Finding Mixed Nash Equilibria of Generative Adversarial Networks

Ya-Ping Hsieh **Chen Liu** **Volkan Cevher**

Laboratory for Information and Inference Systems (LIONS),
EPFL, Lausanne, Switzerland
{ya-ping.hsieh, chen.liu, volkan.cevher}@epfl.ch

Abstract

We reconsider the training objective of Generative Adversarial Networks (GANs) from the *mixed Nash Equilibria* (NE) perspective. Inspired by the classical prox methods, we develop a novel algorithmic framework for GANs via an infinite-dimensional two-player game and prove rigorous convergence rates to the mixed NE. We then propose a principled procedure to reduce our novel prox methods to simple sampling routines, leading to practically efficient algorithms. Finally, we provide experimental evidence that our approach outperforms methods that seek pure strategy equilibria, such as SGD, Adam, and RMSProp, both in speed and quality.

1 Introduction

Training of Generative Adversarial Networks (GANs) are known to be notoriously difficult. In the language of game theory, GAN seeks for a *pure strategy* equilibrium, which is well-known to be ill-posed in many scenarios [6]. Indeed, it is known that a pure strategy equilibrium might not exist [2], might be degenerate [22], or cannot be reliably reached by existing algorithms [16]. These theoretical barriers are corroborated with abundant empirical evidence, where popular algorithms such as SGD or Adam lead to unstable training.

In this work, we propose to study the *mixed Nash Equilibrium* (NE) of GANs: Instead of searching for an optimal pure strategy which might not even exist, we optimize over the set of *probability distributions* over pure strategies of the networks. We demonstrate that the prox methods of [19, 17] can be extended to continuously many strategies, and hence applicable to training GANs. We then construct a principled procedure to reduce our novel prox methods to certain sampling tasks that were empirically proven easy by recent work [4, 5, 8]. Further, we establish heuristic guidelines to greatly scale down the memory and computational costs. Finally, we experimentally show that our algorithms consistently achieve better or comparable performance than popular baselines.

Related Work: While the literature on training GANs is vast, to our knowledge, there exist only few papers on the mixed NE perspective [2, 10, 20], and they propose only heuristic algorithms. The work [11] 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.

Due to its fundamental role in game theory, many prox methods have been applied to study the training of GANs [7, 9, 15]. However, these work focus on the classical pure strategy equilibria and is hence distinct from our problem formulation.

Extended abstract. Smooth Games Optimization and Machine Learning Workshop (NIPS 2018), Montréal, Canada.

Algorithm 1: Infinite-Dimensional Entropic MD

Input: Initial distributions μ_1, ν_1, learning rate η
for $t = 1, 2, \ldots, T-1$ **do**
$\quad \nu_{t+1} = \mathrm{MD}_\eta\left(\nu_t, -G^\dagger \mu_t\right), \quad \mu_{t+1} = \mathrm{MD}_\eta\left(\mu_t, -g + G\nu_t\right);$
return $\bar{\nu}_T = \frac{1}{T}\sum_{t=1}^T \nu_t$ and $\bar{\mu}_T = \frac{1}{T}\sum_{t=1}^T \mu_t$.

2 Problem Formulation: Mixed Strategy Formulation for GANs

For illustration, let us focus on the Wasserstein GAN [1]:

$$\min_{\boldsymbol{\theta}\in\Theta} \max_{\boldsymbol{w}\in\mathcal{W}} \mathbb{E}_{X\sim\mathbb{P}_{\mathrm{real}}}[f_{\boldsymbol{w}}(X)] - \mathbb{E}_{X\sim\mathbb{P}_{\boldsymbol{\theta}}}[f_{\boldsymbol{w}}(X)], \tag{1}$$

where Θ is the set of parameters for the generator and $\mathcal{W}$ the set of parameters for the discriminator[1] f, typically both taken to be neural nets.

The high-level idea of our approach is, instead of solving (1) directly, we focus on the *mixed strategy* formulation of (1). In other words, letting $\mathcal{M}(\Theta)$ and $\mathcal{M}(\mathcal{W})$ be the set of all probability distributions over Θ and $\mathcal{W}$, we search for the optimal distribution that solves:

$$\min_{\nu\in\mathcal{M}(\Theta)} \max_{\mu\in\mathcal{M}(\mathcal{W})} \mathbb{E}_{\boldsymbol{w}\sim\mu}\mathbb{E}_{X\sim\mathbb{P}_{\mathrm{real}}}[f_{\boldsymbol{w}}(X)] - \mathbb{E}_{\boldsymbol{w}\sim\mu}\mathbb{E}_{\boldsymbol{\theta}\sim\nu}\mathbb{E}_{X\sim\mathbb{P}_{\boldsymbol{\theta}}}[f_{\boldsymbol{w}}(X)]. \tag{2}$$

Define the function $g : \mathcal{W} \to \mathbb{R}$ by $g(\boldsymbol{w}) \coloneqq \mathbb{E}_{X\sim\mathbb{P}_{\mathrm{real}}}[f_{\boldsymbol{w}}(X)]$ and the operator $G : \mathcal{M}(\Theta) \to \mathcal{F}(\mathcal{W})$ as $(G\nu)(\boldsymbol{w}) \coloneqq \mathbb{E}_{\boldsymbol{\theta}\sim\nu,\mathbf{X}\sim\mathbb{P}_{\boldsymbol{\theta}}}[f_{\boldsymbol{w}}(X)]$. Denoting $\langle \mu, h\rangle \coloneqq \mathbb{E}_\mu h$ for any probability measure μ and function h, we may rewrite (2) as

$$\min_{\nu\in\mathcal{M}(\Theta)} \max_{\mu\in\mathcal{M}(\mathcal{W})} \langle \mu, g\rangle - \langle \mu, G\nu\rangle . \tag{3}$$

We have thus shown that the finding the mixed NE of Wasserstein GANs is nothing but solving a bi-affine two-player game with *continuously infinite* strategies, in contrast to the classical finite-strategy setting. Inspired by the **entropic prox methods** [18, 3, 17] for solving finite games, we ask:

> *Can the entropic Mirror Descent and Mirror-Prox be extended to infinite dimension to solve (3)? Can we retain the convergence rates as in the finite-strategy setting?*

3 Infinite-Dimensional Prox Methods and Convergence Rates

The purpose of this section is to answer the above two questions with an affirmative "yes".

Theorem 1 (Infinite-Dimensional Mirror Descent, informal)**.** *1. Let Φ be the negative Shannon entropy, $\Phi^\star$ be its Fenchel dual, and $\mathrm{d}\Phi$ be the Fréchet derivative. For any probability measure μ on a set $\mathcal{Z}$, we may define*

$$\mu_+ = \mathrm{MD}_\eta(\mu, h) \quad \equiv \quad \mu_+ = \mathrm{d}\Phi^\star\left(\mathrm{d}\Phi(\mu) - \eta h\right) \equiv \quad \mathrm{d}\mu_+ = \frac{e^{-\eta h}\mathrm{d}\mu}{\int e^{-\eta h}\mathrm{d}\mu}. \tag{4}$$

*2. Assume that we have access to the deterministic derivatives $\left\{-G^\dagger\mu_t\right\}_{t=1}^T$ and $\{g - G\nu\}_{t=1}^T$, then **Algorithm 1** achieves $O\left(T^{-1/2}\right)$-NE. If we only have access to unbiased stochastic derivatives, then **Algorithm 1** achieves $O\left(T^{-1/2}\right)$-NE in expectation.*

Remark. The case for entropic Mirror-Prox can be similarly derived.

Notice that **Algorithm 1** iterates over the space of probability measures, which we cannot compute. We hence need to an algorithm to approximate these probability updates, which is the purpose of the next section.

[1] Also known as "critic" in Wasserstein GAN literature.

Algorithm 2: MIRROR-GAN: APPROXIMATE MIRROR DECENT FOR GANS

Input: $\bar{\boldsymbol{w}}_1, \bar{\boldsymbol{\theta}}_1 \leftarrow$ random initialization, $\{\gamma_t\}_{t=1}^T, \{\epsilon_t\}_{t=1}^T, \{K_t\}_{t=1}^{T-1}, \beta$.

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

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

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

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

Generate $A = \{X_1, \ldots, X_n\} \sim \mathbb{P}_{\boldsymbol{\theta}_t^{(k)}}$;

$\boldsymbol{\theta}_t^{(k+1)} = \boldsymbol{\theta}_t^{(k)} + \frac{\gamma_t}{n} \nabla_{\boldsymbol{\theta}} \sum_{X_i \in A} f_{\boldsymbol{w}_t}(X_i) + \sqrt{2\gamma_t}\epsilon_t \mathcal{N}(0, I)$;

Generate $B = \{X_1^{\text{real}}, \ldots, X_n^{\text{real}}\} \sim \mathbb{P}_{\text{real}}$;

Generate $B' = \{X_1', \ldots, X_n'\} \sim \mathbb{P}_{\boldsymbol{\theta}_t}$;

$$\boldsymbol{w}_t^{(k+1)} = \boldsymbol{w}_t^{(k)} + \frac{\gamma_t}{n} \nabla_{\boldsymbol{w}} \sum_{X_i^{\text{real}} \in B} f_{\boldsymbol{w}_t^{(k)}}(X_i^{\text{real}}) - \frac{\gamma_t}{n} \nabla_{\boldsymbol{w}} \sum_{X_i' \in B'} f_{\boldsymbol{w}_t^{(k)}}(X_i') + \sqrt{2\gamma_t}\epsilon_t \mathcal{N}(0, I);$$

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

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

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

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

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

4 From Theory to Practice

Section 4.1 reduces **Algorithm 1** to a sampling routine [23] that has widely been used in machine learning. Section 4.2 proposes to further simplify the algorithm by summarizing a batch of samples by their mean.

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

Step 1: Reformulating Entropic Mirror Descent Iterates

Our first step is to express (4) in a more tractable form. Using properties of the (negative) Shannon entropy, we may express the probability measures of **Algorithm 1** in terms of the history: $\mathrm{d}\mu_T = \frac{\exp\left\{(T-1)g - G\sum_{s=1}^{T-1}\nu_s\right\}\mathrm{d}\boldsymbol{w}}{\int \exp\left\{(T-1)g - G\sum_{s=1}^{T-1}\nu_s\right\}\mathrm{d}\boldsymbol{w}}$ and $\mathrm{d}\nu_T = \frac{\exp\left\{G^\dagger \sum_{s=1}^{T-1}\mu_s\right\}\mathrm{d}\boldsymbol{\theta}}{\int \exp\left\{G^\dagger \sum_{s=1}^{T-1}\mu_s\right\}\mathrm{d}\boldsymbol{\theta}}$.

Step 2: Empirical Approximation for Stochastic Derivatives

The derivatives of (3) involve the function g and operator G, which involve taking expectation over distributions we do not have access to. However, if we are able to draw samples from μ_t and ν_t, then we can approximate the expectation via the empirical average, which leads to unbiased stochastic derivatives.

Step 3: Sampling by Stochastic Gradient Langevin Dynamics

Now, assuming that we have obtained unbiased stochastic derivatives $-\sum_{s=1}^t \hat{G}^\dagger \mu_s$ and $\sum_{s=1}^t \left(-\hat{g} + \hat{G}\nu_s\right)$. We may draw samples from the updated probability measures (μ_{t+1}, ν_{t+1}) by using the *Stochastic Gradient Langevin Dynamics* (SGLD) [23] as follows. For any probability distribution with density function $e^{-h}\mathrm{d}\boldsymbol{z}$, the SGLD iterates as

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \gamma \hat{\nabla} h(\boldsymbol{z}_k) + \sqrt{2\gamma}\epsilon \xi_k, \tag{5}$$

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. The

(a) RMSProp (b) Adam (c) Mirror-GAN

Figure 1: Dataset `LSUN bedroom`, 10^5 iterations.

theory of [23] states that, for large enough k, the iterates of SGLD above (approximately) generate samples according to the probability measures (μ_{t+1}, ν_{t+1}). Recursively applying Steps 1-3, we can then acquire approximate samples from (μ_T, ν_T).

4.2 Summarizing Samples by Averaging: A Simple yet Effective Heuristic

Although the algorithm in Section 4.1 is implementable, the resulting computational complexity is $O(T^2)$, and is hence too extensive for practical use.

An intuitive approach to alleviate the computational issue is to summarize each distribution by only *one* parameter. To this end, the mean of the distribution is the most natural candidate, as it not only stablizes the algorithm, but also is often easier to acquire than the actual samples. In this paper, we adopt the same approach as in [4] where we use exponential damping (the β term in **Algorithm 2)** to increase stability. **Algorithm 2**, termed the Mirror-GAN, shows how to encompass this idea into entropic MD.

5 Experimental Evidence

We first repeat the synthetic setup as in [12]. On synthetic data, including 8 Gaussian mixtures, 25 Gaussian mixtures and swissroll, our proposed methods obtain better results than its baseline counterparts such as SGD and Adam.

For real images, we use `MNIST` and `LSUN` as the dataset. We use the same architecture (DCGAN) as in [21] with batch normalization. As the networks become deeper in this case, the gradient magnitudes differ significantly across different layers. To alleviate such issues, we replace SGLD by the RMSProp-preconditioned SGLD [14] for our sampling routines. For baselines, we consider two adaptive gradient methods: RMSProp and Adam.

Figure 1 shows the results at the 10^5th iteration. The RMSProp and Mirror-GAN produce images with reasonable quality, while Adam outputs random noise. The visual quality of Mirror-GAN is better than RMSProp, as RMSProp sometimes generates blurry images (the $(3, 3)$- and $(1, 5)$-th entry of Figure 1(a)).

6 Conclusions

Our goal of systematically understanding and expanding on the game theoretic perspective of mixed NE along with stochastic Langevin dynamics for training GANs is a promising research vein. While simple in retrospect, we provide guidelines in developing approximate infinite-dimensional prox methods that probably learn the mixed NE of GANs. Our proposed Mirror- and Mirror-Prox-GAN algorithm feature cheap per-iteration complexity while rapidly converging to solutions of good quality.

Acknowledgments

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

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, pages 224–232, 2017.

[3] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[4] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations*, 2017.

[5] Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, and Guillaume Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks. *Research in the Mathematical Sciences*, 5(3):30, Jun 2018.

[6] Partha Dasgupta and Eric Maskin. The existence of equilibrium in discontinuous economic games, i: Theory. *The Review of economic studies*, 53(1):1–26, 1986.

[7] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 2018.

[8] Gintare Karolina Dziugaite and Daniel Roy. 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*, pages 1376–1385, 2018.

[9] Gauthier Gidel, Hugo Berard, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial nets. *arXiv preprint arXiv:1802.10551*, 2018.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] Paulina Grnarova, Kfir Y Levy, Aurelien Lucchi, Thomas Hofmann, and Andreas Krause. An online learning approach to generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[12] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[13] Paul R Halmos. *Measure theory*, volume 18. Springer, 2013.

[14] Chunyuan Li, Changyou Chen, David E Carlson, and Lawrence Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *AAAI*, 2016.

[15] Panayotis Mertikopoulos, Houssam Zenati, Bruno Lecouat, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.

[16] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. In *Advances in Neural Information Processing Systems*, pages 1825–1835, 2017.

[17] Arkadi Nemirovski. 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.

[18] Arkadii Nemirovskii and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

[19] AS Nemirovsky and DB Yudin. Problem complexity and method efficiency in optimization. 1983.

[20] Frans A Oliehoek, Rahul Savani, Jose Gallego, Elise van der Pol, and Roderich Groß. Beyond local nash equilibria for adversarial networks. *arXiv preprint arXiv:1806.07268*, 2018.

[21] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[22] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. 2017.

[23] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.