
Finding Mixed Nash Equilibria of Generative Adversarial Networks

Ya-Ping Hsieh¹ Chen Liu¹ Volkan Cevher¹

Abstract

Generative adversarial networks (GANs) are known to achieve the state-of-the-art performance on various generative tasks, but these results come at the expense of a notoriously difficult training phase. Current training strategies typically draw a connection to optimization theory, whose scope is restricted to *local* convergence due to the presence of non-convexity. In this work, we tackle the training of GANs by rethinking the problem formulation from the *mixed Nash Equilibria* (NE) perspective. Via a classical lifting trick, we show that essentially all existing GAN objectives can be relaxed into their mixed strategy forms, whose *global* optima can be solved via *sampling*, in contrast to the exclusive use of optimization framework in previous work. We further propose a mean-approximation sampling scheme, which allows to systematically exploit methods for bi-affine games to delineate novel, practical training algorithms of GANs. Finally, we provide experimental evidence that our approach yields comparable or superior results to contemporary training algorithms, and outperforms classical methods such as SGD, Adam, and RMSProp.

1. Introduction

Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) achieve the state-of-the-art for learning real-world images (Brock et al., 2018; Karras et al., 2018), as well as a number of applications including image translation (Isola et al., 2017; Kim et al., 2017; Zhu et al., 2017), super-resolution imaging (Wang et al., 2015; Ledig et al., 2017), pose editing (Pumarola et al., 2018b), and facial animation (Pumarola et al., 2018a).

A major obstacle blocking their full impact in machine learning is its notoriously difficult training phase (Goodfellow,

¹LIONS, EPFL, Switzerland. Correspondence to: Ya-Ping Hsieh <ya-ping.hsieh@epfl.ch>.

2016). Indeed, the stochastic gradient descent (SGD), along with its variants, remain the method of choice for training GANs. In order to understand the training dynamics of gradient-based algorithms, a popular approach is to cast the training into a min-max program, and then invoke results in the mathematical programming or algorithmic game theory literature, with the caveat that the theory applies almost exclusively to convex-concave objectives.

Despite having inspired many novel algorithms, the major deviations from theory met several limitations. On one hand, in the absence of convexity, the theory is bound to focus only on the *local* convergence. Even under such a strong restriction, it is recently shown by Adolfs et al. (2018) that there exist *stable* stationary points that attract gradient-based algorithms, but that are not locally min-max, suggesting that even the local theory can break down if one blindly applies intuitions from convex optimization.

On the other hand, current approaches tend to tackle different issues on a case-by-case basis, rendering the solution of one incompatible to another, and therefore restricting their applicability in new problems. Furthermore, to our knowledge, these theoretical insights do not translate into rigorous convergence of the proposed algorithms, even under as strong of an assumption as that SGD can globally optimize a *single* neural net.

The aim of this paper is to resolve the above issues by making the following contributions:

1. We propose to study the *mixed Nash Equilibria* (NE) of GANs, which are *global optima* of infinite-dimensional bi-affine games. We show that, in principle, any existing GAN objective can be lifted into its mixed NE formulation, thereby admitting the search for a globally optimizing algorithm.
2. We demonstrate how one can consistently derive novel algorithms for finding mixed NE of GANs, via tapping into the abundant literature for solving finite-dimensional bi-affine games. We showcase this by proving rigorous convergence results for simple extensions of two prox methods, namely the Mirror Descent (Nemirovsky & Yudin, 1983) and Mirror-Prox (Nemirovski, 2004), under the empirically-supported assumption that *sampling* succeeds for a single neural

net (Chaudhari et al., 2017; 2018; Dziugaite & Roy, 2018).

3. Although our framework can motivate new convergent algorithms, they require impractical computational resources. To this end, we construct a principled procedure to reduce our novel prox methods to certain simple iterative methods for approximating the mean, whose per-iteration complexity is as cheap as SGD.
4. We experimentally show that our algorithms consistently achieve better or comparable performance than popular baselines such as SGD, Adam, and RMSProp, as well as the recently proposed algorithms (Gidel et al., 2018) that are inspired by the same prox methods, but which are from the pure (local) NE perspective.

An important feature of our framework is its flexibility, in that we can essentially take any GAN objective, any method for solving finite-dimensional bi-affine games, and return a novel algorithm that is as scalable as SGD. See Section 6 for concrete examples.

Related Work: While the literature on training GANs is vast, to our knowledge, there exist only few papers on the mixed NE perspective. The notion of mixed NE is already present in (Goodfellow et al., 2014), but is stated only as an existential result. The authors of (Arora et al., 2017) advocate the mixed strategies, but do not provide a provably convergent algorithm. Oliehoek et al. (2018) also consider mixed NE, but only with countably many parameters, and is under the unconventional assumption of finite floating point numbers to represent GANs. The work (Grnarova et al., 2018) proposes a provably convergent algorithm for finding the mixed NE of GANs under the unrealistic assumption that the discriminator is a single-layered neural network. In contrast, our results are applicable to arbitrary architectures, including popular ones (Arjovsky et al., 2017; Gulrajani et al., 2017).

There exists many novel algorithms that are inspired by optimization or algorithmic game theory (Balduzzi et al., 2018; Daskalakis et al., 2018; Gidel et al., 2018; Mertikopoulos et al., 2018), and each of them tackles different problems of classical gradient methods. However, these works focus on the classical pure strategy equilibria, and are hence distinct from our problem formulation. In particular, they give rise to drastically different algorithms.

In terms of analysis techniques, our framework is related to learning games in Banach spaces (Sridharan & Tewari, 2010; Srebro et al., 2011), and is closest to that of (Balandat et al., 2016), but with several important distinctions. First, the analysis of (Balandat et al., 2016) is based on dual averaging (Nesterov, 2009), while we consider Mirror Descent and also the more sophisticated Mirror-Prox (see Section 3).

Second, unlike our work, (Balandat et al., 2016) do not provide any convergence rate for learning mixed NE of two-player games. Finally, (Balandat et al., 2016) is only of theoretical interest with no practical algorithm.

Notation: Throughout the paper, we use z to denote a generic variable and $\mathcal{Z} \subseteq \mathbb{R}^d$ its domain. We denote the set of all (sufficiently regular) Borel probability measures on \mathcal{Z} by $\mathcal{M}(\mathcal{Z})$, and the set of all (sufficiently regular)¹ functions on \mathcal{Z} by $\mathcal{F}(\mathcal{Z})$. We write $d\mu = \rho dz$ to mean that the density function of $\mu \in \mathcal{M}(\mathcal{Z})$ with respect to the Lebesgue measure is ρ . All integrals without specifying the measure are understood to be with respect to Lebesgue. For any objective of the form $\min_{\mathbf{x}} \max_{\mathbf{y}} F(\mathbf{x}, \mathbf{y})$, we say that $(\mathbf{x}_T, \mathbf{y}_T)$ is an $O(T^{-1/2})$ -NE if $\max_{\mathbf{x}, \mathbf{y}} \{F(\mathbf{x}_T, \mathbf{y}) - F(\mathbf{x}, \mathbf{y}_T)\} = O(T^{-1/2})$. Similarly we can define $O(T^{-1})$ -NE. The symbol $\|\cdot\|_{\mathbb{L}^\infty}$ denotes the \mathbb{L}^∞ -norm of functions, and $\|\cdot\|_{\text{TV}}$ denotes the total variation norm of probability measures.

2. Mixed Nash Equilibria and Infinite-Dimensional Bi-Affine Games

We review standard results in game theory in Section 2.1, whose proof can be found in (Bubeck, 2013a;b;c). Section 2.2 performs a lifting trick to transform GAN objectives into the mixed NE form, and then relates the training of GANs to the two-player game in Section 2.1, thereby suggesting to generalize the prox methods to infinite dimension.

2.1. Preliminary: Finite Bi-Affine Games

Consider the classical formulation of a two-player game with *finitely* many strategies:

$$\min_{\mathbf{p} \in \Delta_m} \max_{\mathbf{q} \in \Delta_n} \langle \mathbf{q}, \mathbf{a} \rangle - \langle \mathbf{q}, A\mathbf{p} \rangle, \quad (1)$$

where A is a payoff matrix, \mathbf{a} is a vector, and $\Delta_d := \left\{ \mathbf{z} \in \mathbb{R}^d \mid \sum_{i=1}^d z_i = 1 \right\}$ is the probability simplex, representing the *mixed strategies* (i.e., probability distributions) over d pure strategies. A pair $(\mathbf{p}_{\text{NE}}, \mathbf{q}_{\text{NE}})$ achieving the min-max value in (1) is called a mixed NE.

Assume that the matrix A is too expensive to evaluate whereas the (stochastic) gradients of (1) are easy to obtain. Under such settings, a celebrated algorithm, the so-called **entropic Mirror Descent** (entropic MD), learns an $O(T^{-1/2})$ -NE: Let $\phi(\mathbf{z}) := \sum_{i=1}^d z_i \log z_i$ be the entropy function and $\phi^*(\mathbf{y}) := \log \sum_{i=1}^d e^{y_i} = \sup_{\mathbf{z} \in \Delta_d} \{ \langle \mathbf{z}, \mathbf{y} \rangle - \phi(\mathbf{z}) \}$ be its Fenchel dual. For a learning rate η and an

¹See (A.1) and (A.2) for precise definitions.

arbitrary vector $\mathbf{b} \in \mathbb{R}^d$, define the MD iterates as

$$\begin{aligned} \mathbf{z}' &= \text{MD}_\eta(\mathbf{z}, \mathbf{b}) \equiv \mathbf{z}' = \nabla \phi^*(\nabla \phi(\mathbf{z}) - \eta \mathbf{b}) \\ &\equiv z'_i = \frac{z_i e^{-\eta b_i}}{\sum_{i=1}^d z_i e^{-\eta b_i}}, \quad \forall 1 \leq i \leq d. \end{aligned} \quad (2)$$

The update rule takes linear time in dimension, which is highly scalable.

Denote by $\bar{\mathbf{p}}_T := \frac{1}{T} \sum_{t=1}^T \mathbf{p}_t$ and $\bar{\mathbf{q}}_T := \frac{1}{T} \sum_{t=1}^T \mathbf{q}_t$ the ergodic average of two sequences $\{\mathbf{p}_t\}_{t=1}^T$ and $\{\mathbf{q}_t\}_{t=1}^T$. Then, with a properly chosen step-size η , the iterates

$$\begin{cases} \mathbf{p}_{t+1} = \text{MD}_\eta(\mathbf{p}_t, -A^\top \mathbf{q}_t) \\ \mathbf{q}_{t+1} = \text{MD}_\eta(\mathbf{q}_t, -\mathbf{a} + A\mathbf{p}_t) \end{cases}$$

come with the guarantee that $(\bar{\mathbf{p}}_T, \bar{\mathbf{q}}_T)$ is an $O(T^{-1/2})$ -NE. Moreover, a slightly more complicated algorithm, called the **entropic Mirror-Prox** (entropy MP) (Nemirovski, 2004), achieves faster rate than the entropic MD:

$$\begin{cases} \mathbf{p}_t = \text{MD}_\eta(\tilde{\mathbf{p}}_t, -A^\top \tilde{\mathbf{q}}_t) \\ \mathbf{q}_t = \text{MD}_\eta(\tilde{\mathbf{q}}_t, -\mathbf{a} + A\tilde{\mathbf{p}}_t) \\ \tilde{\mathbf{p}}_{t+1} = \text{MD}_\eta(\tilde{\mathbf{p}}_t, -A^\top \mathbf{q}_t) \\ \tilde{\mathbf{q}}_{t+1} = \text{MD}_\eta(\tilde{\mathbf{q}}_t, -\mathbf{a} + A\mathbf{p}_t) \end{cases}$$

implies that $(\bar{\mathbf{p}}_T, \bar{\mathbf{q}}_T)$ is an $O(T^{-1})$ -NE. If, instead of deterministic gradients, one uses unbiased stochastic gradients for entropic MD and MP, then both algorithms achieve $O(T^{-1/2})$ -NE in expectation.

2.2. Mixed Strategy Formulation for Generative Adversarial Networks

For illustration, let us focus on the Wasserstein GAN (Arjovsky et al., 2017), whereas the derivation in this section applies to any min-max objective. We perform a common bilinearization trick that dates back at least to the early game theory literature (Glicksberg, 1952), which is also well-known in optimal transport theory (Villani, 2003).

The training objective of Wasserstein GAN is

$$\min_{\theta \in \Theta} \max_{w \in \mathcal{W}} \mathbb{E}_{X \sim \mathbb{P}_{\text{real}}}[f_w(X)] - \mathbb{E}_{X \sim \mathbb{P}_\theta}[f_w(X)], \quad (3)$$

where Θ is the set of parameters for the generator and \mathcal{W} the set of parameters for the discriminator f , typically both taken to be neural nets.

The high-level idea of our approach is, instead of solving (3) directly, we focus on the *mixed strategy* formulation of (3). In other words, we consider the set of all probability distributions over Θ and \mathcal{W} , and we search for the optimal distribution that solves the following program:

$$\begin{aligned} \min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} & \mathbb{E}_{w \sim \mu} \mathbb{E}_{X \sim \mathbb{P}_{\text{real}}}[f_w(X)] \\ & - \mathbb{E}_{w \sim \mu} \mathbb{E}_{\theta \sim \nu} \mathbb{E}_{X \sim \mathbb{P}_\theta}[f_w(X)]. \end{aligned} \quad (4)$$

Define the function $g : \mathcal{W} \rightarrow \mathbb{R}$ by $g(w) := \mathbb{E}_{X \sim \mathbb{P}_{\text{real}}}[f_w(X)]$ and the operator $G : \mathcal{M}(\Theta) \rightarrow \mathcal{F}(\mathcal{W})$ as $(G\nu)(w) := \mathbb{E}_{\theta \sim \nu, X \sim \mathbb{P}_\theta}[f_w(X)]$. Denoting $\langle \mu, h \rangle := \mathbb{E}_\mu h$ for any probability measure μ and function h , we may rewrite (4) as

$$\min_{\nu \in \mathcal{M}(\Theta)} \max_{\mu \in \mathcal{M}(\mathcal{W})} \langle \mu, g \rangle - \langle \mu, G\nu \rangle. \quad (5)$$

Furthermore, the derivative (the analogue of gradient in infinite dimension) of (5) with respect to μ is simply $g - G\nu$, and the derivative of (5) with respect to ν is $-G^\dagger \mu$, where $G^\dagger : \mathcal{M}(\mathcal{W}) \rightarrow \mathcal{F}(\Theta)$ is the adjoint operator of G defined via the relation

$$\forall \mu \in \mathcal{M}(\mathcal{W}), \nu \in \mathcal{M}(\Theta), \quad \langle \mu, G\nu \rangle = \langle \nu, G^\dagger \mu \rangle. \quad (6)$$

One can easily check that $(G^\dagger \mu)(\theta) := \mathbb{E}_{X \sim \mathbb{P}_\theta, w \sim \mu}[f_w(X)]$ achieves the equality in (6).

To summarize, the mixed strategy formulation of Wasserstein GAN is (5), whose derivatives can be expressed in terms of g and G .

Now, observe that (5) is exactly the infinite-dimensional analogue of (1): The distributions over finite strategies are replaced with probability measures over a continuous parameter set, the vector \mathbf{a} is replaced with a function g , the matrix A is replaced with a linear operator² G , and the gradients are replaced with derivatives. Based on Section 2.1, it is then natural to ask:

Can the entropic Mirror Descent and Mirror-Prox be extended to infinite dimension to solve (5)? Are there scalable implementations of these algorithms, at least approximately?

We provide an affirmative answer to the first question in Section 3. The so obtained algorithms, nonetheless, are infinite-dimensional and requires infinite computational power to implement. For practical interest, in Section 4 we propose a sampling framework to approximate the infinite-dimensional prox methods in Section 3.

3. Infinite-Dimensional Prox Methods

This section builds a rigorous infinite-dimensional formalism in parallel to the finite-dimensional prox methods and proves their convergence rates. We remark that these results are folklore among optimization experts and hence we adopt an informal presentation here, deferring all the technical details to the appendix. However, to our knowledge, they are not published.

We first recall the notion of derivative in infinite-dimensional spaces. A (nonlinear) functional $\Phi : \mathcal{M}(\mathcal{Z}) \rightarrow \mathbb{R}$ is said to

²The linearity of G trivially follows from the linearity of expectation.

Algorithm 1 INFINITE-DIMENSIONAL ENTROPIC MD

Input: Initial distributions μ_1, ν_1 , learning rate η

for $t = 1, 2, \dots, T-1$ **do**

$$\begin{cases} \nu_{t+1} = \text{MD}_\eta(\nu_t, -G^\dagger \mu_t) \\ \mu_{t+1} = \text{MD}_\eta(\mu_t, -g + G\nu_t) \end{cases}$$

return $\bar{\nu}_T = \frac{1}{T} \sum_{t=1}^T \nu_t$ and $\bar{\mu}_T = \frac{1}{T} \sum_{t=1}^T \mu_t$.

possess a derivative at $\mu \in \mathcal{M}(\mathcal{Z})$ if there exists a function $d\Phi(\mu) \in \mathcal{F}(\mathcal{Z})$ such that, for all $\mu' \in \mathcal{M}(\mathcal{Z})$, we have

$$\Phi(\mu + \epsilon\mu') = \Phi(\mu) + \epsilon \langle \mu', d\Phi(\mu) \rangle + o(\epsilon).$$

Similarly, a (nonlinear) functional $\Phi^* : \mathcal{F}(\mathcal{Z}) \rightarrow \mathbb{R}$ is said to possess a derivative at $h \in \mathcal{F}(\mathcal{Z})$ if there exists a measure $d\Phi^*(h) \in \mathcal{M}(\mathcal{Z})$ such that, for all $h' \in \mathcal{F}(\mathcal{Z})$, we have

$$\Phi^*(h + \epsilon h') = \Phi^*(h) + \epsilon \langle d\Phi^*(h), h' \rangle + o(\epsilon).$$

The most important functionals in this paper are the (negative) Shannon entropy

$$\mu \in \mathcal{M}(\mathcal{Z}), \quad \Phi(\mu) := \int d\mu \log \frac{d\mu}{dz}$$

and its Fenchel dual

$$h \in \mathcal{F}(\mathcal{Z}), \quad \Phi^*(h) := \log \int e^h dz.$$

The first result of our paper is to show that, in direct analogy to (2), the infinite-dimensional MD iterates can be expressed as:

Theorem 1 (Infinite-Dimensional Mirror Descent, informal). *For a learning rate η and an arbitrary function h , we can equivalently define*

$$\begin{aligned} \mu_+ &= \text{MD}_\eta(\mu, h) \equiv \mu_+ = d\Phi^*(d\Phi(\mu) - \eta h) \\ &\equiv d\mu_+ = \frac{e^{-\eta h} d\mu}{\int e^{-\eta h} d\mu}. \end{aligned} \quad (7)$$

Moreover, most of the essential ingredients in the analysis of finite-dimensional prox methods can be generalized to infinite dimension.

See **Theorem 4** of Appendix A for precise statements and a long list of “essential ingredients of prox methods” generalizable to infinite dimension.

We are now ready introduce two “conceptual” algorithms for solving the mixed NE of Wasserstein GANs: The infinite-dimensional entropic MD in **Algorithm 1** and MP in **Algorithm 2**.

Theorem 2 (Convergence Rates, informal). *Let $\Phi(\mu) = \int d\mu \log \frac{d\mu}{dz}$, and let $D(\cdot, \cdot)$ be the relative entropy. Then, with a properly chosen step-size η , we have*

Algorithm 2 INFINITE-DIMENSIONAL ENTROPIC MP

Input: Initial distributions $\tilde{\mu}_1, \tilde{\nu}_1$, learning rate η

for $t = 1, 2, \dots, T$ **do**

$$\begin{cases} \nu_t = \text{MD}_\eta(\tilde{\nu}_t, -G^\dagger \tilde{\mu}_t) \\ \mu_t = \text{MD}_\eta(\tilde{\mu}_t, -g + G\tilde{\nu}_t) \\ \tilde{\nu}_{t+1} = \text{MD}_\eta(\tilde{\nu}_t, -G^\dagger \mu_t) \\ \tilde{\mu}_{t+1} = \text{MD}_\eta(\tilde{\mu}_t, -g + G\nu_t) \end{cases}$$

return $\bar{\nu}_T = \frac{1}{T} \sum_{t=1}^T \nu_t$ and $\bar{\mu}_T = \frac{1}{T} \sum_{t=1}^T \mu_t$.

1. Assume that we have access to the deterministic derivatives. Then **Algorithm 1** achieves $O(T^{-1/2})$ -NE, and **Algorithm 2** achieves $O(T^{-1})$ -NE.

2. Assume that we have access to stochastic derivatives such that the bias and the variance are small. Then **Algorithm 1** with stochastic derivatives achieves $O(T^{-1/2})$ -NE in expectation, and **Algorithm 2** with stochastic derivatives achieves $O(T^{-1/2})$ -NE in expectation.

The precise statements of **Theorem 2** and their proofs can be found in Appendix B.

4. A Sampling Framework for Approximate Infinite-Dimensional Prox Methods

Section 4.1 reduces **Algorithm 1** and **Algorithm 2** to a sampling routine (Welling & Teh, 2011) that has widely been used in machine learning. Section 4.2 proposes to further simplify the algorithms by summarizing a batch of samples by their mean.

For simplicity, we will only derive the algorithm for entropic MD; the case for entropic MP is similar but requires more computation. To ease the notation, we assume $\eta = 1$ throughout this section as η does not play an important role in the derivation below.

4.1. Implementable Entropic MD: From Probability Measure to Samples

We demonstrate how **Algorithm 1** with stochastic derivatives can be reduced to simple sampling tasks. Consider **Algorithm 1**. The reduction consists of three steps.

Step 1: Reformulating Entropic Mirror Descent Iterates

The definition of the MD iterate (7) relates the updated probability measure μ_{t+1} to the current probability measure μ_t , but it tells us nothing about the density function of μ_{t+1} , from which we want to sample. Our first step is to express (7) in a more tractable form. By recursively applying (7) and using **Theorem 4.10** in Appendix A, we have, for some

constants C_1, \dots, C_{T-1} ,

$$\begin{aligned} d\Phi(\mu_T) &= d\Phi(\mu_{T-1}) - (-g + G\nu_{T-1}) + C_{T-1} \\ &= d\Phi(\mu_{T-2}) - (-g + G\nu_{T-2}) \\ &\quad - (-g + G\nu_{T-1}) + C_{T-1} + C_{T-2} \\ &= \dots \\ &= d\Phi(\mu_1) - \left(-(T-1)g + G \sum_{s=1}^{T-1} \nu_s \right) + \sum_{s=1}^{T-1} C_s. \end{aligned}$$

For simplicity, assume that μ_1 is uniform so that $d\Phi(\mu_1)$ is a constant function. Then, by (A.7) and that $d\Phi^*(d\Phi(\mu_T)) = d\mu_T$, we see that the density function of μ_T is simply $d\mu_T = \frac{\exp\{(T-1)g - G \sum_{s=1}^{T-1} \nu_s\} d\mathbf{w}}{\int \exp\{(T-1)g - G \sum_{s=1}^{T-1} \nu_s\} d\mathbf{w}}$. Similarly, we have $d\nu_T = \frac{\exp\{G^\dagger \sum_{s=1}^{T-1} \mu_s\} d\boldsymbol{\theta}}{\int \exp\{G^\dagger \sum_{s=1}^{T-1} \mu_s\} d\boldsymbol{\theta}}$.

Step 2: Empirical Approximation for Stochastic Derivatives

The derivatives of (5) involve the function g and operator G . Recall that g requires taking expectation over the real data distribution, which we do not have access to. A common approach is to replace the true expectation with its empirical average:

$$g(\mathbf{w}) = \mathbb{E}_{X \sim \mathbb{P}_{\text{real}}} [f_{\mathbf{w}}(X)] \simeq \frac{1}{n} \sum_{i=1}^n f_{\mathbf{w}}(X_i^{\text{real}}) \triangleq \hat{g}(\mathbf{w})$$

where X_i 's are real data and n is the batch size. Clearly, \hat{g} is an unbiased estimator of g .

On the other hand, $G\nu_t$ and $G^\dagger \mu_t$ involve expectation over ν_t and μ_t , respectively, and also over the fake data distribution $\mathbb{P}_{\boldsymbol{\theta}}$. Therefore, if we are able to draw samples from μ_t and ν_t , then we can again approximate the expectation via the empirical average:

$$\begin{aligned} \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots, \boldsymbol{\theta}^{(n')} &\sim \nu_t, \left\{ X_i^{(j)} \right\}_{i=1}^n \sim \mathbb{P}_{\boldsymbol{\theta}^{(j)}}, \\ \hat{G}\nu_t(\mathbf{w}) &\simeq \frac{1}{nn'} \sum_{i=1}^n \sum_{j=1}^{n'} f_{\mathbf{w}}(X_i^{(j)}), \end{aligned}$$

and similarly,

$$\begin{aligned} \mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \dots, \mathbf{w}^{(n')} &\sim \mu_t, \{X_i\}_{i=1}^n \sim \mathbb{P}_{\boldsymbol{\theta}}, \\ \hat{G}^\dagger \mu_t(\boldsymbol{\theta}) &\simeq \frac{1}{nn'} \sum_{i=1}^n \sum_{j=1}^{n'} f_{\mathbf{w}^{(j)}}(X_i). \end{aligned}$$

Now, assuming that we have obtained unbiased stochastic derivatives $-\sum_{s=1}^t \hat{G}^\dagger \mu_s$ and $\sum_{s=1}^t (-\hat{g} + \hat{G}\nu_s)$, how do we actually draw samples from μ_{t+1} and ν_{t+1} ? Provided we can answer this question, then we can start with two

easy-to-sample distributions (μ_1, ν_1) , and then we will be able to draw samples from (μ_2, ν_2) . These samples in turn will allow us to draw samples from (μ_3, ν_3) , and so on. Therefore, it only remains to answer the above question. This leads us to:

Step 3: Sampling by Stochastic Gradient Langevin Dynamics

For any probability distribution with density function $e^{-h} d\mathbf{z}$, the Stochastic Gradient Langevin Dynamics (SGLD) (Welling & Teh, 2011) iterates as

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \gamma \hat{\nabla} h(\mathbf{z}_k) + \sqrt{2\gamma\epsilon} \xi_k, \quad (8)$$

where γ is the step-size, $\hat{\nabla} h$ is an unbiased estimator of ∇h , ϵ is the thermal noise, and $\xi_k \sim \mathcal{N}(0, I)$ is a standard normal vector, independently drawn across different iterations.

Suppose we start at (μ_1, ν_1) . Plugging $h \leftarrow -\hat{G}^\dagger \mu_1$ and $h \leftarrow -\hat{g} + \hat{G}\nu_1$ into (8), we obtain, for $\{X_i\}_{i=1}^n \sim \mathbb{P}_{\boldsymbol{\theta}_k}$, $\{\mathbf{w}^{(j)}\}_{j=1}^{n'} \sim \mu_1$, standard normal ξ_k, ξ'_k , and $X_i^{\text{real}} \sim \mathbb{P}_{\text{real}}$, $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{n'} \sim \nu_1$, $\{X_i^{(j)}\} \sim \mathbb{P}_{\boldsymbol{\theta}^{(j)}}$, the following update rules:

$$\begin{aligned} \boldsymbol{\theta}_{k+1} &= \boldsymbol{\theta}_k + \gamma \nabla_{\boldsymbol{\theta}} \left(\frac{1}{nn'} \sum_{i=1}^n \sum_{j=1}^{n'} f_{\mathbf{w}^{(j)}}(X_i) \right) + \sqrt{2\gamma\epsilon} \xi_k, \\ \mathbf{w}_{k+1} &= \mathbf{w}_k + \gamma \nabla_{\mathbf{w}} \left(\frac{1}{n} \sum_{i=1}^n f_{\mathbf{w}_k}(X_i^{\text{real}}) \right. \\ &\quad \left. - \frac{1}{nn'} \sum_{i=1}^n \sum_{j=1}^{n'} f_{\mathbf{w}_k}(X_i^{(j)}) \right) + \sqrt{2\gamma\epsilon} \xi'_k. \end{aligned}$$

The theory of (Welling & Teh, 2011; Teh et al., 2016) states that, for large enough k , the iterates of SGLD above (approximately) generate samples according to the probability measures (μ_2, ν_2) . We can then apply this process recursively to obtain samples from $(\mu_3, \nu_3), (\mu_4, \nu_4), \dots, (\mu_T, \nu_T)$. Finally, since the entropic MD and MP output the averaged measure $(\bar{\mu}_T, \bar{\nu}_T)$, it suffices to pick a random index $\hat{t} \in \{1, 2, \dots, T\}$ and then output samples from $(\mu_{\hat{t}}, \nu_{\hat{t}})$.

Putting Step 1-3 together, we obtain **Algorithm 4** and **5** in Appendix C.

Remark. In principle, any first-order sampling method is valid above. In the experimental section, we also use a RMSProp-preconditioned version of the SGLD (Li et al., 2016).

4.2. Summarizing Samples by Averaging: A Simple yet Effective Heuristic

Although **Algorithm 4** and **5** are implementable, they are quite complicated and resource-intensive, as the total computu-

Algorithm 3 MIRROR-GAN: APPROXIMATE MIRROR DECENT FOR GANS

Input: $\bar{\mathbf{w}}_1, \bar{\boldsymbol{\theta}}_1 \leftarrow$ random initialization, $\{\gamma_t\}_{t=1}^T, \{\epsilon_t\}_{t=1}^T, \{K_t\}_{t=1}^{T-1}, \beta$ (see Appendix C for meaning of the hyperparameters), standard normal noise ξ_k, ξ'_k .

for $t = 1, 2, \dots, T-1$ **do**

$\bar{\mathbf{w}}_t, \mathbf{w}_t^{(1)} \leftarrow \mathbf{w}_t, \bar{\boldsymbol{\theta}}_t, \boldsymbol{\theta}_t^{(1)} \leftarrow \boldsymbol{\theta}_t$

for $k = 1, 2, \dots, K_t$ **do**

Generate $A = \{X_1, \dots, X_n\} \sim \mathbb{P}_{\boldsymbol{\theta}_t^{(k)}}, \boldsymbol{\theta}_t^{(k+1)} = \boldsymbol{\theta}_t^{(k)} + \frac{\gamma_t}{n} \nabla_{\boldsymbol{\theta}} \sum_{X_i \in A} f_{\mathbf{w}_t}(X_i) + \sqrt{2\gamma_t} \epsilon_t \xi_k$

Generate $B = \{X_1^{\text{real}}, \dots, X_n^{\text{real}}\} \sim \mathbb{P}_{\text{real}}$ Generate $B' = \{X'_1, \dots, X'_n\} \sim \mathbb{P}_{\boldsymbol{\theta}_t}$

$\mathbf{w}_t^{(k+1)} = \mathbf{w}_t^{(k)} + \frac{\gamma_t}{n} \nabla_{\mathbf{w}} \sum_{X_i^{\text{real}} \in B} f_{\mathbf{w}_t^{(k)}}(X_i^{\text{real}}) - \frac{\gamma_t}{n} \nabla_{\mathbf{w}} \sum_{X'_i \in B'} f_{\mathbf{w}_t^{(k)}}(X'_i) + \sqrt{2\gamma_t} \epsilon_t \xi'_k;$

$\bar{\mathbf{w}}_t \leftarrow (1 - \beta)\bar{\mathbf{w}}_t + \beta \mathbf{w}_t^{(k+1)}$ $\bar{\boldsymbol{\theta}}_t \leftarrow (1 - \beta)\bar{\boldsymbol{\theta}}_t + \beta \boldsymbol{\theta}_t^{(k+1)}$

$\mathbf{w}_{t+1} \leftarrow (1 - \beta)\mathbf{w}_t + \beta \bar{\mathbf{w}}_t$ $\boldsymbol{\theta}_{t+1} \leftarrow (1 - \beta)\boldsymbol{\theta}_t + \beta \bar{\boldsymbol{\theta}}_t$

return $\mathbf{w}_T, \boldsymbol{\theta}_T$.

tational complexity is $O(T^2)$. This high complexity comes from the fact that, when computing the stochastic derivatives, we need to store all the historical samples and evaluate new gradients at these samples.

An intuitive approach to alleviate the above issue is to try to summarize each distribution by only *one* parameter. To this end, the mean of the distribution is the most natural candidate, which has also proven effective in practice. Moreover, the mean is often easier to acquire than the actual samples. For instance, computing the mean of distributions of the form $e^{-h}dz$, where h is a loss function defined by deep neural networks, has been empirically proven successful in (Chaudhari et al., 2017; 2018; Dziugaite & Roy, 2018) via SGLD. In this paper, we adopt the same approach as in (Chaudhari et al., 2017) where we use exponential damping (the β term in **Algorithm 4.1**) to increase stability. **Algorithm 4.1**, dubbed the *Mirror-GAN*, shows how to encompass this idea into entropic MD; the pseudocode for the similar *Mirror-Prox-GAN* can be found in **Algorithm 6** of Appendix C.

5. Experimental Evidence

The purpose of our experiments is to use established baselines to demonstrate that Mirror- and Mirror-Prox-GAN consistently achieve better or comparable performance than common algorithms. We also incorporate the comparison to recently work (Daskalakis et al., 2018; Gidel et al., 2018), whose algorithms are also motivated by the same prox methods as Mirror-GAN and Mirror-Prox-GAN, but is however from the pure NE perspective.

We use visual quality of the generated images to evaluate different algorithms. We avoid reporting numerical metrics, as recent studies (Barratt & Sharma, 2018; Borji, 2018; Lucic et al., 2018) suggest that these metrics might be flawed.

Setting of the hyperparameters and more auxiliary results can be found in Appendix D.

5.1. Synthetic Data

We repeat the synthetic setup as in (Gulrajani et al., 2017). The tasks include learning the distribution of 8 Gaussian mixtures, 25 Gaussian mixtures, and the Swiss Roll. For both the generator and discriminator, we use two MLPs with three hidden layers of 512 neurons. We choose SGD and Adam as baselines, and we compare them to Mirror- and Mirror-Prox-GAN. We also incorporate two contemporary algorithms, namely the Optimistic Adam (Daskalakis et al., 2018) and (Simultaneous) Extra-Adam (Gidel et al., 2018). The step-sizes for all algorithms are determined via parameter sweeping.

All algorithms are run up to 10^5 iterations³. The results of 25 Gaussian mixtures are shown in Figure 1; An enlarged figure of 25 Gaussian Mixtures and other cases can be found in Appendix D.1.

As Figure 1 shows, SGD performs poorly in this task, while the other algorithms yield reasonable results. However, compared to Adam, Mirror- and Mirror-Prox-GAN fit the true distribution better in two aspects. First, the modes found by Mirror- and Mirror-Prox-GAN are more accurate than the ones by Adam, Optimistic Adam, and Extra-Adam, which are perceptibly biased. Second, Mirror- and Mirror-Prox-GAN perform much better in capturing the variance (how spread the blue dots are), while Adam-based algorithms tend to collapse to modes. These observations are consistent throughout the synthetic experiments; see Appendix D.1.

³One iteration here means using one mini-batch of data. It does not correspond to the T in our algorithms, as there might be multiple SGLD iterations within each time step t .

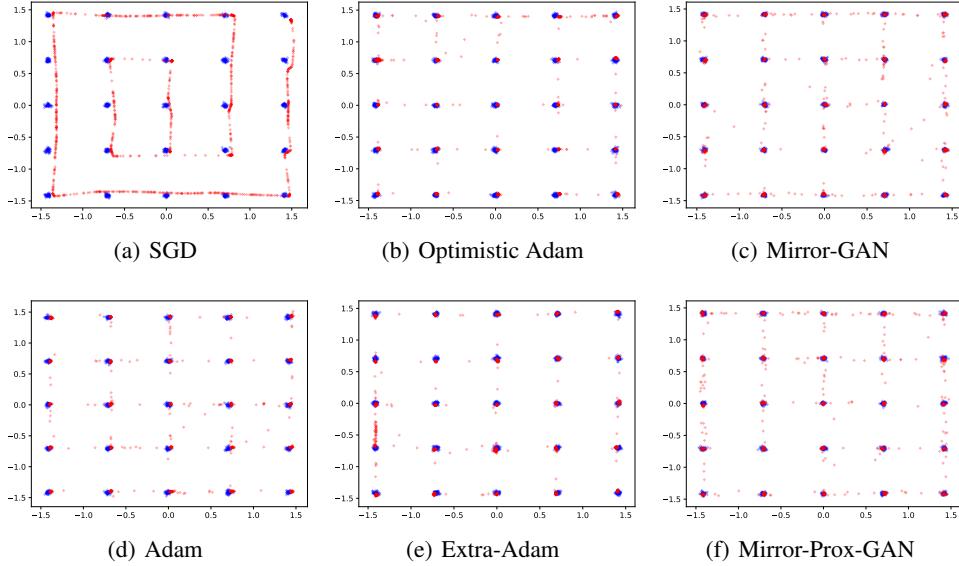


Figure 1. Fitting 25 Gaussian mixtures up to 10^5 iterations. Blue dots represent the true distribution and red ones are from the trained generator.

We also report that Mirror-GAN and Mirror-Prox-GAN are not only better in terms of solution quality, but also in speed: see Figure 6 in Appendix D.1.

5.2. Real Data

For real images, we use the LSUN bedroom dataset (Yu et al., 2015). We have also conducted a similar study with MNIST; more results can be found in Appendix D.2.

We use the same architecture (DCGAN) as in (Radford et al., 2015) with batch normalization. As the networks become deeper in this case, the gradient magnitudes differ significantly across different layers. As a result, non-adaptive methods such as SGD or SGLD do not perform well in this scenario. To alleviate such issues, we replace SGLD by the RMSProp-preconditioned SGLD (Li et al., 2016) for our sampling routines. For baselines, we consider two adaptive gradient methods: RMSprop and Adam.

We also include the Extra-Adam, along with its alternated version (Gidel et al., 2018). However, we remark that the theory of (Gidel et al., 2018) only provides motivations for simultaneous updates, and Alternated Extra-Adam should be considered as a heuristics. We drop Optimistic Adam in the this experiment since it is reported by Gidel et al. (2018) to be outperformed by Extra-Adam.

Figure 2 shows the results at the 10^5 th iteration, where step-sizes for all algorithms are determined by parameter sweeping. The RMSProp, Alternated Extra-Adam and Mirror-GAN produce images with reasonable quality, while Adam and simultaneous Extra-Adam fail to learn the distributions.

The visual quality of Alternated Extra-Adam and Mirror-GAN are comparable, and are better than RMSProp, as RMSProp sometimes generates blurry images (the (3, 3)- and (1, 5)-th entry of Figure 8.(b)).

It is worth mentioning that Adam can learn the true distribution at intermediate iterations, but later on suffers from mode collapse and finally degenerates to noise; see Appendix D.2.2.

6. Discussion and Future Work

The non-convexity arising in GANs remains poorly understood to date, and necessarily restricts the scope of optimization-based frameworks to local theory. Moreover, recent studies (Adolphs et al., 2018) suggest that even the local theory is surprisingly brittle, which possibly explains the instability of existing training methods.

Our mixed NE perspective can be considered as a first affirmative step towards a *global* theory of training GANs, under the practically-supported premise that *sampling* for a single neural net succeeds. Through experiments on both synthetic and real data, we have shown that even the very primitive products of our framework, the Mirror-GAN and Mirror-Prox-GAN, already achieve state-of-the-art in training GANs. Furthermore, thanks for the global guarantees, these algorithms are stable in the sense that they perform well in learning all the distributions we tested, while existing methods tend to work on one while fail on the others.

Our framework suggests several immediate research directions. For instance, it is evident that the deriva-



Figure 2. Dataset LSUN bedroom, 10^5 iterations.

tion in Section 2.2 applies to any objective of the form $\min_{\theta} \max_{\mathbf{w}} h(\theta, \mathbf{w})$, and hence any GAN. Besides, we have only extended two simple methods for solving bi-affine games, while there exists a rich body of literature, including the celebrated Chambolle-Pock algorithm (Chambolle & Pock, 2011), that one can import to attack the training of GANs via our framework. Finally, we have only tested the simultaneous updates for our algorithms, while one can apply the practically useful hack of alternation. Overall, we expect to see a more in-depth theoretical and empirical investigation for mixed NE of GANs.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement n° 725594 - time-data), Microsoft Research through its PhD scholarship program, and 2019 Google Faculty Research Award.

References

- Adolphs, L., Daneshmand, H., Lucchi, A., and Hofmann, T. Local saddle point optimization: A curvature exploitation approach. *arXiv preprint arXiv:1805.05751*, 2018.
- Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan.
- arXiv preprint arXiv:1701.07875, 2017.
- Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, pp. 224–232, 2017.
- Balandat, M., Krichene, W., Tomlin, C., and Bayen, A. Minimizing regret on reflexive banach spaces and nash equilibria in continuous zero-sum games. In *Advances in Neural Information Processing Systems*, pp. 154–162, 2016.
- Baldazzi, D., Racaniere, S., Martens, J., Foerster, J., Tuyls, K., and Graepel, T. The mechanics of n-player differentiable games. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 354–363, Stockholm, Sweden, 10–15 Jul 2018. PMLR.
- Barratt, S. and Sharma, R. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.
- Borji, A. Pros and cons of gan evaluation measures. *arXiv preprint arXiv:1802.03446*, 2018.
- Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

- Bubeck, S. Orf523: Mirror descent, part i/ii, 2013a. URL <https://blogs.princeton.edu/imabandit/2013/04/16/orf523-mirror-descent-part-iii/>.
- Bubeck, S. Orf523: Mirror descent, part ii/ii, 2013b. URL <https://blogs.princeton.edu/imabandit/2013/04/18/orf523-mirror-descent-part-iiii/>.
- Bubeck, S. Orf523: Mirror prox, 2013c. URL <https://blogs.princeton.edu/imabandit/2013/04/23/orf523-mirror-prox/>.
- Chambolle, A. and Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.
- Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-sgd: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations*, 2017.
- Chaudhari, P., Oberman, A., Osher, S., Soatto, S., and Carlier, G. Deep relaxation: partial differential equations for optimizing deep neural networks. *Research in the Mathematical Sciences*, 5(3):30, Jun 2018.
- Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training GANs with optimism. In *International Conference on Learning Representations*, 2018.
- Dziugaite, G. K. and Roy, D. Entropy-sgd optimizes the prior of a pac-bayes bound: Generalization properties of entropy-sgd and data-dependent priors. In *International Conference on Machine Learning*, pp. 1376–1385, 2018.
- Gibbs, J. W. *Elementary principles in statistical mechanics*. Yale University Press, 1902.
- Gidel, G., Berard, H., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial nets. *arXiv preprint arXiv:1802.10551*, 2018.
- Glicksberg, I. L. A further generalization of the kakutani fixed point theorem, with application to nash equilibrium points. *Proceedings of the American Mathematical Society*, 3(1):170–174, 1952.
- Goodfellow, I. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.
- Gray, R. M. *Entropy and information theory*. Springer Science & Business Media, 2011.
- Grnarova, P., Levy, K. Y., Lucchi, A., Hofmann, T., and Krause, A. An online learning approach to generative adversarial networks. In *International Conference on Learning Representations*, 2018.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.
- Halmos, P. R. *Measure theory*, volume 18. Springer, 2013.
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.
- Juditsky, A. and Nemirovski, A. First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. *Optimization for Machine Learning*, pp. 149–183, 2011.
- Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. 2018.
- Kim, T., Cha, M., Kim, H., Lee, J. K., and Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pp. 1857–1865, 2017.
- Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114. IEEE, 2017.
- Li, C., Chen, C., Carlson, D. E., and Carin, L. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *AAAI*, 2016.
- Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. Are gans created equal? a large-scale study. In *Advances in neural information processing systems*, 2018.
- Mertikopoulos, P., Zenati, H., Lecouat, B., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. Mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.
- Nemirovski, A. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

- Nemirovsky, A. and Yudin, D. Problem complexity and method efficiency in optimization. 1983.
- Nesterov, Y. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.
- Oliehoek, F. A., Savani, R., Gallego, J., van der Pol, E., and Groß, R. Beyond local nash equilibria for adversarial networks. *arXiv preprint arXiv:1806.07268*, 2018.
- Pumarola, A., Agudo, A., Martínez, A. M., Sanfeliu, A., and Moreno-Noguer, F. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 818–833, 2018a.
- Pumarola, A., Agudo, A., Sanfeliu, A., and Moreno-Noguer, F. Unsupervised person image synthesis in arbitrary poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8620–8628, 2018b.
- Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- Srebro, N., Sridharan, K., and Tewari, A. On the universality of online mirror descent. In *Advances in neural information processing systems*, pp. 2645–2653, 2011.
- Sridharan, K. and Tewari, A. Convex games in banach spaces. In *COLT*, pp. 1–13. Citeseer, 2010.
- Teh, Y. W., Thiery, A. H., and Vollmer, S. J. Consistency and fluctuations for stochastic gradient langevin dynamics. *The Journal of Machine Learning Research*, 17(1):193–225, 2016.
- Villani, C. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.
- Wang, Z., Liu, D., Yang, J., Han, W., and Huang, T. Deep networks for image super-resolution with sparse prior. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 370–378, 2015.
- Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 681–688, 2011.
- Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017.

A. A Framework for Infinite-Dimensional Mirror Descent

A.1. A note on the regularity

It is known that the (negative) Shannon entropy is *not* Fréchet differentiable in general. However, below we show that the Fréchet derivative can be well-defined if we restrict the probability measures to within the set

$$\mathcal{M}(\mathcal{Z}) := \{ \text{all probability measures on } \mathcal{Z} \text{ that admit densities w.r.t. the Lebesgue measure, and the density is continuous and positive almost everywhere on } \mathcal{Z} \}. \quad (\text{A.1})$$

We will also restrict the set of functions to be bounded and integrable:

$$\mathcal{F}(\mathcal{Z}) := \left\{ \text{all bounded continuous functions } f \text{ on } \mathcal{Z} \text{ such that } \int e^{-f} < \infty \right\}. \quad (\text{A.2})$$

These assumptions on the probability measures and functions are sufficient for most practical applications.

It is important to notice that $\mu \in \mathcal{M}(\mathcal{Z})$ and $h \in \mathcal{F}(\mathcal{Z})$ implies $\mu' = \text{MD}_\eta(\mu, h) \in \mathcal{M}(\mathcal{Z})$; this readily follows from the formula (7).

A.2. Properties of Entropic Mirror Map

The total variation of a (possibly non-probability) measure $\mu \in \mathcal{M}(\mathcal{Z})$ is defined as (Halmos, 2013)

$$\|\mu\|_{\text{TV}} = \sup_{\|h\|_{\mathbb{L}^\infty} \leq 1} \int h d\mu = \sup_{\|h\|_{\mathbb{L}^\infty} \leq 1} \langle \mu, h \rangle.$$

Recall the standard topology induced by $\|\cdot\|_{\text{TV}}$ and $\|\cdot\|_{\mathbb{L}^\infty}$ for measures and functions (Halmos, 2013), respectively. Whenever we speak about continuity or differentiability below, it is understood to be w.r.t. to the standard topology. Notice also that the G operator defined in (5) is bounded if the discriminator f_w is bounded, and hence continuous (Halmos, 2013).

We depart from the fundamental *Gibbs Variational Principle*, which dates back to the earliest work of statistical mechanics (Gibbs, 1902). For two probability measures μ, μ' , denote their relative entropy by (the reason for this notation will become clear in (A.8))

$$D_\Phi(\mu, \mu') := \int_{\mathcal{Z}} d\mu \log \frac{d\mu}{d\mu'}.$$

By the definition of $\mathcal{M}(\mathcal{Z})$, it is clear that the relative entropy is well-defined for any $\mu, \mu' \in \mathcal{M}(\mathcal{Z})$.

Theorem 3 (Gibbs Variation Principle). *Let $h \in \mathcal{F}(\mathcal{Z})$ and $\mu' \in \mathcal{M}(\mathcal{Z})$ be a reference measure. Then*

$$\log \int_{\mathcal{Z}} e^h d\mu' = \sup_{\mu \in \mathcal{M}(\mathcal{Z})} \langle \mu, h \rangle - D_\Phi(\mu, \mu'), \quad (\text{A.3})$$

and equality is achieved by $d\mu^* = \frac{e^h d\mu'}{\int_{\mathcal{Z}} e^h d\mu'}$.

Part of the following theorem is folklore in the mathematics and learning community. However, to the best of our knowledge, the relation to the entropic MD has not been systematically studied before, as we now do.

Theorem 4. *For a probability measure $d\mu = \rho dz$, let $\Phi(\mu) = \int \rho \log \rho dz$ be the negative Shannon entropy, and let $\Phi^*(h) = \log \int_{\mathcal{Z}} e^h dz$. Then*

1. Φ^* is the Fenchel conjugate of Φ :

$$\Phi^*(h) = \sup_{\mu \in \mathcal{M}(\mathcal{Z})} \langle \mu, h \rangle - \Phi(\mu); \quad (\text{A.4})$$

$$\Phi(\mu) = \sup_{h \in \mathcal{F}(\mathcal{Z})} \langle \mu, h \rangle - \Phi^*(h). \quad (\text{A.5})$$

2. The derivatives admit the expression

$$d\Phi(\mu) = 1 + \log \rho = \arg \max_{h \in \mathcal{F}(\mathcal{Z})} \langle \mu, h \rangle - \Phi^*(h); \quad (\text{A.6})$$

$$d\Phi^*(h) = \frac{e^h dz}{\int_{\mathcal{Z}} e^h dz} = \arg \max_{\mu \in \mathcal{M}(\mathcal{Z})} \langle \mu, h \rangle - \Phi(\mu). \quad (\text{A.7})$$

3. The Bregman divergence of Φ is the relative entropy:

$$D_\Phi(\mu, \mu') = \Phi(\mu) - \Phi(\mu') - \langle \mu - \mu', d\Phi(\mu') \rangle = \int_{\mathcal{Z}} d\mu \log \frac{d\mu}{d\mu'}. \quad (\text{A.8})$$

4. Φ is 4-strongly convex with respect to the total variation norm: For all $\lambda \in (0, 1)$,

$$\Phi(\lambda\mu + (1 - \lambda)\mu') \leq \lambda\Phi(\mu) + (1 - \lambda)\Phi(\mu') - \frac{1}{2} \cdot 4\lambda(1 - \lambda)\|\mu - \mu'\|_{\text{TV}}^2. \quad (\text{A.9})$$

5. The following duality relation holds: For any constant C , we have

$$\forall \mu, \mu' \in \mathcal{M}(\mathcal{Z}), \quad D_\Phi(\mu, \mu') = D_{\Phi^*}(d\Phi(\mu'), d\Phi(\mu)) = D_{\Phi^*}(d\Phi(\mu') + C, d\Phi(\mu)). \quad (\text{A.10})$$

6. Φ^* is $\frac{1}{4}$ -smooth with respect to $\|\cdot\|_{\mathbb{L}^\infty}$:

$$\forall h, h' \in \mathcal{F}(\mathcal{Z}), \quad \|d\Phi^*(h) - d\Phi^*(h')\|_{\text{TV}} \leq \frac{1}{4} \|h - h'\|_{\mathbb{L}^\infty}. \quad (\text{A.11})$$

7. Alternative to (A.11), we have the equivalent characterization of Φ^* :

$$\forall h, h' \in \mathcal{F}(\mathcal{Z}), \quad \Phi^*(h) \leq \Phi^*(h') + \langle d\Phi^*(h'), h - h' \rangle + \frac{1}{2} \cdot \frac{1}{4} \|h - h'\|_{\mathbb{L}^\infty}^2. \quad (\text{A.12})$$

8. Similar to (A.10), we have

$$\forall h, h', \quad D_{\Phi^*}(h, h') = D_\Phi(d\Phi^*(h'), d\Phi^*(h)). \quad (\text{A.13})$$

9. The following three-point identity holds for all $\mu, \mu', \mu'' \in \mathcal{M}(\mathcal{Z})$:

$$\langle \mu'' - \mu, d\Phi(\mu') - d\Phi(\mu) \rangle = D_\Phi(\mu, \mu') + D_\Phi(\mu'', \mu) - D_\Phi(\mu'', \mu'). \quad (\text{A.14})$$

10. Let the Mirror Descent iterate be defined as in (7). Then the following statements are equivalent:

- (a) $\mu_+ = \text{MD}_\eta(\mu, h)$.
- (b) There exists a constant C such that $d\Phi(\mu_+) = d\Phi(\mu) - \eta h + C$.

In particular, for any $\mu', \mu'' \in \mathcal{M}(\mathcal{Z})$ we have

$$\text{Let } \langle \mu' - \mu'', \eta h \rangle = \langle \mu' - \mu'', d\Phi(\mu) - d\Phi(\mu_+) \rangle. \quad (\text{A.15})$$

Proof.

1. Equation (A.4) is simply the Gibbs variational principle (A.3) with $d\mu \leftarrow dz$.

By (A.4), we know that

$$\forall h \in \mathcal{F}(\mathcal{Z}), \quad \Phi(\mu) \geq \langle \mu, h \rangle - \log \int_{\mathcal{Z}} e^h dz. \quad (\text{A.16})$$

But for $d\mu = \rho dz$, the function $h := 1 + \log \rho$ saturates the equality in (A.16).

2. We prove a more general result on the Bregman divergence D_Φ in (A.17) below.

Let $d\mu = \rho d\mathbf{z}$, $d\mu' = \rho' d\mathbf{z}$, and $d\mu'' = \rho'' d\mathbf{z} \in \mathcal{M}(\mathcal{Z})$. Let $\epsilon > 0$ be small enough such that $(\rho + \epsilon\rho'')d\mathbf{z}$ is absolutely continuous with respect to $d\mu'$; note that this is possible because $\mu, \mu',$ and $\mu'' \in \mathcal{M}(\mathcal{Z})$. We compute

$$\begin{aligned} D_\Phi(\rho + \epsilon\rho'', \rho') &= \int_{\mathcal{Z}} (\rho + \epsilon\rho'') \log \frac{\rho + \epsilon\rho''}{\rho'} \\ &= \int_{\mathcal{Z}} \rho \log \frac{\rho}{\rho'} + \int_{\mathcal{Z}} \rho \log \left(1 + \epsilon \frac{\rho''}{\rho}\right) + \epsilon \int_{\mathcal{Z}} \rho'' \log \frac{\rho}{\rho'} + \epsilon \int_{\mathcal{Z}} \rho'' \log \left(1 + \epsilon \frac{\rho''}{\rho}\right) \\ &\stackrel{(i)}{=} \int_{\mathcal{Z}} \rho \log \frac{\rho}{\rho'} + \epsilon \int_{\mathcal{Z}} \rho'' + \epsilon \int_{\mathcal{Z}} \rho'' \log \frac{\rho}{\rho'} + \epsilon^2 \int_{\mathcal{Z}} \frac{\rho''^2}{\rho} + o(\epsilon) \\ &= D_\Phi(\rho, \rho') + \epsilon \int_{\mathcal{Z}} \rho'' \left(1 + \log \frac{\rho}{\rho'}\right) + o(\epsilon), \end{aligned}$$

where (i) uses $\log(1 + t) = t + o(t)$ as $t \rightarrow 0$. In short, for all $\mu', \mu'' \in \mathcal{M}(\mathcal{Z})$,

$$d_\mu D_\Phi(\mu, \mu')(\mu'') = \left\langle \mu'', 1 + \log \frac{\rho}{\rho'} \right\rangle \quad (\text{A.17})$$

which means $d_\mu D_\Phi(\mu, \mu') = 1 + \log \frac{\rho}{\rho'}$. The formula (A.6) is the special case when $d\mu' \leftarrow d\mathbf{z}$.

We now turn to (A.7). For every $h \in \mathcal{F}(\mathcal{Z})$, we need to show that the following holds for every $h' \in \mathcal{F}(\mathcal{Z})$:

$$\Phi^*(h + \epsilon h') - \Phi^*(h) = \log \int_{\mathcal{Z}} e^{h+\epsilon h'} d\mathbf{z} - \log \int_{\mathcal{Z}} e^h d\mathbf{z} = \epsilon \int_{\mathcal{Z}} h' \frac{e^h}{\int_{\mathcal{Z}} e^h} d\mathbf{z} + o(\epsilon). \quad (\text{A.18})$$

Define an auxiliary function

$$T(\epsilon) := \log \int_{\mathcal{Z}} \frac{e^h}{\int_{\mathcal{Z}} e^h} e^{\epsilon h'} d\mathbf{z}.$$

Notice that $T(0) = 0$ and T is smooth as a function of ϵ . Thus, by the Intermediate Value Theorem,

$$\begin{aligned} \Phi^*(h + \epsilon h') - \Phi^*(h) &= T(\epsilon) - T(0) \\ &= (\epsilon - 0) \cdot \left. \frac{d}{d\epsilon} T(\cdot) \right|_{\epsilon'}. \end{aligned}$$

for some $\epsilon' \in [0, \epsilon]$. A direct computation shows

$$\left. \frac{d}{d\epsilon} T(\cdot) \right|_{\epsilon'} = \int_{\mathcal{Z}} h' \frac{e^{h+\epsilon' h'}}{\int_{\mathcal{Z}} e^{h+\epsilon' h'}} d\mathbf{z}.$$

Hence it suffices to prove $\frac{e^{h+\epsilon' h'}}{\int_{\mathcal{Z}} e^{h+\epsilon' h'}} = \frac{e^h}{\int_{\mathcal{Z}} e^h} + o(1)$ in ϵ . To this end, let $C = \sup |h'| < \infty$. Then

$$\frac{e^h}{\int_{\mathcal{Z}} e^h} e^{-2\epsilon' C} \leq \frac{e^{h+\epsilon' h'}}{\int_{\mathcal{Z}} e^{h+\epsilon' h'}} \leq \frac{e^h}{\int_{\mathcal{Z}} e^h} e^{2\epsilon' C}.$$

It remains to use $e^t = 1 + t + o(t)$ and $\epsilon' \leq \epsilon$.

3. Let $d\mu = \rho d\mathbf{z}$ and $d\mu' = \rho' d\mathbf{z}$. We compute

$$\begin{aligned} D_\Phi(\mu, \mu') &= \Phi(\mu) - \Phi(\mu') - \langle \mu - \mu', d\Phi(\mu') \rangle \\ &= \int_{\mathcal{Z}} \rho \log \rho d\mathbf{z} - \int_{\mathcal{Z}} \rho' \log \rho' d\mathbf{z} - \langle \mu - \mu', 1 + \log \rho' \rangle \quad \text{by (A.6)} \\ &= \int_{\mathcal{Z}} \rho \log \frac{\rho}{\rho'} d\mathbf{z} \\ &= \int_{\mathcal{Z}} d\mu \log \frac{d\mu}{d\mu'}. \end{aligned}$$

4. Define $\mu_\lambda = \lambda\mu + (1 - \lambda)\mu'$. By (A.8) and the classical Pinsker's inequality (Gray, 2011), we have

$$\Phi(\mu) \geq \Phi(\mu_\lambda) + \langle (1 - \lambda)(\mu - \mu'), d\Phi(\mu_\lambda) \rangle + 2\|(1 - \lambda)(\mu - \mu')\|_{\text{TV}}^2, \quad (\text{A.19})$$

$$\Phi(\mu') \geq \Phi(\mu_\lambda) + \langle \lambda(\mu' - \mu), d\Phi(\mu_\lambda) \rangle + 2\|\lambda(\mu - \mu')\|_{\text{TV}}^2. \quad (\text{A.20})$$

Equation (A.9) follows by multiplying with λ and $1 - \lambda$ respectively and summing the two inequalities up.

5. Let $\mu = \rho d\mathbf{z}$ and $\mu' = \rho' d\mathbf{z}$. Then, by the definition of Bregman divergence and (A.6), (A.7),

$$\begin{aligned} D_{\Phi^*}(d\Phi(\mu'), d\Phi(\mu)) &= \Phi^*(d\Phi(\mu')) - \Phi^*(d\Phi(\mu)) - \left\langle \frac{e^{1+\log \rho} d\mathbf{z}}{\int_{\mathcal{Z}} e^{1+\log \rho}}, 1 + \log \rho' - 1 - \log \rho \right\rangle \\ &= \log \int_{\mathcal{Z}} e^{1+\log \rho'} - \log \int_{\mathcal{Z}} e^{1+\log \rho} + \int_{\mathcal{Z}} \rho \log \frac{\rho}{\rho'} \\ &= \int_{\mathcal{Z}} \rho \log \frac{\rho}{\rho'} = D_{\Phi}(\mu, \mu') \end{aligned}$$

since $\int_{\mathcal{Z}} \rho d\mathbf{z} = \int_{\mathcal{Z}} \rho' d\mathbf{z} = 1$. This proves the first equality.

For the second equality, we write

$$\begin{aligned} D_{\Phi^*}(d\Phi(\mu') + C, d\Phi(\mu)) &= \Phi^*(d\Phi(\mu') + C) - \Phi^*(d\Phi(\mu)) - \left\langle \frac{e^{1+\log \rho} d\mathbf{z}}{\int_{\mathcal{Z}} e^{1+\log \rho}}, 1 + \log \rho' + C - 1 - \log \rho \right\rangle \\ &= \log \int_{\mathcal{Z}} e^{1+\log \rho' + C} - \log \int_{\mathcal{Z}} e^{1+\log \rho} + \int_{\mathcal{Z}} \rho \log \frac{\rho}{\rho'} - C \\ &= \int_{\mathcal{Z}} \rho \log \frac{\rho}{\rho'} \\ &= D_{\Phi}(\mu, \mu') = D_{\Phi^*}(d\Phi(\mu'), d\Phi(\mu)) \end{aligned}$$

where we have used the first equality in the last step.

6. Let $\mu_h = d\Phi^*(h)$, $\mu_{h'} = d\Phi^*(h')$, and $\mu_\lambda = \lambda\mu_h + (1 - \lambda)\mu_{h'}$ for some $\lambda \in (0, 1)$. By Pinsker's inequality and (A.8), we have

$$\Phi(\mu_\lambda) \geq \Phi(\mu_h) + \langle \mu_\lambda - \mu_h, d\Phi(\mu_h) \rangle + 2\|\mu_\lambda - \mu_h\|_{\text{TV}}^2, \quad (\text{A.21})$$

$$\Phi(\mu_\lambda) \geq \Phi(\mu_{h'}) + \langle \mu_\lambda - \mu_{h'}, d\Phi(\mu_{h'}) \rangle + 2\|\mu_\lambda - \mu_{h'}\|_{\text{TV}}^2. \quad (\text{A.22})$$

Now, notice that

$$\begin{aligned} \langle \mu_\lambda - \mu_h, d\Phi(\mu_h) \rangle &= \langle \mu_\lambda - \mu_h, d\Phi(d\Phi^*(h)) \rangle \\ &= \left\langle \mu_\lambda - \mu_h, d\Phi \left(\frac{e^h d\mathbf{z}}{\int_{\mathcal{Z}} e^h} \right) \right\rangle \quad \text{by (A.7)} \\ &= \left\langle \mu_\lambda - \mu_h, 1 + h - \log \int_{\mathcal{Z}} e^h \right\rangle \quad \text{by (A.6)} \\ &= \langle \mu_\lambda - \mu_h, h \rangle \end{aligned}$$

and, similarly, we have $\langle \mu_\lambda - \mu_{h'}, d\Phi(\mu_{h'}) \rangle = \langle \mu_\lambda - \mu_{h'}, h' \rangle$. Multiplying (A.21) by λ and (A.22) by $1 - \lambda$, summing the two up, and using the above equalities, we get

$$\Phi(\mu_\lambda) - \left(\lambda\Phi(\mu_h) + (1 - \lambda)\Phi(\mu_{h'}) \right) + \lambda(1 - \lambda) \langle \mu_h - \mu_{h'}, h - h' \rangle \geq 2\lambda(1 - \lambda) \|\mu_h - \mu_{h'}\|_{\text{TV}}^2.$$

By (A.9), we know that

$$\Phi(\mu_\lambda) - \left(\lambda\Phi(\mu_h) + (1 - \lambda)\Phi(\mu_{h'}) \right) \leq -2\lambda(1 - \lambda) \|\mu_h - \mu_{h'}\|_{\text{TV}}^2.$$

Moreover, by definition of the total variation norm, it is clear that

$$\langle \mu_h - \mu_{h'}, h - h' \rangle \leq \|\mu_h - \mu_{h'}\|_{\text{TV}} \|h - h'\|_{\mathbb{L}^\infty}. \quad (\text{A.23})$$

Combing the last three inequalities gives (A.11).

7. Let K be a positive integer and $k \in \{0, 1, 2, \dots, K\}$. Set $\lambda_k = \frac{k}{K}$ and $h'' = h - h'$. Then

$$\begin{aligned}\Phi^*(h) - \Phi^*(h') &= \Phi^*(h' + \lambda_K h'') - \Phi^*(h' + \lambda_0 h'') \\ &= \sum_{k=0}^{K-1} (\Phi^*(h' + \lambda_{k+1} h'') - \Phi^*(h' + \lambda_k h'')).\end{aligned}\tag{A.24}$$

By convexity of Φ^* , we have

$$\begin{aligned}\Phi^*(h' + \lambda_{k+1} h'') - \Phi^*(h' + \lambda_k h'') &\leq \langle d\Phi^*(h' + \lambda_{k+1} h''), (\lambda_{k+1} - \lambda_k) h'' \rangle \\ &= \frac{1}{K} \langle d\Phi^*(h' + \lambda_{k+1} h''), h'' \rangle.\end{aligned}\tag{A.25}$$

By (A.23) and (A.11), we may further upper bound (A.25) as

$$\begin{aligned}\Phi^*(h' + \lambda_{k+1} h'') - \Phi^*(h' + \lambda_k h'') &\leq \frac{1}{K} \left(\langle d\Phi^*(h'), h'' \rangle + \langle d\Phi^*(h' + \lambda_{k+1} h'') - d\Phi^*(h'), h'' \rangle \right) \\ &\leq \frac{1}{K} \left(\langle d\Phi^*(h'), h'' \rangle + \|d\Phi^*(h' + \lambda_{k+1} h'') - d\Phi^*(h')\|_{TV} \|h''\|_{\mathbb{L}^\infty} \right) \\ &\leq \frac{1}{K} \left(\langle d\Phi^*(h'), h'' \rangle + \frac{\lambda_{k+1}}{4} \|h''\|_{\mathbb{L}^\infty}^2 \right).\end{aligned}\tag{A.26}$$

Summing up (A.26) over k , we get, in view of (A.24),

$$\begin{aligned}\Phi^*(h) - \Phi^*(h') &\leq \langle d\Phi^*(h'), h'' \rangle + \frac{1}{4} \|h''\|_{\mathbb{L}^\infty}^2 \sum_{k=0}^{K-1} \lambda_{k+1} \\ &= \langle d\Phi^*(h'), h'' \rangle + \frac{1}{4} \cdot \frac{K+1}{2K} \|h''\|_{\mathbb{L}^\infty}^2.\end{aligned}\tag{A.27}$$

Since K is arbitrary, we may take $K \rightarrow \infty$ in (A.27), which is (A.12).

8. Straightforward calculation shows

$$D_{\Phi^*}(h, h') = \log \int_{\mathcal{Z}} e^h - \log \int_{\mathcal{Z}} e^{h'} - \int_{\mathcal{Z}} \frac{e^{h'}}{\int_{\mathcal{Z}} e^{h'}} (h - h').$$

On the other hand, by definition of the Bregman divergence and (A.6), (A.7), we have

$$\begin{aligned}D_{\Phi}(d\Phi^*(h'), d\Phi^*(h)) &= \int_{\mathcal{Z}} \frac{e^{h'}}{\int_{\mathcal{Z}} e^{h'}} h' - \log \int_{\mathcal{Z}} e^{h'} - \int_{\mathcal{Z}} \frac{e^h}{\int_{\mathcal{Z}} e^h} h + \log \int_{\mathcal{Z}} e^h \\ &\quad - \int_{\mathcal{Z}} \left(1 + h - \log \int_{\mathcal{Z}} e^h \right) \left(\frac{e^{h'}}{\int_{\mathcal{Z}} e^{h'}} - \frac{e^h}{\int_{\mathcal{Z}} e^h} \right) \\ &= \int_{\mathcal{Z}} \frac{e^{h'}}{\int_{\mathcal{Z}} e^{h'}} (h' - h) - \log \int_{\mathcal{Z}} e^{h'} + \log \int_{\mathcal{Z}} e^h \\ &= \Phi^*(h) - \Phi^*(h') - \langle d\Phi^*(h'), h - h' \rangle \\ &= D_{\Phi^*}(h, h').\end{aligned}$$

9. By definition of the Bregman divergence, we have

$$\begin{aligned}D_{\Phi}(\mu, \mu') &= \Phi(\mu) - \Phi(\mu') - \langle \mu - \mu', d\Phi(\mu') \rangle, \\ D_{\Phi}(\mu'', \mu) &= \Phi(\mu'') - \Phi(\mu) - \langle \mu'' - \mu, d\Phi(\mu) \rangle, \\ D_{\Phi}(\mu'', \mu') &= \Phi(\mu'') - \Phi(\mu') - \langle \mu'' - \mu', d\Phi(\mu') \rangle.\end{aligned}$$

Equation (A.14) then follows by straightforward calculations.

10. First, let $\mu_+ = \text{MD}_\eta(\mu, h)$. Then if $\mu_+ = \rho_+ d\mathbf{z}$ and $\mu = \rho d\mathbf{z}$, then (7) implies

$$\rho_+ = \frac{\rho e^{-\eta h}}{\int_{\mathcal{Z}} \rho e^{-\eta h}}.$$

By (A.6), we therefore have

$$\begin{aligned} d\Phi(\mu_+) &= 1 + \log \rho_+ \\ &= 1 + \log \rho - \eta h - \log \int_{\mathcal{Z}} \rho e^{-\eta h} \end{aligned}$$

whence (A.15) holds with $C = -\log \int_{\mathcal{Z}} \rho e^{-\eta h}$.

Conversely, assume that $d\Phi(\mu_+) = d\Phi(\mu) - \eta h + C$ for some constant C , and apply $d\Phi^*$ to both sides. The left-hand side becomes

$$\begin{aligned} d\Phi^*(d\Phi(\mu_+)) &= d\Phi^*(1 + \log \rho_+) \\ &= \frac{\rho_+ d\mathbf{z}}{\int \rho_+ d\mathbf{z}} = \rho_+ d\mathbf{z} = d\mu_+, \end{aligned}$$

where as the formula (A.7) implies that

$$\begin{aligned} d\Phi^*(d\Phi(\mu) - \eta h + C) &= \frac{e^{1+\log \rho - \eta h + C}}{\int_{\mathcal{Z}} e^{1+\log \rho - \eta h + C}} d\mathbf{z} \\ &= \frac{\rho e^{-\eta h} d\mathbf{z}}{\int_{\mathcal{Z}} \rho e^{-\eta h}} \\ &= \frac{e^{-\eta h} d\mu}{\int_{\mathcal{Z}} e^{-\eta h} d\mu}. \end{aligned}$$

Combining the two equalities gives $d\mu_+ = \frac{e^{-\eta h} d\mu}{\int_{\mathcal{Z}} e^{-\eta h} d\mu}$ which exactly means $\mu_+ = \text{MD}_\eta(\mu, h)$.

□

B. Convergence Rates for Infinite-Dimensional Prox Methods

B.1. Rigorous Statements

For **Algorithm 1** and **2**, we have the following guarantees:

Theorem 5 (Convergence Rates). Let $\Phi(\mu) = \int d\mu \log \frac{d\mu}{d\mathbf{z}}$. Let M be a constant such that $\max[\|-\mathbf{g} + G\nu\|_{\mathbb{L}^\infty}, \|G^\dagger \mu\|_{\mathbb{L}^\infty}] \leq M$, and L be such that $\|G(\nu - \nu')\|_{\mathbb{L}^\infty} \leq L \|\nu - \nu'\|_{\text{TV}}$ and $\|G^\dagger(\mu - \mu')\|_{\mathbb{L}^\infty} \leq L \|\mu - \mu'\|_{\text{TV}}$. Let $D(\cdot, \cdot)$ be the relative entropy, and denote by $D_0 := D(\mu_{\text{NE}}, \mu_1) + D(\nu_{\text{NE}}, \nu_1)$ the initial distance to the mixed NE. Then

1. Assume that we have access to the deterministic derivatives $\{-G^\dagger \mu_t\}_{t=1}^T$ and $\{g - G\nu\}_{t=1}^T$. Then **Algorithm 1** achieves $O(T^{-1/2})$ -NE with $\eta = \frac{2}{M} \sqrt{\frac{D_0}{T}}$, and **Algorithm 2** achieves $O(T^{-1})$ -NE with $\eta = \frac{4}{L}$.
2. Assume that we have access to stochastic derivatives $\{-\hat{G}^\dagger \mu_t\}_{t=1}^T$ and $\{\hat{g} - \hat{G}\nu\}_{t=1}^T$ such that $\max[\mathbb{E} \|\hat{g} - \hat{G}\nu\|_{\mathbb{L}^\infty}, \mathbb{E} \|\hat{G}^\dagger \mu\|_{\mathbb{L}^\infty}] \leq M'$, and the variance is upper bounded by σ^2 . Assume also that the bias of stochastic derivatives satisfies $\max[\|\mathbb{E}[-\hat{g} + \hat{G}\nu] + g - G\nu\|_{\mathbb{L}^\infty}, \|\mathbb{E}[\hat{G}^\dagger \mu] - G^\dagger \mu\|_{\mathbb{L}^\infty}] \leq \tau$. Then **Algorithm 1** with stochastic derivatives achieves $O(T^{-1/2})$ -NE in expectation with $\eta = \sqrt{\frac{D_0}{T(4\tau + M'/4)}}$, and **Algorithm 2** with stochastic derivatives achieves $(O(T^{-1/2}) + O(\tau))$ -NE in expectation with $\eta = \min\left[\frac{4}{\sqrt{3}L}, \sqrt{\frac{2D_0}{3T\sigma^2}}\right]$.

B.2. Proof of Convergence Rates for Infinite-Dimensional Mirror Descent

B.2.1. MIRROR DESCENT, DETERMINISTIC DERIVATIVES

By the definition of the algorithm, (A.15), and the three-point identity (A.14), we have, for any $\mu \in \mathcal{M}(\mathcal{W})$,

$$\begin{aligned}\langle \mu_t - \mu, -g + G\nu_t \rangle &= \frac{1}{\eta} \langle \mu_t - \mu, d\Phi(\mu_t) - d\Phi(\mu_{t+1}) \rangle \\ &= \frac{1}{\eta} (D_\Phi(\mu, \mu_t) - D_\Phi(\mu, \mu_{t+1}) + D_\Phi(\mu_t, \mu_{t+1})).\end{aligned}\quad (\text{B.1})$$

By item 10 of **Theorem 4**, there exists a constant C_t such that

$$d\Phi(\mu_{t+1}) = d\Phi(\mu_t) - \eta(-g + G\nu_t) + C_t. \quad (\text{B.2})$$

Using (A.10), we see that

$$\begin{aligned}D_\Phi(\mu_t, \mu_{t+1}) &= D_{\Phi^*}(d\Phi(\mu_{t+1}), d\Phi(\mu_t)) \\ &= D_{\Phi^*}(d\Phi(\mu_{t+1}) - C_t, d\Phi(\mu_t)) \\ &\leq \frac{1}{8} \|d\Phi(\mu_{t+1}) - C_t - d\Phi(\mu_t)\|_{\mathbb{L}^\infty}^2 \quad \text{by (A.12)} \\ &= \frac{\eta^2}{8} \|-g + G\nu_t\|_{\mathbb{L}^\infty}^2 \quad \text{by (B.2)} \\ &\leq \frac{\eta^2 M^2}{8}.\end{aligned}$$

Consequently, we have

$$\begin{aligned}\sum_{t=1}^T \langle \mu_t - \mu, -g + G\nu_t \rangle &= \sum_{t=1}^T \frac{1}{\eta} (D_\Phi(\mu, \mu_t) - D_\Phi(\mu, \mu_{t+1}) + D_\Phi(\mu_t, \mu_{t+1})) \\ &\leq \frac{D_\Phi(\mu, \mu_1)}{\eta} + \frac{\eta M^2 T}{8}.\end{aligned}\quad (\text{B.3})$$

Exactly the same argument applied to ν_t 's yields, for any $\nu \in \mathcal{M}(\Theta)$,

$$\sum_{t=1}^T \langle \nu_t - \nu, -G^\dagger \mu_t \rangle \leq \frac{D_\Phi(\nu, \nu_1)}{\eta} + \frac{\eta M^2 T}{8}. \quad (\text{B.4})$$

Summing up (B.3) and (B.4), substituting $\mu \leftarrow \mu_{\text{NE}}$, $\nu \leftarrow \nu_{\text{NE}}$ and dividing by T , we get

$$\frac{1}{T} \sum_{t=1}^T \left(\langle \mu_t - \mu_{\text{NE}}, -g + G\nu_t \rangle + \langle \nu_t - \nu_{\text{NE}}, -G^\dagger \mu_t \rangle \right) \leq \frac{D_0}{\eta T} + \frac{\eta M^2}{4}. \quad (\text{B.5})$$

The left-hand side of (B.5) can be simplified to

$$\begin{aligned}\frac{1}{T} \sum_{t=1}^T \left(\langle \mu_t - \mu_{\text{NE}}, -g + G\nu_t \rangle + \langle \nu_t - \nu_{\text{NE}}, -G^\dagger \mu_t \rangle \right) &= \frac{1}{T} \sum_{t=1}^T \left(\langle \mu_{\text{NE}} - \mu_t, g \rangle - \langle \mu_{\text{NE}}, G\nu_t \rangle + \langle \mu_t, G\nu_{\text{NE}} \rangle \right) \\ &= \langle \mu_{\text{NE}}, g - G\bar{\nu}_T \rangle - \langle \bar{\mu}_T, g - G\nu_{\text{NE}} \rangle.\end{aligned}\quad (\text{B.6})$$

By definition of the Nash Equilibrium, we have

$$\begin{aligned}\langle \bar{\mu}_T, g - G\nu_{\text{NE}} \rangle &\leq \langle \mu_{\text{NE}}, g - G\nu_{\text{NE}} \rangle \leq \langle \mu_{\text{NE}}, g - G\bar{\nu}_T \rangle, \\ \langle \bar{\mu}_T, g - G\nu_{\text{NE}} \rangle &\leq \langle \bar{\mu}_T, g - G\bar{\nu}_T \rangle \leq \langle \mu_{\text{NE}}, g - G\bar{\nu}_T \rangle,\end{aligned}\quad (\text{B.7})$$

which implies

$$|\langle \bar{\mu}_T, g - G\bar{\nu}_T \rangle - \langle \mu_{\text{NE}}, g - G\nu_{\text{NE}} \rangle| \leq \langle \mu_{\text{NE}}, g - G\bar{\nu}_T \rangle - \langle \bar{\mu}_T, g - G\nu_{\text{NE}} \rangle. \quad (\text{B.8})$$

Combining (B.18)-(B.21), we conclude that

$$\eta = \frac{2}{M} \sqrt{\frac{D_0}{T}} \Rightarrow |\langle \bar{\mu}_T, g - G\bar{\nu}_T \rangle - \langle \mu_{\text{NE}}, g - G\nu_{\text{NE}} \rangle| \leq M \sqrt{\frac{D_0}{T}}.$$

B.2.2. MIRROR DESCENT, STOCHASTIC DERIVATIVES

We first write

$$\left\langle \mu_t - \mu, \eta(-\hat{g} + \hat{G}\nu_t) \right\rangle = \langle \mu_t - \mu, \eta(-g + G\nu_t) \rangle + \left\langle \mu_t - \mu, \eta \left[-\hat{g} + \hat{G}\nu_t + g - G\nu_t \right] \right\rangle.$$

Taking conditional expectation and using the bias estimate of stochastic derivatives, we conclude that

$$\begin{aligned} \mathbb{E} \left\langle \mu_t - \mu, \eta(-\hat{g} + \hat{G}\nu_t) \right\rangle &\leq \langle \mu_t - \mu, \eta(-g + G\nu_t) \rangle + \|\mu_t - \mu\|_{\text{TV}} \cdot \eta\tau \\ &\leq \langle \mu_t - \mu, \eta(-g + G\nu_t) \rangle + 2\eta\tau. \end{aligned}$$

Therefore, using exactly the same argument leading to (B.3), we may obtain

$$\mathbb{E} \sum_{t=1}^T \left\langle \mu_t - \mu, -\hat{g} + \hat{G}\nu_t \right\rangle \leq \frac{\mathbb{E} D_\Phi(\mu, \mu_1)}{\eta} + \frac{\eta M'^2 T}{8} + 2\eta T\tau.$$

The rest is the same as with deterministic derivatives.

B.3. Proof of Convergence Rates for Infinite-Dimensional Mirror-Prox

We first need a technical lemma, which is **Lemma 6.2** of (Juditsky & Nemirovski, 2011) tailored to our infinite-dimensional setting. We give a slightly different proof.

Lemma 6. *Given any $\mu \in \mathcal{M}(\mathcal{Z})$ and $h, h' \in \mathcal{F}(\mathcal{Z})$, let $\mu = \text{MD}_\eta(\tilde{\mu}, h)$ and $\tilde{\mu}_+ = \text{MD}_\eta(\tilde{\mu}, h')$. Let Φ be α -strongly convex (recall that $\alpha = 4$ when Φ is the entropy). Then, for any $\mu_* \in \mathcal{M}(\mathcal{Z})$, we have*

$$\langle \mu - \mu_*, \eta h' \rangle \leq D_\Phi(\mu_*, \tilde{\mu}) - D_\Phi(\mu_*, \tilde{\mu}_+) + \frac{\eta^2}{2\alpha} \|h - h'\|_{\mathbb{L}^\infty}^2 - \frac{\alpha}{2} \|\mu - \tilde{\mu}\|_{\text{TV}}^2. \quad (\text{B.9})$$

Proof. Recall from (A.9) that entropy is α -strongly convex with respect to $\|\cdot\|_{\text{TV}}$. We first write

$$\langle \mu - \mu_*, \eta h' \rangle = \langle \tilde{\mu}_+ - \mu_*, \eta h' \rangle + \langle \mu - \tilde{\mu}_+, \eta h \rangle + \langle \mu - \tilde{\mu}_+, \eta(h' - h) \rangle. \quad (\text{B.10})$$

For the first term, (A.14) and (A.15) implies

$$\begin{aligned} \langle \tilde{\mu}_+ - \mu_*, \eta h' \rangle &= \langle \tilde{\mu}_+ - \mu_*, d\Phi(\tilde{\mu}) - d\Phi(\tilde{\mu}_+) \rangle \\ &= -D_\Phi(\tilde{\mu}_+, \tilde{\mu}) - D_\Phi(\mu_*, \tilde{\mu}_+) + D_\Phi(\mu_*, \tilde{\mu}). \end{aligned} \quad (\text{B.11})$$

Similarly, the second term of the right-hand side of (B.10) can be written as

$$\langle \mu - \tilde{\mu}_+, \eta h \rangle = -D_\Phi(\mu, \tilde{\mu}) - D_\Phi(\tilde{\mu}_+, \mu) + D_\Phi(\tilde{\mu}_+, \tilde{\mu}). \quad (\text{B.12})$$

Hölder's inequality for the third term gives

$$\begin{aligned} \langle \mu - \tilde{\mu}_+, \eta(h' - h) \rangle &\leq \|\mu - \tilde{\mu}_+\|_{\text{TV}} \|\eta(h' - h)\|_{\mathbb{L}^\infty} \\ &\leq \frac{\alpha}{2} \|\mu - \tilde{\mu}_+\|_{\text{TV}}^2 + \frac{1}{2\alpha} \|\eta(h' - h)\|_{\mathbb{L}^\infty}^2. \end{aligned} \quad (\text{B.13})$$

Finally, recall that Φ is α -strongly convex, and hence we have

$$-D_\Phi(\tilde{\mu}_+, \mu) \leq -\frac{\alpha}{2} \|\mu - \tilde{\mu}_+\|_{\text{TV}}^2, \quad -D_\Phi(\mu, \tilde{\mu}) \leq -\frac{\alpha}{2} \|\mu - \tilde{\mu}\|_{\text{TV}}^2. \quad (\text{B.14})$$

The lemma follows by combining inequalities (B.11)-(B.14) in (B.10). \square

B.3.1. MIRROR-PROX, DETERMINISTIC DERIVATIVES

Let $\alpha = 4$, $\bar{\mu}_T := \frac{1}{T} \sum_{t=1}^T \mu_t$, and $\bar{\nu}_T := \frac{1}{T} \sum_{t=1}^T \nu_t$.

In **Lemma 6**, substituting $\mu_* \leftarrow \mu_{\text{NE}}$, $\tilde{\mu} \leftarrow \tilde{\mu}_t$, $h \leftarrow -g + G\tilde{\nu}_t$ (so that $\mu = \mu_t$) and $h' \leftarrow -g + G\nu_t$ (so that $\tilde{\mu}_+ = \tilde{\mu}_{t+1}$), we get

$$\langle \mu_t - \mu_{\text{NE}}, \eta(-g + G\nu_t) \rangle \leq D_\Phi(\mu_{\text{NE}}, \tilde{\mu}_t) - D_\Phi(\mu_{\text{NE}}, \tilde{\mu}_{t+1}) + \frac{\eta^2}{2\alpha} \|G(\nu_t - \tilde{\nu}_t)\|_{\mathbb{L}^\infty}^2 - \frac{\alpha}{2} \|\tilde{\mu}_t - \mu_t\|_{\text{TV}}^2. \quad (\text{B.15})$$

Similarly, we have

$$\langle \nu_t - \nu_{\text{NE}}, -\eta G^\dagger \mu_t \rangle \leq D_\Phi(\nu_{\text{NE}}, \tilde{\nu}_t) - D_\Phi(\nu_{\text{NE}}, \tilde{\nu}_{t+1}) + \frac{\eta^2}{2\alpha} \|G^\dagger(\mu_t - \tilde{\mu}_t)\|_{\mathbb{L}^\infty}^2 - \frac{\alpha}{2} \|\tilde{\nu}_t - \nu_t\|_{\text{TV}}^2. \quad (\text{B.16})$$

Since $\|G(\nu_t - \tilde{\nu}_t)\|_{\mathbb{L}^\infty} \leq L \cdot \|\nu_t - \tilde{\nu}_t\|_{\text{TV}}$ and $\|G^\dagger(\mu_t - \tilde{\mu}_t)\|_{\mathbb{L}^\infty} \leq L \cdot \|\mu_t - \tilde{\mu}_t\|_{\text{TV}}$, summing up (B.15) and (B.16) yields

$$\begin{aligned} \langle \mu_t - \mu_{\text{NE}}, \eta(-g + G\nu_t) \rangle + \langle \nu_t - \nu_{\text{NE}}, -\eta G^\dagger \mu_t \rangle &\leq D_\Phi(\mu_{\text{NE}}, \tilde{\mu}_t) - D_\Phi(\mu_{\text{NE}}, \tilde{\mu}_{t+1}) + D_\Phi(\nu_{\text{NE}}, \tilde{\nu}_t) - D_\Phi(\nu_{\text{NE}}, \tilde{\nu}_{t+1}) \\ &\quad + \left(\frac{\eta^2 L^2}{2\alpha} - \frac{\alpha}{2} \right) \left(\|\tilde{\mu}_t - \mu_t\|_{\text{TV}}^2 + \|\tilde{\nu}_t - \nu_t\|_{\text{TV}}^2 \right) \\ &\leq D_\Phi(\mu_{\text{NE}}, \tilde{\mu}_t) - D_\Phi(\mu_{\text{NE}}, \tilde{\mu}_{t+1}) + D_\Phi(\nu_{\text{NE}}, \tilde{\nu}_t) - D_\Phi(\nu_{\text{NE}}, \tilde{\nu}_{t+1}) \end{aligned}$$

if $\eta \leq \frac{\alpha}{L} = \frac{4}{L}$. Summing up the last inequality over t and using $D_\Phi(\cdot, \cdot) \geq 0$, we obtain

$$\frac{1}{T} \sum_{t=1}^T \left(\langle \mu_t - \mu_{\text{NE}}, \eta(-g + G\nu_t) \rangle + \langle \nu_t - \nu_{\text{NE}}, -\eta G^\dagger \mu_t \rangle \right) \leq \frac{D_\Phi(\mu_{\text{NE}}, \tilde{\mu}_1) + D_\Phi(\nu_{\text{NE}}, \tilde{\nu}_1)}{T} = \frac{D_0}{T}. \quad (\text{B.17})$$

The left-hand side of (B.17) can be simplified to

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \left(\langle \mu_t - \mu_{\text{NE}}, \eta(-g + G\nu_t) \rangle + \langle \nu_t - \nu_{\text{NE}}, -\eta G^\dagger \mu_t \rangle \right) &= \frac{\eta}{T} \sum_{t=1}^T \left(\langle \mu_{\text{NE}} - \mu_t, g \rangle - \langle \mu_{\text{NE}}, G\nu_t \rangle + \langle \mu_t, G\nu_{\text{NE}} \rangle \right) \\ &= \eta \left(\langle \mu_{\text{NE}}, g - G\bar{\nu}_T \rangle - \langle \bar{\mu}_T, g - G\nu_{\text{NE}} \rangle \right). \end{aligned} \quad (\text{B.18})$$

By definition of the $(\mu_{\text{NE}}, \nu_{\text{NE}})$, we have

$$\begin{aligned} \langle \bar{\mu}_T, g - G\nu_{\text{NE}} \rangle &\leq \langle \mu_{\text{NE}}, g - G\nu_{\text{NE}} \rangle \leq \langle \mu_{\text{NE}}, g - G\bar{\nu}_T \rangle, \\ \langle \bar{\mu}_T, g - G\nu_{\text{NE}} \rangle &\leq \langle \bar{\mu}_T, g - G\bar{\nu}_T \rangle \leq \langle \mu_{\text{NE}}, g - G\bar{\nu}_T \rangle, \end{aligned} \quad (\text{B.19})$$

which implies

$$|\langle \bar{\mu}_T, g - G\bar{\nu}_T \rangle - \langle \mu_{\text{NE}}, g - G\nu_{\text{NE}} \rangle| \leq \langle \mu_{\text{NE}}, g - G\bar{\nu}_T \rangle - \langle \bar{\mu}_T, g - G\nu_{\text{NE}} \rangle. \quad (\text{B.20})$$

Combining (B.17)-(B.20), we conclude

$$\eta \leq \frac{4}{L} \Rightarrow |\langle \bar{\mu}_T, g - G\bar{\nu}_T \rangle - \langle \mu_{\text{NE}}, g - G\nu_{\text{NE}} \rangle| \leq \frac{D_0}{T\eta}.$$

B.3.2. MIRROR-PROX, STOCHASTIC DERIVATIVES

Let $\alpha = 4$, $\bar{\mu}_T := \frac{1}{T} \sum_{t=1}^T \mu_t$, and $\bar{\nu}_T := \frac{1}{T} \sum_{t=1}^T \nu_t$. Set the step-size to $\eta = \min \left[\frac{\alpha}{\sqrt{3}L}, \sqrt{\frac{\alpha D_0}{6T\sigma^2}} \right]$.

In **Lemma 6**, substituting $\mu_* \leftarrow \mu_{\text{NE}}$, $\tilde{\mu} \leftarrow \tilde{\mu}_t$, $h \leftarrow -\hat{g} + \hat{G}\tilde{\nu}_t$ (so that $\mu = \mu_t$), and $h' \leftarrow -\hat{g} + \hat{G}\nu_t$ (so that $\tilde{\mu}_+ = \tilde{\mu}_{t+1}$), we get

$$\langle \mu_t - \mu_{\text{NE}}, \eta(-\hat{g} + \hat{G}\nu_t) \rangle \leq D_\Phi(\mu_{\text{NE}}, \tilde{\mu}_t) - D_\Phi(\mu_{\text{NE}}, \tilde{\mu}_{t+1}) + \frac{\eta^2}{2\alpha} \|\hat{G}\nu_t - \hat{G}\tilde{\nu}_t\|_{\mathbb{L}^\infty}^2 - \frac{\alpha}{2} \|\tilde{\mu}_t - \mu_t\|_{\text{TV}}^2. \quad (\text{B.21})$$

Note that

$$\begin{aligned}\mathbb{E} \left\| \hat{G}\nu_t - \hat{G}\tilde{\nu}_t \right\|_{\mathbb{L}^\infty}^2 &\leq 3 \left(\mathbb{E} \left\| \hat{G}\nu_t - G\nu_t \right\|_{\mathbb{L}^\infty}^2 + \mathbb{E} \|G\nu_t - G\tilde{\nu}_t\|_{\mathbb{L}^\infty}^2 + \mathbb{E} \left\| G\tilde{\nu}_t - \hat{G}\tilde{\nu}_t \right\|_{\mathbb{L}^\infty}^2 \right) \\ &\leq 6\sigma^2 + 3L^2\mathbb{E} \|\nu_t - \tilde{\nu}_t\|_{\text{TV}}^2.\end{aligned}$$

Therefore, taking expectation conditioned on the history for both sides of (B.21) and using the bias estimates of the stochastic derivatives, we get

$$\begin{aligned}\langle \mu_t - \mu_{\text{NE}}, \eta(-g + G\nu_t) \rangle &\leq \mathbb{E} D_\Phi(\mu_{\text{NE}}, \tilde{\mu}_t) - \mathbb{E} D_\Phi(\mu_{\text{NE}}, \tilde{\mu}_{t+1}) + \frac{3\eta^2\sigma^2}{\alpha} \\ &\quad + \frac{3\eta^2 L^2}{2\alpha} \mathbb{E} \|\nu_t - \tilde{\nu}_t\|_{\text{TV}}^2 - \frac{\alpha}{2} \mathbb{E} \|\tilde{\mu}_t - \mu_t\|_{\text{TV}}^2 + 2\eta\tau.\end{aligned}$$

Similarly, we have

$$\begin{aligned}\langle \nu_t - \nu_{\text{NE}}, -\eta G^\dagger \mu_t \rangle &\leq \mathbb{E} D_\Phi(\nu_{\text{NE}}, \tilde{\nu}_t) - \mathbb{E} D_\Phi(\nu_{\text{NE}}, \tilde{\nu}_{t+1}) + \frac{3\eta^2\sigma^2}{\alpha} \\ &\quad + \frac{3\eta^2 L^2}{2\alpha} \mathbb{E} \|\mu_t - \tilde{\mu}_t\|_{\text{TV}}^2 - \frac{\alpha}{2} \mathbb{E} \|\tilde{\nu}_t - \nu_t\|_{\text{TV}}^2 + 2\eta\tau.\end{aligned}$$

Summing up the last two inequalities over t with $\eta \leq \frac{\alpha}{\sqrt{3}L}$ then yields

$$\begin{aligned}\frac{1}{T} \sum_{t=1}^T \left(\langle \mu_t - \mu_{\text{NE}}, -g + G\nu_t \rangle + \langle \nu_t - \nu_{\text{NE}}, -G^\dagger \mu_t \rangle \right) &\leq \frac{D_0}{\eta T} + \frac{6\eta\sigma^2}{\alpha} + 4\tau \\ &\leq \max \left[2\sqrt{\frac{6\sigma^2 D_0}{\alpha T}}, \frac{2\sqrt{3}LD_0}{\alpha T} \right] + 4\tau.\end{aligned}$$

by definition of η . The rest is the same as with deterministic derivatives.

Algorithm 4 APPROX INF MIRROR DECENT

Input: $W[1], \Theta[1] \leftarrow n'$ samples from random initialization, $\{\gamma_t\}_{t=1}^{T-1}, \{\epsilon_t\}_{t=1}^{T-1}, \{K\}_{t=1}^{T-1}, n, n'$, standard normal noise ξ_k, ξ'_k .

for $t = 1, 2, \dots, T-1$ **do**

- $C \leftarrow \cup_{s=1}^t W[s], \quad D \leftarrow \cup_{s=1}^t \Theta[s]$ $\mathbf{w}_t^{(1)} \leftarrow \text{UNIF}(W[t]), \quad \boldsymbol{\theta}_t^{(1)} \leftarrow \text{UNIF}(\Theta[t])$ **for** $k = 1, 2, \dots, K_t, \dots, K_t + n'$ **do**
- Generate $A = \{X_1, \dots, X_n\} \sim \mathbb{P}_{\boldsymbol{\theta}_t^{(k)}} \quad \boldsymbol{\theta}_t^{(k+1)} = \boldsymbol{\theta}_t^{(k)} + \frac{\gamma_t}{nn'} \nabla_{\boldsymbol{\theta}} \sum_{X_i \in A} \sum_{\mathbf{w} \in C} f_{\mathbf{w}}(X_i) + \sqrt{2\gamma_t} \epsilon_t \xi_k$ Generate $B = \{X_1^{\text{real}}, \dots, X_n^{\text{real}}\} \sim \mathbb{P}_{\text{real}}$ $B' \leftarrow \{\}$ **for each** $\boldsymbol{\theta} \in D$ **do**
- Generate $\tilde{B} = \{X'_1, \dots, X'_n\} \sim \mathbb{P}_{\boldsymbol{\theta}}$ $B' \leftarrow B' \cup \tilde{B}$
-
- $\mathbf{w}_t^{(k+1)} = \mathbf{w}_t^{(k)} + \frac{\gamma_t t}{n} \nabla_{\mathbf{w}} \sum_{\substack{X_i^{\text{real}} \in B \\ X_i' \in B'}} f_{\mathbf{w}_t^{(k)}}(X_i^{\text{real}}) - \frac{\gamma_t}{nn'} \nabla_{\mathbf{w}} \sum_{X_i' \in B'} f_{\mathbf{w}_t^{(k)}}(X_i') + \sqrt{2\gamma_t} \epsilon_t \xi'_k;$
-
- $W[t+1] \leftarrow \{\mathbf{w}_t^{(K+1)}, \dots, \mathbf{w}_t^{(K+n')}\}, \quad \Theta[t+1] \leftarrow \{\boldsymbol{\theta}_t^{(K+1)}, \dots, \boldsymbol{\theta}_t^{(K+n')}\}$

idx $\leftarrow \text{UNIF}(1, 2, \dots, T)$ **return** $W[\text{idx}], \Theta[\text{idx}]$.

Algorithm 5 APPROX INF MIRROR-PROX

Input: $\tilde{W}[1], \tilde{\Theta}[1] \leftarrow n'$ samples from random initialization, $\{\gamma_t\}_{t=1}^T, \{\epsilon_t\}_{t=1}^T, \{K_t\}_{t=1}^T, n, n'$, standard normal noise $\xi_k, \xi'_k, \xi''_k, \xi'''_k$.

for $t = 1, 2, \dots, T$ **do**

$C \leftarrow \tilde{W}[t] \cup (\cup_{s=1}^{t-1} W[s]), D \leftarrow \tilde{\Theta}[t] \cup (\cup_{s=1}^{t-1} \Theta[s])$ $\mathbf{w}_t^{(1)} \leftarrow \text{UNIF}(\tilde{W}[t]), \theta_t^{(1)} \leftarrow \text{UNIF}(\tilde{\Theta}[t])$ **for** $k = 1, 2, \dots, K_t, \dots, K_t + n'$ **do**

Generate $A = \{X_1, \dots, X_n\} \sim \mathbb{P}_{\theta_t^{(k)}} \quad \theta_t^{(k+1)} = \theta_t^{(k)} + \frac{\gamma_t}{nn'} \nabla_{\theta} \sum_{X_i \in A} \sum_{\mathbf{w} \in C} f_{\mathbf{w}}(X_i) + \sqrt{2\gamma_t} \epsilon_t \xi_k$ Generate

$B = \{X_1^{\text{real}}, \dots, X_n^{\text{real}}\} \sim \mathbb{P}_{\text{real}}$ $B' \leftarrow \{\}$ **for each** $\theta \in D$ **do**

Generate $\tilde{B} = \{X'_1, \dots, X'_n\} \sim \mathbb{P}_{\theta}$ $B' \leftarrow B' \cup \tilde{B}$

$\mathbf{w}_t^{(k+1)} = \mathbf{w}_t^{(k)} + \frac{\gamma_t}{n} \nabla_{\mathbf{w}} \sum_{X_i^{\text{real}} \in B} f_{\mathbf{w}_t^{(k)}}(X_i^{\text{real}}) - \frac{\gamma_t}{nn'} \nabla_{\mathbf{w}} \sum_{X'_i \in B'} f_{\mathbf{w}_t^{(k)}}(X'_i) + \sqrt{2\gamma_t} \epsilon_t \xi'_k;$

$W[t] \leftarrow \{\mathbf{w}_t^{(K+1)}, \dots, \mathbf{w}_t^{(K+n')}\}, \quad \Theta[t] \leftarrow \{\theta_t^{(K+1)}, \dots, \theta_t^{(K+n')}\}$

$C' \leftarrow \cup_{s=1}^t W[s], \quad D' \leftarrow \cup_{s=1}^t \Theta[s]$ $\tilde{\mathbf{w}}_{t+1}^{(1)} \leftarrow \text{UNIF}(\tilde{W}[t]), \quad \tilde{\theta}_{t+1}^{(1)} \leftarrow \text{UNIF}(\tilde{\Theta}[t])$ **for**

$k = 1, 2, \dots, K_t, \dots, K_t + n'$ **do**

Generate $A = \{X_1, \dots, X_n\} \sim \mathbb{P}_{\tilde{\theta}_t^{(k)}} \quad \tilde{\theta}_{t+1}^{(k+1)} = \tilde{\theta}_{t+1}^{(k)} + \frac{\gamma_t}{nn'} \nabla_{\theta} \sum_{X_i \in A} \sum_{\mathbf{w} \in C'} f_{\mathbf{w}}(X_i) + \sqrt{2\gamma_t} \epsilon_t \xi''_k$ Generate

$B = \{X_1^{\text{real}}, \dots, X_n^{\text{real}}\} \sim \mathbb{P}_{\text{real}}$ $B' \leftarrow \{\}$ **for each** $\theta \in D'$ **do**

Generate $\tilde{B} = \{X'_1, \dots, X'_n\} \sim \mathbb{P}_{\theta}$ $B' \leftarrow B' \cup \tilde{B}$

$\tilde{\mathbf{w}}_{t+1}^{(k+1)} = \tilde{\mathbf{w}}_{t+1}^{(k)} + \frac{\gamma_t}{n} \nabla_{\mathbf{w}} \sum_{X_i^{\text{real}} \in B} f_{\tilde{\mathbf{w}}_{t+1}^{(k)}}(X_i^{\text{real}}) - \frac{\gamma_t}{nn'} \nabla_{\mathbf{w}} \sum_{X'_i \in B'} f_{\tilde{\mathbf{w}}_{t+1}^{(k)}}(X'_i) + \sqrt{2\gamma_t} \epsilon_t \xi'''_k;$

$\tilde{W}[t+1] \leftarrow \{\tilde{\mathbf{w}}_{t+1}^{(K+1)}, \dots, \tilde{\mathbf{w}}_{t+1}^{(K+n')}\}, \quad \tilde{\Theta}[t+1] \leftarrow \{\tilde{\theta}_{t+1}^{(K+1)}, \dots, \tilde{\theta}_{t+1}^{(K+n')}\}$

`idx` $\leftarrow \text{UNIF}(1, 2, \dots, T)$ **return** $W[\text{idx}], \Theta[\text{idx}]$.

C. Omitted Pseudocodes in the Main Text

We use the following notation for the hyperparameters of our algorithms:

- n : number of samples in the data batch.
- n' : number of samples for each probability measure.
- γ_t : SGLD step-size at iteration t .
- ϵ_t : thermal noise of SGLD at iteration t .
- K_t : warmup steps for SGLD at iteration t .
- β : exponential damping factor in the weighted average.

The approximate infinite-dimensional entropic MD and MP in Section 4.1 are depicted in **Algorithm 4** and **5**, respectively. **Algorithm 6** gives the heuristic version of the entropic Mirror-Prox.

D. Details and More Results of Experiments

This section contains all the details regarding our experiments, as well as more results on synthetic and real datasets.

Network Architectures: For all experiments, we consider the gradient-penalized discriminator (Gulrajani et al., 2017) as a soft constraint alternative to the original Wasserstein GANs, as it is known to achieve much better performance. The gradient penalty parameter is denoted by λ below.

For synthetic data, we use fully connected networks for both the generator and discriminator. They consist of three layers,

Algorithm 6 MIRROR-PROX-GAN: APPROXIMATE MIRROR-PROX FOR GANs

Input: $\tilde{\mathbf{w}}_1, \tilde{\boldsymbol{\theta}}_1 \leftarrow$ random initialization, $\mathbf{w}_0 \leftarrow \tilde{\mathbf{w}}_1, \boldsymbol{\theta}_0 \leftarrow \tilde{\boldsymbol{\theta}}_1, \{\gamma_t\}_{t=1}^T, \{\epsilon_t\}_{t=1}^T, \{K_t\}_{t=1}^T, \beta$, standard normal noise $\xi_k, \xi'_k, \xi''_k, \xi'''_k$.

for $t = 1, 2, \dots, T$ **do**

$\bar{\mathbf{w}}_t, \bar{\mathbf{w}}_{t+1}, \tilde{\mathbf{w}}_t^{(1)}, \tilde{\mathbf{w}}_{t+1}^{(1)} \leftarrow \tilde{\mathbf{w}}_t, \bar{\boldsymbol{\theta}}_t, \bar{\boldsymbol{\theta}}_{t+1}, \tilde{\boldsymbol{\theta}}_t^{(1)}, \tilde{\boldsymbol{\theta}}_{t+1}^{(1)} \leftarrow \tilde{\boldsymbol{\theta}}_t$ **for** $k = 1, 2, \dots, K_t$ **do**

Generate $A = \{X_1, \dots, X_n\} \sim \mathbb{P}_{\tilde{\boldsymbol{\theta}}_t^{(k)}}$ $\boldsymbol{\theta}_t^{(k+1)} = \tilde{\boldsymbol{\theta}}_t^{(k)} + \frac{\gamma_t}{n} \nabla_{\boldsymbol{\theta}} \sum_{X_i \in A} f_{\mathbf{w}_t}(X_i) + \sqrt{2\gamma_t} \epsilon_t \xi_k$ Generate $B = \{X_1^{\text{real}}, \dots, X_n^{\text{real}}\} \sim \mathbb{P}_{\text{real}}$ Generate $B' = \{X'_1, \dots, X'_n\} \sim \mathbb{P}_{\tilde{\boldsymbol{\theta}}_t}$
 $\mathbf{w}_t^{(k+1)} = \mathbf{w}_t^{(k)} + \frac{\gamma_t}{n} \nabla_{\mathbf{w}} \sum_{X_i^{\text{real}} \in B} f_{\mathbf{w}_t^{(k)}}(X_i^{\text{real}}) - \frac{\gamma_t}{n} \nabla_{\mathbf{w}} \sum_{X'_i \in B'} f_{\mathbf{w}_t^{(k)}}(X'_i) + \sqrt{2\gamma_t} \epsilon_t \xi'_k;$
 $\bar{\mathbf{w}}_t \leftarrow (1 - \beta)\bar{\mathbf{w}}_t + \beta \mathbf{w}_t^{(k+1)}$ $\bar{\boldsymbol{\theta}}_t \leftarrow (1 - \beta)\bar{\boldsymbol{\theta}}_t + \beta \boldsymbol{\theta}_t^{(k+1)}$
 $\mathbf{w}_t \leftarrow (1 - \beta)\mathbf{w}_{t-1} + \beta \bar{\mathbf{w}}_t$ $\boldsymbol{\theta}_t \leftarrow (1 - \beta)\boldsymbol{\theta}_{t-1} + \beta \bar{\boldsymbol{\theta}}_t$

for $k = 1, 2, \dots, K_t$ **do**

$\bar{\mathbf{w}}_{t+1}, \bar{\mathbf{w}}_{t+2}, \tilde{\mathbf{w}}_{t+1}^{(k+1)}, \tilde{\mathbf{w}}_{t+2}^{(k+1)} \leftarrow \tilde{\mathbf{w}}_{t+1}, \bar{\boldsymbol{\theta}}_{t+1}, \bar{\boldsymbol{\theta}}_{t+2}, \tilde{\boldsymbol{\theta}}_{t+1}^{(k+1)} = \tilde{\boldsymbol{\theta}}_{t+1}^{(k)} + \frac{\gamma_t}{n} \nabla_{\boldsymbol{\theta}} \sum_{X_i \in A} f_{\mathbf{w}_{t+1}}(X_i) + \sqrt{2\gamma_t} \epsilon_t \xi''_k$ Generate $B = \{X_1^{\text{real}}, \dots, X_n^{\text{real}}\} \sim \mathbb{P}_{\text{real}}$ Generate $B' = \{X'_1, \dots, X'_n\} \sim \mathbb{P}_{\boldsymbol{\theta}_{t+1}}$
 $\mathbf{w}_{t+1}^{(k+1)} = \mathbf{w}_{t+1}^{(k)} + \frac{\gamma_t}{n} \nabla_{\mathbf{w}} \sum_{X_i^{\text{real}} \in B} f_{\mathbf{w}_{t+1}^{(k)}}(X_i^{\text{real}}) - \frac{\gamma_t}{n} \nabla_{\mathbf{w}} \sum_{X'_i \in B'} f_{\mathbf{w}_{t+1}^{(k)}}(X'_i) + \sqrt{2\gamma_t} \epsilon_t \xi'''_k;$
 $\bar{\mathbf{w}}_{t+1} \leftarrow (1 - \beta)\bar{\mathbf{w}}_{t+1} + \beta \mathbf{w}_{t+1}^{(k+1)}$ $\bar{\boldsymbol{\theta}}_{t+1} \leftarrow (1 - \beta)\bar{\boldsymbol{\theta}}_{t+1} + \beta \boldsymbol{\theta}_{t+1}^{(k+1)}$
 $\tilde{\mathbf{w}}_{t+1} \leftarrow (1 - \beta)\tilde{\mathbf{w}}_t + \beta \bar{\mathbf{w}}_{t+1}$ $\tilde{\boldsymbol{\theta}}_{t+1} \leftarrow (1 - \beta)\tilde{\boldsymbol{\theta}}_t + \beta \bar{\boldsymbol{\theta}}_{t+1}$

each of them containing 512 neurons, with ReLU as nonlinearity.

For MNIST, we use convolutional neural networks identical to (Gulrajani et al., 2017) as the generator and discriminator.⁴ The generator uses a sigmoid function to map the output to range $[0, 1]$.

For LSUN bedroom, we use DCGAN (Radford et al., 2015), except that the number of the channels in each layer is half of the original model, and the last sigmoid function of the discriminator is removed. The output of the generator is mapped to $[0, 1]$ by hyperbolic tangent and a linear transformation. The architecture contains batch normalization layer to ensure the stability of the training. For our Mirror- and Mirror-Prox-GAN, the Gaussian noise from SGLD is not added to parameters in batch normalization layers, as the batch normalization creates strong dependence among entries of the weight matrix and was not covered by our theory.

Hyperparameter setting: The hyperparameter setting is summarized in Table 1. For baselines (SGD, RMSProp, Adam), we use the settings identical to (Gulrajani et al., 2017). For our proposed Mirror- and Mirror-Prox-GAN, we set the damping factor β to be 0.9. For K_t, γ_t and ϵ_t , we use the simple exponential scheduling:

$$K_t = \lfloor (1 + 10^{-5})^t \rfloor.$$

$$\gamma_t = \gamma \times (1 - 10^{-5})^t, \quad \gamma \text{ in Table 1.}$$

$$\epsilon_t = \epsilon \times (1 - 5 \times 10^{-5})^t, \quad \epsilon \text{ in Table 1.}$$

The idea is that the initial iterations are very noisy, and hence it makes sense to take less SGLD steps. As the iteration counts grow, the algorithms learn more meaningful parameters, and we should increase the number of SGLD steps as well as decreasing the step-size γ_t and thermal noise ϵ_t to make the sampling more accurate. This is akin to the warmup steps in the sampling literature.

⁴Their code is available on https://github.com/igul222/improved_wgan_training.

Algorithm	SGD		RMSProp	Adam			Entropic MD/MP		
Dataset	S	M	L	S	M	L	S	M	L
Step-size γ	10^{-2}		10^{-4}	10^{-4}			10^{-2}	10^{-4}	
Gradient penalty λ	0.1	10		0.1	10		0.1	10	
Noise ϵ							10^{-2}	10^{-3}	10^{-6}
Batch Size n	1024	50	64	1024	50	64	1024	50	64

Table 1. Hyperparameter setting. “S”, “M”, “L” stands for synthetic data, MNIST and LSUN bedroom, respectively. MD for LSUN bedroom uses a RMSProp preconditioner, so the step-size is the same as one in RMSProp.

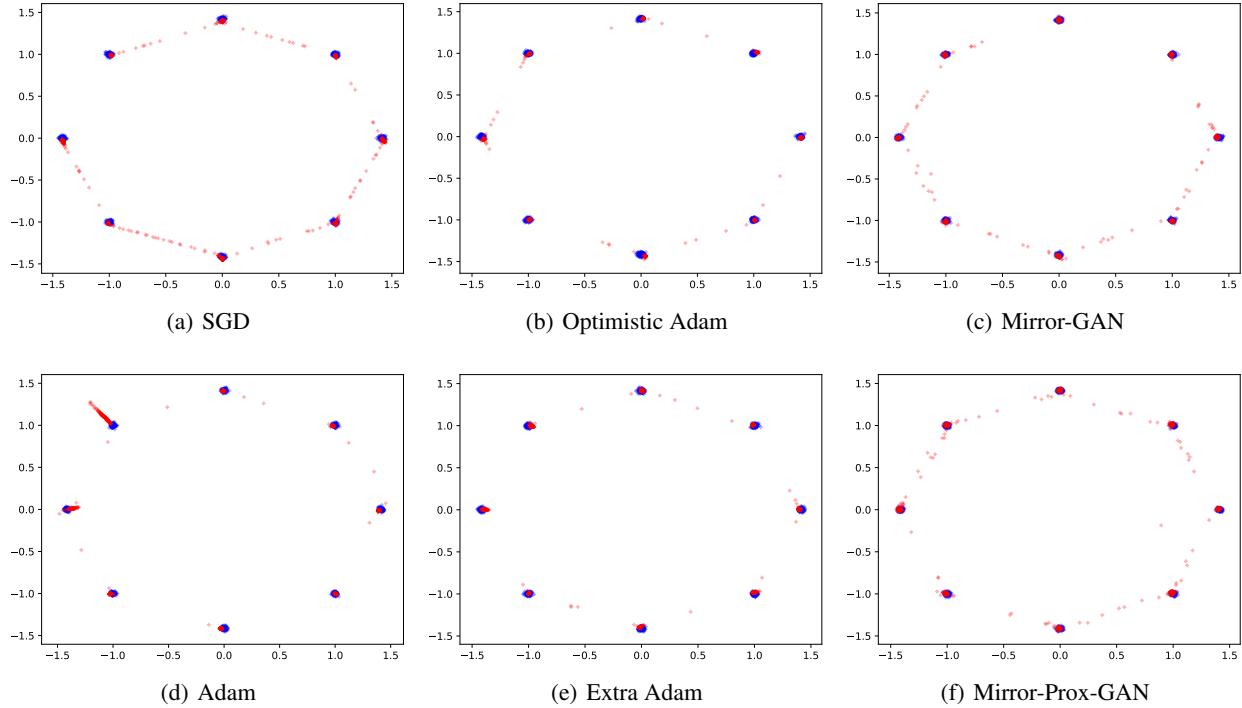


Figure 3. Fitting 8 Gaussian mixtures up to 10^5 iterations.

D.1. Synthetic Data

Figure 3, 4, and 5 show results on learning 8 Gaussian mixtures, 25 Gaussian mixtures, and the Swiss Roll. As in the case for 25 Gaussian mixtures, we find that Mirror- and Mirror-Prox-GAN can better capture the variance of the true distribution, as well as finding the unbiased modes.

In Figure 6, we plot the data generated after 10^4 , 2×10^4 , 5×10^4 , 8×10^4 , and 10^5 iterations by different algorithms for 25 Gaussian mixtures. It is clear that Mirror- and Mirror-Prox-GAN find the modes of the distribution faster. In practice, it was observed that the noise introduced by SGLD quickly drives the iterates to non-trivial parameter regions, whereas SGD tends to get stuck at very bad local minima. Adam, as an adaptive algorithm, is capable of escaping bad local minima, however at a rate slower than Mirror- and Mirror-Prox-GAN. The quality of Adam-based algorithms’ final solutions are also not as good as Mirror- and Mirror-Prox-GAN; see the discussions in Section 5.1.

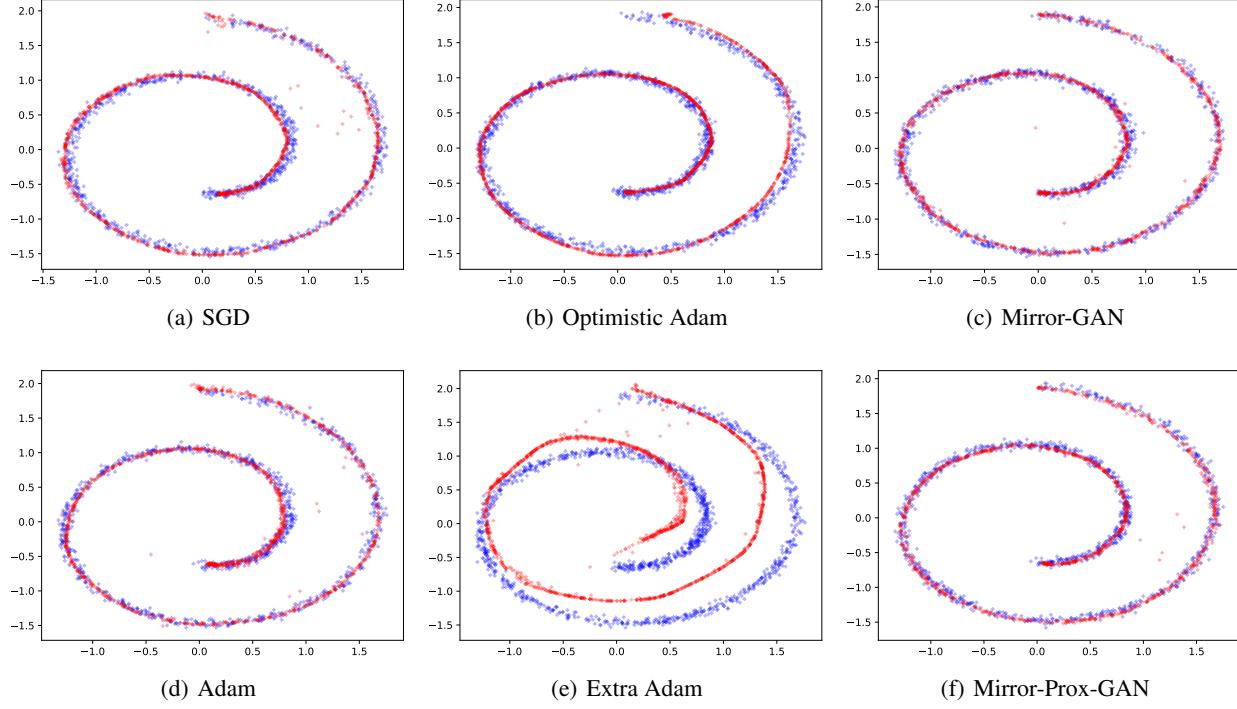


Figure 4. Fitting the ‘Swiss Roll’ up to 10^5 iterations.

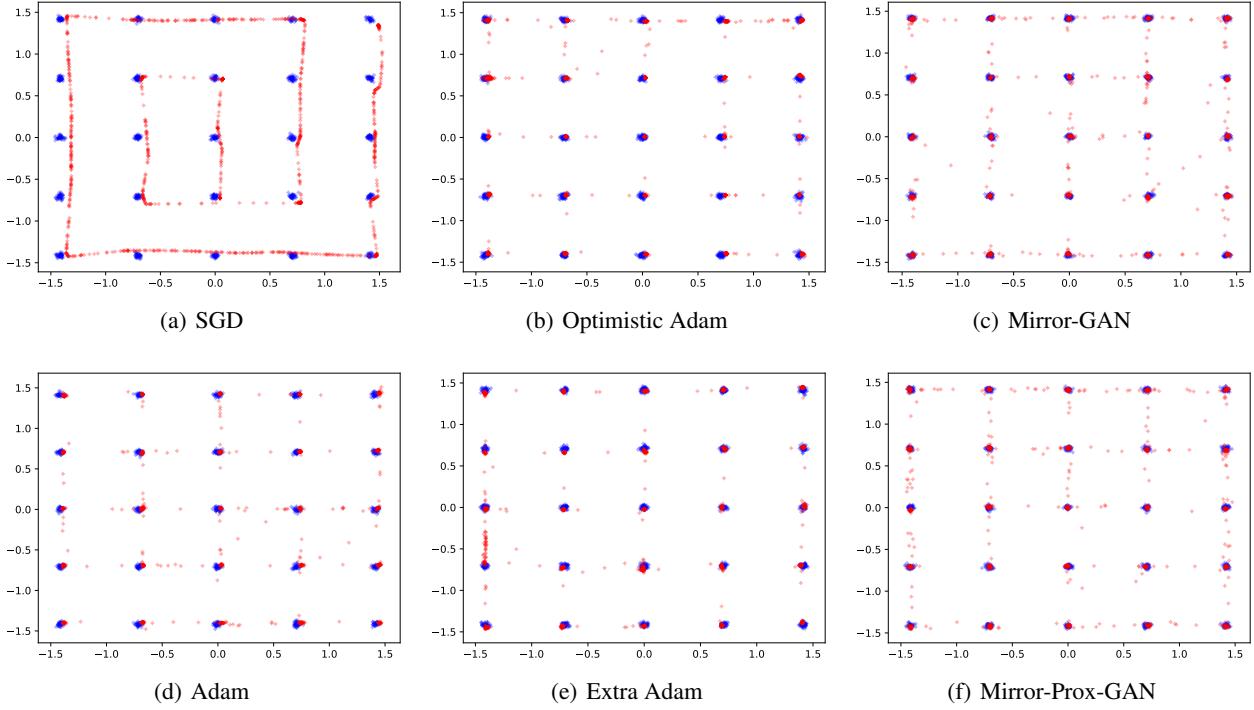


Figure 5. Fitting 25 Gaussian mixtures up to 10^5 iterations.

D.2. Real Data

D.2.1. MNIST

Results on MNIST dataset are shown in Figure 7. The models are trained by each algorithm for 10^5 iterations. We can see that all algorithms achieve comparable performance. Therefore, the dataset seems too weak to be a discriminator for

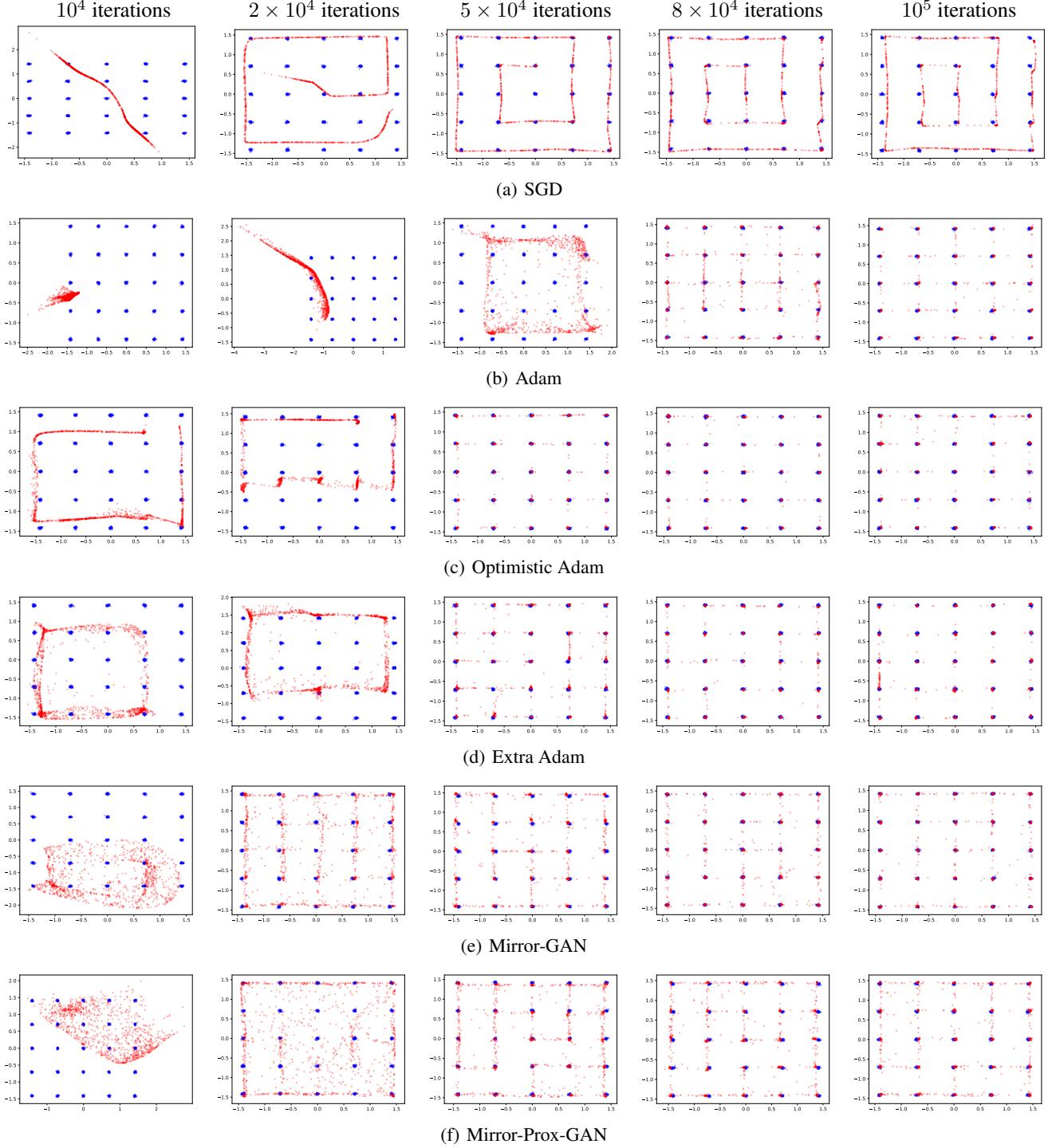


Figure 6. Learning 25 Gaussian mixtures accross different iterations.

different algorithms.

D.2.2. LSUN BEDROOM

More results on the LSUN bedroom dataset are shown in Figure 8. We show images generated after 4×10^4 , 8×10^4 , and 10^5 iterations by each algorithm. We can see that the Mirror-GAN and Alternated Extra-Adam outperform vanilla RMSProp.



(a) True Data



(b) SGD



(c) Adam



(d) Mirror-GAN



(e) Mirror-Prox-GAN

Figure 7. True MNIST images and samples generated by different algorithms.

Adam was able to obtain meaningful images in early stages of training. However, further iterations do not improve the image quality of Adam. In contrast, they lead to severe mode collapse at the 8×10^4 th iteration, and converge to noise later

Algorithm	RMSProp	Adam	Entropic MD	Extra-Adam
Simultaneous	-	-	3.0955	2.0015
Alternated	3.0555	1.3730	-	3.1620

Table 2. Inception Score of generator trained on LSUN dataset. The reported scores are based on the average of 6400 images from each generator.

on. Simultaneous Extra-Adam completely fails in this task.

Finally, for reference, we report the Inception Score in Table 2.

4×10^4 iterations



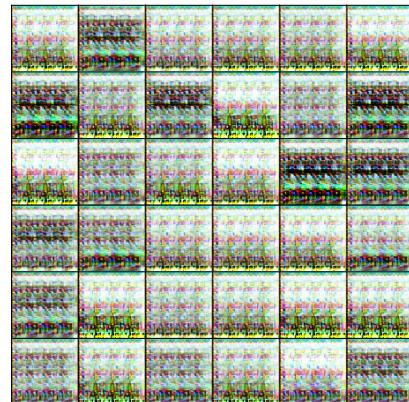
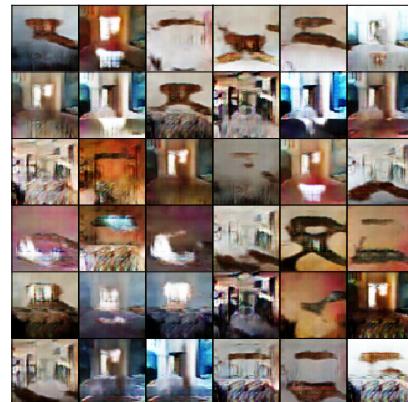
8×10^4 iterations



10^5 iterations



(a) RMSProp



(b) Adam



(c) Mirror-GAN, **Algorithm 4.1**



(d) Simultaneous Extra-Adam



(e) Alternated Extra-Adam

Figure 8. Image generated by RMSProp, Simultaneous and Alternated Extra-Adam, Adam, and Mirror-GAN on the LSUN bedroom dataset.