A Variational Inequality Perspective on Generative Adversarial Networks

Gauthier Gidel[*]
Mila

Hugo Berard[*]
Mila, FAIR[†]

Gaëtan Vignoud
Mila

Pascal Vincent[‡]
Mila, FAIR[†]

Simon Lacoste-Julien[‡]
Mila

DIRO, Université de Montréal, Montreal, Canada

Abstract

Generative adversarial networks (GANs) form a generative modeling approach known for producing appealing samples, but they are notably difficult to train. One common way to tackle this issue has been to propose new formulations of the GAN objective. In this work, we cast GAN optimization problems in the general variational inequality framework. Tapping into the mathematical programming literature, we counter some common misconceptions about the difficulties of saddle point optimization and propose to extend techniques designed for variational inequalities to the training of GANs. We apply *averaging* and *extrapolation* to the stochastic gradient method (SGD) and Adam.

1 Introduction

Generative adversarial networks (GANs) [Goodfellow et al., 2014] form a generative modeling approach known for producing realistic natural images [Karras et al., 2018]. Nevertheless, GANs are also known to be difficult to train, often displaying an unstable behavior [Goodfellow, 2016]. Much recent work has tried to tackle these training difficulties, usually by proposing new formulations of the GAN objective [Nowozin et al., 2016, Arjovsky et al., 2017]. Yet it is known that for some games [Goodfellow, 2016, §8.2] SGD exhibits oscillatory behavior and fails to converge.

We point out that multi-player games can be cast as *variational inequality problems* and consequently the same applies to any GAN formulation posed as a minimax or non-zero-sum game. We present two techniques from this literature, namely *averaging* and *extrapolation*, widely used to solve variational inequality problems (VIP) but which have not been explored in the context of GANs before.[4] We propose to apply these techniques to GAN training methods such as Adam or SGD. Finally, we test these techniques in the context of standard GAN training. We observe a 4%-6% improvement on the inception score [Salimans et al., 2016] of WGAN [Arjovsky et al., 2017] and WGAN-GP [Gulrajani et al., 2017] on the CIFAR-10 dataset.

Related Work The extragradient method is the standard algorithm to optimize variational inequalities. This algorithm has been originally introduced by Korpelevich [1976] and extended by Nesterov [2007] and Nemirovski [2004]. Stochastic versions of the extragradient have been recently analyzed [Juditsky et al., 2011, Yousefian et al., 2014, Iusem et al., 2017] for stochastic

[*] Equal contribution, correspondence to `firstname.lastname@umontreal.ca`.

[†] Facebook Artificial Intelligence Research.

[‡] CIFAR fellow.

[4] Independent works [Mertikopoulos et al., 2018] and [Yazıcı et al., 2018] respectively explored extrapolation and averaging in the context of GANs.

Extended abstract. Smooth Games Optimization and Machine Learning Workshop (NIPS 2018), Montréal, Canada. Longer version available: https://arxiv.org/abs/1802.10551

variational inequalities with *bounded constraints*. A linearly convergent variance reduced version of the stochastic gradient method has been proposed by Palaniappan and Bach [2016] for strongly monotone variational inequalities. Several methods to stabilize GANs consist in transforming a zero-sum formulation into a more general game that can no longer be cast as a saddle point problem. This is the case of the *non-saturating* formulation of GANs [Goodfellow et al., 2014, Fedus et al., 2018], the DCGANs [Radford et al., 2016]. Yadav et al. [2018] propose an optimization method for GANs based on AltSGD using a momentum based step on the generator. Daskalakis et al. [2018] proposed a method inspired from game theory. Li et al. [2017] suggest to dualize the GAN objective to reformulate it as a maximization problem and Mescheder et al. [2017] propose to add the norm of the gradient in the objective and provide an interesting perspective on GANs, interpreting the training as the search of a two-player game equilibrium. Non-convex results were proved, for a new notion of regret minimization, by Hazan et al. [2017] and in the context of GANs by Grnarova et al. [2018].

2 GAN optimization as a variational inequality problem

The generative adversarial network training strategy can be understood as a *game* between two players called *generator* and *discriminator*. The former produces a sample that the latter has to classify between real or fake data. The final goal is to build a generator able to produce sufficiently realistic samples to fool the discriminator. In the original GAN paper [Goodfellow et al., 2014], the GAN objective is formulated as a *zero-sum game* where the cost function of the discriminator $D_{\boldsymbol{\varphi}}$ is given by the negative log-likelihood of the binary classification task between real or fake data generated from $q_{\boldsymbol{\theta}}$ by the generator,

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\varphi}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\varphi}) \overset{\text{def}}{=} -\mathop{\mathbb{E}}_{\mathbf{x}\sim p}[\log D_{\boldsymbol{\varphi}}(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}'\sim q_{\boldsymbol{\theta}}}[\log(1 - D_{\boldsymbol{\varphi}}(\mathbf{x}'))] \,. \tag{1}$$

The minimax formulation (1), is theoretically convenient it is widely studied and provides guarantees on the existence of equilibria. However, practical considerations lead the GAN literature to consider a different objective for each player. The *two-player game problem* [Von Neumann and Morgenstern, 1944] consists in finding the following *Nash equilibrium*:

$$\boldsymbol{\theta}^* \in \mathop{\arg\min}_{\boldsymbol{\theta}\in\Theta} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta}, \boldsymbol{\varphi}^*) \quad \text{and} \quad \boldsymbol{\varphi}^* \in \mathop{\arg\min}_{\boldsymbol{\varphi}\in\Phi} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\theta}^*, \boldsymbol{\varphi}) \,. \tag{2}$$

One important point to notice is that the two optimization problems in (2) are *coupled* and have to be considered *jointly* from an optimization point of view.

We consider the local necessary conditions that characterize the solution of the *smooth* two-player game (2), defining *stationary points*, which will motivate the definition of a variational inequality. In the unconstrained setting, a *stationary point* is a couple $(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)$ with zero gradient:

$$\|\nabla_{\boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)\| = \|\nabla_{\boldsymbol{\varphi}} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)\| = 0 \,. \tag{3}$$

When constraints are present,[5] a *stationary point* $(\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)$ is such that the directional derivative of each cost function is non-negative in any feasible direction (i.e. there is no feasible descent direction). Defining $\boldsymbol{\omega} \overset{\text{def}}{=} (\boldsymbol{\theta}, \boldsymbol{\varphi})$, $\boldsymbol{\omega}^* \overset{\text{def}}{=} (\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)$, $\Omega \overset{\text{def}}{=} \Theta \times \Phi$, it can be compactly formulated as:

$$F(\boldsymbol{\omega}^*)^\top (\boldsymbol{\omega} - \boldsymbol{\omega}^*) \geq 0 \,, \;\; \forall \boldsymbol{\omega} \in \Omega \quad \text{where} \quad F(\boldsymbol{\omega}) \overset{\text{def}}{=} \Big[\nabla_{\boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta}, \boldsymbol{\varphi}) \quad \nabla_{\boldsymbol{\varphi}} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\theta}, \boldsymbol{\varphi})\Big]^\top \,. \tag{4}$$

These stationary conditions can be generalized to any continuous vector field: let $\Omega \subset \mathbb{R}^d$ and $F : \Omega \to \mathbb{R}^d$ be a continuous mapping. The *variational inequality problem* [Harker and Pang, 1990] (depending on F and Ω) is:

$$\text{find } \boldsymbol{\omega}^* \in \Omega \quad \text{such that} \quad F(\boldsymbol{\omega}^*)^\top (\boldsymbol{\omega} - \boldsymbol{\omega}^*) \geq 0 \,, \;\; \forall \boldsymbol{\omega} \in \Omega \,. \tag{VIP}$$

We call *optimal set* the set Ω^* of $\boldsymbol{\omega} \in \Omega$ verifying (VIP). The intuition behind it is that any $\boldsymbol{\omega}^* \in \Omega^*$ is a *fixed point* of the *constrained* dynamic of F (constrained to Ω).

3 Optimization of Variational Inequalities (batch setting)

Let us begin by looking at techniques that were developed in the optimization literature to solve (4).

[5] An example of constraint for GANs is to clip the parameters of the discriminator [Arjovsky et al., 2017].

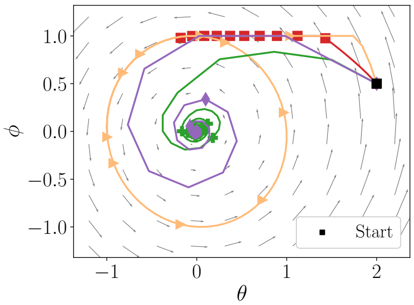

Figure 1: Comparison of the basic gradient method (as well as Adam) with the techniques presented in §3 on the optimization of $\min \max \;\; \theta \cdot \phi$. Only the algorithms advocated in this paper (Averaging, Extrapolation) converge quickly to the solution. Each marker represents 20 iterations.

Averaging We consider first a *weighted averaging* scheme of the *gradient method*,

$$\bar{\boldsymbol{\omega}}_T \overset{\text{def}}{=} \sum_{t=0}^{T-1} \rho_t \boldsymbol{\omega}_t / S_T\,, \quad S_T \overset{\text{def}}{=} \sum_{t=0}^{T-1} \rho_t \quad \text{with} \quad \rho_t \geq 0\,. \tag{5}$$

Averaging schemes can be efficiently implemented in an online fashion noticing that,

$$\bar{\boldsymbol{\omega}}_t = (1 - \tilde{\rho}_t)\bar{\boldsymbol{\omega}}_{t-1} + \tilde{\rho}_t \boldsymbol{\omega}_t \quad \text{where} \quad 0 \leq \tilde{\rho}_t \leq 1\,. \tag{6}$$

For instance, setting $\tilde{\rho}_t = \frac{1}{t}$ provides *uniform averaging* ($\rho_t = 1$) and $\tilde{\rho}_t = 1 - \beta < 1$ provides *geometric averaging* also known as *exponential moving averaging* ($\rho_t = \beta^t$).

Extrapolation Another technique used in the variational inequality literature to prevent oscillations is *extrapolation* [Korpelevich, 1976]. The idea behind this technique is to compute the gradient at an (extrapolated) point different from the current point from which the update is performed, stabilizing the dynamics:

$$\text{Compute extrapolated point:} \;\; \boldsymbol{\omega}_{t+1/2} = P_\Omega[\boldsymbol{\omega}_t - \eta F(\boldsymbol{\omega}_t)]\,, \tag{7}$$

$$\text{Perform update step:} \;\; \boldsymbol{\omega}_{t+1} = P_\Omega[\boldsymbol{\omega}_t - \eta F(\boldsymbol{\omega}_{t+1/2})]\,. \tag{8}$$

Note that, even in the *unconstrained case*, this method is intrinsically different from Nesterov's momentum because of this lookahead step for the gradient computation.

4 Optimization of VIP with stochastic gradients

In this section, we consider extensions of the techniques presented in section §3 for optimizing (VIP), to the context of a *stochastic* operator. In this case, at each time step we no longer have access to the exact gradient $F(\boldsymbol{\omega})$ but to an unbiased *stochastic* estimate of it $F(\boldsymbol{\omega}, \xi)$ where $\xi \sim P$ and $F(\boldsymbol{\omega}) := \mathbb{E}_{\xi \sim P}[F(\boldsymbol{\omega}, \xi)]$.

This is motivated from the GAN formulation where we only have access to a finite sample estimate of the expected gradient, computed on a mini-batch. For GANs, ξ is thus a mini-batch of points coming from the true data distribution p and the generator distribution $q_{\boldsymbol{\theta}}$. For our analysis, we require at least one of the two following assumptions on the stochastic operator:

Assumption 1. *Bounded variance by* σ^2*:* $\mathbb{E}_\xi[\|F(\boldsymbol{\omega}) - F(\boldsymbol{\omega}, \xi)\|^2] \leq \sigma^2\,, \;\; \forall \boldsymbol{\omega} \in \Omega\,.$

Assumption 2. *Bounded expected squared norm by* M^2*:* $\mathbb{E}_\xi[\|F(\boldsymbol{\omega}, \xi)\|^2] \leq M^2$, $\forall \boldsymbol{\omega} \in \Omega$.

To handle constraints such as parameter clipping [Arjovsky et al., 2017], we present a *projected* version of theses algorithms, where $P_\Omega[\boldsymbol{\omega}']$ denotes the projection of $\boldsymbol{\omega}'$ onto Ω. In order to prove the convergence of these three algorithms we will assume that F is monotone:

Assumption 3. F *is* monotone *and* Ω *is a compact convex set, such that* $\max_{\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega} \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|^2 \leq R^2$.

Theorem 1. *Under Assump. 1, 2 and 3, SGD with averaging (5) with a constant step-size gives,*

$$\mathbb{E}[\mathrm{Err}(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R^2}{2\eta T} + \eta \frac{M^2 + \sigma^2}{2}\,, \quad \bar{\boldsymbol{\omega}}_T \overset{\textit{def}}{=} \frac{1}{T} \sum_{t=0}^{T-1} \boldsymbol{\omega}_t\,. \tag{9}$$

Thm. 1 uses a similar proof as [Nemirovski et al., 2009].

Theorem 2. *[Juditsky et al., 2011, Thm. 1] Under Assump. 1 and 3, if* $\mathbb{E}_\xi[F]$ *is L-Lipschitz, then SGD with* extrapolation (7) and averaging (5) *using a constant step-size* $\eta \leq \frac{1}{\sqrt{3}L}$ *gives,*

$$\mathbb{E}[\mathrm{Err}(\bar{\boldsymbol{\omega}}_T)] \leq \frac{R^2}{\eta T} + \frac{7}{2}\eta\sigma^2 \quad \textit{where} \quad \bar{\boldsymbol{\omega}}_T \overset{\textit{def}}{=} \frac{1}{T} \sum_{t=0}^{T-1} \boldsymbol{\omega}'_t\,, \quad \forall T \geq 1\,. \tag{10}$$

Model	WGAN (DCGAN)			WGAN-GP (ResNet)	
Method	no averaging	uniform avg	EMA	no averaging	uniform avg
SimAdam	$\mathit{6.05} \pm \mathit{.12}$	$5.83 \pm .16$	$6.08 \pm .10$	$\mathit{7.54} \pm \mathit{.21}$	$7.74 \pm .27$
AltAdam5	$\mathit{5.45} \pm \mathit{.08}$	$5.72 \pm .06$	$5.49 \pm .05$	$\mathit{7.20} \pm \mathit{.06}$	$7.67 \pm .15$
ExtraAdam	$\mathbf{6.38} \pm \mathbf{.09}$	$\mathbf{6.38} \pm \mathbf{.20}$	$\mathbf{6.37} \pm \mathbf{.08}$	$7.79 \pm .09$	$\mathbf{8.26} \pm \mathbf{.12}$

Table 1: Best inception scores (averaged over 5 runs) achieved on CIFAR10 for every considered Adam variant. EMA denotes *exponential moving average* (with $\beta = 0.999$, see Eq. 6). We see that the techniques of extrapolation and averaging consistently enable improvements over the baselines (in italic).

The extrapolation step reduces the oscillations of the game between the two players compared to simple averaging. A theoretical consequence is that since in practice $\sigma \ll M$, the variance term in (10) is significantly smaller than the one in (9). As discussed previously, Assump. 2 made in Thm. 1 is very strong in the unbounded setting. One advantage of SGD with *averaging* and *extrapolation* is that Thm. 2 does not require this assumption.

These techniques can be combined in practice with existing algorithms. We propose to combine them to two standard algorithms used for training deep neural networks: the Adam optimizer [Kingma and Ba, 2015] and the SGD optimizer [Robbins and Monro, 1951].

5 Experiments

Our goal in this experimental section is not to provide new state-of-the art results with architectural improvements or a new GAN formulation but to show that using the *techniques* (with theoretical guarantees in the monotone case) that we introduced earlier allow us to optimize standard GANs in a better way. We consider the following Adam variants of the different algorithms as it is known to work best in the context of GANs: we consider simultaneous updates on the generator and on the discriminator (**SimAdam**), and k updates on the discriminator alternated with 1 update on the generator (**AltAdam**$\{k\}$)[6]. The variants that use *extrapolation* with Adam is referred to as **ExtraAdam**.

We compare the algorithms on the CIFAR10 dataset [Krizhevsky and Hinton, 2009] for training: a DCGAN architecture [Radford et al., 2016] with a WGAN objective and weight clipping [Arjovsky et al., 2017] (constrained), and a ResNet architecture with the WGAN-GP objective [Gulrajani et al., 2017] (non-zero sum game). Models are evaluated using the inception score [Salimans et al., 2016]. For each algorithm we did an extensive search over the hyperparameters of Adam ($\beta_1 = 0.5$ and $\beta_2 = 0.9$ performed best for all). As proposed in Heusel et al. [2017] we use different learning rates for the generator and discriminator, which proved really important for training the ResNet architecture. We ran each with 5 different random seeds for 500,000 generator updates. Table 1 reports the best inception score achieved on this problem by each considered method. We see that the techniques of *extrapolation* and *averaging* consistently enable improvements over the baselines. On both tasks using an *extrapolation step* and averaging with Adam (ExtraAdam) outperformed all other methods.

6 Conclusion

We newly addressed GAN objectives in the framework of variational inequality. We tapped into the optimization literature to provide more principled sound techniques to optimize such games. We leveraged these techniques to develop practical optimization algorithms suitable for a wide range of GAN training objectives (including non-zero sum games and projections onto constraints). We experimentally verified that this could yield better trained models, achieving to our knowledge the best inception score when optimizing a WGAN objective on the reference unmodified DCGAN architecture [Radford et al., 2016]. The presented techniques address a fundamental problem in GAN training in a principled way, and are orthogonal to the design of new GAN architectures and objectives. They are thus likely to be widely applicable, and benefit future development of GANs.

[6] In the original WGAN [Arjovsky et al., 2017] paper the authors use $k = 5$.

References

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *ICLR*, 2018.

W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and I. Goodfellow. Many paths to equilibrium: GANs do not need to decrease a divergence at every step. In *ICLR*, 2018.

I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv:1701.00160*, 2016.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

P. Grnarova, K. Y. Levy, A. Lucchi, T. Hofmann, and A. Krause. An online learning approach to generative adversarial networks. In *ICLR*, 2018.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein GANs. In *NIPS*, 2017.

P. T. Harker and J.-S. Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 1990.

E. Hazan, K. Singh, and C. Zhang. Efficient regret minimization in non-convex games. In *ICML*, 2017.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NIPS*, 2017.

A. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 2017.

A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 2011.

T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

G. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12, 1976.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, Canada, 2009.

Y. Li, A. Schwing, K.-C. Wang, and R. Zemel. Dualing GANs. In *NIPS*, 2017.

P. Mertikopoulos, H. Zenati, B. Lecouat, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv*, 2018.

L. Mescheder, S. Nowozin, and A. Geiger. The numerics of GANs. In *NIPS*, 2017.

A. 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 J. Optim.*, 2004.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 2009.

Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Math. Program.*, 2007.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *NIPS*, 2016.

B. Palaniappan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In *NIPS*, 2016.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *NIPS*, 2016.

J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior.* Princeton University Press, 1944.

A. Yadav, S. Shah, Z. Xu, D. Jacobs, and T. Goldstein. Stabilizing adversarial nets with prediction methods. In *ICLR*, 2018.

Y. Yazıcı, C.-S. Foo, S. Winkler, K.-H. Yap, G. Piliouras, and V. Chandrasekhar. The unusual effectiveness of averaging in gan training. *arXiv preprint arXiv:1806.04498*, 2018.

F. Yousefian, A. Nedić, and U. V. Shanbhag. Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems. In *CDC*. IEEE, 2014.