Mode-Seeking Divergences: Theory and Applications to GANs

Cheuk Ting Li
The Chinese University of Hong Kong

Farzan Farnia
The Chinese University of Hong Kong

Abstract

Generative adversarial networks (GANs) represent a game between two neural network machines designed to learn the distribution of data. It is commonly observed that different GAN formulations and divergence/distance measures used could lead to considerably different performance results, especially when the data distribution is multi-modal. In this work, we give a theoretical characterization of the mode-seeking behavior of general f-divergences and Wasserstein distances, and prove a performance guarantee for the setting where the underlying model is a mixture of multiple symmetric quasiconcave distributions. This can help us understand the trade-off between the quality and diversity of the trained GANs' output samples. Our theoretical results show the mode-seeking nature of the Jensen-Shannon (JS) divergence over standard KL-divergence and Wasserstein distance measures. We subsequently demonstrate that a hybrid of JS-divergence and Wasserstein distance measures minimized by Lipschitz GANs mimics the mode-seeking behavior of the JS-divergence. We present numerical results showing the mode-seeking nature of the JS-divergence and its hybrid with the Wasserstein distance while highlighting the mode-covering properties of KL-divergence and Wasserstein distance measures. Our numerical experiments indicate the different behavior of several standard GAN formulations in application to benchmark Gaussian mixture and image datasets.

1 INTRODUCTION

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have attained great success in various distribution learning problems. The GAN framework reduces the learning task to a game between the following two machine players that are typically chosen to be deep neural networks: 1) A generator machine trying to map a random noise input to real-like samples that are difficult to distinguish from actual training data, 2) A discriminator function focusing on classifying the generated samples from real collected data.

Nevertheless, standard GAN implementations often struggle in modeling multi-modal distributions comprised of several distinct modes. Two major issues are: 1) over-generalization (Bishop, 2006; Lucas et al., 2019), where low-quality or unrealistic outputs are produced, and 2) mode collapse, where the generator lacks diversity and captures only one or a few of the underlying modes. While such struggles in learning mixture distributions have been reported for various GAN applications, different GAN formulations empirically achieve different diversity and sharpness scores in application to multi-modal data. Such observations highlight the following question:

Do different GAN formulations lead to different underlying solutions in learning mixture models?

In this work, we attempt to address the above question through an information theoretic approach. Different GAN problems are known to minimize different divergence measures between the data and generator's distributions. For example, the vanilla GAN (VGAN) (Goodfellow et al., 2014) targets the Jensen-Shannon (JS) divergence. The f-GANs (Nowozin et al., 2016) generalize the VGAN problem by minimizing a general f-divergence. The Least Square GANs (LSGANs) (Mao et al., 2017) minimize the Pearson χ^2-divergence. The Wasserstein GANs (WGANs) (Arjovsky et al., 2017) target the 1-Wasserstein distance.

The notions of mode-covering divergences and mode-seeking divergences have been introduced to describe the behaviors of different divergence measures (Bishop, 2006). Mode-covering divergences (Bishop, 2006; Poole et al., 2016) result in a fitted model that cover all the modes of the multi-modal data distribution, but may assign mass over the empty space between the modes. An example is the Kullback-Leibler (KL) divergence (Bishop, 2006; Goodfellow, 2016), which arises in the maximum likelihood estimator. On the other hand, mode-seeking divergences (Bishop, 2006; Ke et al., 2020) result in a model that captures a subset

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

of the modes of the data distribution, and tends to avoid assigning masses to empty spaces. An example is the reverse KL divergence (Bishop, 2006; Huszár, 2015).

Mode-covering and mode-seeking divergences have been observed to affect the quality and diversity of the outputs. A common observation is that mode-seeking divergences tend to give a model that produces higher quality outputs (Huszár, 2015; Ghasemipour et al., 2020; Zhang et al., 2019; Ke et al., 2020), whereas mode-covering divergences often produce lower quality or unrealistic samples (Lucas et al., 2019; Williams et al., 2020), and suffer from the problem of over-generalization (Bishop, 2006; Lucas et al., 2019). On the other hand, the use of mode-covering divergences may improve sample diversity of the generative model (Poole et al., 2016), whereas mode-seeking divergences may contribute to the problem of mode collapse (Lucas et al., 2019; Shannon et al., 2020) [1].

Most of the aforementioned works are based on empirical observations of the behaviors of the f-divergences. There has not been a unified framework of the classification of f-divergences based on theoretical guarantees. In particular, even for the popular JS divergence, whether it is mode-seeking or mode-covering is debated (see Section 2).

In this work, we give a theoretical characterization of the mode-seeking behavior of general f-divergences and Wasserstein distances. We study the setting where the generator fits a unimodal symmetric quasiconcave distribution Q_θ to a data distribution P that is a mixture of multiple symmetric quasiconcave components. We demonstrate that an f-divergence with a function f that is strongly-convex in the range $(0, 1+\epsilon]$ and grows at most linearly (e.g. reverse KL, JS, Neyman χ^2 or squared Hellinger distance) is guaranteed to be mode-seeking, in the sense that Q_θ will identify a mode in P. Our theoretical results, therefore, shed light on the mode-seeking nature of VGAN and several other f-GANs under a general theoretical setting.

In addition, we demonstrate that the widely-used Wasserstein distances fail to be mode-seeking, and the trained generator could produce samples not belonging to the existing modes. Subsequently, we analyze a particular hybrid of f-divergence and Wasserstein measures studied in (Farnia and Tse, 2018) which has been shown to be the target divergence metric in Lipschitz GANs (Kodali et al., 2017; Gulrajani et al., 2017; Miyato et al., 2018; Zhou et al., 2019) such as the vanilla GAN with the spectral normalization and with the gradient penalty. We show that the hybrid of a mode-seeking f-divergence and the 1-Wasserstein distance will preserve the mode-seeking nature of the f-divergence, and can provably identify a mode even when only samples from the true data distribution are known. Our analysis therefore proves that the hybrid divergence can provide a mode-seeking distance that retains a major advantage of WGAN that it is continuously changing with the generator's parameters. We summarize the contributions of this paper as follows:

- We develop a unified theoretical framework of classifying mode-seeking f-divergences.
- We prove a theoretical guarantee for mode-seeking f-divergences when the data distribution is a mixture of symmetric quasiconcave distributions.
- We show that a convolutional hybrid of a mode-seeking f-divergences and the 1-Wasserstein distance remains mode-seeking, while retaining the continuity property of the Wasserstein distance.
- We numerically support our theoretical findings on Gaussian mixture and image datasets.

2 RELATED WORKS

Except KL divergence (agreed to be mode-covering (Bishop, 2006; Goodfellow, 2016)) and reverse KL divergence (agreed to be mode-seeking (Bishop, 2006; Huszár, 2015)), there was no clear-cut classification of mode-covering and mode-seeking divergences. For example, JS divergence has been regarded as 1) comparatively mode-seeking / quality-driven (Huszár, 2015; Theis et al., 2015; Lucas et al., 2019), 2) comparatively mode-covering (Poole et al., 2016), 3) neither mode-seeking nor covering (Shannon et al., 2020), and 4) mode-seeking or covering depending on the situation (Ke et al., 2020). All these claims (except (Shannon et al., 2020)) were based on empirical evidence or heuristics rather than theoretical analysis, and hence depends greatly on the setting and various factors other than the choice of divergence. To the best of the authors' knowledge, the only theoretical treatment of mode-covering/seeking divergence is (Shannon et al., 2020), where two quantities about f-divergences – left and right tail weights – were introduced to describe its mode-covering and mode-seeking behaviors respectively. Nevertheless, (Shannon et al., 2020) does not provide any theoretical guarantee on the mode-seeking performance of f-divergences in model fitting.

A closely-related concept is zero-avoiding/forcing divergences (Minka, 2005; Bishop, 2006). When fitting a distribution Q to the data distribution P, a zero-avoiding divergence results in a Q where $Q(x) > 0$ for any x with $P(x) > 0$, whereas a zero-forcing divergences results in a Q where $Q(x) = 0$ for any x with $P(x) = 0$. While zero-avoiding is conceptually almost the same as mode-covering, zero-forcing does not necessarily imply (strongly) mode-seeking in the sense studied in this paper, since we require Q to capture a mode in P accurately. For works on mode-covering/seeking α-divergences, see (Minka, 2005; Hernandez-Lobato et al., 2016; Li and Turner, 2016; Wang et al., 2018). The α-divergence is zero-avoiding when

[1] It was argued in (Goodfellow, 2016) that the choice of divergence is not a major factor in mode collapse.

$\alpha \geq 1$, and zero-forcing when $\alpha \leq -1$ (Bishop, 2006; Minka, 2005). We will prove that α-divergence is mode-seeking when $\alpha < 1$, showing that mode-seeking is not exactly the same as zero-forcing.

Regarding our evaluation of target divergences in GANs, the numerical studies in (Lucic et al., 2018; Kurach et al., 2019) report similar Fréchet inception distance (FID) scores for different GAN formulations. However, as also discussed in (Sajjadi et al., 2018; Borji, 2022), this observation does not indicate the same diversity and quality scores for the learnt distributions, as FID scores lead to a one-dimensional evaluation of GANs. To address this issue, (Sajjadi et al., 2018; Kynkäänniemi et al., 2019) propose the precision and recall scores to contrast different generative models in the 2-dimensional space of diversity and quality of generated data. As a complementary approach, our work focuses on a theoretical framework for mode-seeking divergence measure to demonstrate their power in improving the quality of generated data. Also, we use an information-theoretic decomposition of Inception score (Salimans et al., 2016) to measure the quality and diversity of generated image data in our experiments. Finally, regarding the mode collapse phenomenon in GANs, (Arjovsky et al., 2017) suggests that Wasserstein GANs can resolve the mode collapse issue. Furthermore, (Nagarajan and Kolter, 2017) has included a regularization term to WGAN, and (An et al., 2019) uses the Brenier potential on a latent space via an autoencoder.

3 PRELIMINARIES

3.1 f-divergence measures and f-GANs

Given a convex function $f : [0, \infty) \to \mathbb{R} \cup \{\infty\}$ with $f(1) = 0$, the f-divergence (Csiszár and Shields, 2004) of P from Q (both P, Q are regarded as probability density functions) is defined as

$$D_f(P\|Q) := \int f\left(\frac{P(x)}{Q(x)}\right) Q(x)\mathrm{d}x.$$

f-GAN (Nowozin et al., 2016) attempts to solve the following divergence minimization problem for the f-divergence from the observed data distribution P_X to the generator's model $P_{G(\mathbf{Z})}$: $\min_{G\in\mathcal{G}} \; D_f(P_X\|P_{G(\mathbf{Z})})$. In the above, $\mathcal{G}$ represents the set of generator mappings and $\mathbf{Z}$ denotes the noise random vector input to generator G. f-GAN uses the following variational formulation of f-divergence (Nguyen et al., 2010) to lower-bound the above divergence minimization problem with a minimax optimization problem:

$$D_f(P\,\|\,Q) \geq \sup_{T\in\mathcal{T}} \Big(\mathbb{E}_{\mathbf{x}\sim P}[T(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim Q}[f^*(T(\mathbf{x}))]\Big). \tag{1}$$

Here, f^* denotes f's convex conjugate, $f^*(s) := \sup_t(st - f(t))$, and $\mathcal{T}$ is an arbitrary function set.

3.2 Wasserstein distances and Wasserstein GANs

The ρ-Wasserstein distance (Villani, 2003) with parameter $\rho > 0$ is defiend as:

$$W_\rho(P, Q) := \Big(\inf_{R\in\Gamma(P,Q)} \int \|\mathbf{x}-\mathbf{y}\|^\rho R(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y})\Big)^{1/\max\{\rho,\,1\}},$$

where $\Gamma(P, Q)$ is the set of couplings of P and Q. The Wasserstein GAN (WGAN) problem (Arjovsky et al., 2017) aims to find the generative model with the minimum 1-Wasserstein distance to the data distribution $\min_{G\in\mathcal{G}} \; W_1(P_X, P_{G(\mathbf{Z})})$. To solve the distance minimization problem, WGANs leverage the Kantorovich-Rubinstein duality result (Villani, 2003) revealing that

$$W_1(P, Q) = \sup_{T \text{ 1-Lipschitz}} \mathbb{E}_{\mathbf{x}\sim P}[T(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim Q}[T(\mathbf{x})], \tag{2}$$

where the discriminator function T is constrined to be 1-Lipschitz. Aside from standard WGANs minimizing the 1-Wasserstein distance, the W2GAN problem minimizing the 2-Wasserstein distance has also been studied in (Bousquet et al., 2017; Feizi et al., 2020; Taghvaei and Jalali, 2019).

3.3 The hybrid of f-divergence and Wasserstein distance: Lipschitz GANs

While f-GANs typically lack a stable convergence behavior which may lead to training failures, the f-GAN problems with a regularized discrminator with bounded Lipschitz constant, e.g. under spectral normalization and gradient penalty (Miyato et al., 2018; Kodali et al., 2017), have been empircally observed to enjoy higher training stability. (Farnia and Tse, 2018) theoretically shows that such an f-GAN problem with a $\frac{1}{\lambda}$-Lipschitz discriminator minimizes the following hybrid of the f-divergence and 1-Wasserstein distance:

$$D_{\lambda f, W_1}(P\,\|\,Q) := \inf_{\tilde{P}} \Big(W_1(P, \tilde{P}) + \lambda D_f(\tilde{P}\,\|\,Q)\Big), \tag{3}$$

where the infimum is taken over all distributions. It can be seen that the above divergence measure has the continuous behavior of Wasserstein distances in the input distributions.

4 MODE-SEEKING f-DIVERGENCES

In existing literature, mode-seekingness (Bishop, 2006; Ke et al., 2020) has a purely operational meaning, where a divergence/distance is mode-seeking if minimizing the divergence between a multimodal data distribution and the distribution of the model allows the model to capture one of the modes. Here we give a theoretical characterization of mode-seeking f-divergences, which will be proven in Theorem 4.3 to guarantee the aforementioned operational behavior.

Consider f-divergence D_f, where $f(t)$ is convex with $f(1) = 0$. Consider the following conditions:

Table 1: Mode-seeking and mode-covering f-divergences, ordered loosely in decreasing order of mode-seeking power.

	f-divergence	$f(t)$	Mode-seeking order $\mathring{f}(\gamma)$
Uniformly mode-seeking (MS1-4)	Neyman χ^2-divergence	$t^{-1}-1$	$O(\gamma^{1/3})$
	Softened reverse KL (Shannon et al., 2020)	$2(t+1)\log\frac{t+1}{t}-4\log 2$	$O(\gamma^{1/3})$
	$\mathcal{G}_{\mathrm{ALT}}$ divergence (Poole et al., 2016)	$\log(1+t^{-1})-\log 2$	$O(\gamma^{1/3})$
	Reverse KL divergence	$-\log t$	$O(\gamma^{1/3}\sqrt{-\log\gamma})$
	Jensen-Shannon divergence	$\frac{1}{2}\left(t\log\frac{t}{t+1}-\log\frac{t+1}{4}\right)$	$O(\gamma^{1/3}\sqrt{-\log\gamma})$
	Squared Hellinger distance	$2(1-\sqrt{t})$	$O(\gamma^{1/5})$
	α-divergence for $\alpha<1, \alpha\neq -1$	$\frac{4}{1-\alpha^2}(1-t^{(1+\alpha)/2})$	$O(\gamma^{(1-\alpha)/(5-\alpha)}+\gamma^{1/3})$
Weakly mode-seeking (MS1-2 only)	Total variation distance	$\max\{1-t,\,0\}$	$O(1)$
Mode-covering (none of MS1-4)	KL divergence	$t\log t$	N/A
	Pearson χ^2-divergence	$(t-1)^2$	N/A
	α-divergence for $\alpha>1$	$\frac{4}{1-\alpha^2}(1-t^{(1+\alpha)/2})$	N/A

- **(MS1)** $\lim_{t\to\infty} f(t)/t<\infty$.
- **(MS2)** There is no $s\in(0,1)$ such that $f(t)$ is a straight line (an affine function) for $t\in[s,\infty)$.
- **(MS3)** f is strongly convex for $t\in(0,s]$ for some $s>1$ (i.e., there exists $\beta>0$ such that $t\mapsto f(t)-\beta t^2/2$ is convex for $t\in(0,s]$).
- **(MS4)** There exists $s>1$ such that f is twice continuously differentiable for $t\in(0,s]$, and $f''(t)$ is non-increasing for $t\in(0,s]$.

Definition 4.1. We call D_f *weakly mode-seeking* if it satisfies MS1–2. We call D_f *strongly mode-seeking* if it satisfies MS1–3 (it suffices to check MS1 and MS3). We call D_f *uniformly mode-seeking* if it satisfies MS1–4.

For example, Jensen-Shannon divergence, reverse KL divergence and Neyman χ^2-divergence are uniformly mode-seeking, whereas total variation distance is only weakly mode-seeking. KL divergence and Pearson χ^2-divergence are not mode-seeking. Refer to Table 1 and Figures 1 and 13 for more examples.

To illustrate the behaviors of various f-divergences, consider the data distribution $P=0.75\mathcal{N}(0,1)+0.25\mathcal{N}(\delta,1)$, a mixture of 2 Gaussian distributions ($\delta\geq 0$ is the separation between the two modes), and we fit a Gaussian distribution Q that minimizes $D_f(P\|Q)$. The plots of the center of Q against δ are given in Figures 2 and 14. Observe the following three kinds of behaviors: 1) Uniformly mode-seeking divergences (Neyman χ^2, reverse KL, JS, squared Hellinger) where the center of Q tends to the largest mode 0 as δ increases, correctly identifying a mode with increasing accuracy as the modes become more well-separated. 2) Weakly mode-seeking divergence (TV) where the center stays within a bounded distance from 0, identifying a mode without increasing accuracy. 3) Mode-covering divergences (KL, Pearson χ^2), where Q is centered in the middle of the two modes. The legends of the plot is ordered in decreasing order of mode-seeking power according to the plot.

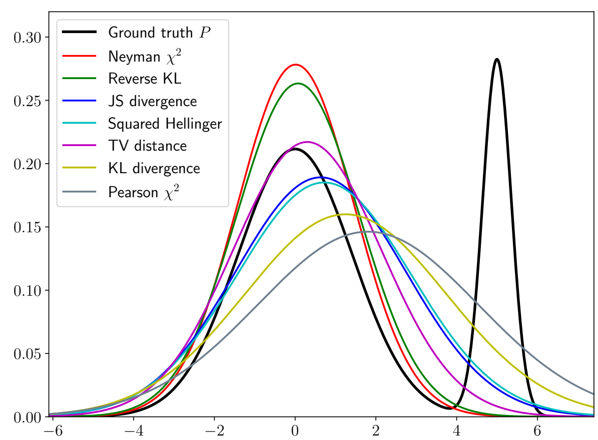

Figure 1: Plot of $\mathrm{argmin}_{Q\text{ Gaussian}} D_f(P\|Q)$ for various f-divergences, where the ground truth $P=0.75\mathcal{N}(0,2)+0.25\mathcal{N}(5,1/8)$ is a mixture of 2 Gaussian distributions. A mode-seeking divergence tends to capture the mode on the left, whereas a mode-covering divergence tends to be closer to the center.

We will give a theoretical justification of the aforementioned

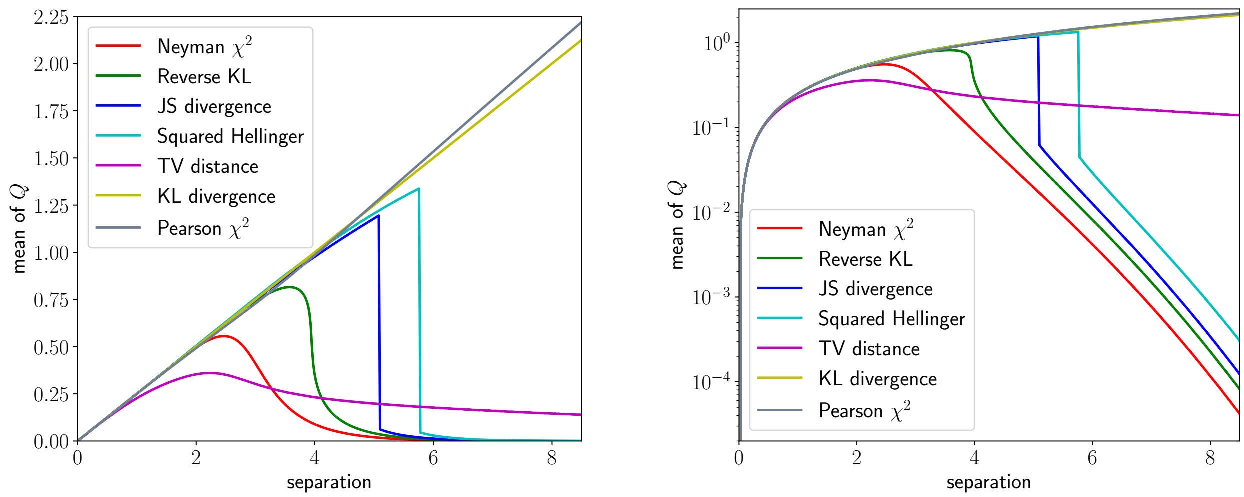

Figure 2: Plot of the center of $\mathrm{argmin}_{Q \text{ Gaussian}} D_f(P\|Q)$ for various f-divergences, where the ground truth $P = 0.75\mathcal{N}(0,1) + 0.25\mathcal{N}(\delta,1)$ is a mixture of 2 Gaussian distributions, where $\delta \geq 0$ is the separation between the two modes. We plot the center of Q against δ (left: linear scale, right: log-scale).

characterization. First, we state the definition of symmetric quasiconcave distributions, which includes Gaussian distributions and Laplace distributions as special cases.

Definition 4.2. A probability density function $p : \mathbb{R}^d \to \mathbb{R}$ is *symmetric quasiconcave* if the superlevel set $\{\mathbf{x} \in \mathbb{R}^d : p(\mathbf{x}) \geq t\}$ is convex for any $t \geq 0$, and there exists $\boldsymbol{\mu} \in \mathbb{R}^d$ (the *center*) such that $p(\boldsymbol{\mu}+\mathbf{x}) = p(\boldsymbol{\mu}-\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$.

Given $\mathcal{P}$ which is an arbitrary set of symmetric quasiconcave distributions with finite second moments over $\mathbb{R}^d$, we consider the setting where the data distribution $P(\mathbf{x}) := \sum_{i=1}^k w_i p_i(\mathbf{x})$ is a mixture of $k \geq 2$ distributions in $\mathcal{P}$, where $p_1, \ldots, p_k \in \mathcal{P}$ with distinct centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^d$ and covariance matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k \in \mathbb{R}^{d\times d}$, and $w_1, \ldots, w_k > 0$ with $\sum_{i=1}^k w_i = 1$. We are going to fit $Q \in \mathcal{P}$ to P according to

$$Q := \underset{Q\in\mathcal{P}}{\mathrm{argmin}}\ D_f(P \,\|\, Q).$$

We will show that, as long as D_f is mode-seeking, Q can identify one of the modes of P. As observed in Figure 2, this works only when the components p_i are sufficiently well-separated. Well-separatedness is measured in terms of

$$\sigma_{\max} := \max_i \lambda_{\max}^{1/2}(\boldsymbol{\Sigma}_i), \quad \delta_{\min} := \min_{i\neq j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2.$$

The components are well-separated if $\sigma_{\max}/\delta_{\min}$ is small. We now state the main result. The proof is in Appendix C.

Theorem 4.3. *Consider the aforementioned setting of fitting $Q := \mathrm{argmin}_{Q\in\mathcal{P}} D_f(P \,\|\, Q)$ to a mixture distribution P of distributions in $\mathcal{P}$. Denote the center of Q as $\boldsymbol{\mu}_Q$. If such minimizer Q exists, then we have:*

- *If D_f is weakly mode-seeking, then there is a constant $C_{f,k} > 0$ (only depends on f, k) such that*
$$\min_i \left\|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_i\right\|_2 \leq C_{f,k}\sigma_{\max}. \tag{4}$$
Hence, a mode is identified without increasing accuracy.

- *If D_f is strongly mode-seeking, then there is a constant $C_{f,k} > 0$ (only depends on f, k) such that*
$$\min_i \left\|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_i\right\|_2 \leq C_{f,k}\sigma_{\max}\mathring{f}\left(\sigma_{\max}/\delta_{\min}\right), \tag{5}$$
where
$$\mathring{f}(\gamma) := \inf_{0<\epsilon<1/2} \left\{ \frac{\gamma}{\epsilon} + \sqrt{-\epsilon f\left(\frac{1}{\epsilon}\right) + (1-\epsilon)\lim_{t\to\infty}\frac{f(t)}{t} + \epsilon} \right\}$$
is called the mode-seeking order *of D_f. Note that $\lim_{\gamma\to 0}\mathring{f}(\gamma) = 0$. A mode is identified with increasing accuracy as the modes become more well-separated.*

- *If D_f is uniformly mode-seeking, then there is a constant $C_f > 0$ (only depends on f) such that*
$$\min_i \left\|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_i\right\|_2 / \sigma_{\max} \leq C_f k \mathring{f}\left(k\sigma_{\max}/\delta_{\min}\right) \tag{6}$$
as long as the right hand side is not greater than 1.

Explicit expressions for the constants in (4), (5), (6) can be found in (10), (25), (28) in the proof respectively. Intuitively, in the bound for uniformly mode-seeking divergences in (6), the order of growth with respect to the number of modes k is stated explicitly, and it is uniform in the sense that the constant C_f does not depend on k.

Note that the result is independent of the dimension d. A limitation of Theorem 4.3 is that it only applies when the true data distribution P is known, whereas, in practice, only a dataset containing samples from P is known. The situation where we are only given samples from P will be discussed in Section 6. Table 1 lists several f-divergences with their mode-seeking orders, ordered loosely in decreasing order of mode-seeking power. We choose $f(t)$ satisfying $\lim_{t\to\infty} f(t)/t = 0$ for mode-seeking divergences. For a weakly mode-seeking divergence, let its mode-seeking order be $O(1)$ so (5) holds.

5 WASSERSTEIN DISTANCE

In this section, we show that the Wasserstein distances W_ρ are not mode-seeking in the operational sense, i.e., it fails to capture a mode in a mixture distribution. We first consider the case where we fit a point mass to a discrete distribution, i.e., $\mathcal{P} = \{\delta_{\mathbf{z}} : \mathbf{z} \in \mathbb{R}^d\}$ (where $\delta_{\mathbf{z}}$ denotes the degenerate distribution at $\mathbf{z}$), the data distribution is $P = \sum_{i=1}^k w_i \delta_{\mathbf{z}_i}$, and we fit $Q \in \mathcal{P}$ that minimizes $W_\rho(P, Q)$, which is equivalent to finding $\mathbf{x}$ that minimizes $\sum_{i=1}^k w_i \|\mathbf{x} - \mathbf{z}_i\|^\rho$. A more general case will be discussed later. We can show by convexity that if $\rho \geq 1$, then the minimizing Q may not coincide with any of the modes $\mathbf{z}_i$'s (if P is symmetric around 0, then $\mathbf{x} = 0$ is a minimizer). For $\rho = 2$, the $\mathbf{x}$ that minimizes $\sum_{i=1}^k w_i \|\mathbf{x} - \mathbf{z}_i\|^\rho$ is the mean of P. For $\rho = 1$, the minimizer is the weighted geometric median, which generally does not coincide with any $\mathbf{z}_i$ when the dimension $d \geq 2$.

Nevertheless, when $0 < \rho < 1$, the Wasserstein distance corresponds to a transportation cost with concave cost function (McCann, 1999; Santambrogio, 2015), which might be mode-seeking. Indeed, for the one-dimensional case $d = 1$, the Q that minimizes W_ρ for $\rho < 1$ must coincide with one of the modes $\mathbf{z}_i$'s. This can be seen by letting $z_1 < \cdots < z_k$, and noting that $W_\rho(P, \delta_x)$ is concave for $x \in [z_i, z_{i+1}]$ for $i = 1, \ldots, k-1$. However, this fails when the dimension $d \geq 2$, as shown in the following lemma. The proof is given in Appendix D.

Lemma 5.1. *Fix $d \geq 2$, $\rho > 0$. There exists k and $\mathbf{z}_1, \ldots, \mathbf{z}_k \in \mathbb{R}^d$ such that the minimizer of $\mathbf{x} \mapsto k^{-1} \sum_{i=1}^k \|\mathbf{x} - \mathbf{z}_i\|^\rho$ is unique and does not belong to $\{\mathbf{z}_1, \ldots, \mathbf{z}_k\}$. Hence the minimizer of $\mathbf{x} \mapsto W_\rho(P, \delta_{\mathbf{x}})$ where $P = k^{-1} \sum_{i=1}^k \delta_{\mathbf{z}_i}$ does not coincide with any of the modes in P.*

Hence, Wasserstein distances fail to be mode-seeking for the more general case where p_i's are symmetric quasiconcave distributions with small variances instead of point masses.

Theorem 5.2. *Fix $d \geq 2$, $0 < \rho \leq 2$, and any class of symmetric quasiconcave distributions $\mathcal{P}$ satisfying that $\sup_{p\in\mathcal{P}} \lambda_{\max}^{1/2}(\mathbf{\Sigma}_p) =: \sigma_{\max} < \infty$ (where $\mathbf{\Sigma}_p$ is the covariance matrix of p), and for each $\mathbf{x} \in \mathbb{R}^d$, there exists $p \in \mathcal{P}$ centered at $\mathbf{x}$. For any $\beta > 0$, there exists $p_1, \ldots, p_k \in \mathcal{P}$ such that $Q := \mathrm{argmin}_{Q\in\mathcal{P}} W_\rho(P, Q)$ (where $P := k^{-1} \sum_{i=1}^k p_i$) satisfies $\min_i \|\mathbb{E}Q - \mathbb{E}p_i\|_2 > \beta$, where we write $\mathbb{E}Q := \mathbb{E}_{\mathbf{x}\sim Q}[\mathbf{x}]$.*

Informally, assuming each $p \in \mathcal{P}$ is sufficiently concentrated around its mean, if we fit a distribution Q to the mixture distribution P using Wasserstein distances, then the mean of Q can be arbitrarily far from the closest mode of P. Refer to Appendix D for the proof. In contrast, Theorem 4.3 showed that a mode-seeking divergence attains a distance from the closest mode in the order $O(\sigma_{\max})$. The only (weakly) mode-seeking Wasserstein distance is W_0, i.e. the total variation distance.

6 HYBRID OF f-DIVERGENCE AND WASSERSTEIN DISTANCE

Theorem 4.3 shows that a mode-seeking f-divergence can identify a mode when the true data distribution P is known. Nevertheless, in practice, we are only given the empirical distribution $\hat{P} := n^{-1} \sum_{i=1}^n \delta_{\mathbf{x}_i}$, where $\{\mathbf{x}_i\}$ is the data set, and $\delta_{\mathbf{x}_i}$ is the degenerate distribution at $\mathbf{x}_i$. Applying Theorem 4.3 on $\hat{P}$ instead of P shows that the optimizer Q would merely be the degenerate distribution at one of the data points, which is in some sense the "intended behavior" of a mode-seeking divergence, since each point $\mathbf{x}_i$ is a mode and is well-separated from other modes. Therefore, being "too mode-seeking" may be detrimental.

The variational formulation of f-divergence (1) in (Nowozin et al., 2016), including the vanilla GAN (Goodfellow et al., 2014), works even on empirical distributions, by restricting the function $T \in \mathcal{T}$ to be representable by a neural network. This approach requires a careful balance in training the generator and the discriminator, and may perform poorly if the discriminator is trained to optimality (Arjovsky and Bottou, 2017; Arjovsky et al., 2017). On the other hand, WGAN (Arjovsky et al., 2017) imposes a Lipschitz condition on the discriminator, allowing the discriminator to be trained to optimality when only the empirical distribution is known.

The hybrid of f-divergence and Wasserstein distance (3) in (Farnia and Tse, 2018) retains the advantage of WGAN. We now show that the hybrid divergence can be applied on empirical distributions while retaining the mode-seeking behavior of f-divergence. We present an informal version of the theorem, which states that as long as $\lambda = O(\sigma_{\max})$, we have $\min_i \|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_i\|_2 = O(\sigma_{\max})$ with high probability. The formal theorem and the proof is in Appendix E.

Theorem 6.1. *(Informal) Consider the hybrid divergence $D_{\lambda f, W_1}$, where the f-divergence D_f is weakly mode-seeking. Let $\mathcal{P}$ be a set of symmetric quasiconcave distributions. Define $P(\mathbf{x}) := \sum_{i=1}^k w_i p_i(\mathbf{x})$ and $\sigma_{\max}$ as in Theorem 4.3. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} P$, and $\hat{P} := n^{-1} \sum_{i=1}^n \delta_{\mathbf{x}_i}$ be the empirical distribution. Let $Q :=$*

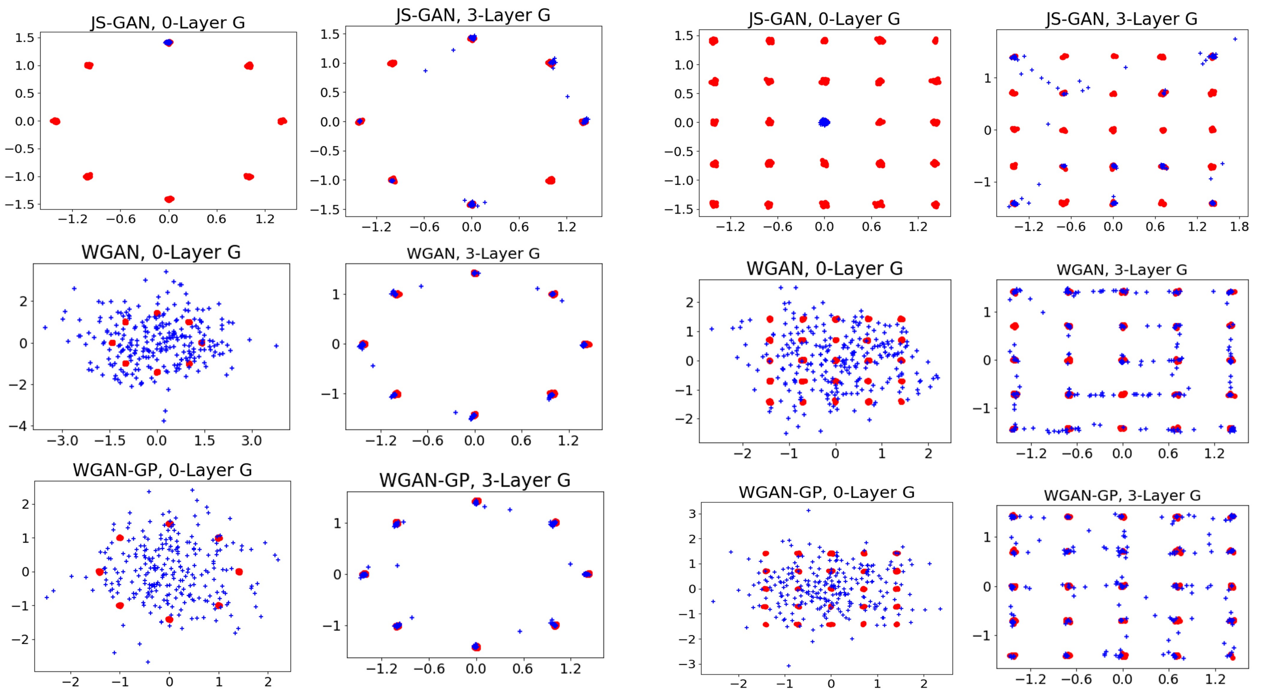

Figure 3: Samples generated by the trained generator (colored blue) and the original training data for the 8 and 25-component Gaussian mixture model (colored red). Rows 1-3 show the samples generated by the VGAN, WGAN-WC, and WGAN-GP.

$\text{argmin}_{Q\in\mathcal{P}} D_{\lambda f,W_1}(\hat{P}\,\|\,Q)$, *and denote its center as* $\boldsymbol{\mu}_Q$. *For fixed* d, k *and* $\zeta > 0$, *if* $\lambda = O(\sigma_{\max})$, *then*

$$\mathbb{P}\left(\min_i \|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_i\|_2 \geq k\sigma_{\max}\zeta\right) \leq O\Big(\lambda^{-1}\left(\mathbb{E}\left[\|\mathbf{x}-\mathbb{E}[\mathbf{x}]\|_2^3\right]\right)^{1/3} G_d(n)\Big),$$

where $\mathbf{x} \sim P$, *and* $G_d(n) := n^{-1/\max\{2,d\}}$ *if* $d \neq 2$, $G_d(n) := n^{-1/2}\log(1+n)$ *if* $d = 2$.

The term $G_d(n)$ comes from the sample size needed to estimate a distribution within a small Wasserstein distance (Fournier and Guillin, 2015), which grows exponentially with the dimension d. The curse of dimensionality is inevitable uness a stronger assumption is made on $\mathcal{P}$.

7 NUMERICAL EXPERIMENTS

In this section, we present the results of our numerical experiments on applying the discussed GAN problems to learn Gaussian mixture models and image data distributions. The numerical experiments have been performed over the following datasets that are used as benchmark cases in the literature: 1) An 8-component Gaussian mixture dataset adapted from Gulrajani et al. (2017) with the modes centered around the vertices of a regular 8-sided polygon. The standard deviation parameter of every isotropic Gaussian mode is set to be the 0.02, 2) A 25-component Gaussian mixture dataset adapted from Gulrajani et al. (2017) with modes centered around a two-dimensional 5×5-grid with a unit column and row size and the standard deviation of 0.05, 3) CIFAR-10 dataset Krizhevsky and Hinton (2009) including 50,000 training samples with ten labels.

We performed the numerical experiments using the following GAN formulations: 1) the Vanilla GAN Goodfellow et al. (2014) with no regularization which targets the JS divergence, 2) the KL-GAN which targets the KL divergence, 3) the ReverseKL-GAN Nowozin et al. (2016) which targets the reverse-KL divergence 4) the Wasserstein GAN implemented via weight clipping Arjovsky et al. (2017) and gradient penalty Gulrajani et al. (2017) targeting the 1-Wasserstein distance, 5) Spectrally-normalized VGAN (SN-GAN) Miyato et al. (2018) which uses the spectral normalization on the discriminator neural net's layers to ensure the discriminator is a K-Lipschitz function with the value of K determined by the product of the spectral norms of the neural net's layers. The Lipschitz VGAN targets the hybrid divergence $D_{(1/K)f,W_1}$. We defer the detailed description of the numerical settings to the Appendix A.

7.1 Different divergence measures in learning multi-modal Gaussian Data

In our numerical experiments for the 8-component and 25-component Gaussian mixture data, we consistently observed that the JS divergence in the VGAN led to a mode seeking fit of the underlying Gaussian mixture in comparison to the

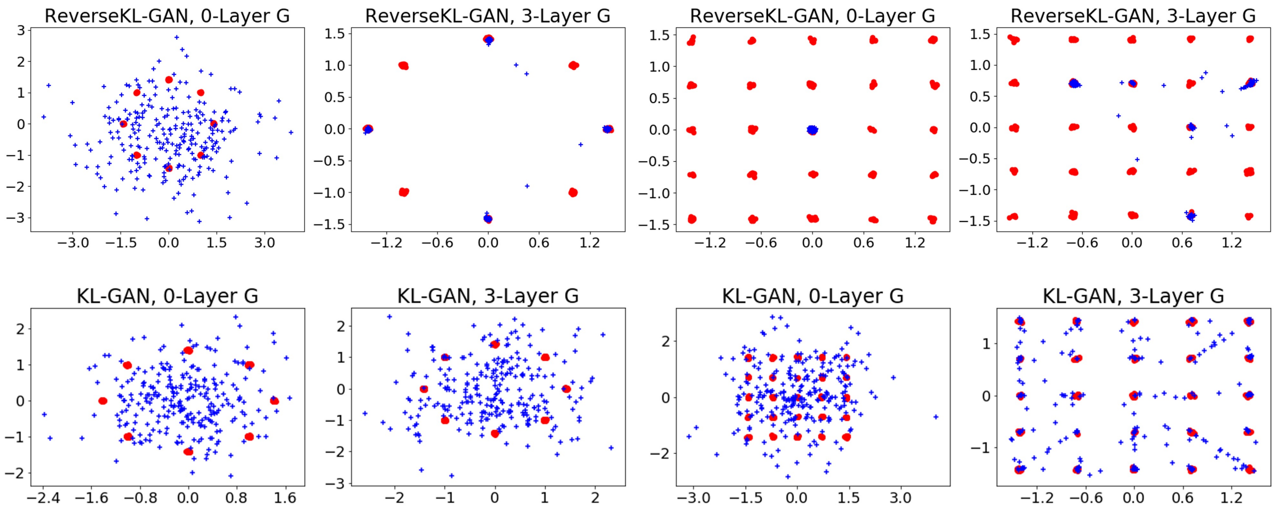

Figure 4: Samples generated by the trained generator (colored blue) and the original training data for the 8 and 25-component Gaussian mixture model (colored red). Rows 1 and 2 show the samples generated by the ReverseKL-GAN and KL-GAN.

1-Wasserstein distance in WGAN and WGAN-GP. Figure 3 shows the samples generated by the trained generator (in blue) and the original training data (in red). The empirical results suggest that the vanilla GAN with no regularization tends to fit only one of the existing Gaussian modes, when the generator is an affine map and produces only one Gaussian mode. The number of captured modes are increasing with the number of layers in the generator network. On the other hand, both of the standard implementations of the Wasserstein GAN displayed a mode covering tendency. In the Appendix Table 2, we report the log-likelihood scores of the generated samples indicating the lower quality of WGAN samples than VGANs. For an affine generator mapping, the trained WGANs covered all the modes which led to lower-quality samples. For the generators with greater depths, although WGANs captured all the modes, they still generated lower quality samples compared to VGAN, suggesting the mode covering nature of Wasserstein distances.

Furthermore, we applied the ReverseKL-GAN minimizing the Reverse-KL divergence and the KL-GAN targeting the KL divergence to the same Gaussian mixture datasets. As shown in Figure 4, the trained KL-GAN did not demonstrate a mode-seeking behavior, while the ReverseKL-GAN behaved in a mode seeking fashion in the numerical experiments. The numerical observations were consistent with our theoretical results in Theorems 4.3 and 5.2. For the complete results of the experiments including the results for different generator network's depth, we refer the readers to the Appendix A – Figures 6, 7, 8, 9.

7.2 Hybrid-divergence in Lipschitz GANs

In another set of experiments, we tested the Lipschitz VGAN problem with different Lipschitz coefficients in fitting mixture models. In our experiments, we simulated different Lipschitz coefficients by altering the spectral norm of the neural net's layers in $\{1, 2, 3, 4\}$. As illustrated in Figure 5, the higher Lipschitz constant 4.0 resulted in a mode-seeking fit of the underlying mixture model, while the lower Lipschitz constant 1.0 led to a mode-covering fit of the underlying distribution. As the experiment suggests, the Lipschitz constant hyperparameter allows the VGAN learner to adjust the mode seeking power of the divergence measure. For the complete set of our numerical results, we refer the reader to Figures 10 and 11 in the Appendix A.

Finally, we trained the VGAN, WGAN, WGAN-GP and SN-VGAN on the CIFAR-10 dataset and measured the sharpness and diversity components of the Inception score (Salimans et al., 2016) to evaluate the effect of the underlying divergence measure on the quality and diversity of the generated samples. Given sample X and the pre-trained Inception-net's output Y, the Inception score is defined as

$$\begin{aligned}\mathrm{IS}(P_{X,Y}) &:= \exp\Big(\mathbb{E}[D_{\mathrm{KL}}(p(Y|X)\|p(Y))]\Big)\\ &= \exp\big(H(Y)\big)\,\exp\big(-H(Y|X)\big),\end{aligned}$$

where $H(\cdot)$ denotes the Shannon entropy. Hence, $\exp(H(Y))$ can be interpreted as a measure of the diversity of generated data, while $\exp(-H(Y|X))$ can be interpreted as a sharpness score. In our CIFAR-10 experiments, the (sharpness,diversity) scores for the generated samples were (0.75, 9.23) for the VGAN, (0.73,9.61) for the Lipschitz VGAN, (0.49,9.72) for the WGAN, and (0.66,9.70) for the WGAN-GP. As suggested by the evaluated scores, the WGAN formulation seems to attain higher diversity at the cost of lower sharpness, while the VGAN achieved higher quality while leading to a lower diversity score. On the other hand, the Lipschitz VGAN managed to balance

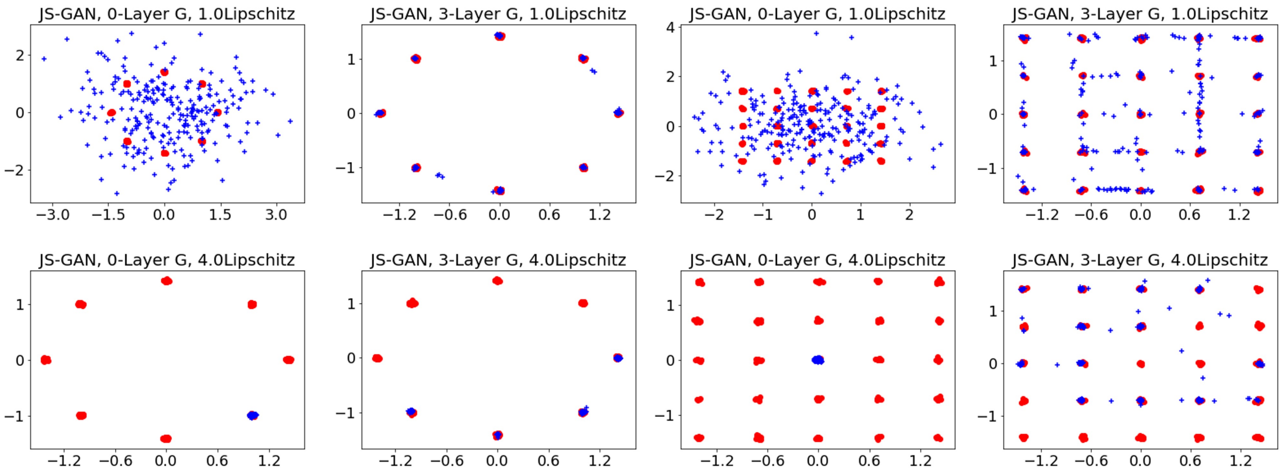

Figure 5: Samples generated by the Lipschitz Vanilla GAN with different Lipschitz constants (colored blue) and the Gaussian mixture data (colored red). Rows 1,2 show the samples generated with the spectral norms 1,4 for the discriminator's layers.

the quality and sharpness scores, which also resulted in the maximum product of the two scores that is the Inception score. We defer the generated CIFAR-10 samples of the trained generators to the Appendix A.

8 CONCLUSION

In this paper, we provided a unified theoretical framework for mode-seeking f-divergences and their hybrid with Wasserstein distances. According to this framework, we analyzed the divergence minimizing solution of fitting a unimodal distribution to a multi-modal underlying model. Our analysis reveals simple conditions on a convex function f, under which the corresponding f-divergence results in fitting an existing mode in the underlying mixture model. In addition, we supported our theoretical findings through several numerical results on standard Gaussian mixture models.

We note that our theoretical and numerical analysis suggests several future directions. Since our analysis focuses on mode-seeking divergence measures, an interesting future direction is to create a similar theoretical framework for mode-covering distances which applies to Wasserstein distances. In addition, our numerical experiments mostly focus on synthetic Gaussian mixture models as it offers prior knowledge of the ground-truth model. However, the multi-modal distribution of standard image datasets is typically unknown and is formed by several unknown hidden factors. A future extension is to develop an empirical methodology for counting the number of existing modes in an image generative model and its dependency on the choice of fitting divergence measure.

Another future direction is to analyze the local optima of the divergence minimization problem. While this paper focuses on the global optimum, local optima are relevant to practical implementations with gradient-based optimization algorithms. We finally note that the theoretical results in this paper focus on fitting a unimodal model distribution to a multimodal data distribution. We may also investigate the implication of mode-seeking divergences in fitting a multimodal model distribution, either in a theoretical setting or in practical algorithms such as (Gurumurthy et al., 2017; Khayatkhoei et al., 2018).

Acknowledgements

The work of Cheuk Ting Li was supported in part by the Hong Kong Research Grant Council Grant ECS No. CUHK 24205621. The work of Farzan Farnia was supported by Hong Kong Research Grant Council Grant GRF No. CUHK 14209920. The authors would also like to thank the anonymous reviewers for their constructive feedback.

References

An, D., Guo, Y., Lei, N., Luo, Z., Yau, S.-T., and Gu, X. (2019). AE-OT: a new generative model based on extended semi-discrete optimal transport. *ICLR 2020*.

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*. Springer Science+Business Media.

Borji, A. (2022). Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329.

Bousquet, O., Gelly, S., Tolstikhin, I., Simon-Gabriel, C.-J., and Schoelkopf, B. (2017). From optimal transport to generative modeling: the VEGAN cookbook. *arXiv preprint arXiv:1705.07642*.

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

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

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.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738.

Ghasemipour, S. K. S., Zemel, R., and Gu, S. (2020). A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, pages 1259–1277. PMLR.

Goodfellow, I. (2016). NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.

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.

Gurumurthy, S., Kiran Sarvadevabhatla, R., and Venkatesh Babu, R. (2017). Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 166–174.

Hernandez-Lobato, J., Li, Y., Rowland, M., Bui, T., Hernández-Lobato, D., and Turner, R. (2016). Black-box alpha divergence minimization. In *International Conference on Machine Learning*, pages 1511–1520. PMLR.

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

Ke, L., Choudhury, S., Barnes, M., Sun, W., Lee, G., and Srinivasa, S. (2020). Imitation learning as f-divergence minimization. In *International Workshop on the Algorithmic Foundations of Robotics*, pages 313–329. Springer.

Khayatkhoei, M., Singh, M. K., and Elgammal, A. (2018). Disconnected manifold learning for generative adversarial networks. *Advances in Neural Information Processing Systems*, 31.

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

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

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

Kurach, K., Lučić, M., Zhai, X., Michalski, M., and Gelly, S. (2019). A large-scale study on regularization and normalization in gans. In *International Conference on Machine Learning*, pages 3581–3590. PMLR.

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.

Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1081–1089.

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.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802.

McCann, R. J. (1999). Exact solutions to the transportation problem on the line. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 455(1984):1341–1380.

Minka, T. (2005). Divergence measures and message passing. Technical report, Microsoft Research.

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

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.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training GANs. *Advances in neural information processing systems*, 29:2234–2242.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94.

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*.

Taghvaei, A. and Jalali, A. (2019). 2-Wasserstein approximation via restricted convex potentials with application to improved training for GANs. *arXiv preprint arXiv:1902.07197*.

Theis, L., van den Oord, A., and Bethge, M. (2015). A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*.

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

Wang, D., Liu, H., and Liu, Q. (2018). Variational inference with tail-adaptive f-divergence. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5742–5752.

Williams, W., Ringer, S., Ash, T., Hughes, J., MacLeod, D., and Dougherty, J. (2020). Hierarchical quantized autoencoders. *arXiv preprint arXiv:2002.08111*.

Zhang, M., Bird, T., Habib, R., Xu, T., and Barber, D. (2019). Variational f-divergence minimization. *arXiv preprint arXiv:1907.11891*.

Zhou, Z., Liang, J., Song, Y., Yu, L., Wang, H., Zhang, W., Yu, Y., and Zhang, Z. (2019). Lipschitz generative adversarial nets. In *International Conference on Machine Learning*, pages 7584–7593. PMLR.

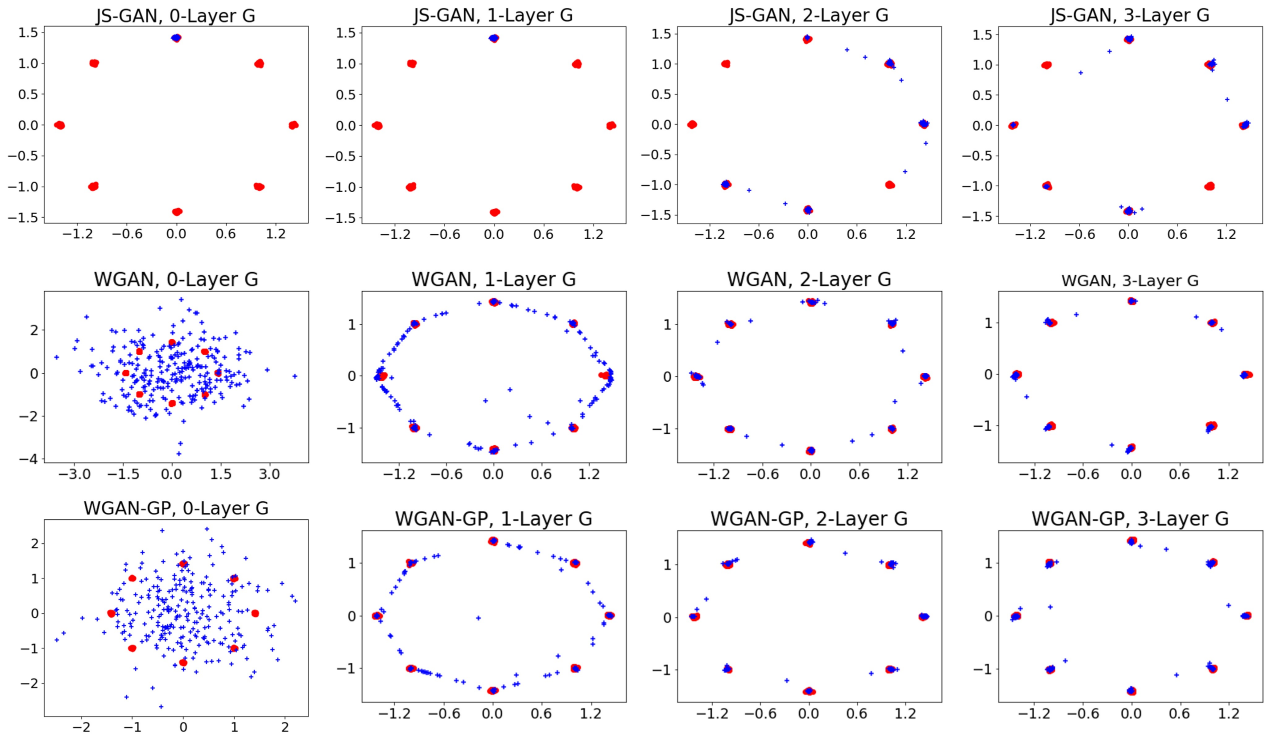

Figure 6: Plot of the samples generated by the trained generator (in blue color) and the original training data for the 8-component Gaussian mixture (in red color). Rows 1,2,3 show samples generated by the VGAN, WGAN-WC, and WGAN-GP respectively, while the generators in columns 1,2,3,4 include 0,1,2,3 hidden layers.

A APPENDIX – DETAILED DESCRIPTION OF THE NUMERICAL SETUP & ADDITIONAL NUMERICAL RESULTS

Regarding the numerical experiments, for the generator and discriminator architectures, in the experiments on the Gaussian mixture data we used a 3 hidden-layer multilayer perceptron (MLP) neural net discriminator with 64 ReLU ($\mathrm{ReLU}(z) = \max\{z, 0\}$) neurons per layer. To simulate generators with different capacities, we experimented four different multilayer perceptron neural networks with the following number of ReLU-based hidden layers: 0,1,2,3. Note that the network with zero layers in fact represents an affine map from the hidden space to the sample space. In all the Gaussian mixture experiments, we used a two-dimensional latent variable $\mathbf{Z} \in \mathbb{R}^2 \sim \mathcal{N}(\mathbf{0}, I_2)$ with an isotropic normal distribution. In the case of CIFAR-10 experiments, we used the standard 4-layer architecture of DCGAN for both the generator and discriminator.

For the optimization of generator and discriminator parameters, we used the ADAM optimizer (Kingma and Ba, 2014) for 200,000 generator iterations. We applied 5 discriminator ADAM updates per generator iteration. For the SN-GAN experiments, we used the standard implementation of spectral normalization in (Miyato et al., 2018) that is based on the power method for computing the layers' operator norm.

For the complete set of the experimental results in Figures 3 and 5, we refer the readers to Figures 6, 7, 8, 9, 10, 11. Also, to have a quantitative comparison between the models learned in the Gaussian mixture settings, we report the averaged log-likelihood of the generated samples based on the true distribution of the 8 and 25 Gaussian mixture models in Table 2. Our numerical results suggest the superiority of the models learnt by minimizing the mode-seeking Reverse-KL, JS and the hybrid divergences in producing higher quality samples.

B APPENDIX – PLOTS FOR VARIOUS f-DIVERGENCES

In Figure 13, we plot the function f for various f-divergences. For the sake of comparison, we plot $\alpha f(t) + \beta(t-1)$ instead of f, where α, β are chosen such that $\lim_{t\to\infty} f(t)/t = 0$ and the left derivative of f is -1 at $t = 1$ for mode-seeking

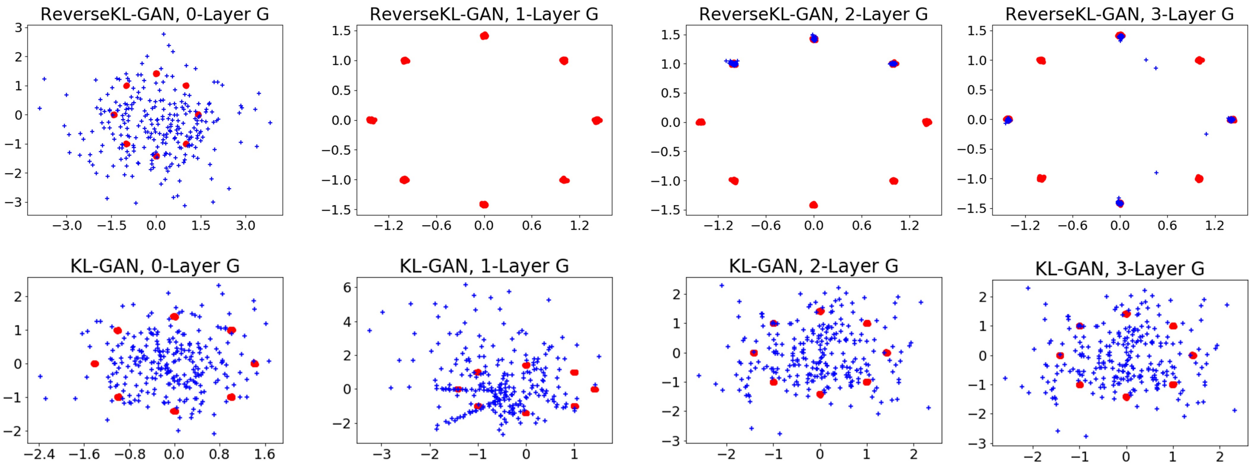

Figure 7: Plot of the samples generated by the trained generator (in blue color) and the original training data for the 8-component Gaussian mixture (in red color). The upper and lower rows show samples generated by the Reverse-KL-GAN and KL-GAN, respectively, while the generators in columns 1,2,3,4 include 0,1,2,3 hidden layers.

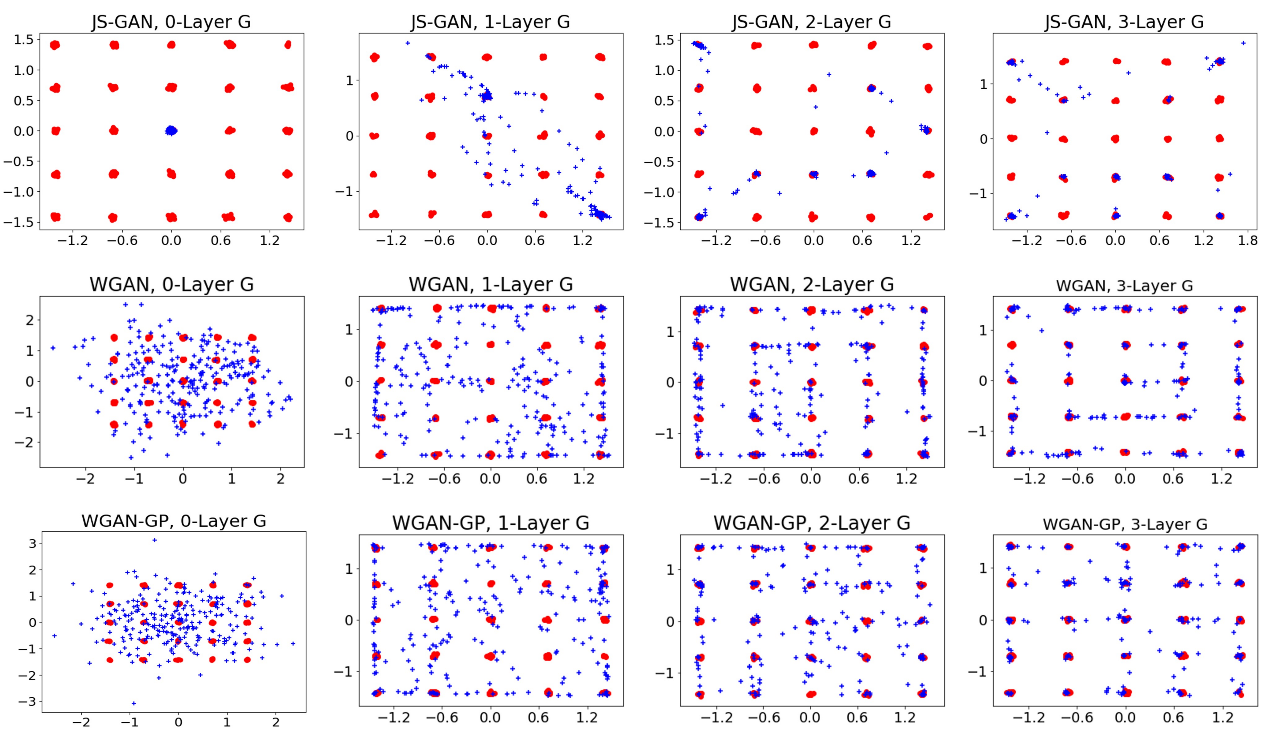

Figure 8: Samples generated by the trained generator (colored blue) and the original training data for the 25-component Gaussian mixture model (colored red). Rows 1,2,3 show samples generated by the VGAN, WGAN-WC, and WGAN-GP, respectively, while the generators in columns 1,2,3,4 include 0,1,2,3 hidden layers.

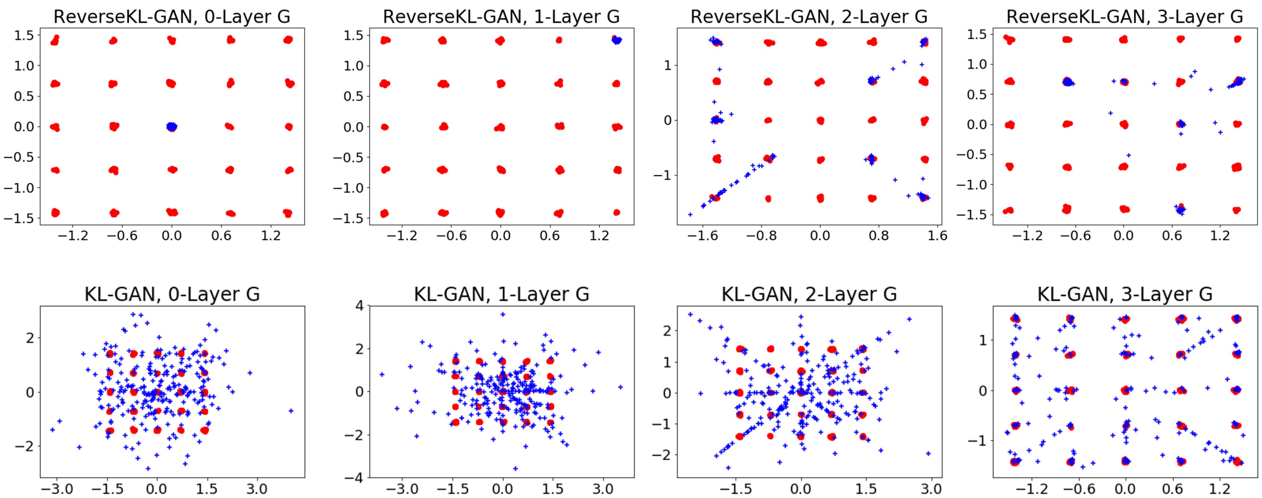

Figure 9: Samples generated by the trained generator (colored blue) and the original training data for the 25-component Gaussian mixture model (colored red). The upper and lower rows show samples generated by the Reverse-KL-GAN and KL-GAN, respectively, while the generators in columns 1,2,3,4 include 0,1,2,3 hidden layers.

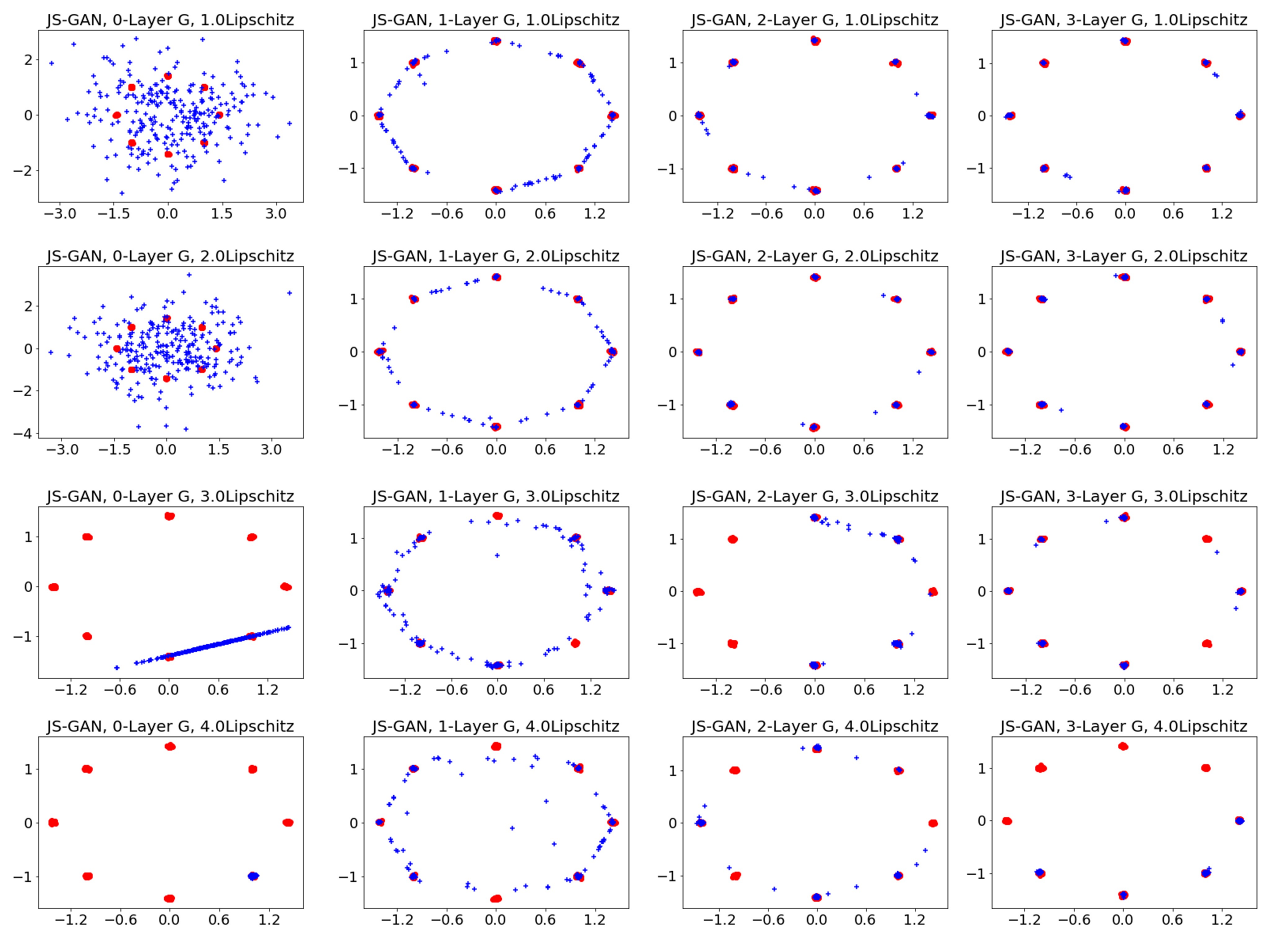

Figure 10: Samples generated by the Lipschitz Vanilla GAN (colored blue) and the training data for the 8-component Gaussian mixture (colored red). The rows show samples generated using the spectral norm values 1,2,3,4 for the discriminator network's layers, and the generators in columns 1,2,3,4 have 0,1,2,3 hidden layers.

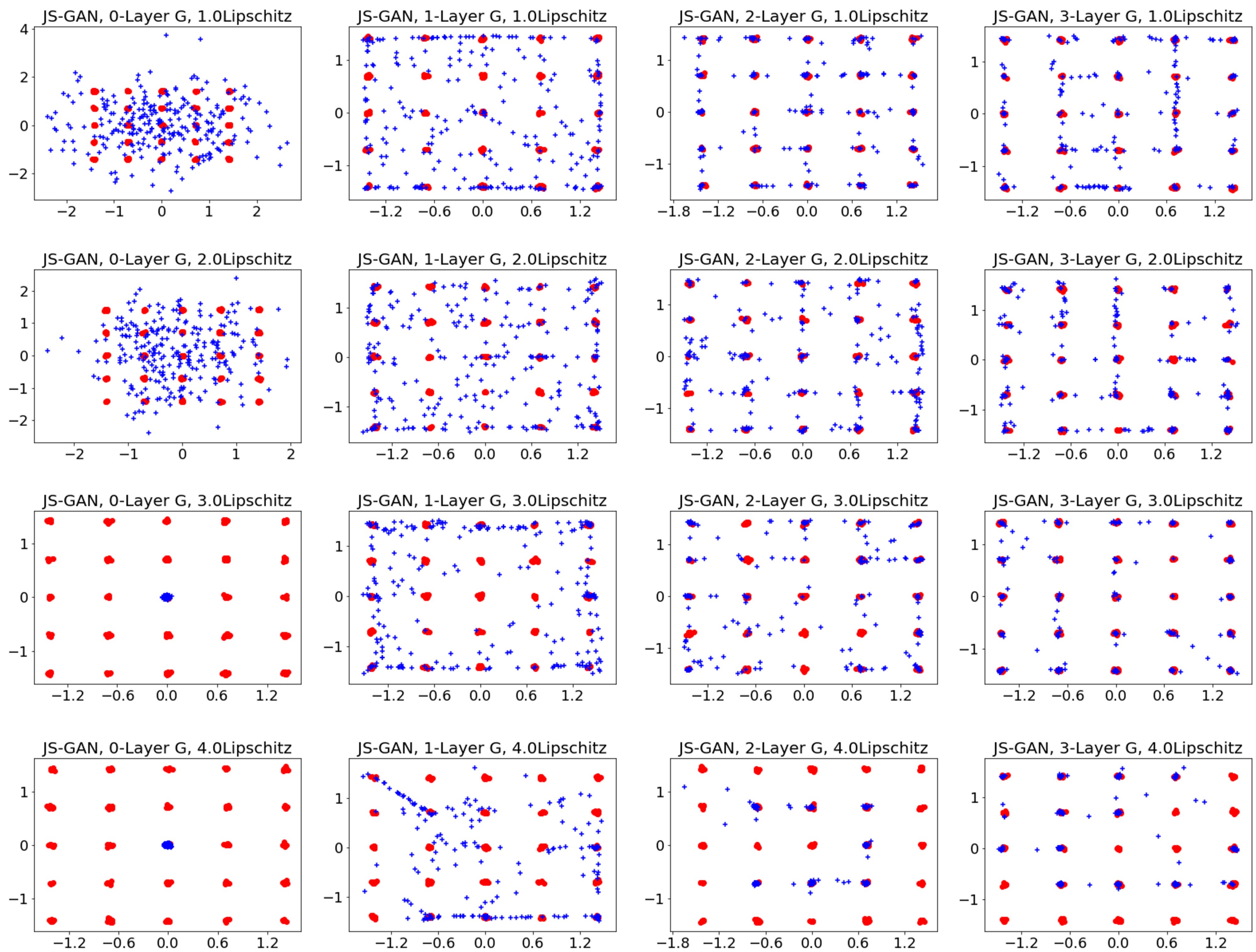

Figure 11: Samples generated by the Lipschitz Vanilla GAN (colored blue) and the training data for the 25-component Gaussian mixture model (colored red). The rows show samples generated using the spectral norm values 1,2,3,4 for the discriminator network's layers, and the generators in columns 1,2,3,4 have 0,1,2,3 hidden layers.

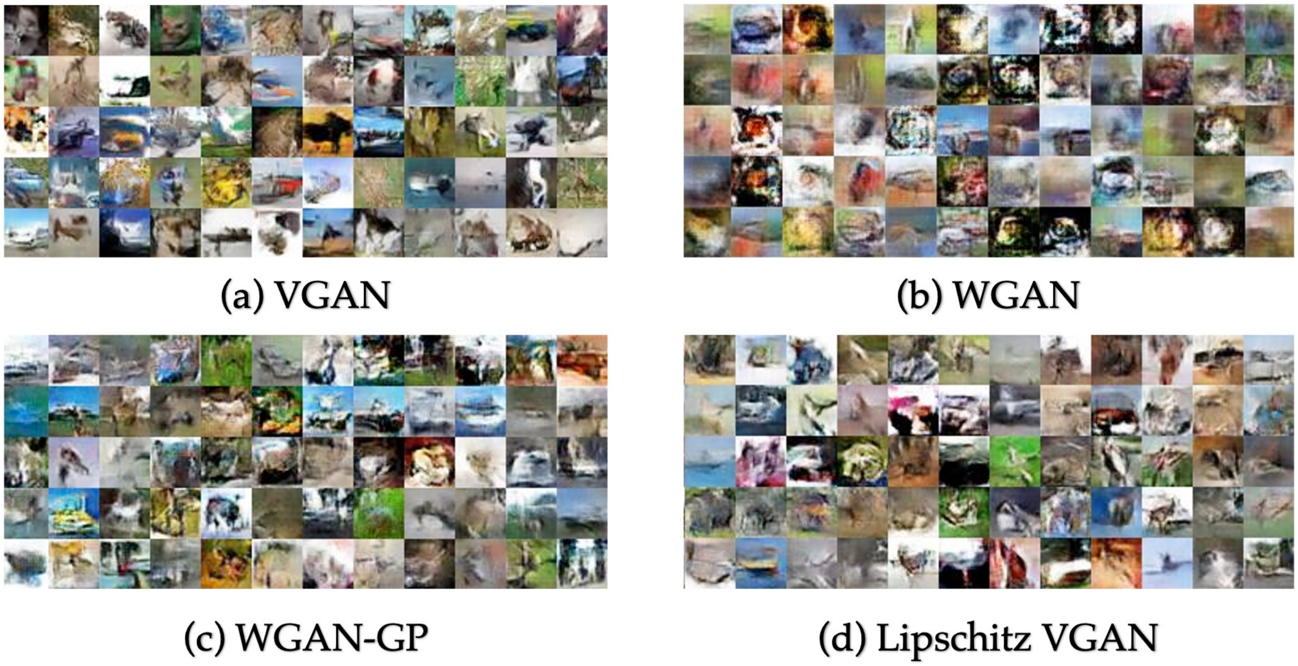

Figure 12: CIFAR-10 samples generated by the trained generator of the VGAN, WGAN, WGAN-GP, spectrally-normalized VGAN.

divergences, and we choose $\alpha = 1$ and β such that the left derivative is -1 for mode-covering divergences. The most mode-seeking divergences are Neyman χ^2, softened reverse KL and $\mathcal{G}_{\mathrm{ALT}}$ divergence, where f is lower-bounded by a constant, and the mode-seeking order is $O(\gamma^{1/3})$. The functions f for mode-covering divergences (KL and Pearson χ^2) grow faster than linearly. While Jeffreys divergence also grows faster than linearly (it does not satisfy the definition of weakly mode-seeking), it is unclear whether it should be considered as mode-covering.

Figure 14 is a more complete version of Figure 2 for the mixture data distribution $P = 0.75\mathcal{N}(0,1) + 0.25\mathcal{N}(\delta, 1)$, where we also include the softened reverse KL divergence (Shannon et al., 2020), $\mathcal{G}_{\mathrm{ALT}}$ divergence (Poole et al., 2016), and Jeffreys divergence. While Jeffreys divergence does not satisfy the definition of weakly mode-seeking in this paper, it appears to have a weakly mode-seeking behavior similar to total variation distance in this example.

C APPENDIX – PROOF OF THEOREM 4.3

Before we prove Theorem 4.3, we show the following results about symmetric quasiconcave distributions.

Proposition C.1. *Let $\mathbf{x} \in \mathbb{R}^d$ be a random vector with a symmetric quasiconcave distribution centered at 0, and $\mathbf{a} \in \mathbb{R}^d \backslash \{0\}$. We have*

1. *$\mathbf{a}^T\mathbf{x}$ also has a symmetric quasiconcave distribution centered at 0 (i.e., $\mathbf{a}^T\mathbf{x}$ is symmetric and unimodal).*
2. *For $t \geq 0$,*
$$\begin{aligned}\mathbb{P}(\mathbf{a}^T\mathbf{x} \geq t) &\leq \frac{1}{2}\max\left\{1 - \frac{t}{9\sqrt{\mathbb{E}[(\mathbf{a}^T\mathbf{x})^2]}}, \frac{1}{3}\right\} \\ &\leq \frac{1}{2}\max\left\{1 - \frac{t}{9\|\mathbf{a}\|_2\sqrt{\lambda_{\max}(\mathbf{\Sigma})}}, \frac{1}{3}\right\},\end{aligned}$$
where $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{x}$.
3. *For $t, r > 0$,*
$$\mathbb{P}(\mathbf{a}^T\mathbf{x} \in [t - r, t + r]) \leq \frac{r}{t}.$$

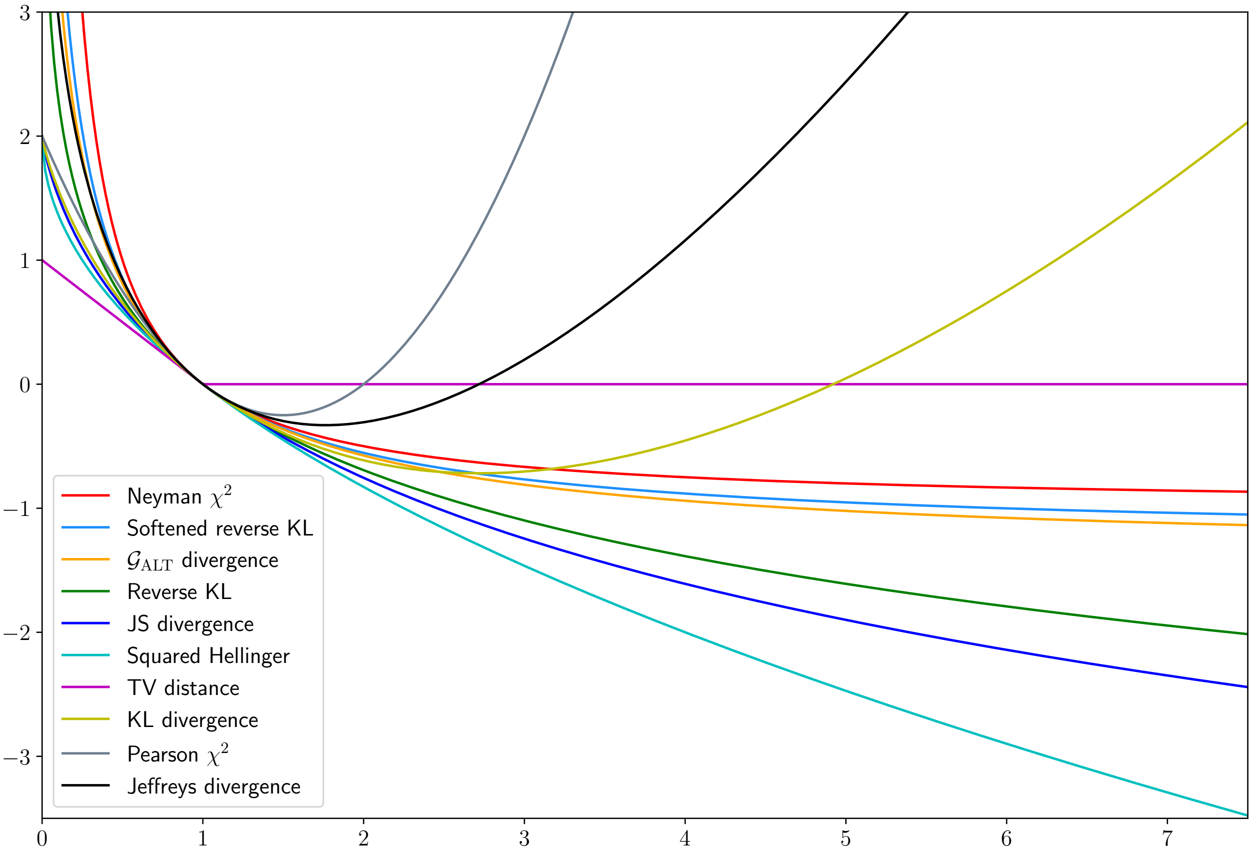

Figure 13: Plot of the function f for various f-divergences.

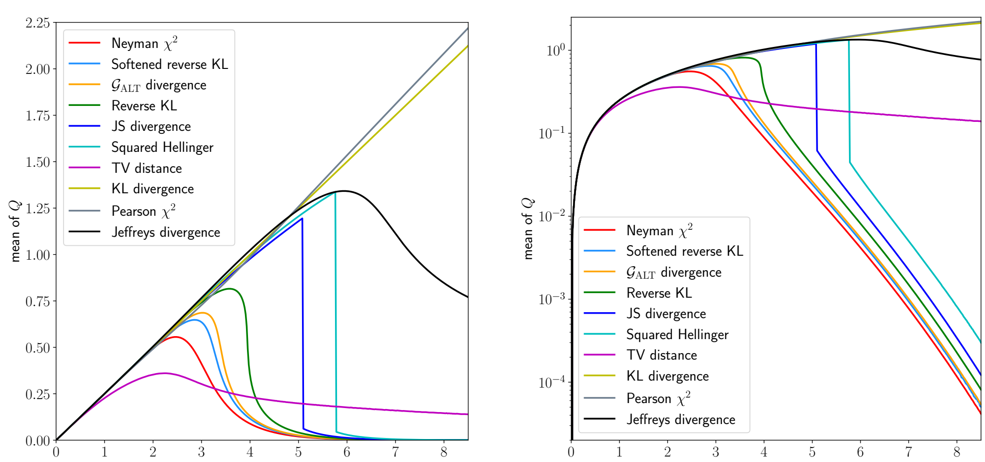

Figure 14: Plot of the center of $\mathrm{argmin}_{Q\ \mathrm{Gaussian}} D_f(P\|Q)$ for various f-divergences, where the ground truth $P = 0.75\mathcal{N}(0,1) + 0.25\mathcal{N}(\delta,1)$ is a mixture of 2 Gaussian distributions, where $\delta \geq 0$ is the separation between the two modes. We plot the center of Q against δ (left: linear scale, right: log-scale).

8-comp GMM Non-Hybrid Divergences				
GAN / Gen. Layers	0	1	2	3
JS-GAN	2.01	2.05	1.92	1.88
WGAN	-0.05	0.78	1.51	1.86
WGAN-GP	0.16	0.89	1.78	1.89
KL-GAN	0.10	-0.4	0.03	0.15
Reverse-KL	0.18	2.02	2.05	1.98
8-comp GMM: JS-W_1 Hybrid Divergence				
Lip. Cons. / Gen. Layers	0	1	2	3
1.0	0.16	0.98	1.56	1.88
2.0	0.28	0.94	1.63	1.92
3.0	0.48	0.90	1.57	1.96
4.0	2.05	1.34	1.78	2.00

25-comp GMM: Non-Hybrid Divergences				
GAN / Gen. Layers	0	1	2	3
JS-GAN	3.15	2.34	2.89	2.78
WGAN	0.31	1.56	1.48	1.89
WGAN-GP	0.24	1.23	1.64	2.27
KL-GAN	0.17	0.28	0.25	1.87
Reverse-KL	3.04	3.01	2.41	2.24
25-comp GMM: JS-W_1 Hybrid Divergence				
Lip. Cons. / Gen. Layers	0	1	2	3
1.0	0.29	1.20	2.21	2.14
2.0	0.12	1.12	2.08	2.43
3.0	3.00	1.41	1.95	2.74
4.0	3.02	1.61	2.78	2.89

Table 2: Averaged normalized log-likelihood of GANs' generated samples

Proof. First we show that $\mathbf{a}^T\mathbf{x}$ has a symmetric quasiconcave distribution centered at 0. We first consider the case $\mathbf{x} \sim \mathrm{Unif}(A)$, where $A \subseteq \mathbb{R}^d$ is a convex set with finite positive volume that is symmetric around 0. We also assume $\mathbf{a} = [1, 0, \ldots, 0]$ without loss of generality. Assume $\mathbf{x} \sim \mathrm{Unif}(A)$. Write $A_t := \{\mathbf{z} \in \mathbb{R}^{d-1} : [t, \mathbf{z}] \in A\}$ for the cross section of A. Since A is convex, we have

$$\frac{t+s}{2t}A_t + \frac{t-s}{2t}A_{-t} \subseteq A_s$$

for $0 \le s < t$, where the "+" stands for Minkowski sum. By Brunn-Minkowski theorem and that $A_{-t} = -A_t$ (since A is symmetric around 0),

$$\begin{aligned}
\mathrm{Vol}(A_s) &\ge \left(\mathrm{Vol}^{1/d}\left(\frac{t+s}{2t}A_t\right) + \mathrm{Vol}^{1/d}\left(\frac{t-s}{2t}A_{-t}\right)\right)^d \\
&= \left(\frac{t+s}{2t}\mathrm{Vol}^{1/d}(A_t) + \frac{t-s}{2t}\mathrm{Vol}^{1/d}(A_t)\right)^d \\
&= \mathrm{Vol}(A_t).
\end{aligned}$$

Therefore $\mathrm{Vol}(A_t)$ is non-increasing for $t \ge 0$. The result follows from that the probability density function of $\mathbf{a}^T\mathbf{x}$ is $t \mapsto \mathrm{Vol}(A_t)/\mathrm{Vol}(A)$.

Consider the general case where $\mathbf{x}$ has a symmetric quasiconcave probability density function p. For $\alpha > 0$, since the superlevel set $L_\alpha^+ := \{\mathbf{x} \in \mathbb{R}^d : p(\mathbf{x}) \ge \alpha\}$ has finite volume and is convex and symmetric around 0, when $\mathbf{x} \sim \mathrm{Unif}(L_\alpha^+)$, the density function of $\mathbf{a}^T\mathbf{x}$ (let it be q_α) is symmetric quasiconcave and centered at 0. Note that we can generate $\mathbf{x} \sim p$ by first generating α according to the probability density function $\alpha \mapsto \mathrm{Vol}(L_\alpha^+)$, and then generating $\mathbf{x} \sim \mathrm{Unif}(L_\alpha^+)$. Therefore, when $\mathbf{x} \sim p$, the density function of $\mathbf{a}^T\mathbf{x}$ is $\int_0^\infty \mathrm{Vol}(L_\alpha^+)q_\alpha(x)\mathrm{d}\alpha$, which is also symmetric quasiconcave (since it is non-increasing for $x \ge 0$) and centered at 0.

For the second claim, let $z = |\mathbf{a}^T\mathbf{x}|$, and let its probability density function be $p : [0, \infty) \to \mathbb{R}$. Then p is a non-increasing function. We have

$$\mathbb{E}[z^2\mathbf{1}\{z < t\}] \ge \frac{t^3}{3}p(t).$$

And

$$\mathbb{E}[z^2\mathbf{1}\{z \ge t\}] \ge \frac{1}{3}\left(\left(\frac{\mathbb{P}(z \ge t)}{p(t)} + t\right)^3 - t^3\right)p(t).$$

Hence,

$$\begin{aligned}\mathbb{E}[z^2] &\geq \frac{1}{3}\left(\frac{\mathbb{P}(z \geq t)}{p(t)} + t\right)^3 p(t) \\ &\geq \frac{1}{3}\frac{(\mathbb{P}(z \geq t))^3}{(p(t))^2} \\ &\geq \frac{1}{3}\frac{(\mathbb{P}(z \geq t))^3}{((1-\mathbb{P}(z \geq t))/t)^2} \\ &= \frac{t^2}{3} \cdot \frac{(\mathbb{P}(z \geq t))^3}{(1-\mathbb{P}(z \geq t))^2}.\end{aligned}$$

If $\mathbb{P}(z \geq t) \geq 1/3$, then

$$\mathbb{E}[z^2] \geq \frac{t^2}{81} \cdot \frac{1}{(1-\mathbb{P}(z \geq t))^2},$$

$$\mathbb{P}(z \geq t) \leq 1 - \frac{t}{9\sqrt{\mathbb{E}[z^2]}}.$$

Hence we have

$$\mathbb{P}(z \geq t) \leq \max\left\{1 - \frac{t}{9\sqrt{\mathbb{E}[z^2]}}, \frac{1}{3}\right\}.$$

For the third claim, assume $\mathbf{a} = [1, 0, \ldots, 0]$ and $t = 1$ without loss of generality. It suffices to consider the case where $\mathbf{x} \sim \mathrm{Unif}(A)$, where $A \subseteq \mathbb{R}^d$ is a convex set with finite positive volume that is symmetric around 0. For $r < 1$,

$$\begin{aligned}&\mathrm{Vol}(A \cap \{\mathbf{z} : 1 - r \leq z_1 \leq 1 + r\}) \\ &= \int_{1-r}^{1+r} \mathrm{Vol}(A_s)\,\mathrm{d}s \\ &\leq \frac{2r}{1+r}\int_0^{1+r} \mathrm{Vol}(A_s)\,\mathrm{d}s \\ &\leq \frac{r}{1+r}\mathrm{Vol}(A) \\ &\leq r\mathrm{Vol}(A).\end{aligned}$$

Clearly $\mathrm{Vol}(A \cap \{\mathbf{z} : 1 - r \leq z_1 \leq 1 + r\}) \leq r\mathrm{Vol}(A)$ also holds for $r \geq 1$. The result follows. $\square$

We now prove Theorem 4.3.

Proof of Theorem 4.3. We use the notation $D_f(s\|t) = D_f(\mathrm{Bern}(s)\|\mathrm{Bern}(t))$ where $\mathrm{Bern}(s)$ denotes the Bernoulli distribution with parameter s. Assume D_f is weakly mode-seeking. By condition MS1, we can let

$$f_2(x) = f(x) - (x-1)\lim_{t\to\infty}\frac{f(t)}{t},$$

which is a nonincreasing function with $\lim_{t\to\infty} f_2(t)/t = 0$. We have $D_f(P\|Q) = D_{f_2}(P\|Q)$. Therefore, without loss of generality, we can assume f is convex and nonincreasing with $\lim_{t\to\infty} f(t)/t = 0$. By condition MS2, there does not exist $s \in (0, 1)$ such that $f(t)$ is constant for $t \in [s, \infty)$. Since f is convex and nonincreasing, we know that $f(t)$ is strictly decreasing for $t \in [0, 1]$. Let the center of Q be $\boldsymbol{\mu} = \boldsymbol{\mu}_Q$. Let

$$w_{\max} := \max_i w_i,$$

and assume $w_{\max} = w_{i^*}$. Note that

$$\begin{aligned}&D_f\big(P \,\big\|\, p_{i^*}\big)\\&= \int f\left(\frac{P(\mathbf{x})}{p_{i^*}(\mathbf{x})}\right) p_{i^*}(\mathbf{x})\mathrm{d}\mathbf{x}\\&\le \int f\left(\frac{w_{i^*} p_{i^*}(\mathbf{x})}{p_{i^*}(\mathbf{x})}\right) p_{i^*}(\mathbf{x})\mathrm{d}\mathbf{x}\\&= f(w_{\max}).\end{aligned}$$

Let Q be a symmetric quasiconcave distribution with center $\boldsymbol{\mu} = \boldsymbol{\mu}_Q$. Let

$$\delta_{\boldsymbol{\mu}} := \min_i \|\boldsymbol{\mu}_i - \boldsymbol{\mu}\|_2.$$

Without loss of generality, assume $\delta_{\boldsymbol{\mu}} := \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}\|_2$. It remains to prove that $D_f(P\|Q) > f(w_{\max})$ whenever $\delta_{\boldsymbol{\mu}}$ is not small.

We first prove the case for weakly mode-seeking. Let $r > 0$, and

$$\begin{aligned}T := \Bigg\{ &\mathbf{x} \in \mathbb{R}^d : \exists i \in \{1, \ldots, k\}.\\ &\left| \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}\|_2}(\mathbf{x} - \boldsymbol{\mu}) - \|\boldsymbol{\mu}_i - \boldsymbol{\mu}\|_2 \right| \le r \Bigg\}.\end{aligned}$$

We have, by Proposition C.1.3,

$$\begin{aligned}Q(T) &\le \sum_{i=1}^k Q\left(\mathbf{x} : \left| \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}\|_2}(\mathbf{x} - \boldsymbol{\mu}) - \|\boldsymbol{\mu}_i - \boldsymbol{\mu}\|_2 \right| \le r\right)\\&= \sum_{i=1}^k Q\left(\mathbf{x} : \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T}{\|\boldsymbol{\mu} - \boldsymbol{\mu}_i\|_2}(\mathbf{x} - \boldsymbol{\mu}) \in [\|\boldsymbol{\mu} - \boldsymbol{\mu}_i\|_2 - r,\, \|\boldsymbol{\mu} - \boldsymbol{\mu}_i\|_2 + r]\right)\\&\le \sum_{i=1}^k \frac{r}{\|\boldsymbol{\mu} - \boldsymbol{\mu}_i\|_2}\\&\le \frac{kr}{\delta_{\boldsymbol{\mu}}}. \qquad (7)\end{aligned}$$

Let $0 < \epsilon < 1/2$, and we choose r such that $Q(T) = \epsilon$ (this is possible since Q has a density, so $Q(T)$ changes continuously from 0 to approach 1 as r increases from 0 to ∞). We have $kr/\delta_{\boldsymbol{\mu}} \ge \epsilon$,

$$r \ge \frac{\delta_{\boldsymbol{\mu}}\epsilon}{k}. \qquad (8)$$

Also, by Chebyshev's inequality,

$$\begin{aligned}p_i(T^c) &\le p_i\left(\mathbf{x} : \left| \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}\|_2}(\mathbf{x} - \boldsymbol{\mu}) - \|\boldsymbol{\mu}_i - \boldsymbol{\mu}\|_2 \right| > r\right)\\&\le p_i\left(\mathbf{x} : \left| \frac{(\boldsymbol{\mu} - \boldsymbol{\mu}_i)^T}{\|\boldsymbol{\mu} - \boldsymbol{\mu}_i\|_2}(\mathbf{x} - \boldsymbol{\mu}_i) \right| > r\right)\\&\le \frac{\lambda_{\max}(\boldsymbol{\Sigma}_i)}{r^2}\\&\le \frac{\sigma_{\max}^2}{r^2}\\&\le \left(\frac{k\sigma_{\max}}{\delta_{\boldsymbol{\mu}}\epsilon}\right)^2, \qquad (9)\end{aligned}$$

and hence

$$P(T^c) \le \left(\frac{k\sigma_{\max}}{\delta_{\boldsymbol{\mu}}\epsilon}\right)^2.$$

Therefore,

$$\begin{aligned}&D_f\big(P\,\big\|\,Q\big)\\&\ge D_f\big(P(T)\,\big\|\,Q(T)\big)\\&\ge D_f\left(\max\left\{1-\Big(\frac{k\sigma_{\max}}{\delta_{\boldsymbol{\mu}}\epsilon}\Big)^2,\,\epsilon\right\}\,\middle\|\,\epsilon\right).\end{aligned}$$

Hence $D_f(P\|Q) > f(1/k) \ge f(w_{\max})$ (and hence Q cannot be the minimizer) whenever $\delta_{\boldsymbol{\mu}}/\sigma_{\max} > \check{f}(k)$, where

$$\check{f}(k) = \inf\left\{\gamma > 0:\, \exists\epsilon > 0.\, D_f\left(\max\left\{1-\left(\frac{k}{\gamma\epsilon}\right)^2,\,\epsilon\right\}\,\middle\|\,\epsilon\right) > f\left(\frac{1}{k}\right)\right\}. \tag{10}$$

Note that $\check{f}(k)$ is well-defined and finite since after substituting $\epsilon = (k/\gamma)^{2/3}$, we have

$$D_f\left(1-\left(\frac{k}{\gamma}\right)^{2/3}\,\middle\|\,\left(\frac{k}{\gamma}\right)^{2/3}\right) \to \lim_{t\to 0} f(t) > f\left(\frac{1}{k}\right)$$

as $\gamma \to \infty$. As a result,

$$\min_i \big\|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_i\big\|_2 \le \check{f}(k)\sigma_{\max}.$$

We now prove the case for strongly and uniformly mode-seeking. Let $r > 0$ (not the same as the previous r). We partition the space into three parts:

$$\begin{aligned}S_\epsilon := \Bigg\{&\mathbf{x}\in\mathbb{R}^d:\,\exists i\in\{2,\ldots,k\}.\\&\left|\left|\frac{(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}\|_2}(\mathbf{x}-\boldsymbol{\mu})\right| - \|\boldsymbol{\mu}_i-\boldsymbol{\mu}\|_2\right| \le r\Bigg\},\end{aligned}$$

$$S_+ := \big\{\mathbf{x}\in\mathbb{R}^d:\,(\boldsymbol{\mu}_1-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu}) \ge 0\big\}\setminus S_\epsilon,$$

$$S_- := \big\{\mathbf{x}\in\mathbb{R}^d:\,(\boldsymbol{\mu}_1-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu}) < 0\big\}\setminus S_\epsilon.$$

Similar to (7), we have

$$\begin{aligned}Q(S_\epsilon) &\le \sum_{i=2}^k Q\left(\mathbf{x}:\,\left|\left|\frac{(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}\|_2}(\mathbf{x}-\boldsymbol{\mu})\right| - \|\boldsymbol{\mu}_i-\boldsymbol{\mu}\|_2\right| \le r\right)\\&= \sum_{i=2}^k Q\left(\mathbf{x}:\,\left|\frac{(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T}{\|\boldsymbol{\mu}-\boldsymbol{\mu}_i\|_2}(\mathbf{x}-\boldsymbol{\mu})\right| \in [\|\boldsymbol{\mu}-\boldsymbol{\mu}_i\|_2 - r,\,\|\boldsymbol{\mu}-\boldsymbol{\mu}_i\|_2 + r]\right)\\&\overset{(a)}{\le} \sum_{i=2}^k \frac{2r}{\|\boldsymbol{\mu}-\boldsymbol{\mu}_i\|_2}\\&\overset{(b)}{\le} \frac{4kr}{\delta_{\min}},\end{aligned} \tag{11}$$

where (a) is by Proposition C.1.3, and (b) is because

$$\begin{aligned}\delta_{\min} &\le \|\boldsymbol{\mu}_1-\boldsymbol{\mu}_i\|_2\\&\le \|\boldsymbol{\mu}-\boldsymbol{\mu}_1\|_2 + \|\boldsymbol{\mu}-\boldsymbol{\mu}_i\|_2\\&\le 2\|\boldsymbol{\mu}-\boldsymbol{\mu}_i\|_2.\end{aligned}$$

Let $0 < \epsilon < 1/2$, and we choose r such that $Q(S_\epsilon) = \epsilon$ (this is possible since Q has a density, so $Q(S_\epsilon)$ changes continuously from 0 to approach 1 as r increases from 0 to ∞). By (11),

$$r \ge \frac{\delta_{\min}\epsilon}{4k}.$$

Also, since Q is symmetric around $\boldsymbol{\mu}$,

$$Q(S_-) = Q(S_+) = \frac{1-\epsilon}{2}.$$

Moreover, by Proposition C.1.2,

$$\begin{aligned}
p_1(S_-) &\le p_1\left(\left\{\mathbf{x} : (\boldsymbol{\mu}_1 - \boldsymbol{\mu})^T(\mathbf{x} - \boldsymbol{\mu}) \le 0\right\}\right) \\
&= p_1\left(\left\{\mathbf{x} : (\boldsymbol{\mu} - \boldsymbol{\mu}_1)^T(\mathbf{x} - \boldsymbol{\mu}_1) \ge \|\boldsymbol{\mu} - \boldsymbol{\mu}_1\|_2^2\right\}\right) \\
&\le \frac{1}{2}\max\left\{1 - \frac{\|\boldsymbol{\mu} - \boldsymbol{\mu}_1\|_2^2}{9\|\boldsymbol{\mu} - \boldsymbol{\mu}_1\|_2\sqrt{\lambda_{\max}(\boldsymbol{\Sigma}_i)}}, \frac{1}{3}\right\} \\
&\le \frac{1}{2}\max\left\{1 - \frac{\delta_{\boldsymbol{\mu}}}{9\sigma_{\max}}, \frac{1}{3}\right\} \\
&= \frac{1}{2}\left(1 - \min\left\{\frac{\delta_{\boldsymbol{\mu}}}{9\sigma_{\max}}, \frac{2}{3}\right\}\right), \qquad (12)
\end{aligned}$$

and by Chebyshev's inequality, for $i \ge 2$,

$$\begin{aligned}
p_i(S_\epsilon^c) &\le p_i\left(\mathbf{x} : \left|\left|\frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}\|_2}(\mathbf{x} - \boldsymbol{\mu})\right| - \|\boldsymbol{\mu}_i - \boldsymbol{\mu}\|_2\right| > r\right) \\
&\le p_i\left(\mathbf{x} : \left|\frac{(\boldsymbol{\mu} - \boldsymbol{\mu}_i)^T}{\|\boldsymbol{\mu} - \boldsymbol{\mu}_i\|_2}(\mathbf{x} - \boldsymbol{\mu}_i)\right| > r\right) \\
&\le \frac{\lambda_{\max}(\boldsymbol{\Sigma}_i)}{r^2} \\
&\le \left(\frac{4k\sigma_{\max}}{\delta_{\min}\epsilon}\right)^2. \qquad (13)
\end{aligned}$$

Assume

$$\left(\frac{4k\sigma_{\max}}{\delta_{\min}\epsilon}\right)^2 \le \frac{1}{6},$$

or equivalently,

$$\epsilon \ge \frac{4\sqrt{6}k\sigma_{\max}}{\delta_{\min}}.$$

We have

$$\begin{aligned}
P(S_-) &\le \frac{w_1}{2}\left(1 - \min\left\{\frac{\delta_{\boldsymbol{\mu}}}{9\sigma_{\max}}, \frac{2}{3}\right\}\right) + (1 - w_1)\left(\frac{4k\sigma_{\max}}{\delta_{\min}\epsilon}\right)^2 \\
&\le \frac{w_{\max}}{2}\left(1 - \min\left\{\frac{\delta_{\boldsymbol{\mu}}}{9\sigma_{\max}}, \frac{2}{3}\right\}\right) + (1 - w_{\max})\left(\frac{4k\sigma_{\max}}{\delta_{\min}\epsilon}\right)^2 \\
&=: \upsilon_-. \qquad (14)
\end{aligned}$$

Also

$$\begin{aligned}
P(S_\epsilon) &\ge (1 - w_1)\left(1 - \left(\frac{4k\sigma_{\max}}{\delta_{\min}\epsilon}\right)^2\right) \\
&\ge (1 - w_{\max})\left(1 - \left(\frac{4k\sigma_{\max}}{\delta_{\min}\epsilon}\right)^2\right) \\
&=: \upsilon_\epsilon. \qquad (15)
\end{aligned}$$

Let

$$\begin{aligned}
\upsilon_+ &:= 1 - \upsilon_- - \upsilon_\epsilon \\
&= w_{\max}\left(1 - \frac{1}{2}\left(1 - \min\left\{\frac{\delta_{\boldsymbol{\mu}}}{9\sigma_{\max}}, \frac{2}{3}\right\}\right)\right) \\
&= \frac{w_{\max}}{2}\left(1 + \min\left\{\frac{\delta_{\boldsymbol{\mu}}}{9\sigma_{\max}}, \frac{2}{3}\right\}\right). \qquad (16)
\end{aligned}$$

We have

$$
\begin{aligned}
&D_f(P \,\|\, Q) \\
&\geq Q(S_\epsilon) f\left(\frac{P(S_\epsilon)}{Q(S_\epsilon)}\right) + Q(S_+) f\left(\frac{P(S_+)}{Q(S_+)}\right) + Q(S_-) f\left(\frac{P(S_-)}{Q(S_-)}\right) \\
&= \epsilon f\left(\frac{P(S_\epsilon)}{\epsilon}\right) + \frac{1-\epsilon}{2} f\left(\frac{2P(S_+)}{1-\epsilon}\right) + \frac{1-\epsilon}{2} f\left(\frac{2P(S_-)}{1-\epsilon}\right) \\
&\overset{(c)}{\geq} \epsilon f\left(\frac{1}{\epsilon}\right) + \frac{1-\epsilon}{2} f\left(\frac{2(1-v_\epsilon - P(S_-))}{1-\epsilon}\right) + \frac{1-\epsilon}{2} f\left(\frac{2P(S_-)}{1-\epsilon}\right) \\
&\overset{(d)}{\geq} \epsilon f\left(\frac{1}{\epsilon}\right) + \frac{1-\epsilon}{2} f\left(\frac{2v_+}{1-\epsilon}\right) + \frac{1-\epsilon}{2} f\left(\frac{2v_-}{1-\epsilon}\right) \\
&\geq \epsilon f\left(\frac{1}{\epsilon}\right) + \frac{1-\epsilon}{2} f\left(\frac{1+\zeta}{1-\epsilon} w_{\max}\right) + \frac{1-\epsilon}{2} f\left(\frac{1-\zeta}{1-\epsilon} w_{\max} + \gamma^2\epsilon^{-2}\right)
\end{aligned}
\tag{17}
$$

where (c) is because f is nonincreasing, and (d) is by the convexity of f, and we let

$$
\begin{aligned}
\zeta &:= \min\left\{\frac{\delta_{\boldsymbol{\mu}}}{9\sigma_{\max}},\, \frac{2}{3}\right\} \\
\gamma &:= \frac{8k\sigma_{\max}}{\delta_{\min}}.
\end{aligned}
$$

Note that as $\epsilon \to 0$ and $\gamma\epsilon^{-1} \to 0$, we have $\epsilon f(1/\epsilon) \to 0$ since $\lim_{t\to\infty} f(t)/t = 0$, and

$$
\begin{aligned}
&\epsilon f\left(\frac{1}{\epsilon}\right) + \frac{1-\epsilon}{2} f\left(\frac{1+\zeta}{1-\epsilon} w_{\max}\right) + \frac{1-\epsilon}{2} f\left(\frac{1-\zeta}{1-\epsilon} w_{\max} + \gamma\right) \\
&\to \frac{1}{2} f\left((1+\zeta) w_{\max}\right) + \frac{1}{2} f\left((1-\zeta) w_{\max}\right) \\
&> f(w_{\max})
\end{aligned}
$$

since f is strictly convex in a neighborhood of $w_{\max}$. Hence, for any fixed $\delta_{\boldsymbol{\mu}}/\sigma_{\max}$, this Q is suboptimal if ϵ and $\gamma\epsilon^{-1}$ are small enough. This shows that if $w_{\max}$ is fixed, then f being strictly convex in $(0, 1]$ and $\lim_{t\to\infty} f(t)/t < \infty$ is sufficient to show that $\delta_{\boldsymbol{\mu}}/\sigma_{\max} \to 0$ as $\epsilon \to 0$ and $\gamma\epsilon^{-1} \to 0$. Nevertheless, the definition of strongly mode-seeking allows us to characterize the mode-seeking order.

Now we prove the mode-seeking order in Theorem 4.3. By strong convexity in MS3, let $\beta > 0$ be such that $t \mapsto f(t) - \beta t^2/2$ is convex for $t \in (0, s]$. Write $f'(t)$ for the left derivative of f. We have, for $w \in (0, 1]$, $t \in (0, s]$,

$$
f(t) \geq f(w) + f'(w)(t-w) + \frac{\beta}{2}(t-w)^2. \tag{18}
$$

Note that

$$
\frac{1+\zeta}{1-\epsilon} w \leq s
$$

as long as $\epsilon \leq 1 - s^{-1/2}$ since $1 + \zeta \leq \sqrt{s}$. Also,

$$
\frac{1-\zeta}{1-\epsilon} w + \gamma^2\epsilon^{-2} \leq s
$$

as long as $\epsilon \leq 1 - s^{-1/2}$ and $\gamma^2\epsilon^{-2} \leq s - \sqrt{s}$. Hence, as long as

$$
\frac{\gamma}{\sqrt{s - \sqrt{s}}} \leq \epsilon \leq 1 - s^{-1/2},
$$

(note that we can assume γ is small enough that the above interval is nonempty, since otherwise (5) is implied by (4)), by

(18),

$$
\begin{aligned}
&\epsilon f\left(\frac{1}{\epsilon}\right)+\frac{1-\epsilon}{2} f\left(\frac{1+\zeta}{1-\epsilon} w\right)+\frac{1-\epsilon}{2} f\left(\frac{1-\zeta}{1-\epsilon} w+\gamma^2 \epsilon^{-2}\right)-f(w) \\
&\geq \epsilon f\left(\frac{1}{\epsilon}\right)-\epsilon f(w)+f^{\prime}(w)\left(\epsilon w+\frac{1-\epsilon}{2} \gamma^2 \epsilon^{-2}\right) \\
&\quad+\frac{1-\epsilon}{2} \frac{\beta}{2}\left(\left(\frac{1+\zeta}{1-\epsilon} w-w\right)^2+\left(\frac{1-\zeta}{1-\epsilon} w+\gamma^2 \epsilon^{-2}-w\right)^2\right) \\
&\geq \epsilon f\left(\frac{1}{\epsilon}\right)-\epsilon f(w)+f^{\prime}(w)\left(\epsilon w+\frac{1}{2} \gamma^2 \epsilon^{-2}\right) \\
&\quad+\frac{\beta(1-\epsilon)}{4}\left(\left(\frac{\zeta+\epsilon}{1-\epsilon} w\right)^2+\left(\frac{\zeta-\epsilon}{1-\epsilon} w-\gamma^2 \epsilon^{-2}\right)^2\right) \\
&\geq \epsilon f\left(\frac{1}{\epsilon}\right)-\epsilon f(w)+f^{\prime}(w)\left(\epsilon w+\frac{1}{2} \gamma^2 \epsilon^{-2}\right) \\
&\quad+\frac{\beta}{4}\left(((\zeta+\epsilon) w)^2+\left((\zeta-\epsilon) w-(1-\epsilon) \gamma^2 \epsilon^{-2}\right)^2\right) \\
&\geq \epsilon f\left(\frac{1}{\epsilon}\right)-\epsilon f(w)+f^{\prime}(w)\left(\epsilon w+\frac{1}{2} \gamma^2 \epsilon^{-2}\right) \\
&\quad+\frac{\beta}{4}\Big(\zeta^2 w^2+\epsilon^2 w^2+2 \zeta \epsilon w^2+\zeta^2 w^2+\epsilon^2 w^2-2 \zeta \epsilon w^2 \\
&\quad+(1-\epsilon)^2 \gamma^4 \epsilon^{-4}-2(1-\epsilon) \zeta w \gamma^2 \epsilon^{-2}+2(1-\epsilon) w \gamma^2 \epsilon^{-1}\Big) \\
&=\epsilon f\left(\frac{1}{\epsilon}\right)-\epsilon f(w)+f^{\prime}(w) w \epsilon+\frac{1}{2} f^{\prime}(w) \gamma^2 \epsilon^{-2} \\
&\quad+\frac{\beta}{4}\Big(2 w^2 \epsilon^2+2 \zeta^2 w^2+2(1-\epsilon) w \gamma^2 \epsilon^{-1} \\
&\quad-2(1-\epsilon) \zeta w \gamma^2 \epsilon^{-2}+(1-\epsilon)^2 \gamma^4 \epsilon^{-4}\Big) \\
&\geq \epsilon f\left(\frac{1}{\epsilon}\right)-\left(f(w)-f^{\prime}(w) w\right) \epsilon-\left(\frac{\beta}{2} \zeta w-\frac{1}{2} f^{\prime}(w)\right) \gamma^2 \epsilon^{-2}+\frac{\beta}{2} \zeta^2 w^2 \\
&\geq \epsilon f\left(\frac{1}{\epsilon}\right)-\left(-f^{\prime}(w)(1-w)-f^{\prime}(w) w\right) \epsilon \\
&\quad-\left(\frac{\beta}{2} \zeta w-\frac{1}{2} f^{\prime}(w)\right) \gamma^2 \epsilon^{-2}+\frac{\beta}{2} \zeta^2 w^2 \\
&\geq \epsilon f\left(\frac{1}{\epsilon}\right)+f^{\prime}(w) \epsilon-\left(\frac{\beta}{2}-\frac{1}{2} f^{\prime}(w)\right) \gamma^2 \epsilon^{-2}+\frac{\beta}{2} \zeta^2 w^2 \\
&\geq-\left(\sqrt{-f^{\prime}(w)+1} \sqrt{-\epsilon f\left(\frac{1}{\epsilon}\right)+\epsilon}+\sqrt{\frac{\beta}{2}-\frac{1}{2} f^{\prime}(w)} \frac{\gamma}{\epsilon}\right)^2+\frac{\beta}{2} \zeta^2 w^2 \\
&\geq-\left(\frac{\beta}{2}-f^{\prime}(w)+1\right)\left(\sqrt{-\epsilon f\left(\frac{1}{\epsilon}\right)+\epsilon}+\frac{\gamma}{\epsilon}\right)^2+\frac{\beta}{2} \zeta^2 w^2 .
\end{aligned} \tag{19}
$$

Let

$$\mathring{f}(\gamma, \epsilon):=\sqrt{-\epsilon f\left(\frac{1}{\epsilon}\right)+\epsilon}+\frac{\gamma}{\epsilon}. \tag{20}$$

Then we take $\mathring{f}(\gamma)=\inf_{0<\epsilon<1/2} \mathring{f}(\gamma, \epsilon)$. We will show that there exists a constant $C_1>0$ (that can depend on f, s) such that

$$\inf_{\frac{\gamma}{\sqrt{s-\sqrt{s}}} \leq \epsilon \leq 1-s^{-1/2}} \mathring{f}(\gamma, \epsilon) \leq C_1 \mathring{f}(\gamma). \tag{21}$$

To show this, note that if $\epsilon \le \gamma/\sqrt{s-\sqrt{s}}$, then

$$\begin{aligned}
&\left(\frac{\sup_{0<\epsilon'<1/2}\mathring{f}(0,\epsilon')}{\sqrt{s-\sqrt{s}}}+1\right)\mathring{f}(\gamma,\epsilon)\\
&\ge \sup_{0<\epsilon'<1/2}\mathring{f}(0,\epsilon')+\sqrt{s-\sqrt{s}}\\
&\ge \mathring{f}\Big(\gamma,\,\gamma/\sqrt{s-\sqrt{s}}\Big).
\end{aligned}$$

If $1-s^{-1/2}\le\epsilon\le 1/2$, then by the convexity of f,

$$\begin{aligned}
&\frac{2}{1-s^{-1/2}}\mathring{f}(\gamma,\epsilon)\\
&\ge\sqrt{-2\epsilon f\left(\frac{1}{\epsilon}\right)+\epsilon+\frac{\gamma}{\epsilon(1-s^{-1/2})}}\\
&\ge\sqrt{-\left(1/\epsilon-1\right)^{-1}f\left(\frac{1}{\epsilon}\right)+1-s^{-1/2}+\frac{\gamma}{1-s^{-1/2}}}\\
&\ge\sqrt{-\left(1/(1-s^{-1/2})-1\right)^{-1}f\left(\frac{1}{1-s^{-1/2}}\right)+1-s^{-1/2}+\frac{\gamma}{1-s^{-1/2}}}\\
&\ge\mathring{f}(\gamma,1-s^{-1/2}).
\end{aligned}$$

Hence (21) holds. Combining (21) with (17), (19), and $w_{\max}\ge 1/k$, we have

$$\begin{aligned}
&D_f(P\,\|\,Q)-f(w_{\max})\\
&\ge -C_1^2\left(\frac{\beta}{2}-f'(w_{\max})+1\right)(\mathring{f}(\gamma))^2+\frac{\beta}{2}\zeta^2w_{\max}^2\\
&\ge -C_1^2\left(\frac{\beta}{2}-f'(k^{-1})+1\right)(\mathring{f}(\gamma))^2+\frac{\beta}{2}\zeta^2k^{-2}\\
&>0
\end{aligned}\tag{22}$$

as long as

$$\zeta=\min\left\{\frac{\delta_{\boldsymbol{\mu}}}{9\sigma_{\max}},\,\frac{2}{3}\right\}\tag{23}$$

$$\ge\frac{2C_1k}{\sqrt{\beta}}\left(\sqrt{\frac{\beta}{2}-f'(k^{-1})+1}\right)\mathring{f}(\gamma)\tag{24}$$

$$=:\tilde{C}_{f,k}\mathring{f}(\gamma).\tag{25}$$

Due to (4), we can assume $\delta_{\boldsymbol{\mu}}/\sigma_{\max}\le C_{f,k}$, and hence $\zeta\ge\delta_{\boldsymbol{\mu}}/(\max\{9,\,3C_{f,k}/2\}\sigma_{\max})$. Therefore, $D_f(P\|Q)>f(w_{\max})$ (and hence Q cannot be the minimizer) whenever $\delta_{\boldsymbol{\mu}}/(\max\{9,\,3C_{f,k}/2\}\sigma_{\max})\ge\tilde{C}_{f,k}\mathring{f}(\gamma)$. The result follows.

For the uniformly mode-seeking case, we first prove the claim that MS1-4 implies that there exist constants $\phi>0$, $s>1$ such that

$$f(t)\ge f(w)+f'(w)(t-w)-\frac{\phi}{2}f'(w)(t-w)^2\tag{26}$$

for any $w\in(0,1]$, $t\in(0,s]$. To prove this, note that by MS3 and MS4, we can let $s>1$ such that $f''(t)$ is non-increasing and $f''(t)\ge\beta$ for $t\in(0,s]$. For any $t\le s$, we have

$$-f'(t)=-f'(s)+\int_t^s f''(\tau)\mathrm{d}\tau.\tag{27}$$

Fix $w \in (0, 1]$. For $t \leq w$,

$$\begin{aligned}
&f(t) - f(w) - f'(w)(t-w) \\
&= \int_t^w (f'(w) - f'(\tau))\mathrm{d}\tau \\
&= \int_t^w (\tau - t) f''(\tau)\mathrm{d}\tau \\
&\geq \frac{(t-w)^2}{2} f''(w) \\
&\geq \frac{(t-w)^2}{2(s-w)} \int_w^s f''(\tau)\mathrm{d}\tau \\
&\geq \frac{1}{2}(t-w)^2 \left(\frac{1}{2}\beta + \frac{1}{2(s-w)} \int_w^s f''(\tau)\mathrm{d}\tau\right) \\
&\geq \frac{1}{2}(t-w)^2(-f'(w)) \min\left\{\frac{\beta}{-2f'(s)}, \frac{1}{2s}\right\},
\end{aligned}$$

where the last line is by (27). For $w < t \leq s$,

$$\begin{aligned}
&f(t) - f(w) - f'(w)(t-w) \\
&= \int_w^t (t-\tau) f''(\tau)\mathrm{d}\tau \\
&\geq \frac{t-w}{2} \int_w^t f''(\tau)\mathrm{d}\tau \\
&\geq \frac{(t-w)^2}{2(s-w)} \int_w^s f''(\tau)\mathrm{d}\tau \\
&\geq \frac{1}{2}(t-w)^2(-f'(w)) \min\left\{\frac{\beta}{-2f'(s)}, \frac{1}{2s}\right\}.
\end{aligned}$$

The claim (26) follows.

By (26) and the same arguments as (19) and (22),

$$\begin{aligned}
&D_f(P \,\|\, Q) - f(w_{\max}) \\
&\geq -\left(-\frac{\phi}{2} f'(w_{\max}) - f'(w_{\max}) + 1\right)(\mathring{f}(\gamma))^2 - \frac{\phi}{2} f'(w_{\max})\zeta^2 w_{\max}^2 \\
&\geq -f'(w_{\max})\Big(-\big(\frac{\phi}{2} + 1 - \frac{1}{f'(1)}\big)(\mathring{f}(\gamma))^2 + \frac{\phi}{2}\zeta^2 k^{-2}\Big) \\
&> 0
\end{aligned}$$

as long as

$$\zeta \geq 2k\mathring{f}(\gamma)\sqrt{\frac{1}{2} + \frac{1}{\phi} - \frac{1}{\phi f'(1)}}.$$

The result follows from $\zeta = \min\{\delta_{\boldsymbol{\mu}}/(9\sigma_{\max}),\, 2/3\} \geq (1/9)\min\{\delta_{\boldsymbol{\mu}}/\sigma_{\max},\, 1\}$, giving a constant

$$C_f = 144\sqrt{\frac{1}{2} + \frac{1}{\phi} - \frac{1}{\phi f'(1)}} \tag{28}$$

for (6). □

D APPENDIX – PROOF OF LEMMA 5.1 AND THEOREM 5.2

We first prove Lemma 5.1. We assume $d = 2$, $0 < \rho < 1$ (the case $\rho \geq 1$ can be proved simply by considering an equilateral triangle). Write $\mathbf{e}_1 := [1, 0]$, $B_r := \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| \leq r\}$, $B + \mathbf{z} := \{\mathbf{x} + \mathbf{z} : \mathbf{x} \in B\}$. Let $0 < \epsilon < 1$. Consider the probability density function $p(\mathbf{x}) := \mathbf{1}\{\mathbf{x} \in B_1 \backslash B_\epsilon\}/(\pi(1-\epsilon^2))$ of the uniform distribution over $B_1 \backslash B_\epsilon$. We will prove that, as long as $\mathbf{a} \notin \mathrm{int}(B_\epsilon)$ (where $\mathrm{int}(B_\epsilon)$ is the interior of B_ϵ which is the open disk), we have

$$\mathbb{E}\left[\|\mathbf{x}-\mathbf{a}\|^\rho\right] > \mathbb{E}\left[\|\mathbf{x}\|^\rho\right],$$

where $\mathbf{x} \sim p$, and hence the $\mathbf{a}$ that minimizes $\mathbb{E}[\|\mathbf{x}-\mathbf{a}\|^\rho]$ must be in $\mathrm{int}(B_\epsilon)$. Without loss of generality, assume $\mathbf{a} = a\mathbf{e}_1$ where $a \in [\epsilon, 1]$ (we can assume $a \le 1$ since $a > 1$ results in a larger average distance than $a = 1$). We have

$$\begin{aligned}
&\pi(1-\epsilon^2)\left(\mathbb{E}\left[\|\mathbf{x}-a\mathbf{e}_1\|^\rho\right] - \mathbb{E}\left[\|\mathbf{x}\|^\rho\right]\right) \\
&= \int_{B_1 - a\mathbf{e}_1} \|\mathbf{x}\|^\rho \mathrm{d}\mathbf{x} - \int_{B_\epsilon - a\mathbf{e}_1} \|\mathbf{x}\|^\rho \mathrm{d}\mathbf{x} - \int_{B_1} \|\mathbf{x}\|^\rho \mathrm{d}\mathbf{x} + \int_{B_\epsilon} \|\mathbf{x}\|^\rho \mathrm{d}\mathbf{x} \\
&\ge \int_{(B_1 - a\mathbf{e}_1)\backslash B_1} \|\mathbf{x}\|^\rho \mathrm{d}\mathbf{x} - \int_{B_1 \backslash (B_1 - a\mathbf{e}_1)} \|\mathbf{x}\|^\rho \mathrm{d}\mathbf{x} - \int_{B_\epsilon - a\mathbf{e}_1} \|\mathbf{x}\|^\rho \mathrm{d}\mathbf{x} \\
&= \int_{(B_1 + a\mathbf{e}_1)\backslash B_1} \left(\|\mathbf{x}\|^\rho - \|\mathbf{x} - a\mathbf{e}_1\|^\rho\right) \mathrm{d}\mathbf{x} - \int_{B_\epsilon - a\mathbf{e}_1} \|\mathbf{x}\|^\rho \mathrm{d}\mathbf{x} \\
&\ge \int_{(B_1 + a\mathbf{e}_1)\backslash B_1} \left(\|\mathbf{x}\|^\rho - \|\mathbf{x} - a\mathbf{e}_1\|^\rho\right) \mathrm{d}\mathbf{x} - 2^\rho \pi \epsilon^2 a^\rho,
\end{aligned}$$

where the last line is because $\|\mathbf{x}\| \le a + \epsilon \le 2a$ for $\mathbf{x} \in B_\epsilon - a\mathbf{e}_1$. Let $g(a) := \int_{(B_1 + a\mathbf{e}_1)\backslash B_1} (\|\mathbf{x}\|^\rho - \|\mathbf{x} - a\mathbf{e}_1\|^\rho)\mathrm{d}\mathbf{x}$ for $a \ge 0$. Note that $g(a)$ is a continuous function with $g(0) = 0$, and $g(a) > 0$ for $a > 0$ since $\|\mathbf{x}\| \ge 1 \ge \|\mathbf{x} - a\mathbf{e}_1\|$ for any $\mathbf{x} \in (B_1 + a\mathbf{e}_1)\backslash B_1$, and the inequality is strict for a set of $\mathbf{x}$ with positive measure. Also $\lim_{a\to\infty} g(a) = \infty$. Hence there exists $a_0 > 0$ such that

$$\begin{aligned}
&\int_{B_1 \backslash (B_1 - a\mathbf{e}_1)} \left(\|\mathbf{x} + a\mathbf{e}_1\|^\rho - \|\mathbf{x}\|^\rho\right) \mathrm{d}\mathbf{x} \\
&\ge \int_{\{\mathbf{x} \in B_1 \backslash (B_1 - a\mathbf{e}_1) :\, x_1 \ge 1/2\}} \left((\|\mathbf{x} + a\mathbf{e}_1\|^2)^{\rho/2} - (\|\mathbf{x}\|^2)^{\rho/2} \right) \mathrm{d}\mathbf{x} \\
&\overset{(a)}{\ge} \int_{\{\mathbf{x} \in B_1 \backslash (B_1 - a\mathbf{e}_1) :\, x_1 \ge 1/2\}} \left(\left(\|\mathbf{x}\|^2 + \left(a + \frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 \right)^{\rho/2} - (\|\mathbf{x}\|^2)^{\rho/2} \right) \mathrm{d}\mathbf{x} \\
&\ge \int_{\{\mathbf{x} \in B_1 \backslash (B_1 - a\mathbf{e}_1) :\, x_1 \ge 1/2\}} \left((\|\mathbf{x}\|^2 + a)^{\rho/2} - (\|\mathbf{x}\|^2)^{\rho/2} \right) \mathrm{d}\mathbf{x} \\
&\overset{(b)}{\ge} \int_{\{\mathbf{x} \in B_1 \backslash (B_1 - a\mathbf{e}_1) :\, x_1 \ge 1/2\}} \left((1+a)^{\rho/2} - 1^{\rho/2} \right) \mathrm{d}\mathbf{x} \\
&\overset{(c)}{\ge} \min\left\{ a,\, \frac{\sqrt{3}-1}{2} \right\} \left((1+a)^{\rho/2} - 1 \right) \\
&\overset{(d)}{\ge} \min\{a,\, 1/4\} 2^{\rho/2-2} \rho a \\
&\ge \min\{\epsilon,\, 1/4\} 2^{\rho/2-2} \rho \epsilon^{1-\rho} a^\rho \\
&\ge 2^{\rho+1} \pi \epsilon^2 a^\rho
\end{aligned}$$

for small enough $\epsilon > 0$ such that $\min\{\epsilon,\, 1/4\}2^{\rho/2-2}\rho\epsilon^{1-\rho} \ge 2^{\rho+1}\pi\epsilon^2$, where (a) is because $\|\mathbf{x} + a\mathbf{e}_1\|^2 - \|\mathbf{x}\|^2 \ge (a + 1/2)^2 - (1/2)^2$ as long as $x_1 \ge 1/2$, (b) is by $\|\mathbf{x}\| \le 1$ and the concavity of $t \mapsto t^{\rho/2}$, (c) is by straightforward geometric arguments on the area of the set $\{\mathbf{x} \in B_1 \backslash (B_1 - a\mathbf{e}_1) :\, x_1 \ge 1/2\}$, and (d) is by $1 + a \le 2$, the concavity of $t \mapsto t^{\rho/2}$ and that $\mathrm{d}t^{\rho/2}/\mathrm{d}t = 2^{\rho/2-2}\rho$ at $t = 2$. Hence there exists $\epsilon > 0$ (that only depends on ρ) such that for any $a \in [\epsilon, 1]$,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{x} - a\mathbf{e}_1\|^\rho\right] - \mathbb{E}\left[\|\mathbf{x}\|^\rho\right] &\ge \frac{2^\rho \pi \epsilon^2 a^\rho}{\pi(1-\epsilon^2)} \\
&\ge \frac{2^\rho \epsilon^{2+\rho}}{1-\epsilon^2} \\
&\ge \epsilon^{2+\rho}.
\end{aligned}$$

Let $\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{iid}{\sim} p$. Let $\hat{p} := n^{-1} \sum_{i=1}^{n} \delta_{\mathbf{y}_i}$ be the empirical distribution. We have $W_1(\hat{p}, p) \to 0$ as $n \to \infty$. Also, for $\mathbf{x} \sim p$, $\mathbf{y} \sim \hat{p}$, and $\mathbf{a}$ such that $p(\mathbf{a}) > 0$, for any $\xi > 0$,

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{y}-\mathbf{a}\|^{\rho}\right] - \mathbb{E}\left[\|\mathbf{y}\|^{\rho}\right] \\
&\geq \mathbb{E}\left[(\max\{\|\mathbf{y}-\mathbf{a}\|, \xi\})^{\rho}\right] - \xi^{\rho} - \mathbb{E}\left[(\max\{\|\mathbf{y}\|, \xi\})^{\rho}\right] \\
&\overset{(a)}{\geq} \mathbb{E}\left[(\max\{\|\mathbf{x}-\mathbf{a}\|, \xi\})^{\rho}\right] - \xi^{\rho} - \mathbb{E}\left[(\max\{\|\mathbf{x}\|, \xi\})^{\rho}\right] - 2\rho\xi^{\rho-1} W_1(\hat{p}, p) \\
&\geq \mathbb{E}\left[\|\mathbf{x}-\mathbf{a}\|^{\rho}\right] - \mathbb{E}\left[\|\mathbf{x}\|^{\rho}\right] - 2\rho\xi^{\rho-1} W_1(\hat{p}, p) - 2\xi^{\rho} \\
&\geq \epsilon^{2+\rho} - 2\rho\xi^{\rho-1} W_1(\hat{p}, p) - 2\xi^{\rho},
\end{aligned}$$

where (a) is because $\mathbf{x} \mapsto (\max\{\|\mathbf{x}\|, \xi\})^{\rho}$ is $(\rho\xi^{\rho-1})$-Lipschitz. Hence we have $\mathbb{E}[\|\mathbf{y}-\mathbf{a}\|^{\rho}] - \mathbb{E}[\|\mathbf{y}\|^{\rho}] > 0$ for any $\mathbf{a} \notin \mathrm{int}(B_{\epsilon})$, by taking ξ small enough such that $2\xi^{\rho} < \epsilon^{2+\rho}/4$, and $\hat{p}$ close enough to p such that $2\rho\xi^{\rho-1} W_1(\hat{p}, p) < \epsilon^{2+\rho}/4$ (which happens with probability approaching 1 as $n \to \infty$).

Finally, for the uniqueness of the minimizer, assume the set of minimizers of $\mathbf{a} \mapsto \mathbb{E}[\|\mathbf{y}-\mathbf{a}\|^{\rho}]$ is $S \subseteq \mathbb{R}^2$, and the minimum is θ. By continuity of $\mathbf{a} \mapsto \mathbb{E}[\|\mathbf{y}-\mathbf{a}\|^{\rho}]$, S is a closed set. We have proved that $S \subseteq \mathrm{int}(B_{\epsilon})$. Let $\mathbf{b} := \mathrm{argmax}_{\mathbf{a} \in S} \|\mathbf{a}\|$ (choose any maximizer if not unique). We have $\|\mathbf{b}\| < \epsilon$. Let $\mathbf{z}_{2i-1} := \mathbf{y}_i - \mathbf{b}$, $\mathbf{z}_{2i} := \mathbf{b} - \mathbf{y}_i$ for $i = 1, \ldots, n$, $\tilde{p} := (2n)^{-1} \sum_{i=1}^{2n} \delta_{\mathbf{z}_i}$, $\mathbf{z} \sim \tilde{p}$. Note that

$$\begin{aligned}
\mathbb{E}[\|\mathbf{z}-\mathbf{a}\|^{\rho}] &= \frac{1}{2}\big(\mathbb{E}[\|\mathbf{y}-(\mathbf{b}-\mathbf{a})\|^{\rho}] + \mathbb{E}[\|\mathbf{y}-(\mathbf{b}+\mathbf{a})\|^{\rho}]\big) \\
&\geq \theta,
\end{aligned} \tag{29}$$

where equality is attained at $\mathbf{a} = 0$. For any $\mathbf{a} \neq 0$, we either have $\|\mathbf{b}-\mathbf{a}\| > \|\mathbf{b}\|$ or $\|\mathbf{b}+\mathbf{a}\| > \|\mathbf{b}\|$, implying that at least one of $\mathbf{b}-\mathbf{a}$, $\mathbf{b}+\mathbf{a}$ is not in S (by the maximality of $\mathbf{b}$), and hence at least one of the two terms in (29) is strictly greater than θ. Therefore, $\mathbf{a} = 0$ is the unique minimizer of $\mathbb{E}[\|\mathbf{z}-\mathbf{a}\|^{\rho}]$, and does not coincide with any $\mathbf{z}_i$ since $\mathbf{z}_i \in ((B_1 \backslash B_{\epsilon}) + \mathbf{b}) \cup ((B_1 \backslash B_{\epsilon}) - \mathbf{b})$ and $\|\mathbf{b}\| < \epsilon$.

We will prove Theorem 5.2 using Lemma 5.1. Since $g(\mathbf{x}) := k^{-1} \sum_{i=1}^{k} \|\mathbf{x} - \mathbf{z}_i\|^{\rho}$ is continuous, if the minimizer $\mathbf{x}^*$ is unique and does not belong to $\{\mathbf{z}_1, \ldots, \mathbf{z}_k\}$, there exists $\epsilon > 0$, $0 < \delta < \min_i \|\mathbf{x}^* - \mathbf{z}_i\|$ such that any $\mathbf{x}$ satisfying $(g(\mathbf{x}))^{1/\max\{\rho,1\}} \leq (g(\mathbf{x}^*))^{1/\max\{\rho,1\}} + \epsilon$ has $\|\mathbf{x} - \mathbf{x}^*\|_2 \leq \delta$ (by Bolzano-Weierstrass theorem since it suffices to consider the compact set $\{\mathbf{x} : \|\mathbf{x}\|_2 \leq 3 \max_i \|\mathbf{z}_i\|_2\}$, as any $\mathbf{x}$ not in this set has $g(\mathbf{x})$ bounded away from the optimum). Take p_i to be centered at $\alpha \mathbf{z}_i$ for a large α. For $Q \in \mathcal{P}$, we have $W_{\rho}(Q, \delta_{\mathbb{E}Q}) \leq \sigma_{\max}^{\min\{\rho,1\}}$, and hence $|\alpha^{-\min\{\rho,1\}} W_{\rho}(P, Q) - (g(\mathbb{E}Q/\alpha))^{1/\max\{\rho,1\}}| = \alpha^{-\min\{\rho,1\}} |W_{\rho}(P, Q) - W_{\rho}(k^{-1} \sum_{i=1}^{k} \delta_{\alpha \mathbf{z}_i}, \delta_{\mathbb{E}Q})| \leq (k+1)(\sigma_{\max}/\alpha)^{\min\{\rho,1\}}$. If α is large enough such that $(k+1)(\sigma_{\max}/\alpha)^{\min\{\rho,1\}} < \epsilon/2$, for $Q = \mathrm{argmin}_{Q \in \mathcal{P}} W_{\rho}(P, Q)$, we must have $\|\mathbb{E}Q/\alpha - \mathbf{x}^*\|_2 \leq \delta$, giving $\min_i \|\mathbb{E}Q - \mathbb{E}p_i\|_2 > \alpha(\min_i \|\mathbf{x}^* - \mathbf{z}_i\| - \delta)$, which can be arbitrarily large.

E APPENDIX – FORMAL VERSION AND PROOF OF THEOREM 6.1

We now state the formal version of Theorem 6.1.

Theorem E.1. *Consider the hybrid divergence $D_{\lambda f, W_1}$, where the f-divergence D_f is weakly mode-seeking. Let $\psi > 2$. Let $\mathcal{P}$ be an arbitrary set of symmetric quasiconcave distributions over $\mathbb{R}^d$ with $\mathbb{E}_{\mathbf{x} \sim p}[\|\mathbf{x}\|_2^{\psi}] < \infty$ for any $p \in \mathcal{P}$. Let $P(\mathbf{x}) := \sum_{i=1}^{k} w_i p_i(\mathbf{x})$ be a mixture of distributions in $\mathcal{P}$, where $p_1, \ldots, p_k \in \mathcal{P}$ with distinct centers. Define $\sigma_{\max}$ as in Theorem 4.3. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} P$, and $\hat{P} := n^{-1} \sum_{i=1}^{n} \delta_{\mathbf{x}_i}$ be the empirical distribution. Let $Q := \mathrm{argmin}_{Q \in \mathcal{P}} D_{\lambda f, W_1}(\hat{P} \,\|\, Q)$, and denote its center as $\boldsymbol{\mu}_Q$. If such minimizer Q always exists, then for any $\zeta > 2\sqrt{2}$, if $\lambda \leq \zeta^{1/3} \sigma_{\max} / (2 \breve{f}(\zeta^{-2/3}))$ where* [2]

$$\breve{f}(t) := -\frac{\mathrm{d}}{\mathrm{d}\tau} D_f(1 - t - \tau \,\|\, t),$$

then we have

$$\begin{aligned}
&\mathbb{P}\left(\min_i \|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_i\|_2 \geq k \sigma_{\max} \zeta\right) \\
&\leq \frac{C_{d,\psi}}{\lambda} \cdot \frac{\left(\mathbb{E}\left[\|\mathbf{x} - \mathbb{E}[\mathbf{x}]\|_2^{\psi}\right]\right)^{1/\psi} G_d(n)}{D_f(\zeta^{-2/3} \,\|\, 1 - \zeta^{-2/3}) - \lim_{t \nearrow 1} D_f(k^{-1} \,\|\, t)}
\end{aligned}$$

[2]We use the notation $D_f(s \| t) = D_f(\mathrm{Bern}(s) \| \mathrm{Bern}(t))$ where $\mathrm{Bern}(s)$ is the Bernoulli distribution with parameter s.

as long as the right hand side above is positive, where $\mathbf{x} \sim P$*, and* $C_{d,\psi} > 0$ *only depends on* d, ψ*, and*

$$G_d(n) := \begin{cases} n^{-1/2} & \text{if } d = 1 \\ n^{-1/2}\log(1+n) & \text{if } d = 2 \\ n^{-1/d} & \text{if } d \geq 3. \end{cases}$$

Loosely speaking, Theorem E.1 implies that, when f, k, d, ψ are fixed, as long as $\lambda = O(\sigma_{\max})$, we have $\min_i \|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_i\|_2 = O(\sigma_{\max})$ with probability $1 - O(\lambda^{-1}(\mathbb{E}[\|\mathbf{x} - \mathbb{E}[\mathbf{x}]\|_2^{\psi}])^{1/\psi} G_d(n))$. To ensure a high probability of success, we propose the following method to select λ:

$$\lambda \propto \sqrt{\lambda_{\max}(\hat{\boldsymbol{\Sigma}})} \cdot G_d(n),$$

where $\hat{\boldsymbol{\Sigma}}$ is the covariance matrix of $\hat{P}$. We use the second moment $\sqrt{\lambda_{\max}(\hat{\boldsymbol{\Sigma}})}$ instead of the ψ-th moment since they are close when $\psi \approx 2$. Note that this λ is approximately the upper bound on $W_1(P, \hat{P})$ given in (Fournier and Guillin, 2015), Theorem 1 (which is used in the proof of Theorem E.1).

Before we prove Theorem E.1, we show the following results about symmetric quasiconcave distributions.

Proposition E.2. *Let* p *be a symmetric quasiconcave distribution over* $\mathbb{R}^d$ *centered at* 0 *with covariance matrix* $\boldsymbol{\Sigma}$*. Let* $\mathbf{x} \sim p$*,* $\tilde{\mathbf{x}} \sim \tilde{p}$*, where* $W_1(p, \tilde{p}) \leq \varpi$*. Let* $\mathbf{a} \in \mathbb{R}^d \backslash \{0\}$*. For* $t > 0$*,*

$$\begin{aligned} \mathbb{P}(|\mathbf{a}^T \tilde{\mathbf{x}}| \geq t) &\leq \frac{\mathbb{E}[(\mathbf{a}^T \mathbf{x})^2] + 2t\varpi\|\mathbf{a}\|_2}{t^2} \\ &\leq \frac{\|\mathbf{a}\|_2^2 \lambda_{\max}(\boldsymbol{\Sigma})}{t^2} + \frac{2\varpi\|\mathbf{a}\|_2}{t}. \end{aligned}$$

Proof. By Proposition C.1.1, it suffices to consider $d = 1$, $\mathbf{a} = [1]$. Consider $h(x) := x^2$ for $|x| \leq t$, $h(x) := 2t|x| - t^2$ for $|x| > t$. By Markov inequality,

$$\begin{aligned} \mathbb{P}(|\tilde{x}| \geq t) &\leq t^{-2}\mathbb{E}[h(\tilde{x})] \\ &\leq t^{-2}\left(\mathbb{E}[h(x)] + 2t\varpi\right) \\ &\leq t^{-2}\left(\mathbb{E}[x^2] + 2t\varpi\right). \end{aligned}$$

□

We now prove Theorem E.1.

Proof of Theorem E.1. Let $\tilde{p}_1, \ldots, \tilde{p}_k$ be distributions such that $\tilde{P} = \sum_{i=1}^k w_i \tilde{p}_i$ and $\varpi := W_1(P, \tilde{P}) = \sum_{i=1}^k w_i \varpi_i$, where $\varpi_i := W_1(p_i, \tilde{p}_i)$. We prove Theorem 6.1 by modifying the proof of Theorem 4.3. Instead of (9), we have, by Proposition E.2,

$$\begin{aligned} \tilde{p}_i(T^c) &\leq \tilde{p}_i\left(\mathbf{x} : \left|\frac{(\boldsymbol{\mu} - \boldsymbol{\mu}_i)^T}{\|\boldsymbol{\mu} - \boldsymbol{\mu}_i\|_2}(\mathbf{x} - \boldsymbol{\mu}_i)\right| > r\right) \\ &\leq \frac{\lambda_{\max}(\boldsymbol{\Sigma}_i)}{r^2} + \frac{2\varpi_i}{r} \\ &\leq \frac{\sigma_{\max}^2}{r^2} + \frac{2\varpi_i}{r} \\ &\leq \left(\frac{k\sigma_{\max}}{\delta_{\boldsymbol{\mu}}\epsilon}\right)^2 + \frac{2k\varpi_i}{\delta_{\boldsymbol{\mu}}\epsilon}, \end{aligned} \tag{30}$$

where the last line is by (8). Hence,

$$\begin{aligned} \tilde{P}(T^c) &= \sum_{i=1}^k w_i \tilde{p}_i(T^c) \\ &\leq \left(\frac{k\sigma_{\max}}{\delta_{\boldsymbol{\mu}}\epsilon}\right)^2 + \frac{2k\varpi}{\delta_{\boldsymbol{\mu}}\epsilon}. \end{aligned}$$

Therefore,

$$
\begin{aligned}
& D_f\big(\tilde{P} \,\big\|\, Q\big) \\
& \geq D_f\big(\tilde{P}(T) \,\big\|\, Q(T)\big) \\
& \geq D_f\left(\max\left\{1 - \left(\frac{k\sigma_{\max}}{\delta_{\boldsymbol{\mu}}\epsilon}\right)^2 - \frac{2k\varpi}{\delta_{\boldsymbol{\mu}}\epsilon},\, \epsilon\right\} \,\middle\|\, \epsilon\right) \\
& \geq D_f\left(\max\left\{1 - \zeta^{-2}\epsilon^{-2} - \frac{2\varpi}{\zeta\epsilon\sigma_{\max}},\, \epsilon\right\} \,\middle\|\, \epsilon\right),
\end{aligned}
$$

where

$$
\zeta := \frac{\delta_{\boldsymbol{\mu}}}{k\sigma_{\max}}.
$$

Substituting $\epsilon = \zeta^{-2/3}$, we have

$$
\begin{aligned}
& W_1(P, \tilde{P}) + \lambda D_f(\tilde{P} \,\|\, Q) - \lambda f(w_{\max}) \\
& \geq \lambda\left(D_f\left(\max\left\{1 - \zeta^{-2}\epsilon^{-2} - \frac{2\varpi}{\zeta\epsilon\sigma_{\max}},\, \epsilon\right\} \,\middle\|\, \epsilon\right) - f\left(\frac{1}{k}\right)\right) + \varpi \\
& = \lambda\left(D_f\left(\max\left\{1 - \zeta^{-2/3} - \frac{2\varpi}{\zeta^{1/3}\sigma_{\max}},\, \zeta^{-2/3}\right\} \,\middle\|\, \zeta^{-2/3}\right) - f\left(\frac{1}{k}\right)\right) + \varpi \\
& \geq \lambda\left(D_f\left(1 - \zeta^{-2/3} \,\middle\|\, \zeta^{-2/3}\right) - f\left(\frac{1}{k}\right)\right) \\
& =: \theta,
\end{aligned}
$$

where the last inequality occurs by convexity of D_f and monotonicity in ϖ if the derivative of the second-to-last line (with respect to ϖ)

$$
\begin{aligned}
& -\lambda\frac{2}{\zeta^{1/3}\sigma_{\max}}\breve{f}(\zeta^{-2/3}) + 1 \geq 0 \\
& \Leftrightarrow\ \lambda \leq \frac{\zeta^{1/3}}{2\breve{f}(\zeta^{-2/3})}\sigma_{\max}.
\end{aligned}
$$

We have shown that any Q with $\min_i \|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_i\|_2 \geq k\sigma_{\max}\zeta$ is suboptimal when the minimization objective is $D_{\lambda f, W_1}(P \,\|\, Q)$, by a gap at least θ. It remains to show that $W_1(P, \hat{P}) < \theta/2$. If this is true, since $|D_{\lambda f, W_1}(P\|Q) - D_{\lambda f, W_1}(\hat{P}\|Q)| < \theta/2$, we know that any Q with $\min_i \|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_i\|_2 \geq k\sigma_{\max}\zeta$ is suboptimal when the minimization objective is $D_{\lambda f, W_1}(\hat{P} \,\|\, Q)$. By Theorem 1 in (Fournier and Guillin, 2015), there exists constant $C_{d,\psi}$ such that

$$
\mathbb{E}\left[W_1(P, \hat{P})\right] \leq \frac{C_{d,\psi}}{2}\left(\mathbb{E}\left[\|\mathbf{x}\|^\psi\right]\right)^{1/\psi} G(n).
$$

By Markov inequality,

$$
\begin{aligned}
& \mathbb{P}\left(W_1(P, \hat{P}) \geq \theta/2\right) \\
& \leq \frac{C_{d,\psi}}{\lambda} \cdot \frac{\left(\mathbb{E}\left[\|\mathbf{x}\|^\psi\right]\right)^{1/\psi} G(n)}{D_f(1 - \zeta^{-2/3} \,\|\, \zeta^{-2/3}) - f(k^{-1})}.
\end{aligned}
$$

The result follows. □

Remark E.3. While it may be possible to extend Theorem E.1 to the strongly and uniformly mode-seeking cases so as to obtain $\min_i \|\boldsymbol{\mu}_Q - \boldsymbol{\mu}_i\|_2 = o(\sigma_{\max})$ instead, we do not consider these cases here due to their complexity.