A Variational Inequality Perspective on Generative Adversarial Networks

Gauthier Gidel[*,1], Hugo Berard[*,1,2], Gaëtan Vignoud[1], Pascal Vincent[1,2,3] and Simon Lacoste-Julien[1,3]

[*]Equal contribution; [1]Mila, U. Montréal ; [2]Facebook AI Research ; [3] CIFAR fellow

Overview

TL;DR

- We survey the *"variational inequality"* framework.
- Encompasses all GAN training methods using gradients.
- Tapping into the mathematical programming literature, we counter some common misconceptions about the difficulties of saddle point optimization.

Contributions & Related Work

Contibutions:

- Extend standard methods designed for variational inequalities to the training of GANs.
- Amongst others, we apply *extrapolation* and *averaging* to the stochastic gradient method (SGD) and Adam, to improve the training of GANs.
- We propose *extrapolation from the past* a cheaper variant of extrapolation.

Related work:

- Extragradient methods have been originally introduced by Korpelevich [5] and extended by Nesterov [7] and Nemirovski [6].
- Recently, Daskalakis et al. [2] proposed a method inspired from game theory related to extrapolation.
- Alternative to extragradient: negative momentum proposed by Gidel et al. [3].

Background

Two-player games Equilibrium

Two-player games *Generalizes* mini-max formulation:

$$\theta^* \in \arg\min_{\theta\in\Theta} \mathcal{L}^G(\theta,\varphi^*)\,, \quad \varphi^* \in \arg\min_{\varphi\in\Phi} \mathcal{L}^D(\theta^*,\varphi)$$

- $\mathcal{L}^G = -\mathcal{L}^D$: *zero-sum game.*
- Otherwise: non zero-sum games.

Examples. *Non-saturating GAN* [4], (*not* zero-sum):

$$\mathcal{L}^G(\theta,\varphi) := -\mathop{\mathbb{E}}_{x'\sim q_\theta} \log f_\varphi(x')$$
$$\mathcal{L}^D(\theta,\varphi) := -\mathop{\mathbb{E}}_{x\sim p} \log f_\varphi(x) - \mathop{\mathbb{E}}_{x'\sim q_\theta} \log(1-f_\varphi(x'))\,.$$

WGAN [1] (zero-sum):

$$\min_{\theta\in\Theta} \max_{\phi\in\Phi, ||f_\phi||_L\leq 1} \mathbb{E}_{x\sim p}[f_\phi(x)] - \mathbb{E}_{x'\sim q_\theta}[f_\phi(x')]. \quad (1)$$

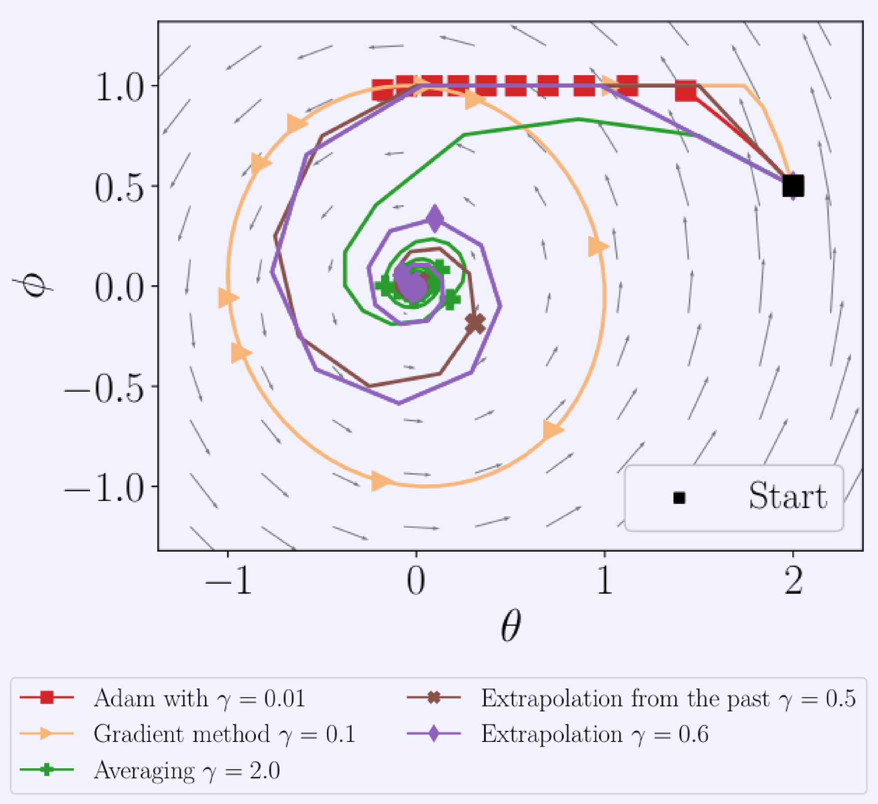
The bilinear WGAN

The discriminator and the generator are linear:

$$D_\varphi(x) = f_\varphi(x) = \varphi^T x\,, \quad G_\theta(z) = \theta z$$

By replacing these expressions in the WGAN objective (1),

$$\min_{\theta\in\Theta} \max_{\varphi\in\Phi, ||\varphi||\leq 1} \varphi^T \mathbb{E}[X] - \varphi^T\theta\mathbb{E}[Z]\,.$$



Problem: $\min_\theta \max_\varphi \ \theta\cdot\varphi$. We have for $N_t := \theta_t^2 + \phi_t^2$,

Simultaneous: $N_{t+1}^2 = (1+\eta^2)N_t^2$ $(Diverges)$,
Alternating: $N_t^2 = \Theta(N_0^2)$ $(Bounded)$,
Extrapolation: $N_{t+1}^2 = (1-\eta^2+\eta^4)N_t^2$ $(Converges)$.

GANs as a Variational inequality

Stationary conditions

Unconstrained: point with zero gradient.

$$\|\nabla_\theta \mathcal{L}^G(\theta^*,\varphi^*)\| = \|\nabla_\varphi \mathcal{L}^D(\theta^*,\varphi^*)\| = 0\,.$$

Constrained: no *feasible* descent directions.

$$\nabla_\theta \mathcal{L}^G(\theta^*,\varphi^*)^\top(\theta-\theta^*) \geq 0\,, \quad \forall\theta\in\Theta$$
$$\nabla_\varphi \mathcal{L}^D(\theta^*,\varphi^*)^\top(\varphi-\varphi^*) \geq 0\,, \quad \forall\varphi\in\Phi.$$

Variational inequality problem (VIP)

Defining $\omega \overset{\text{def}}{=} (\theta,\varphi)$, $\omega^* \overset{\text{def}}{=} (\theta^*,\varphi^*)$, $\Omega \overset{\text{def}}{=} \Theta\times\Phi$, can be compactly formulated as:

$$F(\omega^*)^\top(\omega-\omega^*) \geq 0\,, \quad \forall\omega\in\Omega$$

where $F(\omega) \overset{\text{def}}{=} \left[\nabla_\theta \mathcal{L}^G(\theta,\varphi)\ \nabla_\varphi \mathcal{L}^D(\theta,\varphi)\right]^\top$.

Takeaway

- GAN can be formulated as a Variational Inequality.
- Encompasses most of GANs formulations.
- Standard algorithms from Variational Inequality can be applied to GANs.
- Theoretical Guarantees (for convex and stochastic cost functions).

Algorithms for VIP

Averaging

$$\texttt{return:} \quad \bar\omega_T = \frac{\sum_{t=0}^{T-1}\rho_t\omega_t}{\sum_{t=0}^{T-1}\rho_t}$$

- Converges even for "cycling behavior".
- Easy to implement. • Can combine with any method.
- Can be implemented on an online fashion:

Uniform averaging: $\bar\omega_{T+1} = \frac{T}{T+1}\bar\omega_T + \frac{1}{T+1}\omega_T$

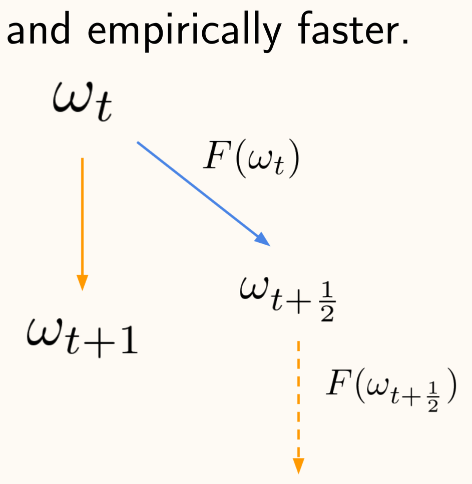
EMA: $\bar\omega_{T+1} = \beta\bar\omega_T + (1-\beta)\omega_T$

Extragradient

$$\begin{cases} \omega_{t+\frac{1}{2}} = \omega_t - \gamma_t F(\omega_t) & \text{(extrapolation step)} \\ \omega_{t+1} = \omega_t - \gamma_t F(\omega_{t+\frac{1}{2}}) & \text{(update step)} \end{cases}$$

Intuition: Look 1 step in the future and anticipate next move of adversary. Close to an *implicit* method.

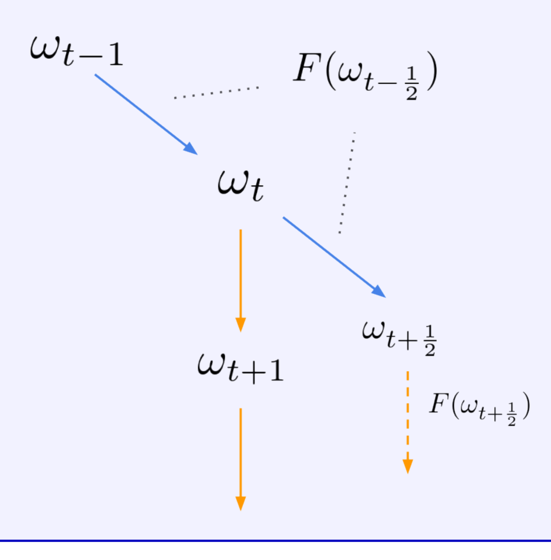
- Does not require averaging.
- Theoretically and empirically faster.



Extrapolation from the past

Problem: Extragradient requires to compute two gradients at each step. (Twice as expensive !)
Solution: Re-use the previous gradient.

$$\begin{cases} \omega_{t+\frac{1}{2}} = \omega_t - \gamma_t F(\omega_{t-\frac{1}{2}}) & \text{(re-use from previous step)} \\ \omega_{t+1} = \omega_t - \gamma_t F(\omega_{t+\frac{1}{2}}) & \text{(same as extragradient)} \end{cases}$$



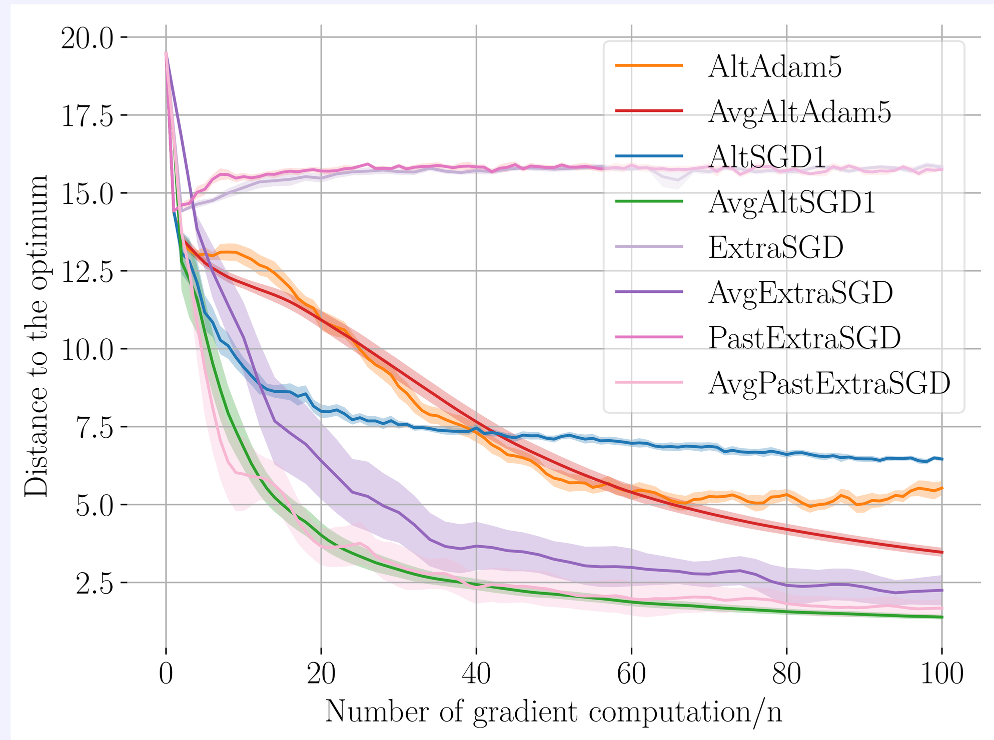
Experiments

Algorithms

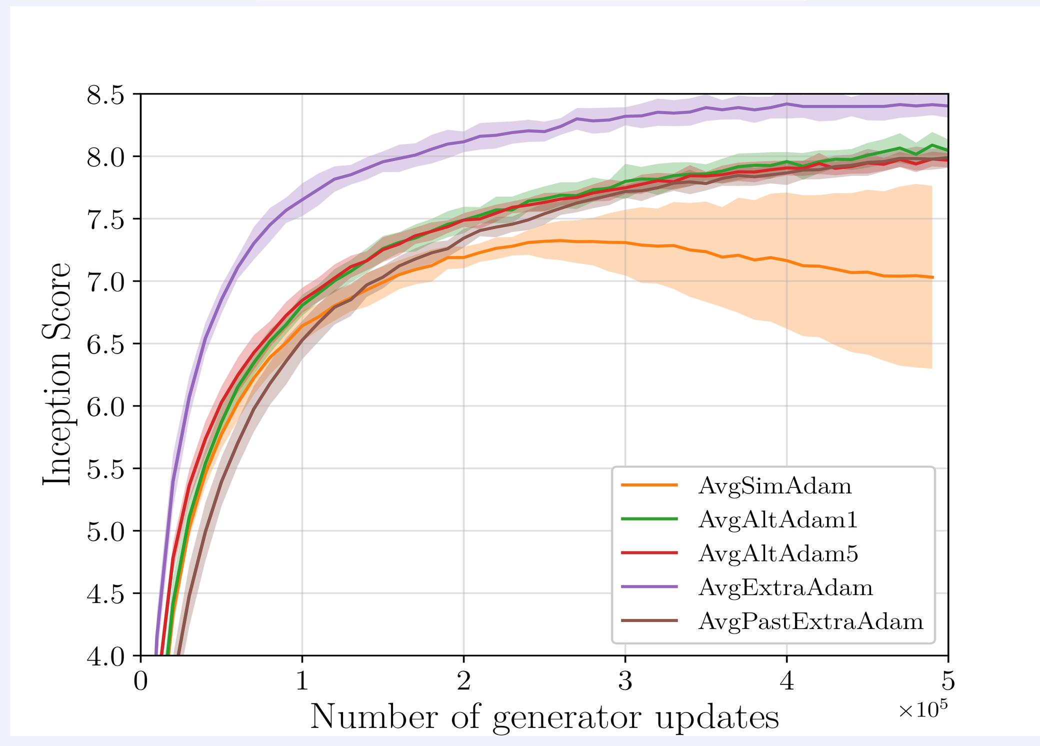
SimSGD: Both parameters are updated simultaneously.
AltSGD: variant of SGD where ϕ is updated before θ.
AvgSGD: return the average of SimSGD iterates.
ExtraSGD: SGD with an extrapolation step.
PastExtraSGD: SGD with extrapolation from the past.

Simple stochastic bilinear example

$$\frac{1}{n}\sum_{i=1}^{n}\left(x^\top M^{(i)} y + x^\top a^{(i)} + y^\top b^{(i)}\right)$$



WGAN-GP on CIFAR-10



Inception Score

Model	WGAN-GP (ResNet)		
Method	no avg	uniform avg	EMA
SimAdam	*7.51 ± .17*	7.68 ± .43	7.60 ± .17
AltAdam5	*7.57 ± .02*	8.01 ± .05	7.66 ± .03
ExtraAdam	7.90 ± .11	**8.47 ± .10**	8.13 ± .07
PastExtraAdam	7.84 ± .06	8.01 ± .09	7.99 ± .03
OptimAdam	*7.98 ± .08*	8.18 ± .09	8.10 ± .06

Frechet Inception Distance

Model	WGAN-GP (ResNet)		
Method	no averaging	uniform avg	EMA
SimAdam	*23.74 ± 2.79*	26.29 ± 5.56	21.89 ± 2.51
AltAdam5	*21.65 ± .66*	19.91 ± .43	20.69 ± .37
ExtraAdam	19.42 ± .15	18.13 ± .51	**16.78 ± .21**
PastEAdam	19.95 ± .38	22.45 ± .93	17.85 ± .40
OptimAdam	*18.88 ± .55*	21.23 ± 1.19	16.91 ± .32

References

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

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

[3] G. Gidel, R. A. Hemmat, M. Pezeshki, G. Huang, R. Lepriol, S. Lacoste-Julien, and I. Mitliagkas. Negative momentum for improved game dynamics. *arXiv preprint arXiv:1807.04740*, 2018.

[4] 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.

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

[6] 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.

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