. Let F be a specified univariate distribution. To test $\mathcal{H}_0 : F_X = F$ vs $\mathcal{H}_1 : F_X \neq F$, the Cramer–Von Mises test statistic is defined as

$$T_{CVM} = \frac{1}{12m} + \sum_{i=1}^m \left(F(X_{(i)}) - \frac{2i-1}{2m}\right)^2$$

and with a weight function $\Psi(\cdot)$, a sample version of the corresponding statistical distance is

$$d_{CVM} = m\int \left(\hat{F}_{X,m} - F\right)^2 \mathbf{\Psi}(F)\,\mathrm{d}F.$$

The Kolmogorov–Smirnov test statistic and related statistical distance are defined respectively as

$$T_{KS} = \max_{1\le i\le m}\left[F(X_{(i)}) - \frac{i-1}{m}, \frac{i}{m} - F(X_{(i)})\right],$$
$$d_{KS} = \sup|\hat{F}_{X,m} - F|.$$

B.3 ECF CLASS: HENZE–ZIRKLER

Suppose X_i are i.i.d. k-dimensional vectors $\sim \mathbb{P}_X$. Following the similar notation in (Henze & Zirkler, 1990), define the scaled residuals $Y_j = S_m^{-1/2}(X_j - \bar{X}), j = 1, \ldots, m$. Let

$$\phi_m(\mathbf{t}) = \frac{1}{m}\sum_{j=1}^m \exp(i\mathbf{t}'Y_j)$$

denote the empirical characteristic function of Y_j. Let $\mathbb{P}_0$ be a specified distribution. To test $\mathcal{H}_0 : \mathbb{P}_X = \mathbb{P}_0$ vs $\mathcal{H}_1 : \mathbb{P}_X \neq \mathbb{P}_0$, the Henze–Zirkler test statistic is defined as

$$T_{HZ} = m\left(4I\{S_m \text{ is singular}\} + D_{m,\beta}I\{S_m \text{ is nonsingular}\}\right),$$

where $\beta > 0$ is a parameter,

$$D_{m,\beta} = \int_{\mathbb{R}^k}\left|\phi_m(\mathbf{t}) - \exp\left(-\frac{1}{2}\|\mathbf{t}\|^2\right)\right|^2 \psi_\beta(t)\,\mathrm{d}\mathbf{t}, \quad \psi_\beta(\mathbf{t}) = (2\pi\beta^2)^{-k/2}\exp\left(-\frac{\|\mathbf{t}\|^2}{2\beta^2}\right).$$

After some simplification Korkmaz et al. (2014), we have

$$D_{m,\beta} = \frac{1}{m^2}\sum_{i,j=1}^{m} e^{-\beta^2\|Y_i-Y_j\|^2/2} - \frac{2}{(1+\beta)^{k/2}m}\sum_{i=1}^{m} e^{-\beta^2\|Y_i\|^2/[2(1+\beta^2)]} + \frac{1}{(1+2\beta^2)^{k/2}}.$$

The value

$$\frac{1}{\sqrt{2}}\left[\frac{m(2k+1)}{4}\right]^{1/(k+4)}$$

is proposed as an optimal choice for β. The corresponding statistical distance is

$$d_{HZ} = \int_{\mathbb{R}^k} |\tilde{\mathbb{P}}_{X,m}(\mathbf{t}) - \tilde{\mathbb{P}}_X(\mathbf{t})|^2 \psi_\beta(\mathbf{t})\,\mathrm{d}\mathbf{t},$$

where

$$\tilde{\mathbb{P}}_{X,m}(\mathbf{t}) = \int_{\mathbb{R}^k} e^{i\mathbf{t}'\mathbf{x}}\mathbb{P}_{X,m}(\mathrm{d}\mathbf{x}), \quad \tilde{\mathbb{P}}_X(\mathbf{t}) = \int_{\mathbb{R}^k} e^{i\mathbf{t}'\mathbf{x}}\mathbb{P}_X(\mathrm{d}\mathbf{x})$$

are the Fourier transforms of $\mathbb{P}_{X,m}$ and $\mathbb{P}_X$, respectively.

B.4 A COMPARISON BETWEEN THE ASSOCIATED STATISTICAL DISTANCES

Proposition S1. *Let X, Y denote two k-variate random variables. Then, (a) $d_{HZ}(\mathbb{P}_X, \mathbb{P}_Y) \le C_1 d_{W_2}(\mathbb{P}_X, \mathbb{P}_Y)$, (b) $d_{KS}(\mathbb{P}_X, \mathbb{P}_Y) \le C_2\sqrt{d_{W_1}(\mathbb{P}_X, \mathbb{P}_Y)}$, (c) if $k = 1$ and the weight function $\Psi(\cdot)$ in the CVM test is bounded, then $d_{CVM}(\mathbb{P}_X, \mathbb{P}_Y) \le C_3 d_{KS}(\mathbb{P}_X, \mathbb{P}_Y)^2$. In these inequalities the constants C_1, C_2 may depend on k, while C_3 may depend on $\Psi(\cdot)$.*

Proof. (a) The Henze–Zirkler test is on the metric

$$D(\mu,\nu) = \left[\int_{\mathbb{R}^d} |\tilde{\mu}(\mathbf{t}) - \tilde{\nu}(\mathbf{t})|^2 \phi(\mathbf{t})\mathrm{d}\mathbf{t}\right]^{1/2},$$

where $\phi = \psi_\beta$. For any $\omega \in \Pi(\mu,\nu)$, by Cauchy–Schwarz inequality and Fubini's theorem

$$\begin{aligned}
D(\mu,\nu)^2 &= \int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d} e^{i\mathbf{t}'\mathbf{x}}\mu(\mathrm{d}\mathbf{x}) - \int_{\mathbb{R}^d} e^{i\mathbf{t}'\mathbf{y}}\nu(\mathrm{d}\mathbf{y})\right|^2 \phi(\mathbf{t})\,\mathrm{d}\mathbf{t} \\
&= \int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d\times\mathbb{R}^d} (e^{i\mathbf{t}'\mathbf{x}} - e^{i\mathbf{t}'\mathbf{y}})\omega(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{y})\right|^2 \phi(\mathbf{t})\,\mathrm{d}\mathbf{t} \\
&\le \int_{\mathbb{R}^d}\left\{\int_{\mathbb{R}^d\times\mathbb{R}^d} |e^{i\mathbf{t}'\mathbf{x}} - e^{i\mathbf{t}'\mathbf{y}}|^2\omega(d\mathbf{x},d\mathbf{y})\right\}\phi(\mathbf{t})\,\mathrm{d}\mathbf{t} \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d}\left\{\int_{\mathbb{R}^d} |e^{i\mathbf{t}'\mathbf{x}-i\mathbf{t}'\mathbf{y}}|^2\phi(\mathbf{t})\,\mathrm{d}\mathbf{t}\right\}\omega(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{y}).
\end{aligned}$$

Using $|e^{ia} - e^{ib}|^2 \le (a-b)^2$ for $a, b \in \mathbb{R}$,

$$\begin{aligned}
D(\mu,\nu)^2 &\le \int_{\mathbb{R}^d\times\mathbb{R}^d}\left\{\int_{\mathbb{R}^d} |\mathbf{t}'\mathbf{x} - \mathbf{t}'\mathbf{y}|^2\phi(\mathbf{t})\,\mathrm{d}\mathbf{t}\right\}\omega(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{y}) \\
&\le \int_{\mathbb{R}^d\times\mathbb{R}^d}\left\{\int_{\mathbb{R}^d} \|\mathbf{t}\|^2\|\mathbf{x}-\mathbf{y}\|^2\phi(\mathbf{t})\,\mathrm{d}\mathbf{t}\right\}\omega(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{y}) \\
&= \int_{\mathbb{R}^d} \|\mathbf{t}\|^2\phi(\mathbf{t})\,\mathrm{d}\mathbf{t}\cdot\int_{\mathbb{R}^d\times\mathbb{R}^d}\|\mathbf{x}-\mathbf{y}\|^2\omega(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{y}) \\
&= C\int_{\mathbb{R}^d\times\mathbb{R}^d}\|\mathbf{x}-\mathbf{y}\|^2\omega(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{y})
\end{aligned}$$

with $C = \int_{\mathbb{R}^d}\|\mathbf{t}\|^2\phi(\mathbf{t})\,\mathrm{d}\mathbf{t} < \infty$. Since the above inequality holds for all $\omega \in \Pi(\mu,\nu)$, then $D(\mu,\nu) \le \sqrt{C}d_{W_2}(\mu,\nu)$. Let $C_1 = \sqrt{C}$, $\mu = \mathbb{P}_X$, and $\nu = \mathbb{P}_Y$. Then (a) follows.

(b) The connection between Kolmogorov–Smirnov distance and L_1-Wasserstain distance under both of the univariate and multivariate cases are well studied, detailed proof can be found in Corollary 3.1 (Koike, 2019), Proposition 1.2 (Ross, 2011) etc.

(c) Letting $C_3 = \sup \Psi$, this inequality is trivial:

$$\int (F_X - F_Y)^2 \Psi(F_Y) \, \mathrm{d}F_B \leq C_3 \int |F_A - F_B|^2 \, \mathrm{d}F_B = C_3 d_{KS}(\mathbb{P}_, \mathbb{P}_Y)^2. \qquad \square$$

C Model Architecture

The architecture for the encoder and decoder of CelebA and MNIST is based in the WAE (Tolstikhin et al., 2017). The discriminator for the WAE-GAN followed Tolstikhin et al. (2017). The architecture for CIFAR10 is based on Lippe (2022). When modeling a specific dataset, the specified encoder and decoder components are the same for all models.

Dataset	Optimizer		Architecture
CelebA	Adam	Input	64x64x3
		Encoder	Conv 128x4x4 (stride 2), 256x4x4 (stride 2), 512x4x4 (stride 2), 1024x4x4 (stride 2), FC 16384, 256. ReLU activation
	RSGD	Input	256
		Stiefel	FC 64 Orthogonal initialization
	Adam	Input	64
		Decoder	FC 65536, ConvT 512x4x4 (stride 2), ConvT 256x4x4 (stride 2), ConvT 128x4x4 (stride 2), ConvT 3x1x1 (stride 1), Sigmoid activation.
MNIST	Adam	Input	28x28x1
		Encoder	Conv 128x4x4 (stride 2), 256x4x4 (stride 2), 512x4x4 (stride 2), 1024x4x4 (stride 2), FC 16384, 256. ReLU activation
	RSGD	Input	256
		Stiefel	FC 10 Orthogonal initialization
	Adam	Input	10
		Decoder	FC 65536, ConvT 512x4x4 (stride 1), ConvT 256x4x4 (stride 1), ConvT 128x4x4 (stride 2), Sigmoid activation.
CIFAR10	Adam	Input	32x32x3
		Encoder	Conv 64x3x3 (stride 2, pad 1), Conv 64x3x3 (stride 1, pad 1) Conv 128x3x3 (stride 2, pad 1), Conv 128x3x3 (stride 1, pad 1) Conv 128x3x3 (stride 2, pad 1), FC 2048, 256, ReLU activation
	RSGD	Input	256
		Stiefel	FC 64 Orthogonal initialization
	Adam	Input	64
		Decoder	FC 2048, ConvT 128x3x3 (stride 2, pad 1, out pad 1) Conv 128x3x3 (stride 1, pad1), ConvT 64x3x3(stride 2, pad 1, out pad 1) Conv 64x3x3 (stride 1, pad 1), ConvT 3x3x3 (stride 2, pad 1, out pad 1), Sigmoid activation

In the above table, Conv $128 \times 4 \times 4$ represents a convolution with 128 filters which are 4×4, FC stands for fully connected layer, ReLU for the rectified linear units, and ConvT for convolution transpose.

D Training Details

D.1 Test Statistic Projections

Once a batch has been encoded, it is projected from 64D to 1D in order to be tested. Instead of randomly sampling directions from the unit sphere to implement the projections, we sample an orthonormal basis and project down each direction, calculating the test statistic along each. This is used in two ways: 1) selecting the most pessimistic direction to optimize, or 2) computing the average direction to optimize.

For example, the SW test statistic value should be large to fail to reject normality. Selecting the direction associated with the smallest statistic value can be used as a new statistic to optimize. Similarly, for tests that fail to reject for small values, selecting the direction associated with the

largest value can be used. We refer to both scenarios as selecting the most pessimistic direction. Alternatively, the mean of *all* the statistics could be used as a new statistic.

D.2 Empirical Distributions

The distribution of a test statistic must be known analytically or estimable in order to compute p-values. There are a few options for calculating the empirical distribution of a test statistic when using GoF tests with projections. The first possibility is to randomly sample from the unit sphere, project the multivariate data, and then use the corresponding distribution to determine the p-value. However, a single projection may not be particularly informative, and early testing indicated this method lead to slower convergence. Another option is to project along multiple directions and calculate the statistics of these directions. It is then possible to create a *new* statistic from these, for example the average, minimum, or maximum. Unfortunately, this method precludes using the original test statistic distribution or calculating p-values. To remedy this, we create an empirical distribution of this new statistic by repeatedly sampling; an empirical p-value are produced from the test statistic samples.

D.3 Mutual Information

The mutual information is estimated after the model has been trained by sampling from $\mathcal{N}_{64}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ where $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ are estimated either during or after training. GoFAE tracks these statistics during training. These samples are decoded, and then encoded. Assuming the encoded data is Gaussian, mutual information may be calculated as $I(\mathbf{Z}; \tilde{\mathbf{Z}}) = \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_Z| \times |\boldsymbol{\Sigma}_{\tilde{Z}}|}{|\boldsymbol{\Sigma}_{Z,\tilde{Z}}|}$.

E Additional Experiments

E.1 Details on Experiments

We trained GoFAE and competing methods on MNIST LeCun et al. (1998) and CelebA Liu et al. (2015). For fair comparisons, we adopted a common architecture for all methods for each dataset as described in Section C.

E.1.1 MNIST

The MNIST dataset was re-scaled to the unit interval. Training proceeded for 50 epochs using mini-batches of size 128. We specified a 10-dimensional code layer. Competitor models include β-VAE Higgins et al. (2017), VAE Kingma & Welling (2013), WAE-GAN Tolstikhin et al. (2017), VAE with learned γ Dai & Wipf (2019), 2-Stage VAE Dai & Wipf (2019), and AE Bengio et al. (2013). Both VAE and β-VAE had an initial learning rate of $1e-4$, while VAE with learnable γ used $1e-3$.

For our models, Adam optimizer Kingma & Ba (2014) was used for the encoder (learning rate $3e-3$, $\beta = (0.5, 0.999)$) and decoder (learning rate $3e-3$, $\beta = (0.5, 0.999)$). Riemannian SGD with learning rate $5e-3$ was used to constrain $\boldsymbol{\Theta}$ to the Stiefel manifold using a one cycle learning rate scheduler with max learning rate as $1e-3$. Singular value decomposition was used for retracting $\boldsymbol{\Theta}$ back to $\mathcal{M}$ after a parameter update.

E.1.2 CelebA

The CelebA dataset was pre-processed following the same procedure in Tolstikhin et al. (2017). First, a 140x140 center crop is taken and the image is resized to 64×64 resolution. For the VAE, the Adam optimizer was used with a learning rate of $1e-4$, and $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The β-VAE used the Adam optimizer with a learning rate of $1e-4$, $\beta_1 = 0.5$, and $\beta_2 = 0.999$. The VAE with learnable γ used the Adam optimizer with a learning rate of $1e-3$, $\beta_1 = 0.5$, and $\beta_2 = 0.999$. The WAE-GAN also used the Adam optimizer where the encoder and decoder parameters had an initial learning rate of $3e-4$ and the discriminator was set to $1e-3$ with $\beta_1 = 0.5$, $\beta_2 = 0.999$. For the two-stage VAE, we used the trained VAE with learned gamma as stage one. A second VAE was trained using the μ, σ from the first stage as inputs, also with a trainable gamma. The architecture

was a three layer dense net with ReLU activations, and the decoder was the inverse, with the last layer being linear.

For our method, which had two components for the encoder, as visualized in Figure (3a) in the main paper. The first component of the encoder used the Adam optimizer with learning rate set to $3e-3$ and $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The decoder also used Adam, with the same learning rates and hyperparameters. The dense layer of the encoding processes is parameterized by $\mathbf{\Theta}$, an element of the Steifel manifold $\mathcal{M}$, and is trained with Riemannian SGD with a learning rate of $5e-3$ while making use of the 1-cycle learning rate scheduler with a maximum learning rate of $1e-3$.

As our methods require projection, a 64×64 dimensional orthonormal basis is sampled and used to project the encoded data.

E.1.3 CIFAR-10

CIFAR10 is the final dataset, and consists of 60 thousand small images. The training set contains 50 thousand with the remaining going to the test set. After setting a seed, the training set was split into a smaller training set, of 45 thousand, and a validation set with the rest. The architecture used follows from the Model Architecture section. A latent dimension size of 64 is used. Models are evaluated on the average reconstruction quality. Here, the mean-squared error of randomly selected batches is computed over over a single pass through the test set, and then averaged. This is repeated 30 times, with the mean and standard deviation reported. FID scores for reconstructed and generated data are also computed.

The Adam optimizer was used for the encoder and decoder of all methods. The learning rate was set at $3e-4$, with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The discriminator for WAE-GAN is the same from the CelebA setup. All models are trained for 200 epochs, and reduce the learning rate when hitting a plateau (***ReduceLROnPlateau*** in PyTorch). For this a minimum learning rate was set as $5e-5$, patience of 10, and a factor of 0.2. For the two-stage VAE, the encoder consisted of 2 dense layers of 32 nodes each with ReLU activation, 64-dimension code layer, and the decoder contained 3 dense layers with ReLU activation except the final layer which was linear. The architecture was explored using the encoded validation set coming from stage 1 (the VAE with learnable gamma). This second stage VAE was trained for 50 epochs, with KL converging to zero quickly. Uniformity for the 2-Stage VAE was assessed using sampling from the test set which was encoded after stage 1 training was complete.

Similar to the models for MNIST and CelebA, the dense layer encoding for the GoFAE models, $\mathbf{\Theta}$, is trained with Riemannian SGD. The learning rate is $5e-4$ and uses a 1-Cycle learning rate scheduler with a maximum learning rate of $1e-4$. Each encoded batch is projected onto a randomly sampled 64×64 orthonormal basis. The test statistic T is applied to each direction and the final statistic (min, average, max) computed.

E.2 MNIST

In the following sections, we visually assess the quality of GoFAE and competing models. For the GoFAE plots, a "good" regularization coefficient $\lambda > 0$ indicates that the model should (i) fail to reject $\mathcal{H}_0$ at an α-level chosen a-priori (close to normal but not overly so), (ii) accurately reconstruct the input, and finally (iii) generate samples which are both qualitatively and quantitatively convincing.

In this section, we present performance of competing and GoFAE methods in Table S1. Reconstruction error, reconstruction FID, generation FID, and the mean (std) of the KS_{unif} p-values are reported.

Additionally, we examine the effects of λ with GoFAE-SW model in Figure S1. The GoF test pushes the transformation to become increasingly singular in the sense that the condition number of the covariance matrix becomes increasingly large (Figure S1). As λ gets larger, the average KS uniformity statistic also increases, indicating the aggregate posterior distribution appears to be indistinguishable from the Gaussian class. However, as λ continues to increase, the condition number of the estimated covariance matrix also increases. Yet, for a small interval of λ, the mutual information has not completely diminished, and the KS uniformity suggests the aggregated posterior is still indistinguishable from Gaussian. This visualization provides insights on how the class of Gaussians allows the model to adapt as needed during training.

Table S1: Evaluation of MNIST by MSE, FID scores Heusel et al. (2017), and samples with p-values from higher criticism.

Algorithm	MSE ↓	FID Score ↓		Kolmogorov–Smirnov Uniformity Test				
		Gen.	Recon.	SW	SF	CVM	KS	EP
AE-Baseline	7.66	62.29	7.39	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
VAE	13.99	16.23	8.33	0.18 (0.18)	0.42 (0.30)	0.07 (0.12)	0.14 (0.18)	0.03 (0.06)
$\beta(2)$-VAE	20.98	20.54	11.52	0.38 (0.31)	0.29 (0.26)	0.41 (0.29)	0.45 (0.31)	0.24 (0.25)
C-β-VAE	20.94	19.21	11.38	0.48 (0.30)	0.37 (0.28)	0.29 (0.26)	0.42 (0.31)	0.19 (0.19)
WAE-GAN	9.49	15.72	**7.28**	0.07 (0.12)	0.49 (0.29)	0.09 (0.15)	0.11 (0.12)	0.06 (0.12)
GoFAE-SW	7.16	16.58	7.57	0.11 (0.16)	0.48 (0.23)	0.06 (0.12)	0.17 (0.24)	0.03 (0.05)
GoFAE-SF	**7.08**	18.35	7.47	0.0 (0.0)	0.26 (0.27)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
GoFAE-CVM	7.67	**15.56**	7.71	0.14 (0.18)	0.46 (0.28)	0.14 (0.22)	0.12 (0.21)	0.05 (0.15)
GoFAE-EP	7.54	15.85	7.84	0.25 (0.27)	0.44 (0.28)	0.12 (0.23)	0.17 (0.20)	0.05 (0.08)

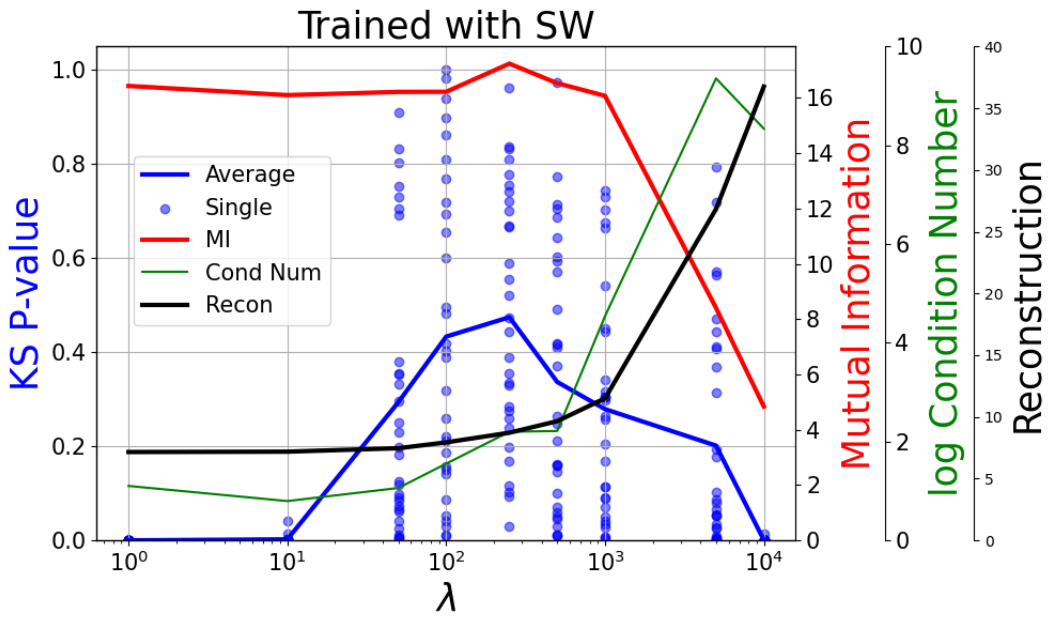


Figure S1: Effects of λ. We evaluated the uniformity KS test p-values (blue dots), average KS p-values (blue line), mutual information (red line), log condition number (green line) and reconstruction error (black line) under different λ for GoFAE-SW.

E.3 Additional Plots

In this section, we include additional experimental results for GoFAE and comparison models.

E.3.1 CelebA

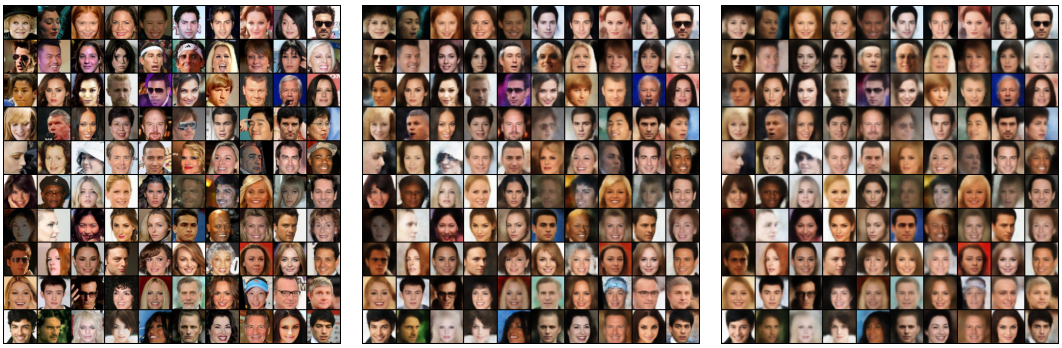
This section contains images produced from GoFAE and competing GAEs on the CelebA dataset with a latent space dimension of 64. Figure S2 depicts reconstructed faces. In Table 1 in the main paper, we evaluated reconstruction quality with mean-squared error and the Frechet-Inception Distance (FID). The visual quality of the images produced by the GoFAE models are consistent with the low MSE and reconstruction FID observed in Table 1. Figure S3 contains faces generated from the model. The GoFAE models again produced competitive FID scores, with GoFAE-SF producing the lowest (best) of the group. Tables S2 and S3 illustrate the effects of λ on KS uniformity test, corresponding p-value, reconstruction error, and test statistics for GoFAE-SW, GoFAE-SF, GoFAE-CVM and GoFAE-KS, respectively.

There are several things to note from the Tables of figures demonstrating the effects of λ on the training of GoFAE architectures (Table S2 and Table S3). The first row shows the mutual information

and GoF p-value distribution computed from the test set as a function of $\log(\lambda)$. As λ increases the models are producing aggregated posteriors that are appearing increasingly indistinguishable from normal. However, beyond a certain λ, the encoding no longer appears normal, GoFAE-SW clearly depicts this occurring in Table (S2a, left). Row (b) tracks the average p-value produced on the test set while training with a particular GoF test statistic. Larger λ can easily be seen to focus more on normality as the p-values tend to start at high levels and stay. The reconstruction error in row (c) converges quickly.

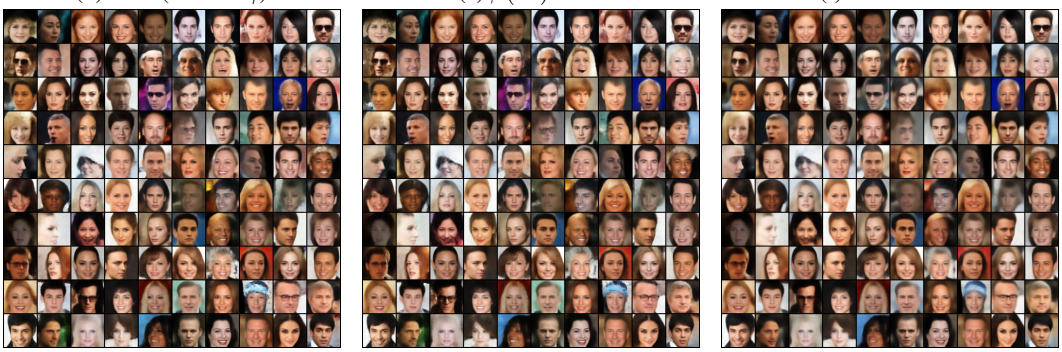
Of particular interest in the $\log(\lambda)$ vs KS p-value plots, is the behavior of the average KS p-value; after the average KS uniformity peaks, the mutual information falls rapidly. This clearly illustrates the relationship posited in the experiments section where the model prioritizes posterior matching.

Similar plots for MNIST can be seen in section E.3.2. with the quantitative information regarding reconstruction error, reconstruction FID and generation FID in Table S1. Once again, the GoFAE models produce overall lower MSE than competitor models, while maintaining competitive generation FID scores.



(a) Ground truth (b) AE (c) VAE (fixed γ)

(d) VAE (learned γ) (e) $\beta(10)$-VAE (f) WAE-GAN



(g) GoFAE-SW (h) GoFAE-SF (i) GoFAE-CVM

(j) GoFAE-KS

Figure S2: Comparison of reconstruction between competing methods: (b) AE, (c) VAE (fixed γ), (d) VAE (learned γ), (e) β-VAE ($\beta = 10$), (f) WAE-GAN, and our framework: (g) GoFAE-SW, (h) GoFAE-SF, (i) GoFAE-CVM, (j) GoFAE-KS with CelebA.

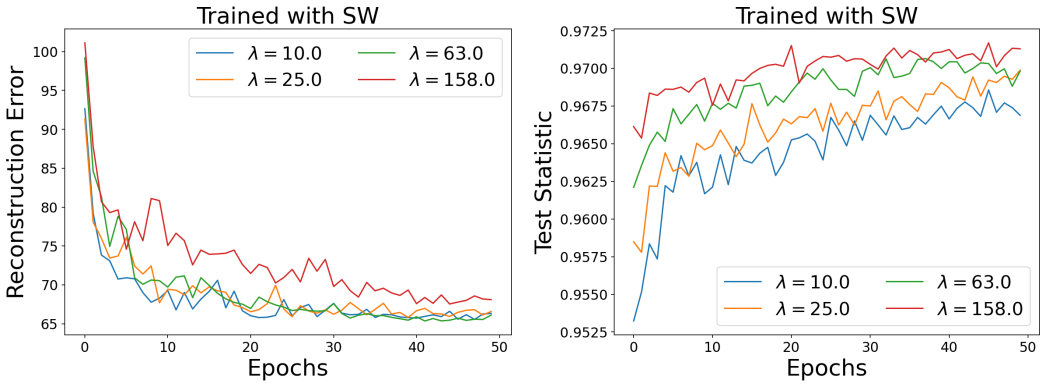
(a) Ground truth (b) AE (c) VAE (fixed γ)

(d) VAE (learned γ) (e) $\beta(10)$-VAE (f) WAE-GAN

(g) 2-Stage VAE (h) GoFAE-SW (i) GoFAE-SF

(j) GoFAE-CVM (k) GoFAE-KS

Figure S3: Comparison of generated samples between competing methods: (b) AE, (c) VAE (fixed γ), (d) VAE (learned γ), (e) β-VAE ($\beta = 10$), (f) WAE-GAN, (g) 2-Stage VAE, and our framework: (h) GoFAE-SW, (i) GoFAE-SF, (j) GoFAE-CVM, (k) GoFAE-KS with CelebA.



(a) Test Reconstructions

(b) T_{SW} Statistics

Figure S4: Convergence of reconstruction error (a) and T_{SW} (b) for multiple λ.

Table S2: The effects of λ with GoFAE-SW (left) and GoFAE-SF (right) for CelebA. From top to bottom, we have: (a) the normality of GoFAE mini-batch encodings using 30 repetitions of Algorithm 1 for different λ. (b) p-value changing from epoch 1-50 under varying λ. (c) Reconstruction error on the testing set with different λ. (d) Test statistics value on the testing set with different λ.

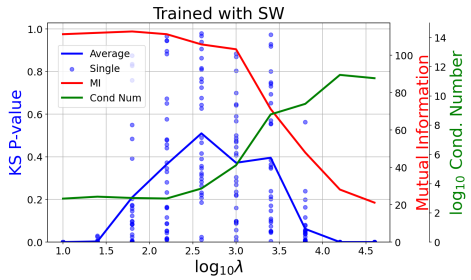
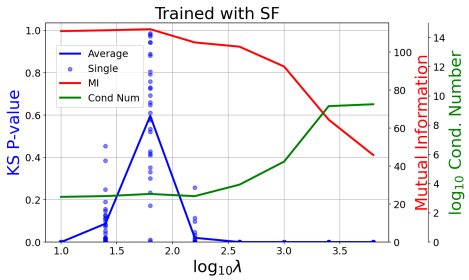
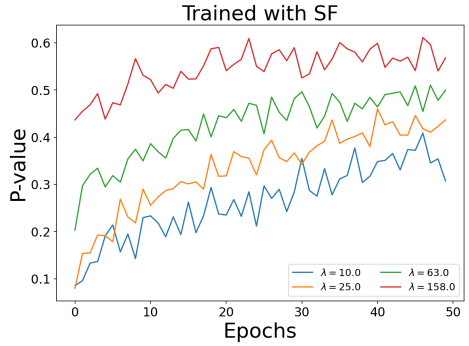
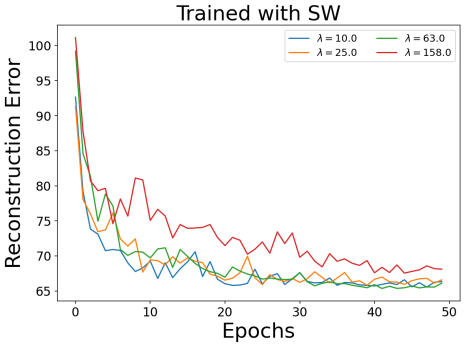
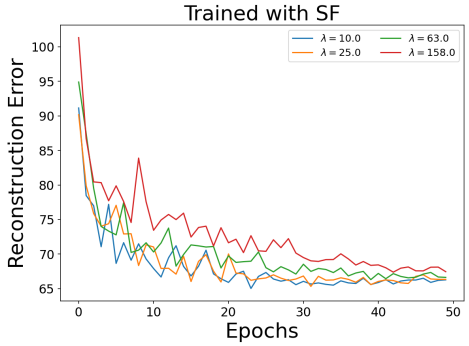
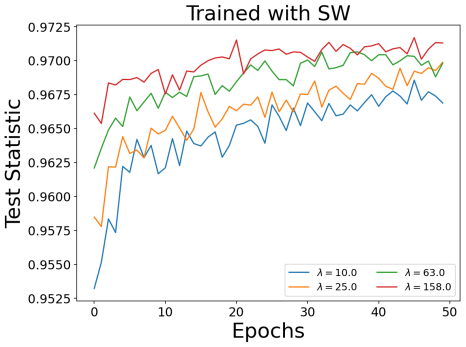
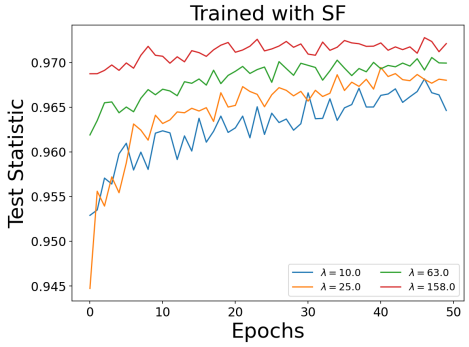
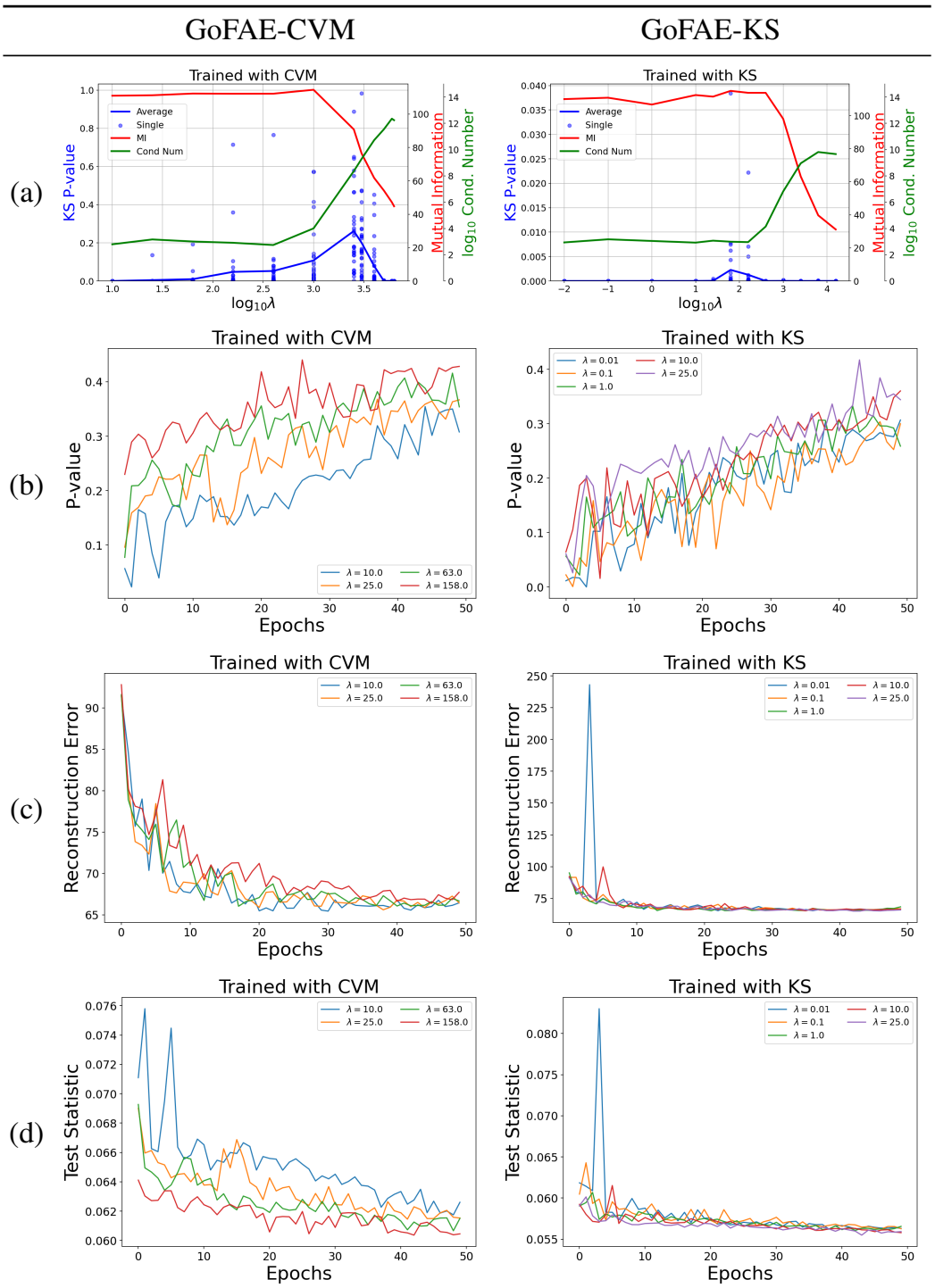
	GoFAE-SW	GoFAE-SF
(a)		
(b)		
(c)		
(d)		

Table S3: The effects of λ with GoFAE-CVM (left) and GoFAE-KS (right) for CelebA. From top to bottom, we have: (a) the normality of GoFAE mini-batch encodings using 30 repetitions of Algorithm 1 for different λ. (b) p-value changing from epoch 1-50 under varying λ. (c) Reconstruction error on the testing set with different λ. (d) Test statistics value on the testing set with different λ.



E.3.2 MNIST

This section includes additional reconstructed and generated samples in Figure S5 and Figure S6. Table S4 and S5 demonstrate the effects of λ on KS uniformity test, corresponding p-value, reconstruction error, and test statistics for GoFAE-SW, GoFAE-SF, GoFAE-CVM and GoFAE-KS, respectively. The experiments on MNIST reach the same conclusion as in the CelebA, that our GoFAE methods is competitive in both reconstruction and generation.

(a) Ground truth

(b) AE

(c) VAE

(d) $\beta(2)$-VAE

(e) C-β-VAE

(f) WAE-GAN

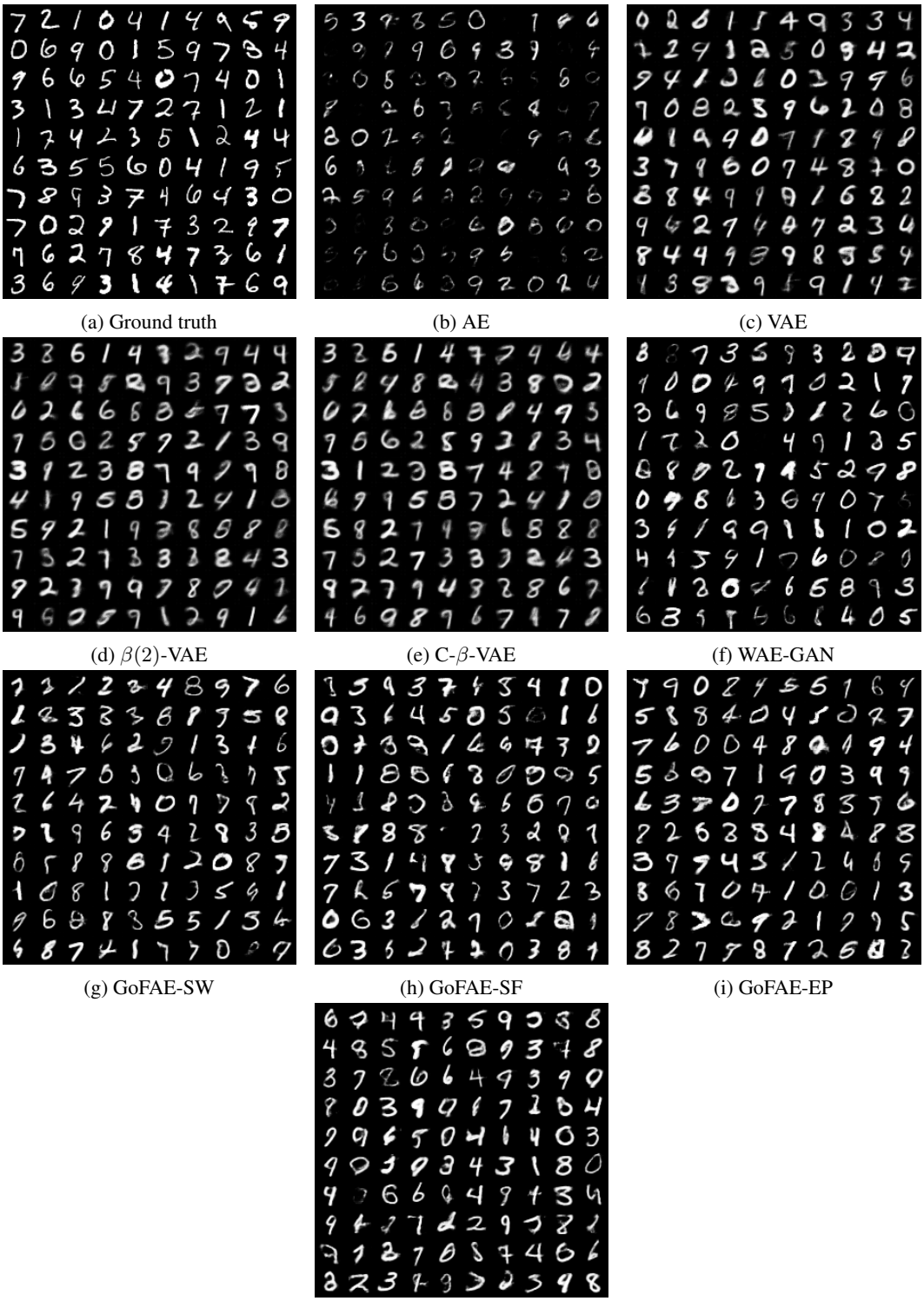
(g) GoFAE-SW

(h) GoFAE-SF

(i) GoFAE-EP

(j) GoFAE-CVM

Figure S5: Comparison of reconstruction between competing methods (b) AE, (c) VAE, (d) β-VAE ($\beta = 2$), (e) C-β-VAE, (f) WAE-GAN, and our framework: (g) GoFAE-SW, (h) GoFAE-SF, (i) GoFAE-EP, (j) GoFAE-CVM with MNIST.



(a) Ground truth (b) AE (c) VAE

(d) $\beta(2)$-VAE (e) C-β-VAE (f) WAE-GAN

(g) GoFAE-SW (h) GoFAE-SF (i) GoFAE-EP

(j) GoFAE-CVM

Figure S6: Comparison of generated samples between competing methods (b) AE, (c) VAE, (d) β-VAE ($\beta = 2$), (e) C-β-VAE, (f) WAE-GAN, and our framework: (g) GoFAE-SW, (h) GoFAE-SF, (i) GoFAE-EP, (j) GoFAE-CVM with MNIST.

Table S4: The effects of λ with GoFAE-SW (left) and GoFAE-SF (right) for MNIST. From top to bottom, we have: (a) p-value changing from epoch 1-50 under varying λ. (b) Reconstruction error on the testing set with different λ. (c) Test statistics value on the testing set with different λ.

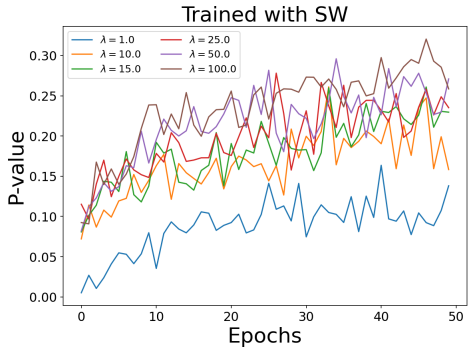
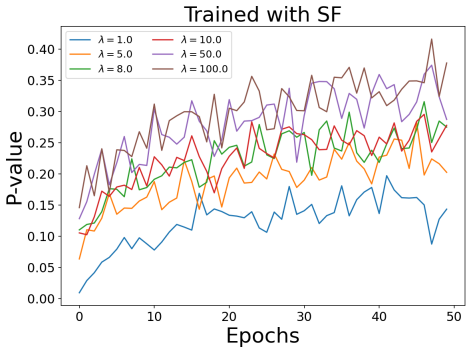
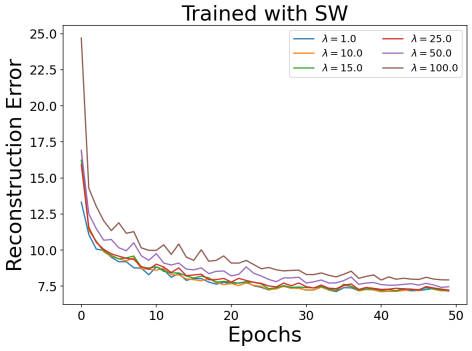
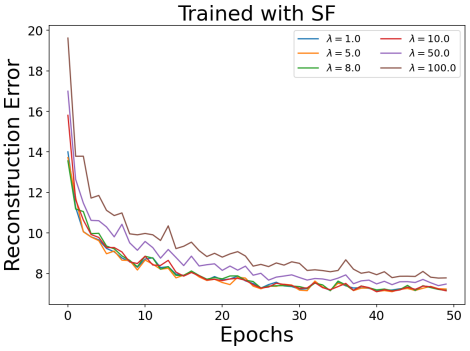
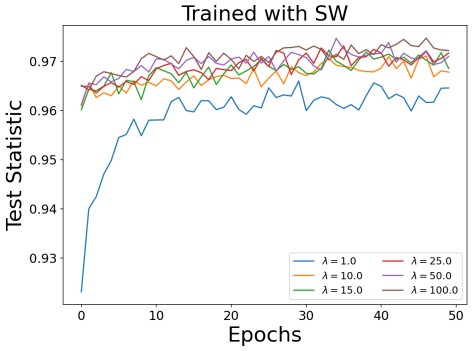
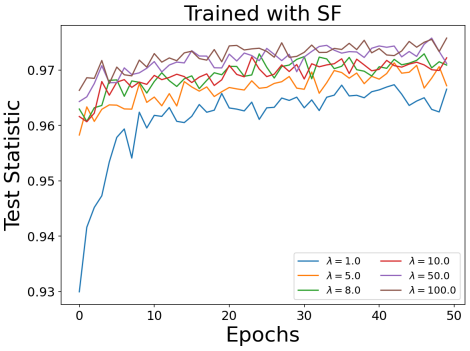
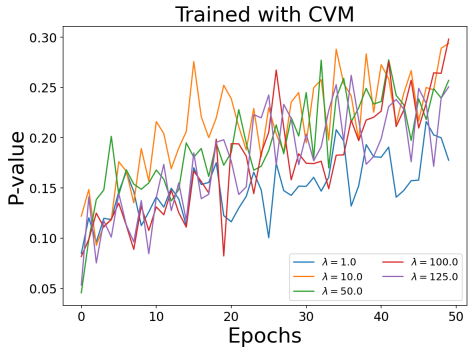
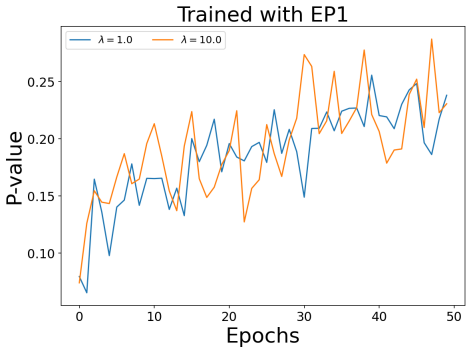
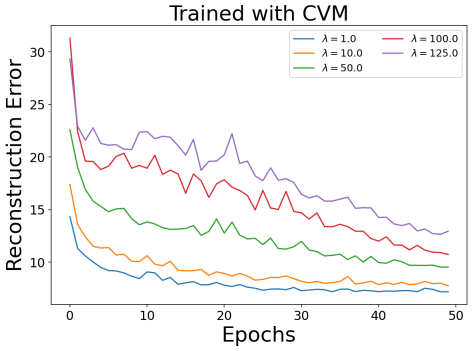
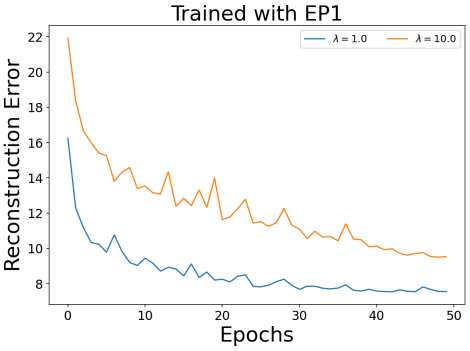
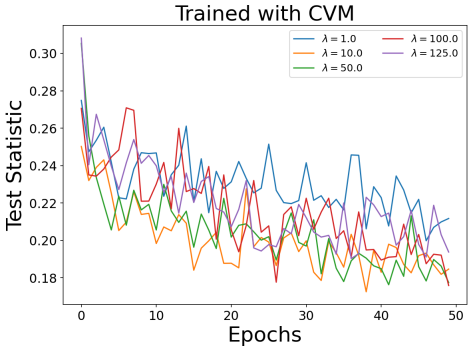
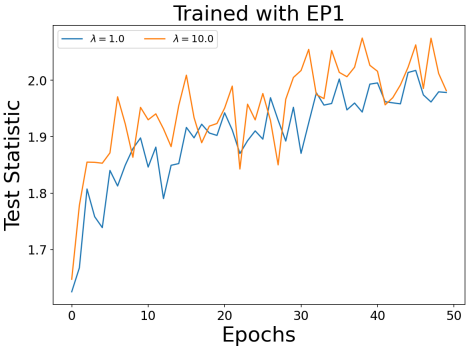
	GoFAE-SW	GoFAE-SF
(a)		
(b)		
(c)		

Table S5: The effects of λ with GoFAE-CVM (left) and GoFAE-EP (right) for MNIST. From top to bottom, we have: (a) p-value changing from epoch 1-50 under varying λ. (b) Reconstruction error on the testing set with different λ. (c) Test statistics value on the testing set with different λ.

	GoFAE-CVM	GoFAE-EP
(a)		
(b)		
(c)		

E.3.3 CIFAR-10

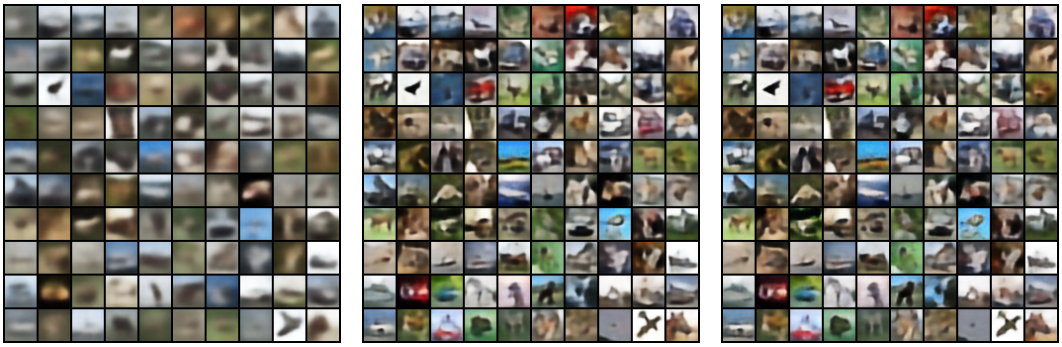
This section includes MSE, FID scores, and p-values from the higher criticism evaluation in Table S6 and reconstructed and generated samples in Figure S7 and Figure S8. Table S7 and S8 demonstrate the effects of λ on KS uniformity test, corresponding p-value, reconstruction error, and test statistics for GoFAE-SW, GoFAE-SF, GoFAE-CVM and GoFAE-KS, respectively. The 2-Stage VAE uses the VAE with learned γ for reconstruction. The experiments on CIFAR-10 reach the same conclusion as in the CelebA and MNIST, that our GoFAE methods is competitive in both reconstruction and generation.

Table S6: Evaluation of CIFAR-10 by MSE, FID scores, and samples with p-values from higher criticism.

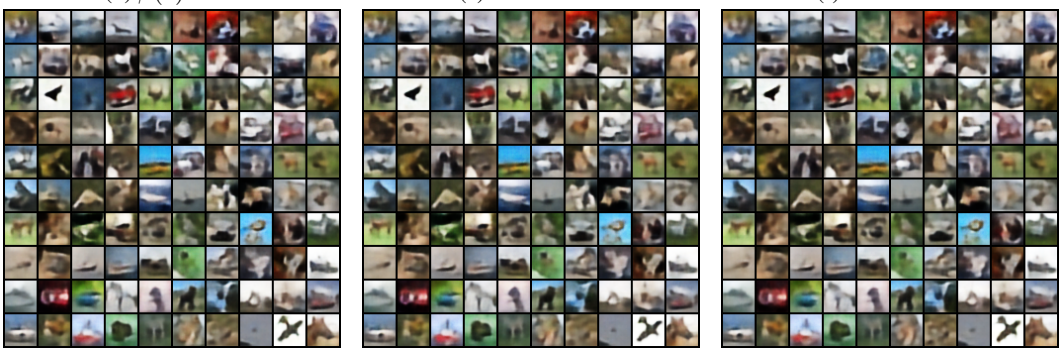
Algorithm	MSE ↓	FID Score ↓	
		Recon.	Gen.
VAE (fixed γ)	51.07 (0.04)	179.75	188.55
VAE (learned γ)	**24.10 (0.01)**	97.72	119.86
2-Stage-VAE	24.10 (0.01)	97.72	117.60
$\beta(2)$-VAE	65.81 (0.07)	252.98	262.55
WAE-GAN	25.53 (0.00)	109.55	120.30
GoFAE-SW	24.57 (0.08)	96.86	115.54
GoFAE-SF	24.69 (0.06)	**97.71**	**115.76**
GoFAE-CVM	24.40(0.06)	99.76	119.51
GoFAE-KS	24.47 (0.05)	102.47	121.97
GoFAE-EP	24.74 (0.06)	106.52	124.79

Algorithm	Kolmogorov–Smirnov Uniformity Test				
	SW	SF	CVM	KS	EP
VAE (fixed γ)	0.29 (0.28)	0.28 (0.24)	0.58 (0.29)	0.54 (0.30)	0.48 (0.28)
VAE (learned γ)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.03)	0.00 (0.00)
2-Stage-VAE	0.44 (0.27)	0.45 (0.28)	0.48 (0.30)	0.49 (0.26)	0.47 (0.28)
$\beta(2)$-VAE	0.43 (0.25)	0.29 (0.26)	0.52 (0.33)	0.46 (0.33)	0.48 (0.31)
WAE-GAN	0.06 (0.18)	0.01 (0.03)	0.41 (0.32)	0.37 (0.32)	0.32 (0.28)
GoFAE-SW(30)	0.47 (0.32)	0.51 (0.28)	0.37 (0.29)	0.44 (0.32)	0.23 (0.22)
GoFAE-SF(12)	0.00 (0.01)	0.10 (0.16)	0.01 (0.05)	0.08 (0.11)	0.01 (0.01)
GoFAE-CVM(5)	0.03 (0.08)	0.00 (0.01)	0.19 (0.21)	0.31 (0.31)	0.12 (0.23)
GoFAE-KS	0.00 (0.00)	0.00 (0.00)	0.02 (0.04)	0.10 (0.18)	0.00 (0.00)
GoFAE-EP(5)	0.43 (0.25)	0.36 (0.26)	0.4 (0.25)	0.53 (0.26)	0.47 (0.27)

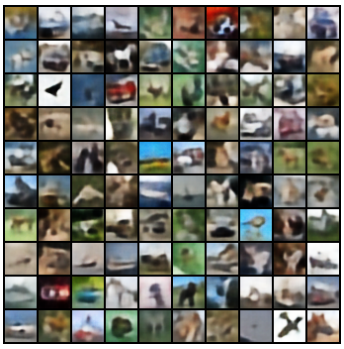
(a) Ground truth (b) VAE (fixed γ) (c) VAE (learned γ)



(d) $\beta(2)$-VAE (e) WAE-GAN (f) GoFAE-SW

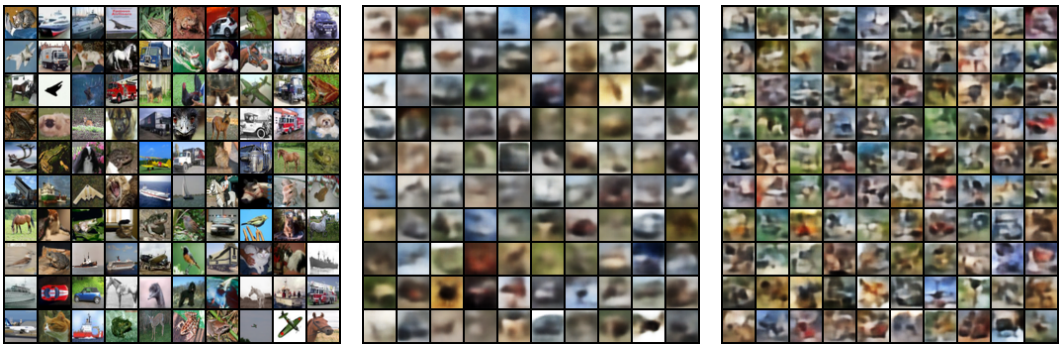


(g) GoFAE-SF (h) GoFAE-CVM (i) GoFAE-KS



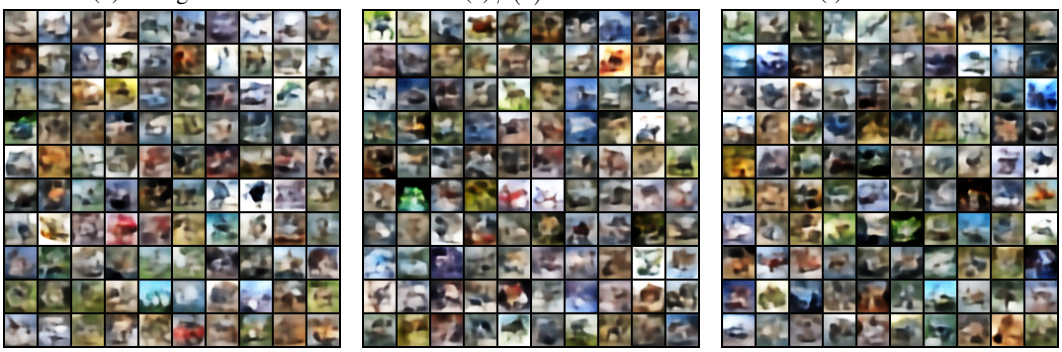
(j) GoFAE-EP

Figure S7: Comparison of reconstructed images between competing methods (b) VAE (fixed γ), (c) VAE (LEARNED γ), (d) β-VAE ($\beta = 2$), (e) WAE-GAN, and our framework: (f) GoFAE-SW, (g) GoFAE-SF, (h) GoFAE-CVM, (i) GoFAE-KS, (j) GoFAE-EP with CIFAR-10.

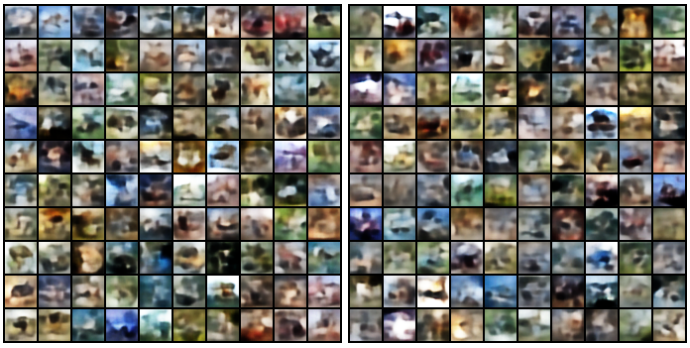


(a) Ground truth (b) VAE (fixed γ) (c) VAE (learned γ)

(d) 2-Stage VAE (e) $\beta(2)$-VAE (f) WAE-GAN



(g) GoFAE-SW (h) GoFAE-SF (i) GoFAE-CVM



(j) GoFAE-KS (k) GoFAE-EP

Figure S8: Comparison of generated samples between competing methods (b) VAE (fixed γ), (c) VAE (LEARNED γ), (d) 2-Stage-VAE, (e) β-VAE ($\beta = 2$), (f) WAE-GAN, and our framework: (g) GoFAE-SW, (h) GoFAE-SF, (i) GoFAE-CVM, (j) GoFAE-KS, (k) GoFAE-EP with CIFAR-10.

Table S7: The effects of λ with GoFAE-SW (left) and GoFAE-SF (right) for CIFAR-10. From top to bottom, we have: (a) p-value changing from epoch 1-50 under varying λ. (b) Reconstruction error on the testing set with different λ. (c) Test statistics value on the testing set with different λ.

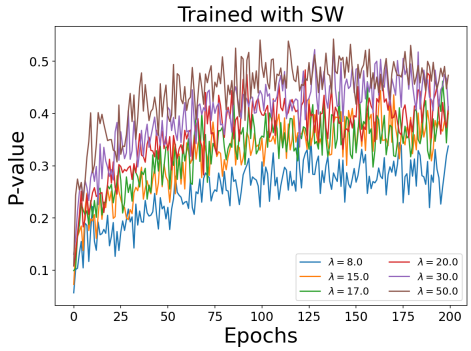
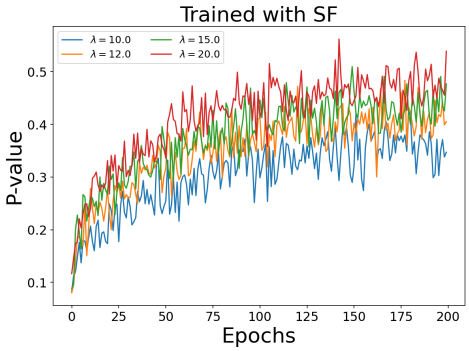
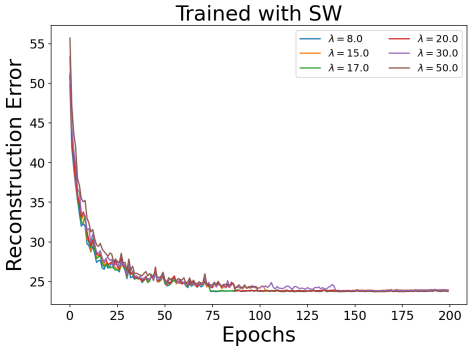
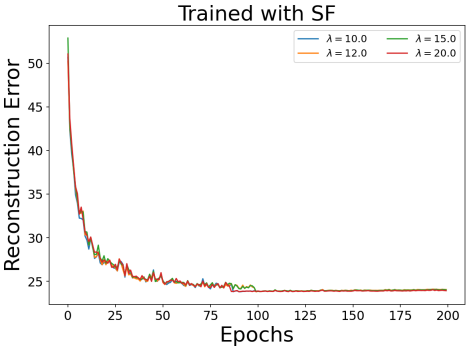
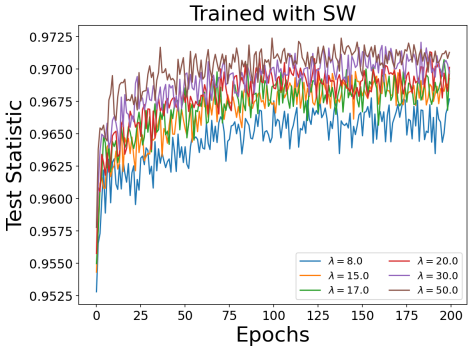
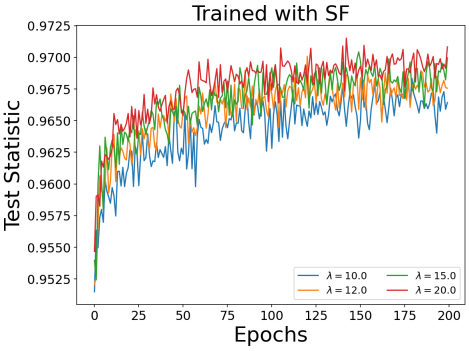
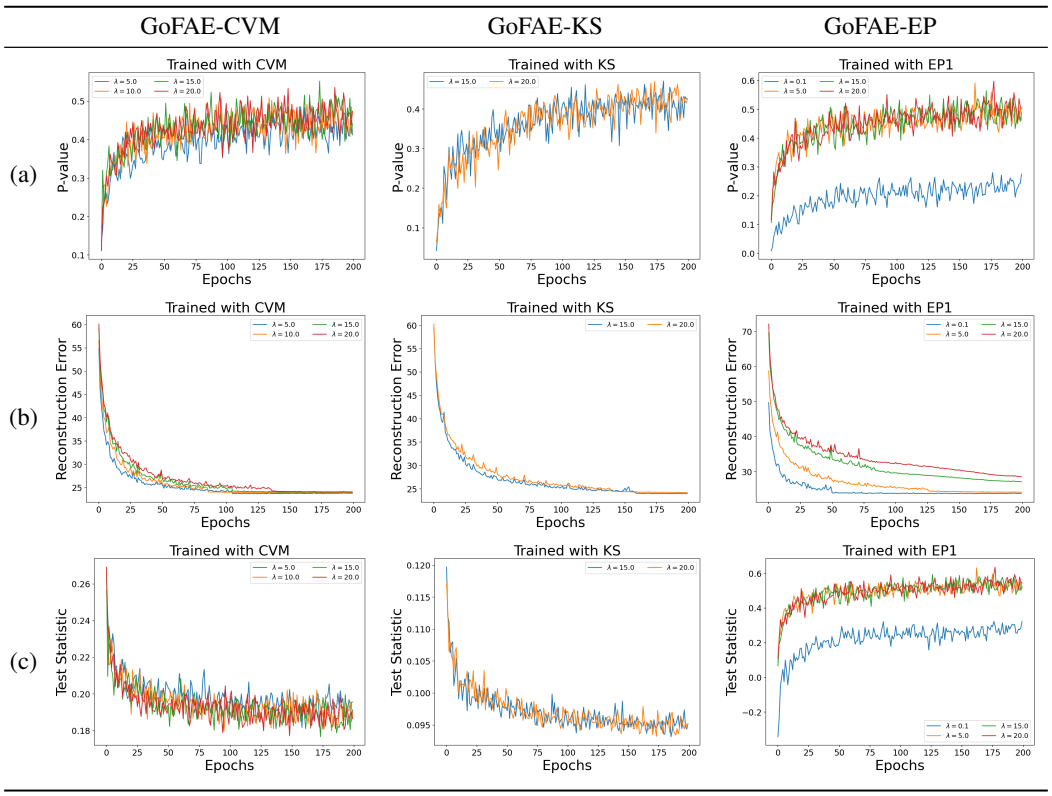
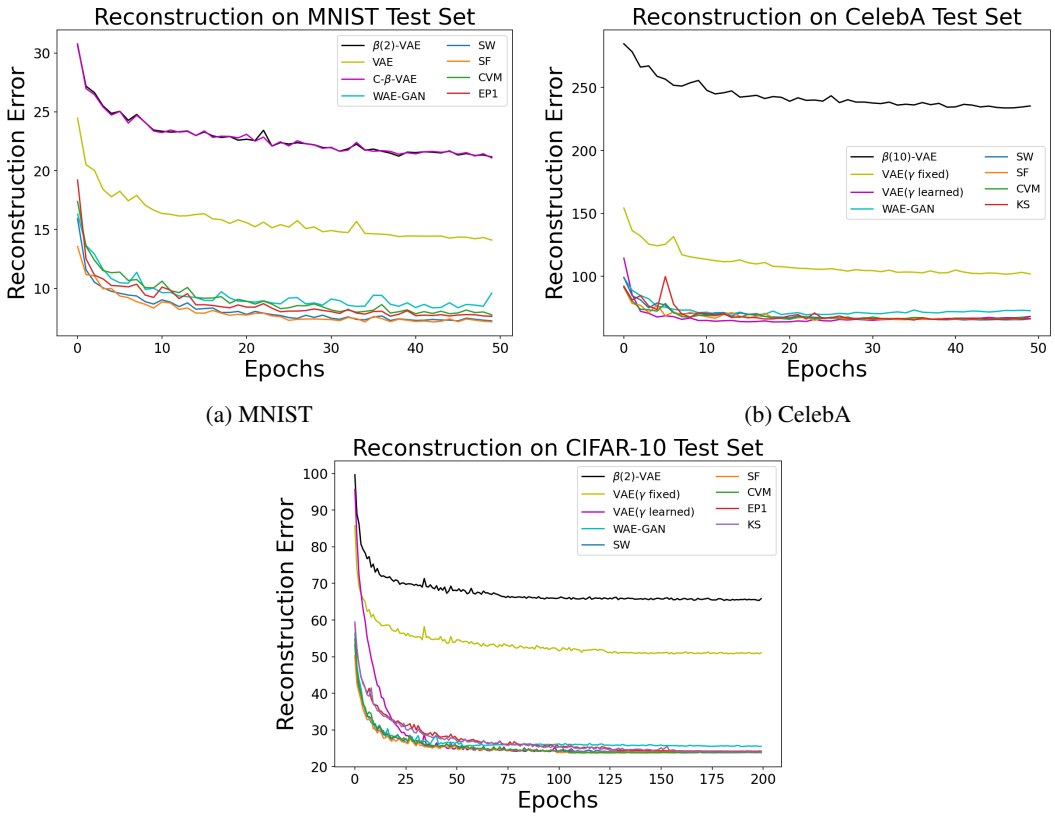
	GoFAE-SW	GoFAE-SF
(a)		
(b)		
(c)		

Table S8: The effects of λ with GoFAE-CVM (left), GoFAE-KS (middle), and GoFAE-EP (right) for CIFAR-10. From top to bottom, we have: (a) p-value changing from epoch 1-50 under varying λ. (b) Reconstruction error on the testing set with different λ. (c) Test statistics value on the testing set with different λ.



E.4 CONVERGENCE ANALYSIS FOR COMPETING METHODS

In this section, we include the test set reconstruction error over training epochs for the competing methods for MNIST, CelebA and CIFAR-10 data. Since we adopted a common autoencoder architecture (Section C), reconstruction error follows a similar trend across all models (Figure S9). Since the 2-Stage VAE shares the VAE with learned γ as the first stage, we was assessed its convergence in the same manner as the VAE with learned γ.



(a) MNIST (b) CelebA

(c) CIFAR-10

Figure S9: Reconstruction error over training epochs for competing methods.

E.5 ABLATION STUDY

We proposed Riemannian SGD as a solution to several problems associated with optimizing GoF test statistics. Theoretically, this approach is necessary to prove convergence. However, it remains to be seen whether there is any impact on model performance when the constraints are removed. To explore this, we considered several GoFAE models trained them without RSGD on MNIST and CIFAR-10. We used traditional SGD in place of Riemannian SGD; $\mathbf{\Theta}$ is no longer retracted to the Stiefel manifold.

GoFAE models with SW, SF, CVM and EP are retrained using the same architecture and hyper-parameter configuration on MNIST. In place of the orthogonal initialization used for $\mathbf{\Theta}$, Kaiming initialization He et al. (2015) was used. Table S1 summarizes the results with the best performing model in bold for each evaluation measure.

GoFAE models with SW, SF, CVM, and EP were also retrained on On CIFAR-10. All architectures and hyper-parameters remained the same as the Riemannian SGD counterparts, except that $\mathbf{\Theta}$ which was initialized following Kaiming initialization. Table S10 summarized the results with the best performing model in bold for each evaluation measure. These results show empirical support for

using Riemannian SGD, particularly for generation. The results for reconstruction performance are less clear and would benefit from further experimentation.

Table S9: Comparison of Riemannian SGD (left) and standard SGD (right) for MNIST.

	Riemannian SGD			Standard SGD		
	MSE	Gen FID	Recon FID	MSE	Gen FID	Recon FID
SW	**7.16**	**16.58**	**7.57**	7.31	19.47	7.71
SF	**7.08**	**18.35**	7.47	7.12	20.36	**7.36**
CVM	**7.67**	15.56	7.71	7.79	**15.12**	**7.60**
EP	**7.54**	**15.85**	7.84	**7.54**	16.59	**7.63**

Table S10: Comparison of Riemannian SGD (left) and standard SGD (right) for CIFAR-10.

	Riemannian SGD			Standard SGD		
	MSE	Gen FID	Recon FID	MSE	Gen FID	Recon FID
SW	24.58	**117.05**	**100.11**	**23.94**	119.48	101.27
SF	24.69	**115.76**	**97.71**	**23.92**	118.05	101.37
CVM	24.40	**119.51**	**99.76**	**23.68**	120.48	100.22
KS	24.47	**121.97**	**102.47**	**24.03**	122.23	103.22
EP	24.74	**124.79**	106.52	**24.05**	124.88	**105.19**