On Convergence in Wasserstein Distance and f-divergence Minimization Problems

Cheuk Ting Li* Chinese University of Hong Kong
Jingwei Zhang* Chinese University of Hong Kong
Farzan Farnia Chinese University of Hong Kong
* Equal Contribution (alphabetically-ordered)

Abstract

The zero-sum game in generative adversarial networks (GANs) for learning the distribution of observed data is known to reduce to the minimization of a divergence measure between the underlying and generative models. However, the current theoretical understanding of the role of the target divergence in the characteristics of GANs' generated samples remains largely inadequate. In this work, we aim to analyze the influence of the divergence measure on the local optima and convergence properties of divergence minimization problems in learning a multi-modal data distribution. We show a mode-seeking f-divergence, e.g. the Jensen-Shannon (JS) divergence in the vanilla GAN, could lead to poor locally optimal solutions missing some underlying modes. On the other hand, we demonstrate that the optimization landscape of 1-Wasserstein distance in Wasserstein GANs does not suffer from such suboptimal local minima. Furthermore, we prove that a randomly-initialized gradient-based optimization of the Wasserstein distance will, with high probability, capture all the existing modes. We present numerical results on standard image datasets, revealing the success of Wasserstein GANs compared to JS-GANs in avoiding suboptimal local optima under a mixture model.

1 INTRODUCTION

Generative adversarial networks (GANs) (Goodfellow et al., 2014) have achieved remarkable results in various tasks in computer vision (Brock et al., 2018), speech processing (Donahue et al., 2018), and computational biology (Gupta and Zou, 2019). The GAN framework is based on a zero-sum game between a generator G aiming to map a random input Z to a real-like sample and discriminator D distinguishing generated samples from real data X. GANs are usually formulated as a min-max optimization problem to optimize the two involved players:

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{D}} V(G,D). \tag{1}$$

Here $\mathcal{G}$, $\mathcal{D}$ represent the function sets for the generator and discriminator players, respectively, and the min-max objective function $V(G,D)$ is designed to be D's assigned dissimilarity score between G's generated samples and training data.

A standard approach to formulate the GAN objective function $V(G,D)$ is to utilize the dual representation of a divergence function $d(P_X, P_{G(Z)})$ between the distribution of real data P_X and generative model $P_{G(Z)}$. For example, the vanilla GAN (VGAN) (Goodfellow et al., 2014) is based on the Jensen-Shannon (JS) divergence; f-GAN (Nowozin et al., 2016) optimizes a general f-divergence; Wasserstein GAN (WGAN) (Arjovsky et al., 2017; Gulrajani et al., 2017) minimizes the 1-Wasserstein distance; Lipschitz GANs such as DRAGAN (Kodali et al., 2017) and SN-GAN (Miyato et al., 2018) minimize a hybrid of Wasserstein distance and JS-divergence according to (Farnia and Tse, 2018).

Considering the GAN formulations optimizing different divergence scores, a natural question is what the effects of a target divergence are on the properties of the trained generative model. In other words, we seek to understand the consequences of optimizing different divergence measures and study the induced properties of the trained generative model. Addressing this question is the key to the proper selection of a GAN formulation for a specific application of interest.

The above question has been numerically investigated in (Lucic et al., 2018; Sajjadi et al., 2018;

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

Kynkäänniemi et al., 2019; Naeem et al., 2020). These studies propose a set of evaluation metrics to assess the quality and diversity of the GAN's generated data, and compare different GAN formulations based on their achieved empirical scores. On the other hand, the theoretical studies of the question have mostly focused on the effect of divergence measures on the training stability and generalization in GAN min-max optimization (Arjovsky et al., 2017; Kodali et al., 2017; Arjovsky and Bottou, 2017; Feizi et al., 2020) and have not adequately compared the properties of the optimal generator following from different divergence measures.

In this work, we carry out a theoretical study on the influence of divergence measures on the *local optima* of divergence minimization problems. In our theoretical analysis, we consider the local minima of the divergence cost function and attempt to evaluate the diversity performance of the trained generative model. Since GANs are commonly trained via first-order optimization methods, the existence of poor local optima would lead to suboptimal generator functions with unsatisfactory diversity and quality performance.

Specifically, we consider *multi-modal underlying distributions* comprised of multiple separate modes, where the learner searches for the center points of the modes. Our first result shows that under bounded and well-separated modes, the optimization landscape of the Wasserstein distance does not suffer from poor locally optimal solutions for gradient-based methods missing any mode present in the data distribution. Additionally, we prove that a randomly-initialized gradient-based minimization of the 1-Wasserstein distance over a multi-modal generative model is expected to discover all the underlying mode centers. This theoretical result suggests the power of Wasserstein GANs in capturing the diversity of a multi-modal distribution.

On the other hand, we prove that for a class of f-divergence measures characterized as "mode-seeking" in the recent paper (Li and Farnia, 2023), such as the JS-divergence, the divergence minimization problem for finding the modes' center will suffer from poor local optima missing one or multiple center points in the underlying distribution. This result not only indicates the possibility of mode collapse in mode-seeking f-GANs, but also indicates that a trained generator via mode-seeking f-GANs may produce high-quality samples from a subset of modes while missing some modes in the underlying mixture distribution. Therefore, our theoretical study implies the challenges of applying mode-seeking f-GANs, e.g. VGAN minimizing the JS-divergence, to learn a multi-modal distribution, a phenomenon that has also been numerically observed in (Lucic et al., 2018; Li and Farnia, 2023).

Finally, we present the results of several numerical experiments to validate our theoretical findings. In our experiments, we consider synthetic Gaussian mixture data as well as standard image datasets to support the theoretical statements. Our empirical results suggest that Wasserstein GANs manage to capture the existing modes in the dataset, while the vanilla GAN minimizing the JS-divergence could fail in detecting all the existing modes and in multiple cases could completely ignore some of the underlying components. In contrast, we numerically show that the Lipschitz GANs minimizing the hybrid divergence of Wasserstein distance and JS-divergence can successfully capture the variety in the training set. In summary, the followings are this work's main contributions:

- A theoretical study of the influence of a divergence measure on the local minima in divergence minimization problems.
- Proving a theoretical guarantee for the convergence of gradient-based Wasserstein distance minimization under a multi-modal distribution.
- Demonstrating the existence of poor locally optimal solutions in mode-seeking f-divergence minimization problems.
- Providing a numerical evaluation of GAN models with different divergence measures in application to multi-modal underlying distributions.

2 RELATED WORK

Convergence and Stability in GAN Training. Multiple studies have been conducted to investigate the convergence properties of different GAN formulations. Generally, there is no guarantee that the nonconvex-nonconcave min-max optimization of GANs will converge to a global saddle point (Nash equilibrium) (Farnia and Ozdaglar, 2020), and the existing results only guarantee convergence to local min-max optima (Heusel et al., 2017). Also, it was shown in (Nagarajan and Kolter, 2017) that a simplified linearized GAN optimization via gradient descent is locally stable. It was hypothesized in (Kodali et al., 2017) that suboptimal local equilibria in GANs are responsible for mode collapse. In another related work (Liu et al., 2017), the relative strength of the convergence of various adversarial formulations of divergence measures is investigated. We note that unlike our analysis, the mentioned works do not focus on the local optima of the GAN's target divergence measure, which is in general different from the set of local min-max optima in the GAN optimization.

A discussed advantage of Wasserstein GANs in the literature is their training stability (Arjovsky and Bottou, 2017; Arjovsky et al., 2017). Nevertheless, it was shown in (Nagarajan and Kolter, 2017; Mescheder et al., 2018) that a simplified linearized formulation of the Wasserstein GAN can have non-convergent limit cycles in the min-max optimization. We remark that the non-convergent cycles in the linearized min-max GAN problem do not correspond to the critical points of the Wasserstein distance that are analyzed in our work, and also the cycles may not appear in a more precise formulation based on the entire set of 1-Lipschitz discriminators. Regarding the stability of Wasserstein GANs, (Feizi et al., 2020) shows that training of the 2-Wasserstein GAN enjoys global stability in a simplified setting with a linear generator and a quadratic discriminator. (Farnia et al., 2023) shows convergence to local stationary minimax solutions for 2-Wasserstein GANs in learning a mixture of two symmetric Gaussians. For other relevant works, a smoothed Wasserstein GAN that is guaranteed to converge to a stationary point was proposed in (Sanjabi et al., 2018). The convergence properties of sliced Wasserstein distance minimization were studied in (Kolouri et al., 2018; Nadjahi et al., 2019).

Mode-seeking and Mode-covering Divergence Measures. Under a multi-modal data distribution, different choices of divergence measures in the GAN optimization objective could lead to highly different fitted distributions. Divergence measures can be divided into two classes: mode-seeking divergence measures (e.g. reverse KL-divergence) which tend to favor the quality of the samples over their diversity (Bishop, 2006; Huszár, 2015), and mode-covering divergence measures (e.g. KL-divergence) which tend to favor diversity over quality (Bishop, 2006; Poole et al., 2016; Lucas et al., 2019). Theoretical aspects of mode-seeking and mode-covering divergence measures were studied in (Shannon et al., 2020; Li and Farnia, 2023). In (Li and Farnia, 2023), precise theoretical characterizations of mode-seeking f-divergences and performance guarantees are given. This will be elaborated in Section 3.3. The 1-Wasserstein distance often exhibits mode-covering behavior, and is shown not to be mode-seeking (Li and Farnia, 2023).

3 PRELIMINARIES

3.1 Wasserstein Distance and WGANs

A family of GAN problems was studied in (Arjovsky et al., 2017), where the 1-Wasserstein distance between the data and generative models is minimized. The definition of the q-Wasserstein distance between distributions P and Q is the following for every $q \geq 1$ (Villani, 2003):

$$W_q\big(P,Q\big) := \inf_{M\in\Pi(P,Q)} \mathbb{E}_{(X,X')\sim M}\Big[\big\|\mathbf{X}-\mathbf{X}'\big\|^q\Big]^{1/q}, \tag{2}$$

where $\Pi(P,Q)$ is the set of all couplings on (X,X') marginally distributed as $X\sim P$ and $X'\sim Q$. Applying the Kantorovich-Rubinstein duality (Villani, 2003), (Arjovsky et al., 2017) formulates the min-max Wasserstein GAN (WGAN) problem that is equivalent to minimizing W_1-distance between P_X and $P_{G(Z)}$ when $\mathcal{D}$ is the set of all 1-Lipschitz functions:

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{D}_{\text{1-Lipschitz}}} \mathbb{E}\big[D(X)\big]-\mathbb{E}\big[D(G(Z))\big].$$

3.2 f-GANs and f-divergence Minimization

For a convex function $f:[0,\infty)\to\mathbb{R}$ with $f(1)=0$, the f-divergence (Csiszár and Shields, 2004) between distributions P,Q with density functions p,q is defined as

$$d_f(P,Q) := \mathbb{E}_{X\sim Q}\Big[f\Big(\frac{p(X)}{q(X)}\Big)\Big] = \int q(x) f\Big(\frac{p(x)}{q(x)}\Big)\mathrm{d}x. \tag{3}$$

Well-known examples of f-divergence are the KL-divergence with $f_{\mathrm{KL}}(t)=t\log t$ and the JS-divergence with $f_{\mathrm{JS}}(t)=t\log\frac{2t}{t+1}+\log\frac{2}{t+1}$.

Using the variational representation of f-divergence scores (Nguyen et al., 2010), (Nowozin et al., 2016) proposes the following GAN min-max optimization problem called f-GAN, which for a set of all functions $\mathcal{D}$ is equivalent to the f-divergence minimization problem $d_f(P_X,P_{G(Z)})$ between the distributions of data X and generator's output $G(Z)$:

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{D}} \mathbb{E}\big[D(X)\big]-\mathbb{E}\big[f^*\big(D(G(Z))\big)\big]. \tag{4}$$

In the above, $f^*(s):=\sup_t\ st-f(t)$ is the Fenchel-conjugate of function f. The above formulation generalizes the vanilla GAN (VGAN) problem targeting the JS-divergence.

3.3 Mode-seeking Divergence Measures

One operational setting for classifying mode-seeking and mode-covering divergence measures is to fit a Gaussian distribution (or another unimodal distribution) to a mixture of Gaussian distributions, where a mode-seeking divergence measure will fit one of the components (or modes), and a mode-covering divergence measure will give a distribution that cover all the modes (Bishop, 2006). The recent work (Li and Farnia, 2023) offers a theoretical characterization of mode-seeking f-divergence measure, called *strongly mode-seeking divergence measure*, by the following conditions:

- **Conditon I**: $\lim_{t\to\infty} f(t)/t < \infty$.
- **Conditon II**: $f(t)$ is strongly convex for $t \in (0, s]$ for some $s > 1$.

It was shown in (Li and Farnia, 2023) that a strongly mode-seeking divergence measure guarantees that, when a symmetric quasiconcave distribution is fitted to a mixture of symmetric quasiconcave distributions, the fitted distribution will coincide with one of the modes.[1] JS-divergence with $f(t) = t\log\frac{2t}{t+1} + \log\frac{2}{t+1}$ and reverse-KL divergence with $f(t) = -\log(t)$ satisfy the mentioned conditions and are strongly mode-seeking. On the other hand, the first condition does not hold for KL-divergence with $f(t) = t\log t$, which is consistent with the mode-covering behavior that is empirically observed for KL-divergence.

4 THEORETICAL RESULTS ON WASSERSTEIN DISTANCE MINIMIZATION

In this section, we prove the main result showing that for fitting a mixture model distribution to a mixture data distribution, the gradient descent algorithm applied on the loss function given by the 1-Wasserstein distance will converge to the global optimum (i.e., the model distribution matches the data distribution) with high probability, when the initial positions of the modes are initialized at random.

To describe the theoretical setup, consider the setting where the data distribution P is a mixture of n shifted modes of the probability density function p over $\mathbb{R}^m$

$$P(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} p(\mathbf{x} - \boldsymbol{\mu}_i), \tag{5}$$

where $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n \in \mathbb{R}^m$ are distinct. To study the optima and convergence properties of the optimization landscape for divergence scores, we will fit the mixture $Q_{\boldsymbol{\nu}_{1:n}}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} p(\mathbf{x} - \boldsymbol{\nu}_i)$ parameterized by center variables $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ to P by minimizing a target divergence measure d:

$$\min_{\boldsymbol{\nu}_1,\ldots,\boldsymbol{\nu}_n \in \mathbb{R}^m} \; d\big(P, Q_{\boldsymbol{\nu}_{1:n}}\big). \tag{6}$$

While it is clear that the divergence is minimized when $Q_{\boldsymbol{\nu}_{1:n}} = P$, i.e., $\{\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n\} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n\}$, whether this optimum can be found via gradient descent depends on the choice of the divergence measure d. First, we study the convergence behavior when d is chosen to be the 1-Wasserstein distance $W_1(P, Q_{\boldsymbol{\nu}_{1:n}})$. We regard the Wasserstein distance as a function of $\{\boldsymbol{\nu}_j\}$, and optimize it via the gradient descent algorithm as a standard first-order method over the variables $\{\boldsymbol{\nu}_j\}$. The gradient descent update rule for $1 \le j \le n$ is

$$\boldsymbol{\nu}_j^{(t+1)} = \boldsymbol{\nu}_j^{(t)} - \alpha^{(t)} \nabla_{\boldsymbol{\nu}_j} W_1(P, Q_{\boldsymbol{\nu}_{1:n}}), \tag{7}$$

where $\alpha^{(t)} > 0$ is the learning rate at iteration t. The following theorem shows that as long as the distribution p is supported within a ball of radius small enough compared to the distances between $\boldsymbol{\mu}_i$'s, i.e., the components of P are well-separated, then gradient descent will converge to (close to) the optimal solution. Note that for a fixed learning rate, we can only guarantee that gradient descent will eventually stay close to the optimum, but not exactly converging to the optimum, due to the inherent limitation of gradient descent applied on minimizing a function with discontinuous gradient. This limitation can be circumvented by having a variable learning rate that tends to 0.[2]

Theorem 1. *Fix $m \ge 3$ and $R, r > 0$. Assume the distribution p is supported within a ball of radius r centered at the origin. Fix any $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n \in B_R^m$ (the m-dimensional ball of radius R centered at 0) with $\delta_{\min} := \min_{i\neq j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$. Assume we initialize the variables $\boldsymbol{\nu}_1^{(0)}, \ldots, \boldsymbol{\nu}_n^{(0)}$ i.i.d. at random uniformly distributed over B_R^m. We have:*

- *If the learning rate $\alpha^{(t)} = \alpha$ is fixed, then as long as $\alpha \le nr$, gradient descent (7) will eventually stay within a distance α/n from an optimal solution, i.e., there exists a permutation σ over $\{1, \ldots, n\}$ such that the set $\{(i,t) : \|\boldsymbol{\nu}_{\sigma(i)}^{(t)} - \boldsymbol{\mu}_i\| > \alpha/n\}$ is finite, with probability at least*

  $$1 - C_{m,n} r/\delta_{\min},$$

  where $C_{m,n} > 0$ only depends on m, n.

- *If the learning rates $\alpha^{(t)}$ satisfy $\alpha^{(t)} \le nr$ for all t, $\sum_{t=0}^{\infty} \alpha^{(t)} = \infty$, and $\lim_{t\to\infty} \alpha^{(t)} = 0$, then gradient descent (7) will converge to an optimal solution, i.e., there exists a permutation σ such that $\lim_{t\to\infty} \boldsymbol{\nu}_{\sigma(i)}^{(t)} = \boldsymbol{\mu}_i$ for all i, with probability at least $1 - C_{m,n} r/\delta_{\min}$.*

Moreover, if $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$ are i.i.d. uniform over B_R^m as well, then the probabilities in the two cases above can be lower-bounded by $1 - C_{m,n} r/R$ instead.

[1] Three definitions of mode-seeking divergence measures, namely weakly, strongly and uniformly mode-seeking, in increasing order of stringency and strength of guarantees, were studied in (Li and Farnia, 2023).

[2] Regarding the case where the learning rate is fixed, note that for the case $n = 1$, we have $W_1(P, Q) = |\nu_1 - \mu_1|$, and hence gradient descent will give $\nu_1^{(t)}$ that oscillates around μ_1, and will never converge. We can only hope for having the iterations eventually stay within a small distance from the optimal solution. This limitation can be circumvented by having a variable learning rate satisfying $\sum_{t=0}^{\infty} \alpha^{(t)} = \infty$ and $\lim_{t\to\infty} \alpha^{(t)} = 0$.

Proof. We defer the proof to the Appendix. □

As the above result suggests, under random initialization of the centers, a gradient-based optimization of the Wasserstein distance will manage to find all the underlying center points.

We note that Li and Farnia (2023) show that when Wasserstein distance is used to fit a *unimodal* model distribution to a multi-modal data distribution, the global minimum may fail to capture any mode. On the other hand, we show that a *multi-modal* model distribution can be fitted to a multi-modal data distribution using Wasserstein distance minimization via gradient descent, indicating that Wasserstein distance is suitable as long as the model is rich enough to capture the multi-modal structure of the data distribution. Next, we will study if these desired properties also hold for f-GANs minimizing an f-divergence measure.

5 THEORETICAL RESULTS ON MODE-SEEKING f-DIVERGENCE

Here, we show that for fitting a mixture model distribution to a mixture data distribution, if the loss function is given by a strongly mode-seeking divergence measure (Li and Farnia, 2023), then there are local optima that are globally suboptimal. More specifically, if the model distribution is initialized at a subset of the modes of the data distribution (e.g. if the data contains images of red, green and blue objects, but the model is initialized such that it captures only the red and green objects), then any local search algorithm will not be able to find the global optimum, i.e., there is no continuous path from the initial distribution to the optimal distribution where the divergence measure is nonincreasing.

Similar to the problem setup in the previous section, consider the setting where the data distribution $P(\mathbf{x}) = n^{-1}\sum_{i=1}^n p(\mathbf{x}-\boldsymbol{\mu}_i)$ is a mixture of n shifted versions of the distribution p. We will fit the mixture $Q_{\boldsymbol{\nu}_{1:n}}(\mathbf{x}) = n^{-1}\sum_{i=1}^n p(\mathbf{x}-\boldsymbol{\nu}_i)$ to P by minimizing the f-divergence $d_f(P, Q_{\boldsymbol{\nu}_{1:n}})$ via gradient descent (or any other local search algorithms) on the variables $\{\boldsymbol{\nu}_j\}$. For gradient descent with a small learning rate, or a local search algorithm with a small step size, the trajectory of the variables $\{\boldsymbol{\nu}_j\}$ can be regarded as a continuous path where $d_f(P, Q_{\boldsymbol{\nu}_{1:n}})$ is monotonically decreasing.

The following theorem shows that, if d_f is a strongly mode-seeking divergence measure (Li and Farnia, 2023) such as Jensen-Shannon divergence,[3] as long as $\{\boldsymbol{\mu}_i\}$ are well-separated, if we initialize $\{\boldsymbol{\nu}_j\}$ such that each $\boldsymbol{\nu}_j$ belongs to the set $\{\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_n\}$, but there is no one-to-one correspondence between $\{\boldsymbol{\nu}_j\}$ and $\{\boldsymbol{\mu}_i\}$ (i.e., we do not initialize at the optimal solution), then there is no continuous path from $\{\boldsymbol{\nu}_j\}$ to the optimal solution where $d_f(P, Q_{\boldsymbol{\nu}_{1:n}})$ is monotonically decreasing, and hence gradient descent fails to converge to the optimal solution for small learning rate.[4]

Theorem 2. *Suppose $f:\mathbb{R}\to\mathbb{R}$ is strictly convex over $(0,1)$ and satisfies $\lim_{t\to\infty} f(t)/t < \infty$. Also, assume that for the density function p, the super-level set $\{\mathbf{x}\in\mathbb{R}^m : p(\mathbf{x})\ge c\}$ is bounded for all $c>0$. We suppose that initial values of $\{\boldsymbol{\nu}_j\}$ satisfy $\{\boldsymbol{\nu}_1,\dots,\boldsymbol{\nu}_n\}\subsetneq\{\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_n\}$. Then, there exists a constant $C_{f,p,n}$ (which only depends on f,p,n) such that as long as $\min_{i\ne j}\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|\ge C_{f,p,n}$, there does not exist any continuous path $\tilde{\boldsymbol{\nu}}_1(t),\dots,\tilde{\boldsymbol{\nu}}_n(t)$ (where $\tilde{\boldsymbol{\nu}}_j:[0,1]\to\mathbb{R}^m$ is a continuous function) such that $d_f(P, n^{-1}\sum_{j=1}^n p(\mathbf{x}-\tilde{\boldsymbol{\nu}}_j(t)))$ is nonincreasing in t, $\tilde{\boldsymbol{\nu}}_j(0)=\boldsymbol{\nu}_j$ for all j, and $\{\tilde{\boldsymbol{\nu}}_1(1),\dots,\tilde{\boldsymbol{\nu}}_n(1)\}=\{\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_n\}$.*

Proof. We defer the proof to the Appendix. □

The above theorem highlights the challenges of f-GANs minimizing mode-seeking f-divergences in the search for the complete set of modes in an underlying multi-modal distribution. This is in contrast to the behavior of Wasserstein distance in Theorem 1, which implies that when some $\boldsymbol{\nu}_i$'s are coinciding with the $\boldsymbol{\mu}_i$'s, as long as no three $\boldsymbol{\nu}_i$'s are located at the same point, this will not result in a poor local optimum for Wasserstein distance, as shown in the following corollary of Theorem 1.

Corollary 3. *Fix $m\ge 3$, $0\le n'\le n$ and $R,r>0$. Assume the distribution p is supported within a ball of radius r centered at the origin. Assume $\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_n$ and $\boldsymbol{\nu}_1,\dots,\boldsymbol{\nu}_{n'}$ are i.i.d. uniformly distributed over B_R^m. With probability at least $1-C_{m,n}r/R$, where $C_{m,n}>0$ only depends on m,n, for every $\boldsymbol{\nu}_{n'+1},\dots,\boldsymbol{\nu}_n\in\{\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_n\}$ such that no three of $\boldsymbol{\nu}_{n'+1},\dots,\boldsymbol{\nu}_n$ are the same, there exists a continuous path $\tilde{\boldsymbol{\nu}}_1(t),\dots,\tilde{\boldsymbol{\nu}}_n(t)$ (where $\tilde{\boldsymbol{\nu}}_j:[0,1]\to\mathbb{R}^m$ is a continuous function) such that $W_1(P, n^{-1}\sum_{j=1}^n p(\mathbf{x}-\tilde{\boldsymbol{\nu}}_j(t)))$ is nonincreasing in t, $\tilde{\boldsymbol{\nu}}_j(0)=\boldsymbol{\nu}_j$ for all j, and $\{\tilde{\boldsymbol{\nu}}_1(1),\dots,\tilde{\boldsymbol{\nu}}_n(1)\}=\{\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_n\}$.*

[3] Here we require that $f(t)$ is strictly convex for $0<t<1$ and $\lim_{t\to\infty} f(t)/t<\infty$. This is a weaker condition compared to the definition of strongly mode-seeking divergence in (Li and Farnia, 2023).

[4] Li and Farnia (2023) showed that if we fit a *unimodal* model distribution to a multi-modal data distribution via a mode-seeking divergence, the *global minimum* will coincide with one of the modes. Nevertheless, when a *multi-modal* model distribution is fitted to a multi-modal data distribution, Theorem 2 in this paper shows that local search algorithms may fail to find the global optimum.

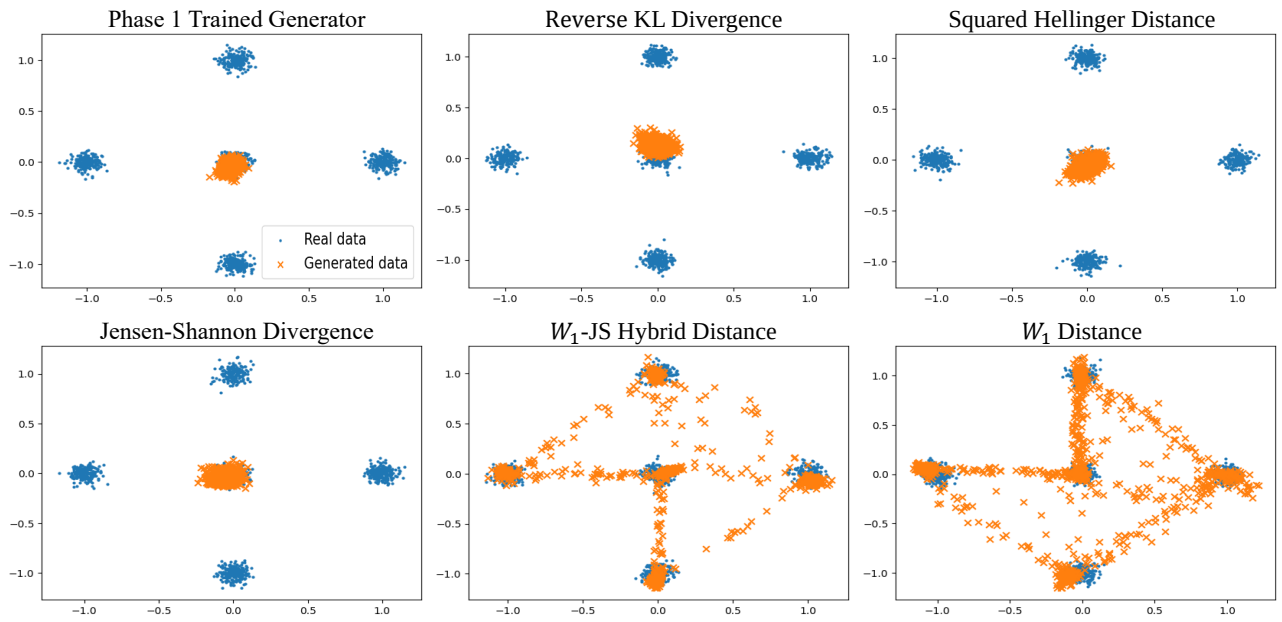

Figure 1: Multi-modal Gaussian mixture data (in blue) and GANs' generated data (in orange). Mode-seeking f-divergences were trapped in unimodal local optima, while the Wasserstein distances covered all the modes.

Proof. We defer the proof to the Appendix. □

Next section, we empirically show how poor local minima could affect the minimization of mode-seeking f-divergences.

6 NUMERICAL RESULTS

In this section, we present the results of our numerical experiments validating the theoretical findings on the local optima and convergence properties of different divergence minimization problems. We performed the numerical experiments on benchmark synthetic Gaussian mixture models and real image datasets including MNIST (LeCun, 1998), CIFAR-10 (Krizhevsky et al., 2009), and CelebA (Liu et al., 2015). To simulate the results of minimizing f-divergence and Wasserstein distances, we considered the min-max optimization in f-GANs (Nowozin et al., 2016) and Wasserstein GAN (Arjovsky et al., 2017) as the dual formulation of the divergence minimization problems. In our experiments, we considered multi-layer perceptions (MLPs) for the discriminator architecture to simulate the space of all functions as needed for the dual formulation of f-divergence measures. We defer the experiments' details to the Appendix.

To validate our theoretical result on the existence of poor locally optimal solutions in mode-seeking f-GANs, we constructed datasets with a pair of clearly separable modes. Then, in Phase 1 of training, we trained the GAN networks on samples coming from the first mode for 50 epochs. Subsequently, in Phase 2 we initialized the GAN networks using their weights at the end of Phase 1 and trained them for another 50 epochs using training data from all of the modes with equal probabilities. We monitored if the generator network in Phase 2 would manage to escape the initial generative model that only captures one of the modes and would be able to generate samples from all the modes by the end of Phase 2.

In the case of Gaussian mixtures, we defined a symmetric mixture of five Gaussians with opposite means at $[0, 0]$, $[\pm 1, 0]$, and $[0, \pm 1]$ as illustrated in Figure 1. In Phase 1 of our GAN training, all the GAN formulations managed to converge to the introduced mode as depicted in the upper-left figure of Figure 1. In the second training phase, we used training data sampled from both of the modes with equal frequencies. The f-GANs formulated by the mode-seeking divergence measures, including Reverse-KL, squared Hellinger distance, and JS-divergence as suggested by (Li and Farnia, 2023) were all trapped around the initial generator capturing only the mode centered at $[0, 0]$. In contrast, the Wasserstein GAN and the hybrid-divergence-based SN-GAN (Miyato et al., 2018) could escape the mode collapsed generator and outputted samples covering all five modes by the end of Phase 2.

For image datasets, due to the complex multi-modal distribution of hidden variables across samples, we attempted to create two well-separated modes with sufficient dissimilarity. Therefore, we conducted experi-

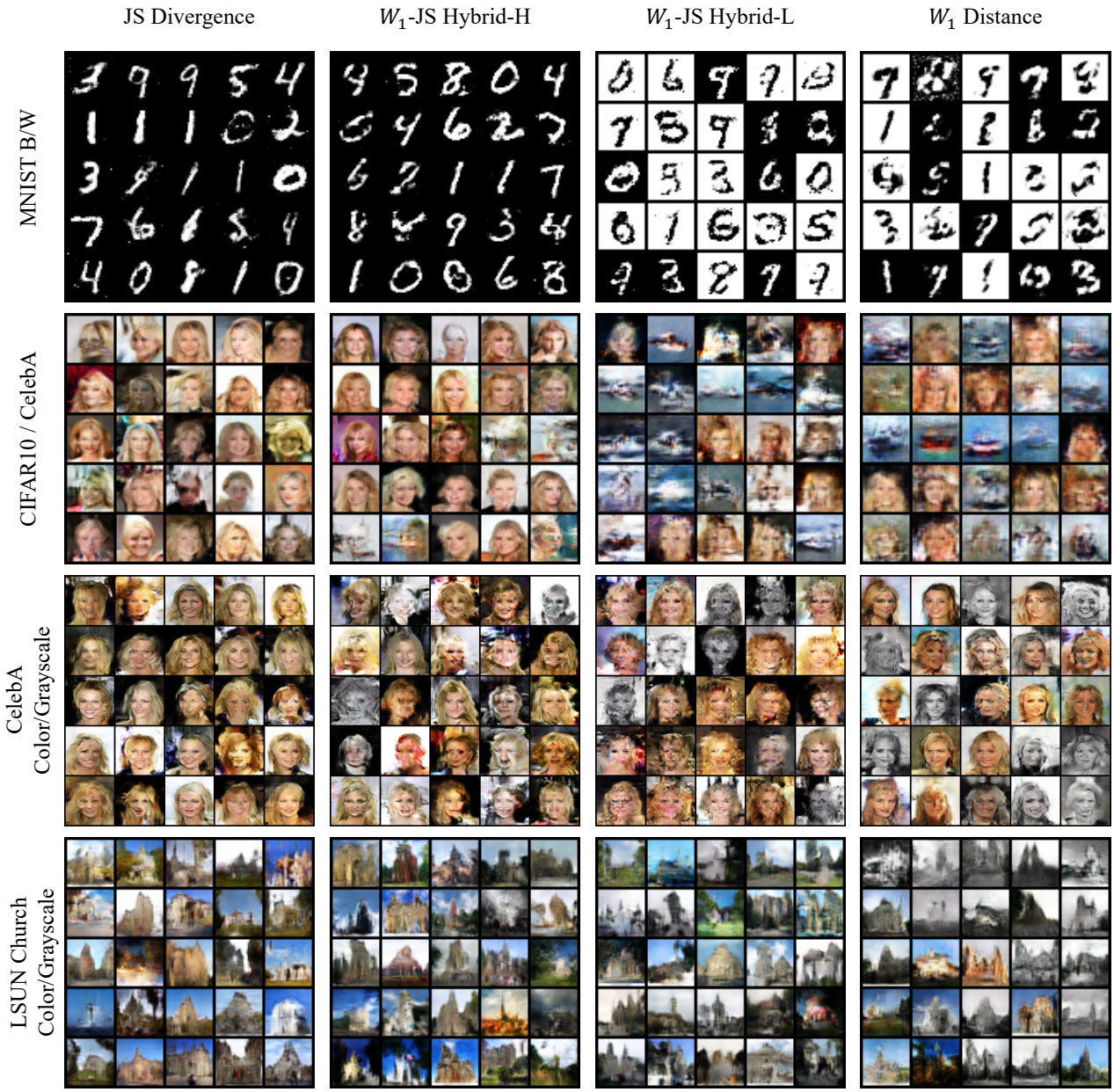

Figure 2: Qualitative results of GAN-based divergence minimization on bimodal image data initialized at a unimodal point, GANs targeting W_1 and W_1-JS-hybrid distances could generate samples from both modes, while JS-divergence led to a generator trapped at a unimodal point.

ments on the following three bimodal image datasets, each containing two distinguishable modes. In every case, we trained a CNN classifier to distinguish samples from the two modes of the bimodal dataset in order to estimate the number of generated data belonging to each mode:

1. **MNIST B/W**: We used one-half (25,000 samples) of the MNIST training data as the first mode and constructed the second mode by flipping the other half of the MNIST samples that had white backgrounds and black digits post-flipping.
2. **CIFAR10/CelebA**: The two identifiable modes are the 5000 ship samples from CIFAR-10 and 5000 randomly selected CelebA samples.
3. **CelebA Color/Grayscale**: The dataset contains half of the original CelebA samples as well as the grayscale version of the other half of CelebA data.
4. **LSUN Church Color/Grayscale**: The dataset contains half of the original LSUN Church samples as well as the grayscale version of the other half of LSUN Church data.

As shown in Figure 2, after training the GANs over Phases 1 and 2, we observed that the vanilla GAN minimizing the mode-seeking JS-divergence was trapped around the suboptimal initial point missing the second mode entirely. In contrast, minimizing the 1-Wasserstein distance in WGAN and its hybrid with JS-divergence in SN-GAN resulted in escaping from the unimodal initial point. The generated samples of WGAN and SN-GAN suggest their capability in discovering both modes. Additionally, we analyzed two cases of the W_1-JS hybrid distance with different Lipschitz constants. In the reported plots, we use the letters "H" and "L" to indicate the higher and lower Lipschitz coefficients, respectively. The Lipschitz constant was controlled by adjusting the spectral normalization coefficient. Precise Lipschitz constants are reported in Figure 3 by finding the maximum ℓ_2-norm of the discriminator's gradient over training data. We note that the W_1-JS hybrid-H is expected to behave similarly to JS-divergence, while the W_1-JS hybrid-L is expected to be similar to W_1-distance.

Moreover, we evaluated the mode ratio statistic, as the ratio of samples generated from the second mode introduced in Phase 2. We used a CNN classifier for the estimation of the mode ratio. Figure 3 plots the mode ratio over Phase 2 of training in the case of real image experiments. Note that VGAN targetting mode-seeking JS-divergence continued to generate all its images from the original mode, while Wasserstein GAN and SN-GAN could successfully capture both the modes. We also note that that the W_1-JS hybrid-H showed weaker mode-seeking property compared to the W_1-JS hybrid-L as the W_1-JS hybrid-H reached a lower mode ratio statistic.

We also analyzed the convergence of divergence measures over the two training phases. Figure 4 displays the convergence behavior of each divergence measure in the experiments for the case of CelebA Color/Grayscale. We highlight Phases 1 and 2 with the blue and orange curves, respectively. In the plots, a plateau can be observed towards the latter stages of the initial training phase in the divergence curves, indicating convergence to a local optimum at the end of Phase 1. Subsequently in Phase 2, the JS-divergence minimized by the VGAN remained almost constant suggesting a unimodal local optimum in the divergence minimziation, while the remaining GANs targeting the W_1-JS hybrid and W_1 distances could converge toward a bimodal solution.

Overall, the experimental results in this section support our theoretical findings on the superior performance of Wasserstein GANs in avoiding poor local optima missing modes of an underlying multimodal distribution. It is also noteworthy that minimizing JS-divergence in VGAN resulted in image samples with visually higher quality, while the Wasserstein GAN often generated noisy images which seem to combine the two modes in a visually imperfect way. This numerical observation could hint at a trade-off between diversity and quality in targeting different divergence measures.

7 CONCLUSION

In this paper, we studied the role of divergence measures in the local optima of divergence minimization problems. We especially focused on multi-modal underlying distributions learned by minimizing a mode-seeking f-divergence, where we theoretically and numerically demonstrated the existence of suboptimal local optima that miss one or several modes. On the other hand, we showed that the Wasserstein distance will not lead to such poor locally optimal solutions, and a gradient-based method could solve the distance minimization problem to discover the center points of the existing modes. We also numerically observed that the hybrid of Wasssretin distance and JS-divergence seem to address the suboptimal local optima in GAN training problems. A future direction for our work is to theoretically analyze the local minima of the minimization of hybrid divergences. Another interesting topic for future exploration is to design regularization methods for avoiding poor local optima of mode-seeking divergence measures. Such regularization schemes would be useful to improve mode diversity in GAN models.

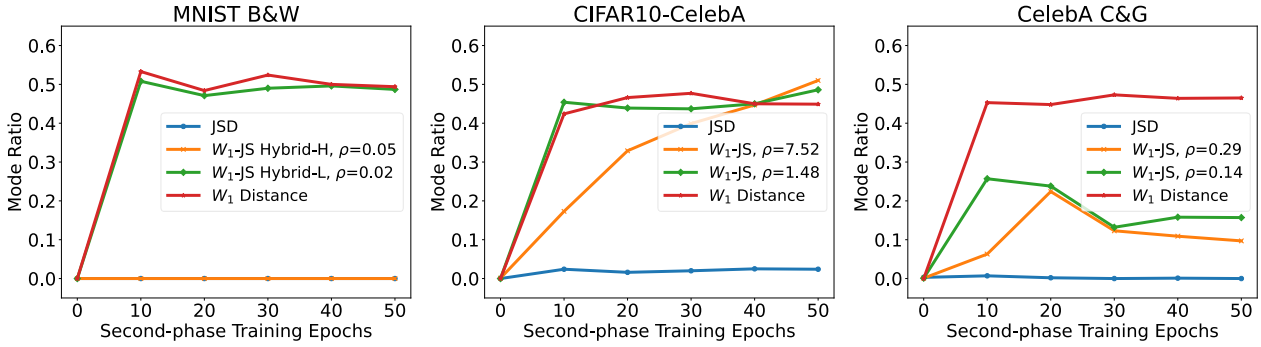

Figure 3: Mode ratio during Phase 2 of GAN training over bimodal image data, An α mode ratio indicates an α-fraction of generated images from the second mode. GANs with W_1 and JS-W_1 hybrid distances led to balanced modes, while VGAN (JS-divergence) led to a unimodal point.

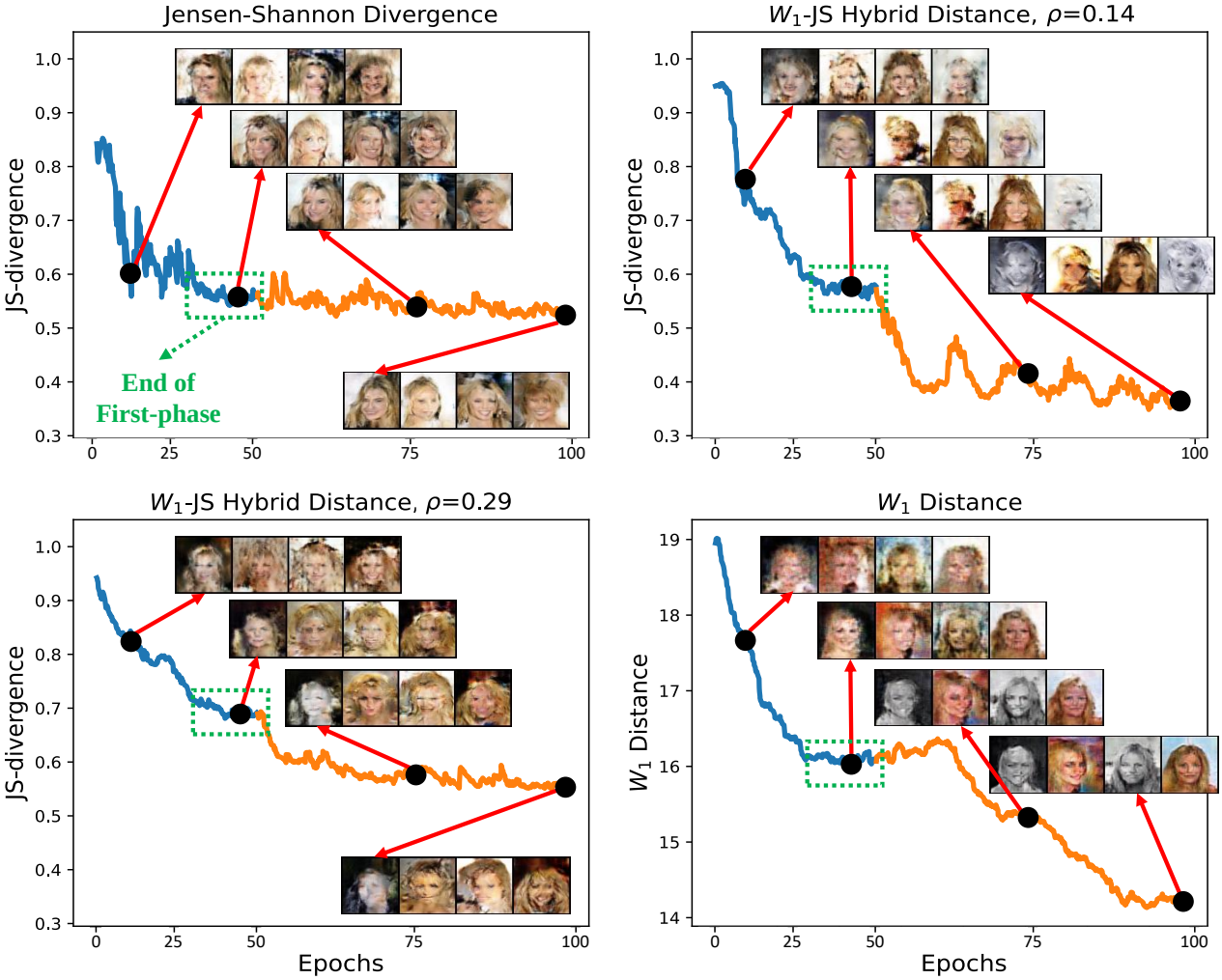

Figure 4: Divergence optimization progress in the CelebA Color&Grayscale bimodal dataset. VGAN minimizing the JS-divergence failed to escape the suboptimal unimodal local optimum, while GANs targeting W_1 distance converged to the bimodal optimum.

Acknowledgements

The work of Cheuk Ting Li was partially supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No.s: CUHK 24205621 (ECS), CUHK 14209823 (GRF)]. The work of Farzan Farnia is partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, Project 14209920, and is partially supported by a CUHK Direct Research Grant with CUHK Project No. 4055164. Also, the authors would like to thank the anonymous reviewers for their constructive feedback and suggestions.

References

Arjovsky, M. and Bottou, L. (2017). Towards principled methods for training generative adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France*.

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

Bishop, C. M. (2006). Pattern recognition and machine learning.

Brock, A., Donahue, J., and Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Csiszár, I. and Shields, P. C. (2004). Information theory and statistics: A tutorial.

Donahue, C., Li, B., and Prabhavalkar, R. (2018). Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5024–5028. IEEE.

Farnia, F. and Ozdaglar, A. (2020). Do gans always have nash equilibria? In *International Conference on Machine Learning*, pages 3029–3039. PMLR.

Farnia, F. and Tse, D. (2018). A convex duality framework for GANs. *Advances in Neural Information Processing Systems*, 31:5248–5258.

Farnia, F., Wang, W. W., Das, S., and Jadbabaie, A. (2023). Gat–gmm: Generative adversarial training for gaussian mixture models. *SIAM Journal on Mathematics of Data Science*, 5(1):122–146.

Feizi, S., Farnia, F., Ginart, T., and Tse, D. (2020). Understanding gans in the lqg setting: Formulation, generalization and stability. *IEEE Journal on Selected Areas in Information Theory*, 1(1):304–311.

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

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017). Improved training of Wasserstein GANs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5769–5779, Red Hook, NY, USA. Curran Associates Inc.

Gupta, A. and Zou, J. (2019). Feedback GAN for DNA optimizes protein functions. *Nature Machine Intelligence*, 1(2):105–111.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in neural information processing systems*, 30.

Huszár, F. (2015). How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*.

Kodali, N., Abernethy, J., Hays, J., and Kira, Z. (2017). On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*.

Kolouri, S., Rohde, G. K., and Hoffmann, H. (2018). Sliced wasserstein distance for learning gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3427–3436.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. (2019). Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32.

LeCun, Y. (1998). The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Li, C. T. and Farnia, F. (2023). Mode-seeking divergences: Theory and applications to GANs. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8321–8350. PMLR.

Liu, S., Bousquet, O., and Chaudhuri, K. (2017). Approximation and convergence properties of generative adversarial learning. *Advances in Neural Information Processing Systems*, 30.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lucas, T., Shmelkov, K., Alahari, K., Schmid, C., and Verbeek, J. (2019). Adaptive density estimation for generative models. *Advances in Neural Information Processing Systems*, 32:12016–12026.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. (2018). Are GANs created equal? a large-scale study. *Advances in neural information processing systems*, 31.

Mescheder, L., Geiger, A., and Nowozin, S. (2018). Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

Nadjahi, K., Durmus, A., Simsekli, U., and Badeau, R. (2019). Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. *Advances in Neural Information Processing Systems*, 32.

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. (2020). Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR.

Nagarajan, V. and Kolter, J. Z. (2017). Gradient descent gan optimization is locally stable. *Advances in neural information processing systems*, 30.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861.

Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-GAN: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 271–279.

Poole, B., Alemi, A. A., Sohl-Dickstein, J., and Angelova, A. (2016). Improved generator objectives for GANs. *arXiv preprint arXiv:1612.02780*.

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. (2018). Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31.

Sanjabi, M., Ba, J., Razaviyayn, M., and Lee, J. D. (2018). On the convergence and robustness of training gans with regularized optimal transport. *Advances in Neural Information Processing Systems*, 31.

Shannon, M., Poole, B., Mariooryad, S., Bagby, T., Battenberg, E., Kao, D., Stanton, D., and Skerry-Ryan, R. (2020). Non-saturating GAN training as divergence minimization. *arXiv preprint arXiv:2010.08029*.

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al. (2021). Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272.

Villani, C. (2003). *Topics in optimal transportation*, volume 58. American Mathematical Soc.

Wendel, J. G. (1948). Note on the Gamma function. *American Mathematical Monthly*, 55:563.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes, the Python codes specifying imported libraries are in the supplementary materials]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes, in the supplemental material]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, they are listed in Section 6 and Section B.1]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes, JS-divergence and Wasserstein distance are defined in Section 3, the measure of Fig.3 is defined in its caption.]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes, in section B.1]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes, we have used public datasets and network architectures proposed by others, we have cited the origins when we mention them in the paper]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes, our codes are provided in the supplemental material]

   (d) Information about consent from data providers/curators. [Yes, all asset owners allow usage of their datasets/network implementations for academic research purpose]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Proofs

A.1 Proof of Theorem 1

Consider a coupling γ between P and Q, i.e., if $(\mathbf{x},\mathbf{y})\sim\gamma$, then the marginal distributions are $\mathbf{x}\sim P$ and $\mathbf{y}\sim Q$. Since $P(\mathbf{x})=n^{-1}\sum_{i=1}^n p(\mathbf{x}-\boldsymbol{\mu}_i)$, we can let $I\sim\mathrm{Unif}\{1,\dots,n\}$ be a random variable such that $\mathbf{x}-\boldsymbol{\mu}_I\sim p$ is independent of I. Equivalently, conditional on $I=i$, $\mathbf{x}$ is distributed according to p shifted by $\boldsymbol{\mu}_i$. Similarly, let $J\sim\mathrm{Unif}\{1,\dots,n\}$ such that $\mathbf{y}-\boldsymbol{\nu}_J\sim p$ is independent of J. We call $\mathbf{G}\in\mathbb{R}^{n\times n}$ an *assignment matrix* corresponding to γ if there exists I,J satisfying the aforementioned conditions such that $G_{i,j}=\mathbb{P}(J=j\,|\,I=i)$. Note that $\mathbf{G}$ is doubly stochastic. We call a coupling γ *pure* if it has a unique assignment matrix that is a permutation matrix.

Consider a permutation σ over $\{1,\dots,n\}$. Its *cost* is defined as

$$c_\sigma=c_\sigma(\{\boldsymbol{\mu}_i\},\{\boldsymbol{\nu}_j\}):=\frac{1}{n}\sum_{i=1}^n\|\boldsymbol{\nu}_{\sigma(i)}-\boldsymbol{\mu}_i\|.$$

We consider the optimal $\sigma^*=\arg\min_\sigma c_\sigma$. Define the *second optimal gap* as

$$\min_{\sigma\neq\sigma^*}c_\sigma-c_{\sigma^*},$$

which measures how close the optimal c_{σ^*} is from the second optimal.

Lemma 4. *Fix $\rho\geq 1$. If the second optimal gap is greater than $4r$, then the coupling γ^* attaining the optimum in $W_\rho(P,Q)$ is pure, and its assignment matrix is the permutation matrix of σ^*.*

Proof. Assume the contrary that γ^* is impure. Then we can find an assignment matrix $\mathbf{G}$ that is not a permutation matrix (if assignment matrices are not unique, then any convex combination of them must also be an assignment matrix, so we can find an assignment matrix that is not a permutation matrix). Let $(\mathbf{x},\mathbf{y})\sim\gamma^*$, and I,J be the random variables corresponding to $\mathbf{G}$. By Birkhoff-von Neumann theorem, assume $\mathbf{G}=\sum_\sigma w_\sigma\mathbf{P}_\sigma$, where the summation is over all permutations over $\{1,\dots,n\}$, $\mathbf{P}_\sigma$ is the permutation matrix of σ, and $w_\sigma\geq 0$ with $\sum_\sigma w_\sigma=1$. Since $\mathbf{G}$ is not a permutation matrix, there exists $\sigma\neq\sigma^*$ with $w_\sigma>0$. Let S be a random permutation with $\mathbb{P}(S=\sigma)=w_\sigma$, and $(I,J)|\{S=\sigma\}\sim\mathrm{Unif}(\{(i,\sigma(i)):i=1,\dots,n\})$. Define a new coupling $(\mathbf{x}',\mathbf{y}')\sim\gamma'$ as follows. If $S=\sigma^*$, then $(\mathbf{x}',\mathbf{y}')=(\mathbf{x},\mathbf{y})$. If $S=\sigma$, $\sigma\neq\sigma^*$, take $\mathbf{x}'=\mathbf{x}$, and generate $\mathbf{y}'$ following $\mathbf{y}'|\{S=\sigma,I=i\}\sim\mathbb{P}_{\mathbf{y}|S=\sigma,\,J=\sigma^*(i)}$ (i.e., the conditional distribution of $\mathbf{y}$ conditional on $S=\sigma$, $J=\sigma^*(i)$). We have

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}-\mathbf{y}\|\right]-\mathbb{E}\left[\|\mathbf{x}'-\mathbf{y}'\|\right]\\
&=\sum_{\sigma\neq\sigma^*}w_\sigma\mathbb{E}\left[\|\mathbf{x}-\mathbf{y}\|-\|\mathbf{x}-\mathbf{y}'\|\,\middle|\,S=\sigma\right]\\
&\overset{(a)}{\geq}\sum_{\sigma\neq\sigma^*}w_\sigma\mathbb{E}\left[\|\boldsymbol{\mu}_I-\boldsymbol{\nu}_{\sigma(I)}\|-\|\boldsymbol{\mu}_I-\boldsymbol{\nu}_{\sigma^*(I)}\|-4r\,\middle|\,S=\sigma\right]\\
&=\sum_{\sigma\neq\sigma^*}w_\sigma\left(c_\sigma-c_{\sigma^*}-4r\right)\\
&>0,
\end{aligned}$$

where (a) is because $\|\mathbf{x}-\boldsymbol{\mu}_I\|$, $\|\mathbf{y}-\boldsymbol{\nu}_{\sigma(I)}\|$, $\|\mathbf{y}'-\boldsymbol{\nu}_{\sigma^*(I)}\|\leq r$. Hence γ^* cannot be optimal, which leads to a contradiction. □

If the second optimal gap is large enough, not only will the optimal assignment be given by σ^*, but also gradient descent will eventually stay within a close distance to the optimal solution.

Lemma 5. *Let $\epsilon>0$. Assume $\alpha^{(t)}\leq nr$, $\sum_{t=0}^\infty\alpha^{(t)}=\infty$, $|\{t:\alpha^{(t)}>\epsilon\}|<\infty$. If the second optimal gap is greater than $6r$, then gradient descent will give $|\{(i,t):\|\boldsymbol{\nu}_{\sigma^*(i)}^{(t)}-\boldsymbol{\mu}_i\|>\epsilon/n\}|<\infty$.*

Proof. We will prove by induction that the iterations of gradient descent is given by

$$\boldsymbol{\nu}_{\sigma^*(i)}^{(t+1)} = \boldsymbol{\nu}_{\sigma^*(i)}^{(t)} + \frac{\alpha^{(t)}}{n} \cdot \frac{\boldsymbol{\mu}_i - \boldsymbol{\nu}_{\sigma^*(i)}^{(t)}}{\|\boldsymbol{\mu}_i - \boldsymbol{\nu}_{\sigma^*(i)}^{(t)}\|}. \tag{8}$$

This implies that $\boldsymbol{\nu}_{\sigma^*(i)}^{(t)}$ always lies on the line connecting $\boldsymbol{\nu}_{\sigma^*(i)}^{(0)}$ and $\boldsymbol{\mu}_i$, and at each iteration t, $\boldsymbol{\nu}_{\sigma^*(i)}^{(t+1)}$ moves towards $\boldsymbol{\mu}_i$ by an amount $\alpha^{(t)}/n$. Since $\sum_{t=0}^{\infty} \alpha^{(t)} = \infty$, $\boldsymbol{\nu}_{\sigma^*(i)}^{(t)}$ will eventually move past $\boldsymbol{\mu}_i$ and then move back and forth around $\boldsymbol{\mu}_i$. Since $|\{t : \alpha^{(t)} > \epsilon\}| < \infty$, we have $\alpha^{(t)} \le \epsilon$ for all large enough t, and hence $\boldsymbol{\nu}_{\sigma^*(i)}^{(t)}$ will stay within a distance ϵ/n from $\boldsymbol{\mu}_i$ for large enough t.

We now prove (8) by induction. First, if at iteration t, the second optimal gap is greater than $4r$ at $\{\boldsymbol{\nu}_j^{(t)}\}$, then the gap is still greater than $4r$ within a neighborhood of $\{\boldsymbol{\nu}_j^{(t)}\}$ since c_σ is a continuous function of $\{\boldsymbol{\nu}_j\}$. Hence by Lemma 4, the assignment matrix is the permutation matrix of σ^* within a neighborhood of $\{\boldsymbol{\nu}_j^{(t)}\}$, and we can assume the assignment matrix is fixed in order to compute the gradient at the point $\{\boldsymbol{\nu}_j^{(t)}\}$. Within the neighborhood, we have $W_1(P, Q) = c_{\sigma^*}(\{\boldsymbol{\mu}_i\}, \{\boldsymbol{\nu}_j\})$, and hence the gradient is

$$\nabla_{\boldsymbol{\nu}_j} c_{\sigma^*}(\{\boldsymbol{\mu}_i\}, \{\boldsymbol{\nu}_j\}) = \frac{1}{n} \cdot \frac{\boldsymbol{\nu}_j - \boldsymbol{\mu}_{(\sigma^*)^{-1}(j)}}{\|\boldsymbol{\nu}_j - \boldsymbol{\mu}_{(\sigma^*)^{-1}(j)}\|},$$

and the next iteration is given by (8). We then prove that the second optimal gap is still greater than $4r$ at $\{\boldsymbol{\nu}_j^{(t+1)}\}$. From (8), since $\boldsymbol{\nu}_{\sigma^*(i)}^{(t+1)}$ lies on the line (not line segment) connecting $\boldsymbol{\nu}_{\sigma^*(i)}^{(0)}$ and $\boldsymbol{\mu}_i$, and its distance to the line segment connecting $\boldsymbol{\nu}_{\sigma^*(i)}^{(0)}$ and $\boldsymbol{\mu}_i$ is at most $\max_t \alpha^{(t)}/n \le r$, we have

$$\|\boldsymbol{\nu}_{\sigma^*(i)}^{(t+1)} - \boldsymbol{\mu}_i\| + \|\boldsymbol{\nu}_{\sigma^*(i)}^{(t+1)} - \boldsymbol{\nu}_{\sigma^*(i)}^{(0)}\| \le \|\boldsymbol{\nu}_{\sigma^*(i)}^{(0)} - \boldsymbol{\mu}_i\| + 2r.$$

Assume the contrary that there exists $\sigma \neq \sigma^*$ with $c_\sigma(\{\boldsymbol{\mu}_i\}, \{\boldsymbol{\nu}_j^{(t+1)}\}) - c_{\sigma^*}(\{\boldsymbol{\mu}_i\}, \{\boldsymbol{\nu}_j^{(t+1)}\}) \le 4r$. We have

$$\begin{aligned}
&c_\sigma(\{\boldsymbol{\mu}_i\}, \{\boldsymbol{\nu}_j^{(0)}\}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{\nu}_{\sigma(i)}^{(0)} - \boldsymbol{\mu}_i\| \\
&\le \frac{1}{n} \sum_{i=1}^{n} \left(\|\boldsymbol{\nu}_{\sigma(i)}^{(t+1)} - \boldsymbol{\mu}_i\| + \|\boldsymbol{\nu}_{\sigma(i)}^{(t+1)} - \boldsymbol{\nu}_{\sigma(i)}^{(0)}\| \right) \\
&\le \frac{1}{n} \sum_{i=1}^{n} \left(\|\boldsymbol{\nu}_{\sigma^*(i)}^{(t+1)} - \boldsymbol{\mu}_i\| + \|\boldsymbol{\nu}_{\sigma(i)}^{(t+1)} - \boldsymbol{\nu}_{\sigma(i)}^{(0)}\| \right) + 4r \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\|\boldsymbol{\nu}_{\sigma^*(i)}^{(t+1)} - \boldsymbol{\mu}_i\| + \|\boldsymbol{\nu}_{\sigma^*(i)}^{(t+1)} - \boldsymbol{\nu}_{\sigma^*(i)}^{(0)}\| \right) + 4r \\
&\le \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{\nu}_{\sigma^*(i)}^{(0)} - \boldsymbol{\mu}_i\| + 6r \\
&= c_{\sigma^*}(\{\boldsymbol{\mu}_i\}, \{\boldsymbol{\nu}_j^{(0)}\}) + 6r,
\end{aligned}$$

violating the assumption that the second optimal gap is greater than $6r$. □

The final step is to show that the second optimal gap is greater than $6r$ with high probability.

Lemma 6. *Let $g > 0$. Generate $\{\boldsymbol{\nu}_i\}$ i.i.d. at random according to $\boldsymbol{\nu}_i \sim \mathrm{Unif}(B_R^m)$. With probability at least*

$$1 - 2^{2m-1} \sqrt{m} (n!)^2 g \delta_{\min}^{-1},$$

the second optimal gap is greater than g.

Proof. Fix any pair of permutations $\sigma_1 \neq \sigma_2$. We have

$$c_{\sigma_2} - c_{\sigma_1} = \sum_{j=1}^{n} \left(\|\boldsymbol{\nu}_j - \boldsymbol{\mu}_{\sigma_2^{-1}(j)}\| - \|\boldsymbol{\nu}_j - \boldsymbol{\mu}_{\sigma_1^{-1}(j)}\| \right). \tag{9}$$

Our goal is to bound the probability density function of $c_{\sigma_2} - c_{\sigma_1}$. Let

$$\mathbf{x}(\theta) := [\cos\theta, \sin\theta],$$

$$\begin{aligned} h_a(\theta) &:= \|a\mathbf{x}(\theta) - [1,0]\| - \|a\mathbf{x}(\theta) + [1,0]\| \\ &= \left(a^2 + 1 - 2a\cos\theta\right)^{1/2} - \left(a^2 + 1 + 2a\cos\theta\right)^{1/2}. \end{aligned}$$

Note that $h_a(\theta)$ is strictly increasing in $0 \leq \theta \leq \pi$ if $a > 0$, and hence its inverse h_a^{-1} exists. If $0 \leq \theta \leq \pi$, the derivative of h_α is

$$\begin{aligned} h_a'(\theta) &= a(\sin\theta)\Big(\left(a^2 + 1 - 2a\cos\theta\right)^{-1/2} \\ &\qquad + \left(a^2 + 1 + 2a\cos\theta\right)^{-1/2} \Big) \\ &\geq a(\sin\theta)\left(\left(a^2 + 1 + 2a\right)^{-1/2} + \left(a^2 + 1 + 2a\right)^{-1/2} \right) \\ &= \frac{2a}{a+1}(\sin\theta) \end{aligned}$$

Let $\mathbf{x}$ be uniformly distributed over $S_a^{m-1} \subseteq \mathbb{R}^m$, the $(m-1)$-dimensional sphere with radius a. Let $H := \|\mathbf{x} - [1,0,\ldots,0]\| - \|\mathbf{x} + [1,0,\ldots,0]\|$. We are interested in upper bounding the probability density function f_H of H. Let Θ be the angle between the ray from the origin to $\mathbf{x}$ and the ray from the origin to $[1,0,\ldots,0]$. We have $H = h_a(\Theta)$. Hence,

$$\begin{aligned} f_H(h) &= \frac{\mathrm{d}}{\mathrm{d}h}\mathbb{P}\left(H \leq h\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}h}\frac{1}{\frac{2\pi^{m/2}}{\Gamma(m/2)}}\int_0^{h_a^{-1}(h)} \frac{2\pi^{(m-1)/2}}{\Gamma((m-1)/2)}(\sin\theta)^{m-2}\mathrm{d}\theta \\ &= \frac{\mathrm{d}}{\mathrm{d}h}\frac{\Gamma(m/2)}{\pi^{1/2}\Gamma((m-1)/2)}\int_0^{h_a^{-1}(h)} (\sin\theta)^{m-2}\mathrm{d}\theta \\ &= \frac{\Gamma(m/2)}{\pi^{1/2}\Gamma((m-1)/2)} \cdot \frac{(\sin h_a^{-1}(h))^{m-2}}{h_a'(h_a^{-1}(h))} \\ &\leq \frac{\Gamma(m/2)}{\pi^{1/2}\Gamma((m-1)/2)} \cdot \frac{(\sin h_a^{-1}(h))^{m-2}}{2a(a+1)^{-1}(\sin h_a^{-1}(h))} \\ &= \frac{\Gamma(m/2)}{\pi^{1/2}\Gamma((m-1)/2)} \cdot \frac{(\sin h_a^{-1}(h))^{m-3}}{2a(a+1)^{-1}} \\ &\leq \frac{\Gamma(m/2)(a+1)}{2\pi^{1/2}\Gamma((m-1)/2)a} \\ &\leq \frac{\sqrt{m/2}(a+1)}{2\pi^{1/2}a}, \end{aligned}$$

where the last inequality is by Wendel's inequality (Wendel, 1948). Now assume $\mathbf{x}$ is uniformly distributed over $B_a^m \subseteq \mathbb{R}^m$, the m-dimensional ball with radius $a \geq 1$. We have

$$f_H(h)$$

$$
\begin{aligned}
&\leq \frac{1}{\frac{\pi^{m/2}a^m}{\Gamma(m/2+1)}} \int_0^a \frac{\sqrt{m/2}(t+1)}{2\pi^{1/2}t} \cdot \frac{2\pi^{m/2}}{\Gamma(m/2)} t^{m-1} \mathrm{d}t \\
&= \frac{m}{a^m} \int_0^a \frac{\sqrt{m/2}(t+1)t^{m-2}}{2\pi^{1/2}} \mathrm{d}t \\
&\leq \frac{m\sqrt{m/2}}{2\pi^{1/2}} \cdot \frac{1}{a^m} \int_0^{a+1} t^{m-1} \mathrm{d}t \\
&= \frac{m\sqrt{m/2}}{2\pi^{1/2}} \cdot \frac{(a+1)^m}{da^m} \\
&\leq 2^{m-2}\sqrt{m}.
\end{aligned}
$$

Now we consider the distribution of $H = \|\boldsymbol{\nu} - \boldsymbol{\mu}_2\| - \|\boldsymbol{\nu} - \boldsymbol{\mu}_1\|$ where $\boldsymbol{\nu}$ is uniformly distributed over B_R^m. By an appropriate shifting and rotation, this is the same as the distribution of $2^{-1}\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|(\|\boldsymbol{\nu} - [1, 0, \ldots, 0]\| - \|\boldsymbol{\nu} + [1, 0, \ldots, 0]\|)$ where $\boldsymbol{\nu}$ is uniformly distributed over $B^m_{2R/\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|}(\tilde{\boldsymbol{\mu}})$, the ball with radius $2R/\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|$ centered at a point $\tilde{\boldsymbol{\mu}}$ with $\|\tilde{\boldsymbol{\mu}}\| = \|\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\|/\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|$. Since

$$
B^m_{2R/\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|}(\tilde{\boldsymbol{\mu}}) \subseteq B^m_{(2R + \|\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\|)/\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|},
$$

we have

$$
\begin{aligned}
&f_H(h) \\
&\leq \frac{1}{2^{-1}\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|} \left(\frac{2R + \|\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\|}{2R}\right)^m 2^{m-2}\sqrt{m} \\
&\leq \frac{2^{2m-1}\sqrt{m}}{\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|}.
\end{aligned}
$$

Consider $c_{\sigma_2} - c_{\sigma_1}$ in (9), which is the sum of at least one i.i.d. random variables in the same form as H above. Since $\sup_h f_{H_1+H_2}(h) \leq \sup_h f_{H_1}(h)$ for independent H_1, H_2, the probability density function of $c_{\sigma_2} - c_{\sigma_1}$ is bounded as

$$
f_{c_{\sigma_2} - c_{\sigma_1}}(h) \leq \frac{2^{2m-1}\sqrt{m}}{\delta_{\min}}.
$$

Note that the probability that $\sigma^* = \arg\min_\sigma c_\sigma$ does not satisfy $c_\sigma > c_{\sigma^*} + g$ for all $\sigma \neq \sigma^*$ is upper bounded by the probability that there exists $\sigma_1 \neq \sigma_2$ with $c_{\sigma_2} - c_{\sigma_1} \in [0, g]$. By union bound, this probability is upper bounded by

$$
2^{2m-1}\sqrt{m}(n!)^2 g \delta_{\min}^{-1}.
$$

□

The theorem follows from Lemma 5, Lemma 6 (on $g = 6r$), and taking $C_{m,n} = 6 \cdot 2^{2m-1}\sqrt{m}(n!)^2$.

Finally, we study the case where $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_n$ are i.i.d. uniform over B_R^m. In this case, the probability of failure is upper-bounded by $\mathbb{E}\left[C_{m,n} r/\delta_{\min}\right]$, where $\delta_{\min} := \min_{1 \leq i < j \leq n} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$ is random. We have

$$
\begin{aligned}
\mathbb{E}\left[1/\delta_{\min}\right] &= \int_0^\infty \mathbb{P}\left(1/\delta_{\min} \geq t\right) \mathrm{d}t \\
&\leq \sum_{1 \leq i < j \leq n} \int_0^\infty \mathbb{P}\left(\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| \leq 1/t\right) \mathrm{d}t \\
&\leq \frac{n^2}{2} \int_0^\infty \mathbb{P}\left(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| \leq 1/t\right) \mathrm{d}t \\
&\leq \frac{n^2}{2} \int_0^\infty \min\left\{\frac{|B^m_{1/t}|}{|B^m_R|}, 1\right\} \mathrm{d}t \\
&= \frac{n^2}{2} \int_0^\infty \min\left\{(tR)^{-m}, 1\right\} \mathrm{d}t
\end{aligned}
$$

$$= \frac{n^2}{2} \cdot \frac{m}{R(m-1)}.$$

Hence the probability of failure is upper-bounded by $C'_{m,n} r/R$, where $C'_{m,n} := \frac{C_{m,n} n^2 m}{2(m-1)}$.

A.2 Proof of Theorem 2

Without loss of generality, assume f is convex and nonincreasing with $\lim_{t\to\infty} f(t)/t = 0$ (we can make this assumption since adding $\alpha(x-1)$ to $f(t)$ for any $\alpha \in \mathbb{R}$ does not change d_f). Let $a_i := |\{j \in \{1,\dots,n\} : \boldsymbol{\nu}_j = \boldsymbol{\mu}_i\}|$. Let $0 < \epsilon < 1$. Write $L_c^+(p) := \{\mathbf{x} \in \mathbb{R}^m : p(\mathbf{x}) \ge c\}$ for the superlevel set of p. Let $c > 0$ such that $\int_{L_c^+(p)} p(\mathbf{x})\mathrm{d}\mathbf{x} \ge 1-\epsilon$. Let $\delta_0 > 0$ such that $L_{\epsilon c}^+(p) \subseteq B_{\delta_0/2}(0)$. Assume $\min_{i\ne j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| \ge \delta_0$. Write $S_i := L_c^+(p) + \boldsymbol{\mu}_i$. We have

$$\begin{aligned}
&d_f(P,Q)\\
&\le \frac{1}{n}\sum_{i=1}^n a_i \left(\int_{S_i} f\left(\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right) p(\mathbf{x}-\boldsymbol{\mu}_i)\mathrm{d}\mathbf{x} + \int_{\mathbb{R}^m\setminus S_i} f\left(\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right) p(\mathbf{x}-\boldsymbol{\mu}_i)\mathrm{d}\mathbf{x}\right)\\
&\le \frac{1}{n}\sum_{i=1}^n a_i \Bigg(\int_{S_i} f\left(\frac{n^{-1}p(\mathbf{x}-\boldsymbol{\mu}_i) + \sum_{j\ne i} n^{-1}p(\mathbf{x}-\boldsymbol{\mu}_j)}{n^{-1}a_i p(\mathbf{x}-\boldsymbol{\mu}_i) + \sum_{j\ne i} n^{-1}a_j p(\mathbf{x}-\boldsymbol{\mu}_j)}\right) p(\mathbf{x}-\boldsymbol{\mu}_i)\mathrm{d}\mathbf{x}\\
&\quad + \int_{\mathbb{R}^m\setminus S_i} f\left(\frac{1}{\max_j a_j}\right) p(\mathbf{x}-\boldsymbol{\mu}_i)\mathrm{d}\mathbf{x}\Bigg)\\
&\overset{(a)}{\le} \frac{1}{n}\sum_{i=1}^n a_i \left(\int_{S_i} f\left(\frac{n^{-1}p(\mathbf{x}-\boldsymbol{\mu}_i) + \sum_{j\ne i} n^{-1}\epsilon c}{n^{-1}a_i p(\mathbf{x}-\boldsymbol{\mu}_i) + \sum_{j\ne i} n^{-1} n\epsilon c}\right) p(\mathbf{x}-\boldsymbol{\mu}_i)\mathrm{d}\mathbf{x} + f(n^{-1})\epsilon\right)\\
&\le \frac{1}{n}\sum_{i=1}^n a_i \int_{S_i} f\left(\frac{n^{-1}c + \epsilon c}{n^{-1}a_i c + n\epsilon c}\right) p(\mathbf{x}-\boldsymbol{\mu}_i)\mathrm{d}\mathbf{x} + f(n^{-1})\epsilon\\
&\le \frac{1}{n}\sum_{i=1}^n a_i f\left(\frac{1+n\epsilon}{a_i + n^2\epsilon}\right) + f(n^{-1})\epsilon\\
&\to \frac{1}{n}\sum_{i=1}^n a_i f\left(a_i^{-1}\right) \qquad (10)
\end{aligned}$$

as $\epsilon \to 0$, where (a) is because $\mathbf{x} \in S_i \subseteq B_{\delta_0/2}(\boldsymbol{\mu}_i)$, $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| \ge \delta_0$, and hence $\|\mathbf{x}-\boldsymbol{\mu}_j\| \ge \delta_0/2$, $\mathbf{x}-\boldsymbol{\mu}_j \notin L_{\epsilon c}^+(p)$, and $p(\mathbf{x}-\boldsymbol{\mu}_j) < \epsilon c$.

Let $r > 0$, $B_i := B_r(\boldsymbol{\mu}_i) = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}-\boldsymbol{\mu}_i\| \le r\}$, and $\bar{B} := \mathbb{R}^m \setminus (B_1 \cup \cdots \cup B_n)$. Note that

$$\mathbb{P}(\|\mathbf{x}\| \ge r) \le \frac{\mathbb{E}\left[\|\mathbf{x}\|^2\right]}{r^2} = \frac{\mathrm{tr}(\boldsymbol{\Sigma})}{r^2} = \kappa$$

where $\kappa := r^{-2}\mathrm{tr}(\boldsymbol{\Sigma})$. Assume $\min_{i\ne j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| \ge 4r$. Assume the contrary that there exists a continuous path $\tilde{\boldsymbol{\nu}}_1(t), \dots, \tilde{\boldsymbol{\nu}}_n(t)$ (where $\tilde{\boldsymbol{\nu}}_j : [0,1] \to \mathbb{R}^m$ is a continuous function) such that $d_f(P, n^{-1}\sum_{j=1}^n p(\mathbf{x}-\tilde{\boldsymbol{\nu}}_j(t)))$ is nonincreasing in t, $\tilde{\boldsymbol{\nu}}_j(0) = \boldsymbol{\nu}_j$, and $\{\tilde{\boldsymbol{\nu}}_1(1), \dots, \tilde{\boldsymbol{\nu}}_n(1)\} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_n\}$. Let $t_0 := \inf\{t \ge 0 : \max_j \min_i \|\tilde{\boldsymbol{\nu}}_j(t) - \boldsymbol{\mu}_i\| \ge 2r\}$, and let j_0 satisfies $\min_i \|\tilde{\boldsymbol{\nu}}_{j_0}(t_0) - \boldsymbol{\mu}_i\| = 2r$. Let $\tilde{Q}(\mathbf{x}) := n^{-1}\sum_{j=1}^n p(\mathbf{x} - \tilde{\boldsymbol{\nu}}_j(t_0))$. Note that $\tilde{Q}(\bar{B}) \ge n^{-1}\int_{\bar{B}} p(\mathbf{x} - \tilde{\boldsymbol{\nu}}_{j_0}(t_0))\mathrm{d}\mathbf{x} \ge n^{-1}(1-\kappa)$. We have

$$\begin{aligned}
&d_f(P,\tilde{Q})\\
&\ge \sum_{i=1}^n f\left(\frac{P(B_i)}{\tilde{Q}(B_i)}\right)\tilde{Q}(B_i) + f\left(\frac{P(\bar{B})}{\tilde{Q}(\bar{B})}\right)\tilde{Q}(\bar{B})\\
&\ge \sum_{i=1}^n f\left(\frac{n^{-1}+\kappa}{\tilde{Q}(B_i)}\right)\tilde{Q}(B_i) + f\left(\frac{\kappa}{\tilde{Q}(\bar{B})}\right)\tilde{Q}(\bar{B})
\end{aligned}$$

$$
\begin{aligned}
&\geq \sum_{i=1}^{n} f\left(\frac{n^{-1}+\kappa}{n^{-1}a_i}\right) n^{-1}a_i - \sum_{i=1}^{n}\Bigg(f\left(\frac{n^{-1}+\kappa}{n^{-1}a_i}\right) n^{-1}a_i \\
&\quad - f\left(\frac{n^{-1}+\kappa}{n^{-1}a_i - \max\{n^{-1}a_i - \tilde{Q}(B_i),\, 0\}}\right)\left(n^{-1}a_i - \max\{n^{-1}a_i - \tilde{Q}(B_i),\, 0\}\right)\Bigg) \\
&\quad + f\left(\frac{\kappa}{\tilde{Q}(\bar{B})}\right)\tilde{Q}(\bar{B}) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} f\left(\frac{n^{-1}+\kappa}{n^{-1}a_i}\right) n^{-1}a_i - n\Bigg(f\left(\frac{n^{-1}+\kappa}{n^{-1}n}\right) n^{-1}n \\
&\quad - f\left(\frac{n^{-1}+\kappa}{1-n^{-1}\sum_{i=1}^{n}\max\{n^{-1}a_i - \tilde{Q}(B_i), 0\}}\right)\left(1 - n^{-1}\sum_{i=1}^{n}\max\{n^{-1}a_i - \tilde{Q}(B_i), 0\}\right)\Bigg) \\
&\quad + f\left(\frac{\kappa}{\tilde{Q}(\bar{B})}\right)\tilde{Q}(\bar{B}) \\
&\overset{(c)}{\geq} \sum_{i=1}^{n} f\left(\frac{n^{-1}+\kappa}{n^{-1}a_i}\right) n^{-1}a_i - n\Bigg(f\left(n^{-1}+\kappa\right) - f\left(\frac{n^{-1}+\kappa}{1-n^{-1}\left(\tilde{Q}(\bar{B})+\kappa\right)}\right)\left(1 - \frac{\tilde{Q}(\bar{B})+\kappa}{n}\right)\Bigg) \\
&\quad + f\left(\frac{\kappa}{\tilde{Q}(\bar{B})}\right)\tilde{Q}(\bar{B}) \\
&\overset{(d)}{\geq} \sum_{i=1}^{n} f\left(\frac{n^{-1}+\kappa}{n^{-1}a_i}\right) n^{-1}a_i - n\Bigg(f\left(n^{-1}+\kappa\right) - f\left(\frac{n^{-1}+\kappa}{1-n^{-1}\left(n^{-1}(1-\kappa)+\kappa\right)}\right) \\
&\quad \cdot \left(1 - n^{-1}\left(n^{-1}(1-\kappa)+\kappa\right)\right)\Bigg) + f\left(\frac{\kappa}{n^{-1}(1-\kappa)}\right) n^{-1}(1-\kappa) \\
&\to \frac{1}{n}\sum_{i=1}^{n} a_i f\left(a_i^{-1}\right) - nf\left(n^{-1}\right) + n\left(1-n^{-2}\right) f\left(\frac{n^{-1}}{1-n^{-2}}\right) + n^{-1}\lim_{t\to 0} f(t) \\
&\overset{(e)}{>} \frac{1}{n}\sum_{i=1}^{n} a_i f\left(a_i^{-1}\right)
\end{aligned}
$$

as $r \to \infty$ (which gives $\kappa \to 0$), where (b) is because $t \mapsto tf((n^{-1}+\kappa)/t)$ is convex and $a_i \leq n$, (c) is because

$$
\begin{aligned}
&\sum_{i=1}^{n} \max\{n^{-1}a_i - \tilde{Q}(B_i),\, 0\} - \tilde{Q}(\bar{B}) \\
&\leq \sum_{i=1}^{n} \max\left\{n^{-1}a_i - n^{-1}\sum_{j:\,\boldsymbol{\nu}_j = \boldsymbol{\mu}_i}\int_{B_i} p(\mathbf{x} - \tilde{\boldsymbol{\nu}}_j(t_0))\mathrm{d}\mathbf{x},\, 0\right\} - \tilde{Q}(\bar{B}) \\
&\leq \sum_{i=1}^{n} \max\left\{n^{-1}a_i - n^{-1}\sum_{j:\,\boldsymbol{\nu}_j = \boldsymbol{\mu}_i}\int_{B_i \cup \bar{B}} p(\mathbf{x} - \tilde{\boldsymbol{\nu}}_j(t_0))\mathrm{d}\mathbf{x},\, 0\right\} \\
&= n^{-1}\sum_{i=1}^{n} \max\left\{\sum_{j:\,\boldsymbol{\nu}_j = \boldsymbol{\mu}_i}\int_{\mathbb{R}^m \setminus (B_i \cup \bar{B})} p(\mathbf{x} - \tilde{\boldsymbol{\nu}}_j(t_0))\mathrm{d}\mathbf{x},\, 0\right\} \\
&\leq n^{-1}\sum_{i=1}^{n} \max\left\{a_i\kappa,\, 0\right\} \\
&= \kappa
\end{aligned}
$$

since $\|\boldsymbol{\mu}_i - \tilde{\boldsymbol{\nu}}_j(t_0)\| \leq 2r$, (d) is because $\tilde{Q}(\bar{B}) \geq n^{-1}(1-\kappa)$, and (e) is because $f(t)$ is strictly convex for $0 < t < 1$. Combining this with (10) shows that $d_f(P,\, n^{-1}\sum_{j=1}^{n} p(\mathbf{x} - \tilde{\boldsymbol{\nu}}_j(t)))$ cannot be nonincreasing in t as long as $\min_{i \neq j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$ is large enough.

A.3 Proof of Corollary 3

Wasserstein distance has the following "cancellation property": for two mixture distributions $P = (1-\lambda)P_0+\lambda P_1$ and $Q = (1-\lambda)P_0 + \lambda P_2$ that shares a component P_0, we have

$$W_1(P,Q) = \lambda W_1(P_1, P_2).$$

This follows from the dual representation of W_1:

$$\begin{aligned} W_1(P,Q) &= \sup_{f\colon \mathrm{Lip}(f)\le 1} \Big(\int f(x)P(\mathrm{d}x) - \int f(x)Q(\mathrm{d}x)\Big) \\ &= \sup_{f\colon \mathrm{Lip}(f)\le 1} \Big(\lambda \int f(x)P_1(\mathrm{d}x) - \lambda \int f(x)P_2(\mathrm{d}x)\Big) \\ &= \lambda W_1(P_1,P_2). \end{aligned}$$

Assume we fix a way to assign $\boldsymbol{\nu}_{n'+1},\dots,\boldsymbol{\nu}_n$ to $\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_n$, where each of $\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_{n_0}$ has no $\boldsymbol{\nu}_i$'s assigned to it, each of $\boldsymbol{\mu}_{n_0+1},\dots,\boldsymbol{\mu}_{n_0+n_2}$ has two $\boldsymbol{\nu}_i$'s assigned to it (assume $\boldsymbol{\nu}_{n'+i},\boldsymbol{\nu}_{n'+n_2+i}$ are assigned to $\boldsymbol{\mu}_{n_0+i}$ for $i=1,\dots,n_2$), and each of $\boldsymbol{\mu}_{n_0+n_2},\dots,\boldsymbol{\mu}_n$ has one $\boldsymbol{\nu}_i$ assigned to it, where $n_0, n_2 \ge 0$, $n_0+n_2\le n$. We can obtain a continuous path to the optimum by moving $\boldsymbol{\nu}_1,\dots,\boldsymbol{\nu}_{n_0}$ to $\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_{n_0}$ (note that $n' = n-(n-n_0-n_2)-2n_2 = n_0-n_2$ so $n'+n_2 = n_0$), while leaving $\boldsymbol{\nu}_{n_0+1},\dots,\boldsymbol{\nu}_n$ (which are already one-to-one matched to $\boldsymbol{\mu}_{n_0+1},\dots,\boldsymbol{\mu}_n$) unchanged. Also note that $\boldsymbol{\nu}_1,\dots,\boldsymbol{\nu}_{n_0}$ and $\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_{n_0}$, are i.i.d. uniformly distributed over B_R^m. By the cancellation property, as long as $W_1(n_0^{-1}\sum_{i=1}^{n_0} p(\mathbf{x}-\boldsymbol{\mu}_i),\, n_0^{-1}\sum_{i=1}^{n_0} p(\mathbf{x}-\boldsymbol{\nu}_i))$ is nonincreasing, the overall $W_1(n^{-1}\sum_{i=1}^{n} p(\mathbf{x}-\boldsymbol{\mu}_i),\, n^{-1}\sum_{i=1}^{n} p(\mathbf{x}-\boldsymbol{\nu}_i))$ is nonincreasing as well.

Invoking Theorem 1 on $\boldsymbol{\nu}_1,\dots,\boldsymbol{\nu}_{n_0}$, $\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_{n_0}$ and $\alpha\to 0$ (also note that the proof of Theorem 1 explicitly constructs a continuous path to the optimum as long as the conditions in Theorem 1 are satisfied), the probability that there is no continuous path to the optimum with nonincreasing Wasserstein distance is upper-bounded by $C_{m,n_0}r/R$. Applying union bound on all possible ways to assign $\boldsymbol{\nu}_{n'+1},\dots,\boldsymbol{\nu}_n$ to $\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_n$ (there are at most n^n ways), the probability that there exists an assignment such that there is no such path is upper-bounded by

$$n^n \max_{1\le n_0\le n} \frac{C_{m,n_0}r}{R} = \Big(n^n \max_{1\le n_0\le n} C_{m,n_0}\Big)\frac{r}{R}.$$

B Experiments

B.1 Details of Experiment Setting

Datasets: In the case of Gaussian mixtures, we defined a symmetric mixture of five Gaussians with opposite means at $[0,0]$, $[\pm 1,0]$, and $[0,\pm 1]$ as illustrated in Figure 1. The covariance matrix is $\sigma^2 I_2$ where $\sigma = 0.05$. The sampling is independent in every training iteration. For the real image datasets, the formulations are as follows:

1. **MNIST B/W**: We used one-half (25,000 samples) of the MNIST training data as the first mode and constructed the second mode by flipping the other half of the MNIST samples that had white backgrounds and black digits post-flipping.
2. **CIFAR10/CelebA**: The two identifiable modes are the 5000 ship samples from CIFAR-10 and 5000 randomly selected CelebA samples.
3. **CelebA Color/Grayscale**: The dataset contains half of the original CelebA samples as well as the grayscale version of the other half of CelebA data.
4. **LSUN Church Color/Grayscale**: The dataset contains half of the original LSUN Church samples as well as the grayscale version of the other half of LSUN Church data.

Network Architecture: For MNIST series experiments, we utilized a Multi-Layer Perceptron (MLP) architecture for both the generator and discriminator neural nets, with 5 and 3 MLP blocks, respectively. In the

	Density/Quality	Coverage/Diversity
JS-divergence	1.1563	0.4891
W_1 distance	0.7801	0.9094
dynamic training	0.9163	0.8375

Table 1: Sample quality/diversity statistics of different GAN training. Density and coverage scores are introduced in (Naeem et al, 2020) evaluating the quality and diversity of generated samples, respectively.

rest of the experiments, to enhance visual quality we alternatively used a CNN architecture for the generator, and the empowered MLP architecture introduced in (Tolstikhin et al., 2021) for discriminators. As commonly applied in training GANs, we also used batch normalization and nearest-upsampling for training the generators. We considered three convolutional blocks in the experiments.

Optimizers and Hyperparameters: We use the AdamW (Loshchilov and Hutter, 2017) optimizer implemented in PyTorch, configured with weight decay rate at 1e-4, default beta parameters 0.5 and 0.999, for both the generator and discriminator networks in the experiments. We used different learning rates for the generator and discriminator as suggested in (Heusel et al., 2017), which are 1e-4 and 4e-4, respectively. All experiments are conducted in a server configured with 4 RTX 3090s.

B.2 Additional Numerical Results

In this subsection, we present the complete set of our visualization results. In addition to the datasets listed in the main text, here we report qualitative and quantitative results on LSUN (Yu et al., 2015) dataset with two induced modes for Color/Grayscale images. Figure 6 is the complete version of Figure 3 in the text. Moreover, similar to Figure 4 of the main text, we present convergence plots for bimodal datasets CIFAR10/CelebA and LSUN Church Color/Grayscale in Figure 8 and Figure 9. Besides the GMM with 5 components in the main text, we also present the results of bi-modal GMM in Figure 5.

We also consider Lipschitz-dynamic training, where the objective is initialized as JS divergence and gradually adjusts the Lipschitz constraint during the training. Figure 7 shows samples from CelebA color&grayscale dataset. Table 1 demonstrates evaluation metrics, density and coverage, which is proposed in Naeem et al. (2020), of different training schemes. We observe that dynamically adjusting Lipschitz constraint during training could prevent the GAN from getting stuck in the unimodal local optimum, and help to improve the diversity and maintain moderate quality. Dynamic training scheme achieves balancing density and coverage scores compared with JS divergence and Wasserstein distance.

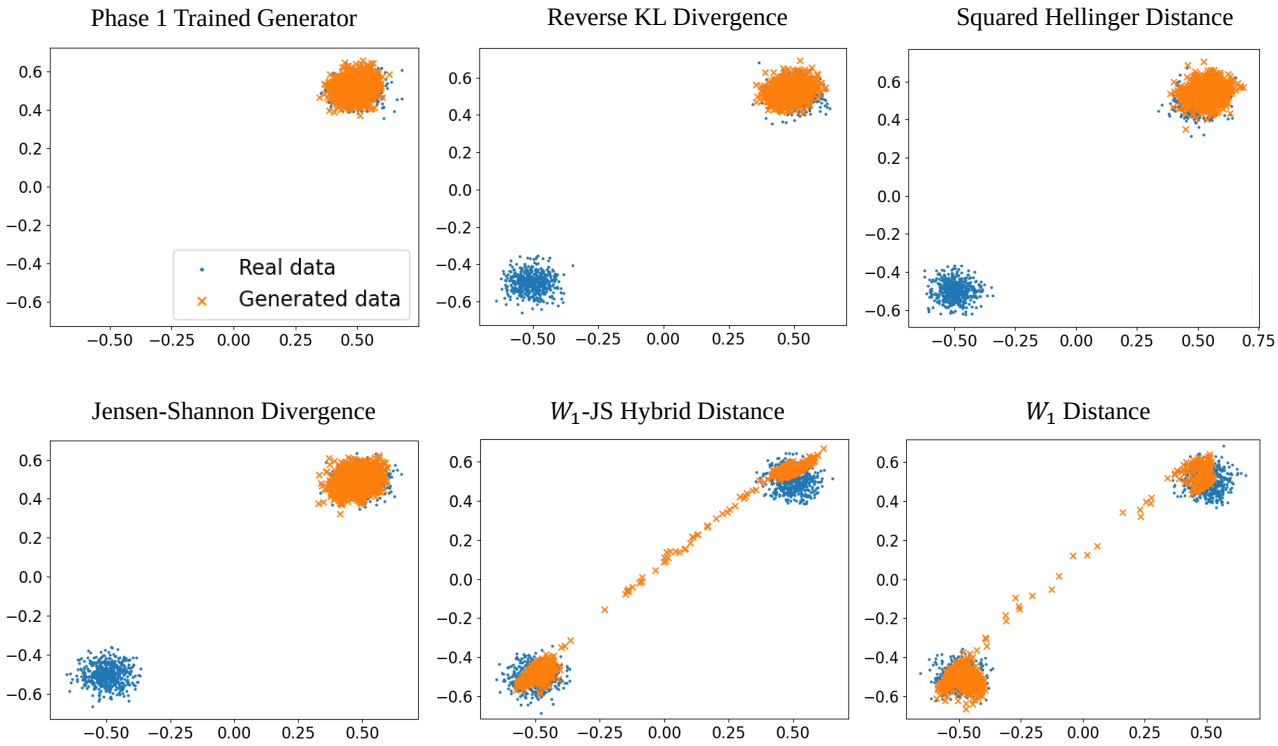

Figure 5: Bi-modal Gaussian mixture data (in blue) and GANs' generated data (in orange). Mode-seeking f-divergences were trapped in unimodal local optima, while the Wasserstein and W_1-JS-hybrid distances resulted in capturing both modes.

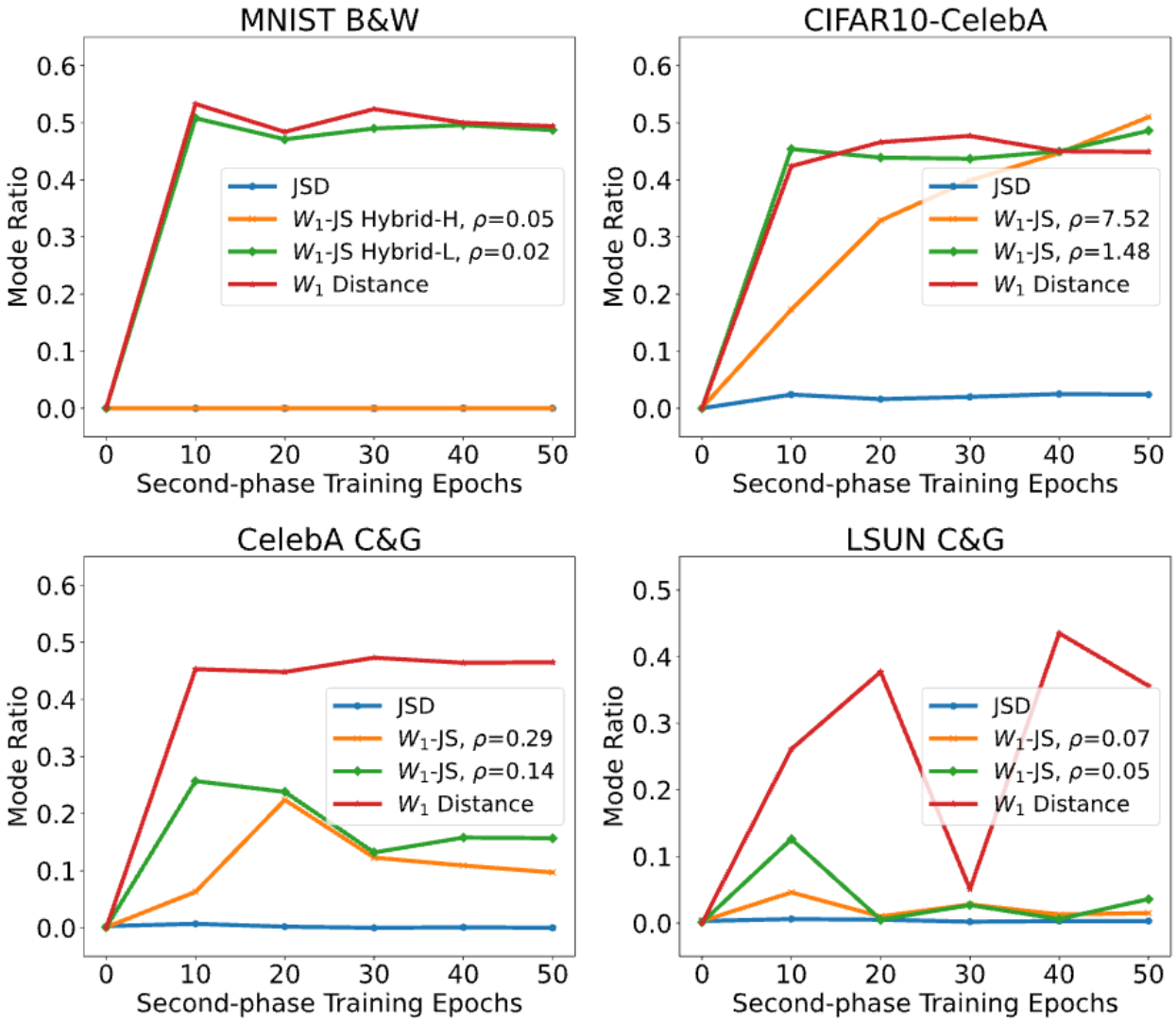

Figure 6: Mode ratio during Phase 2 of GAN training over bimodal image data, An α mode ratio indicates an α-fraction of generated images from the second mode. GANs with W_1 and JS-W_1 hybrid distances led to balanced modes, while VGAN (JS-divergence) led to a unimodal point.

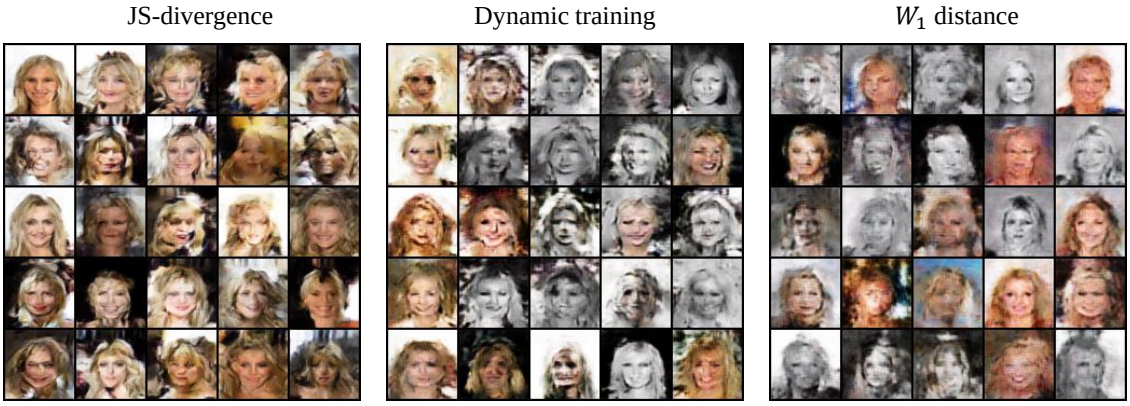

Figure 7: Samples generated from GANs trained by minimizing JS-divergence, hybrid divergence with dynamic Lipschitz coefficient, W_1-distance measures.

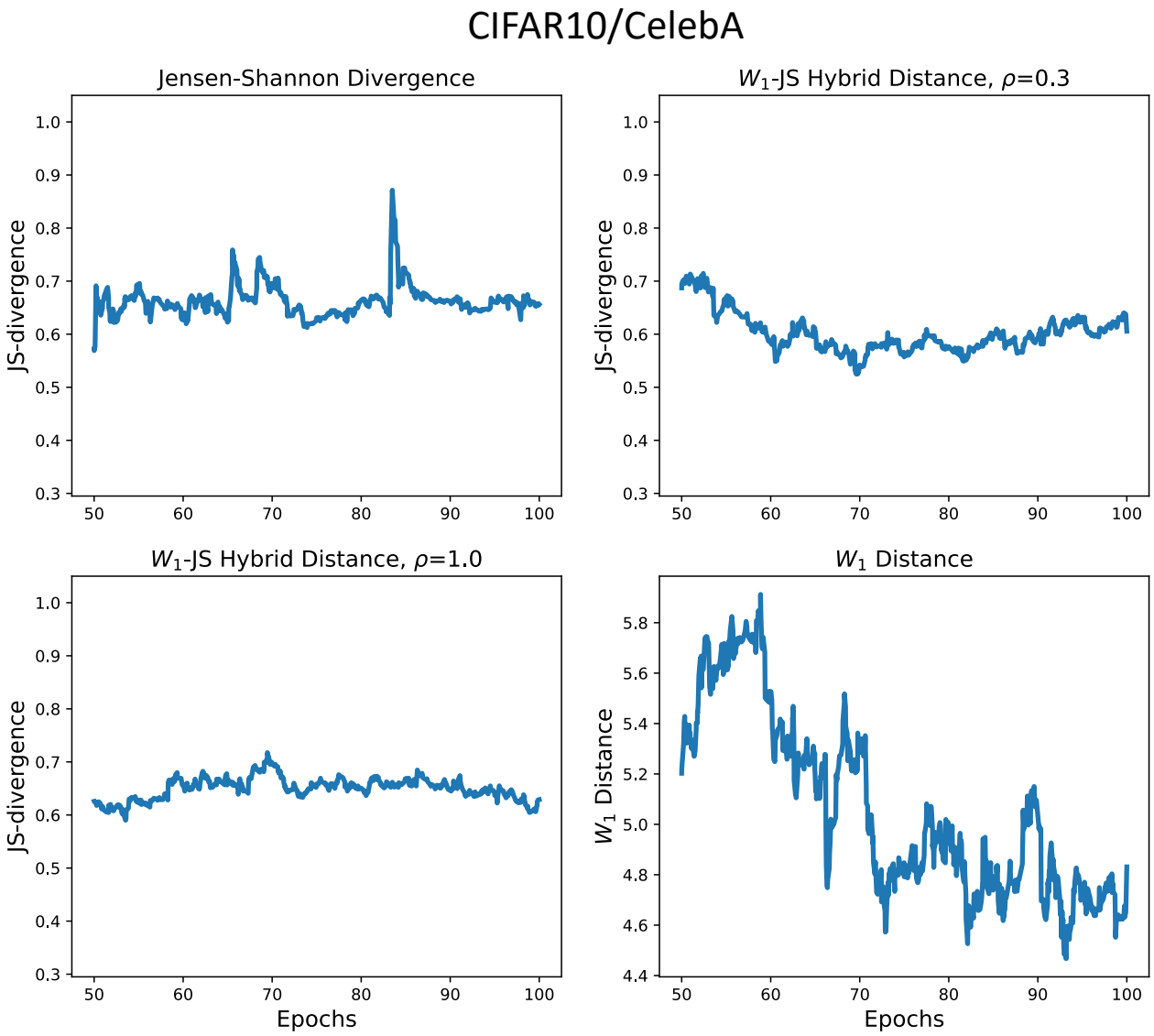

Figure 8: Divergence optimization progress in second-phase of the CIFAR10&CelebA bimodal dataset.

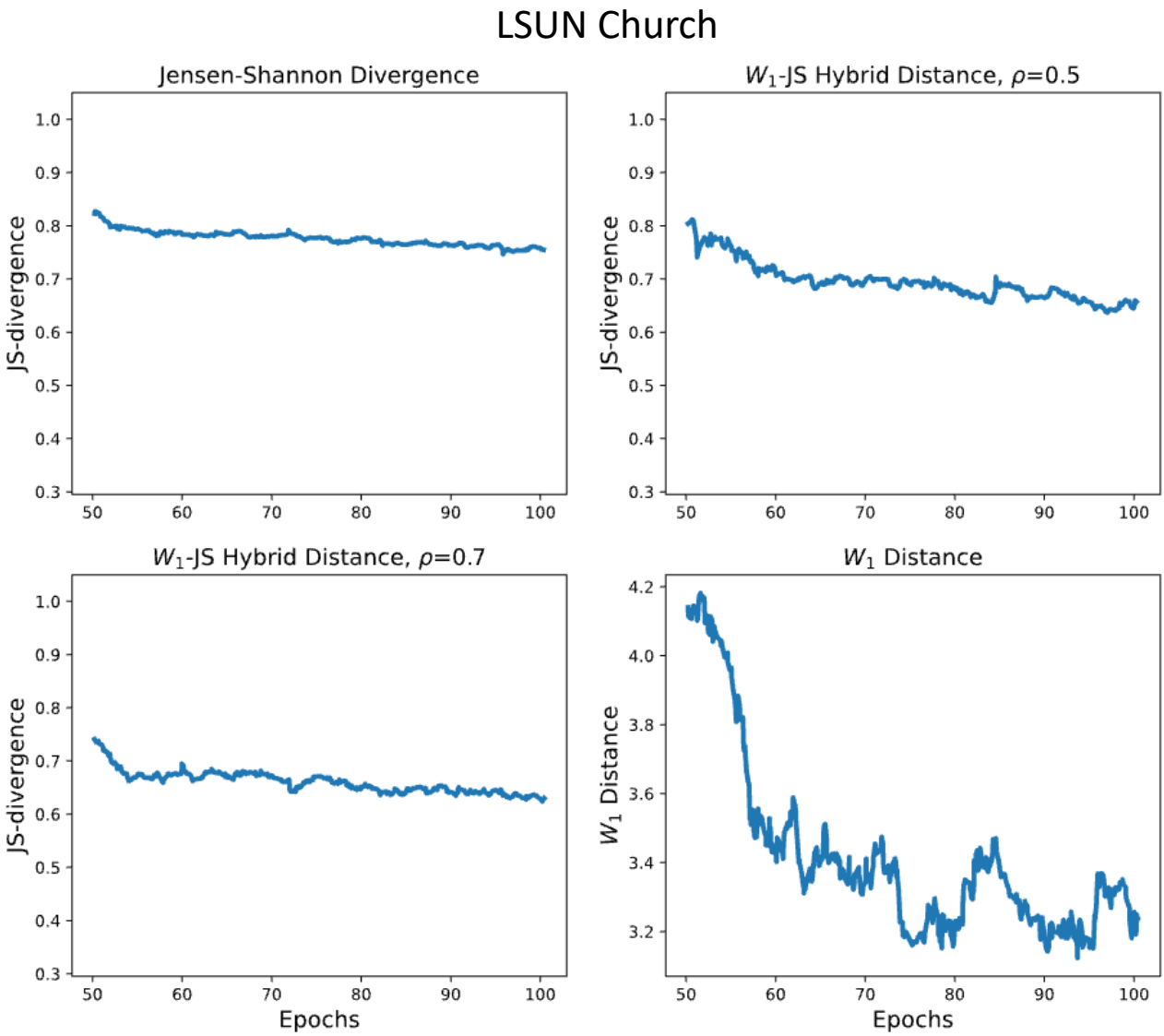

Figure 9: Divergence optimization progress in Phase 2 (epochs 50-100) of the GANs training on LSUN Church Color&Grayscale bimodal dataset.