Reducing Noise in GAN Training with Variance Reduced Extragradient

Tatjana Chavdarova[*,1,2], **Gauthier Gidel**[*,1,3], **François Fleuret**[2] and **Simon Lacoste-Julien**[1,3]

[*]Equal contribution; [1]Mila, Université de Montréal; [2]EPFL, Idiap; [3] Element AI

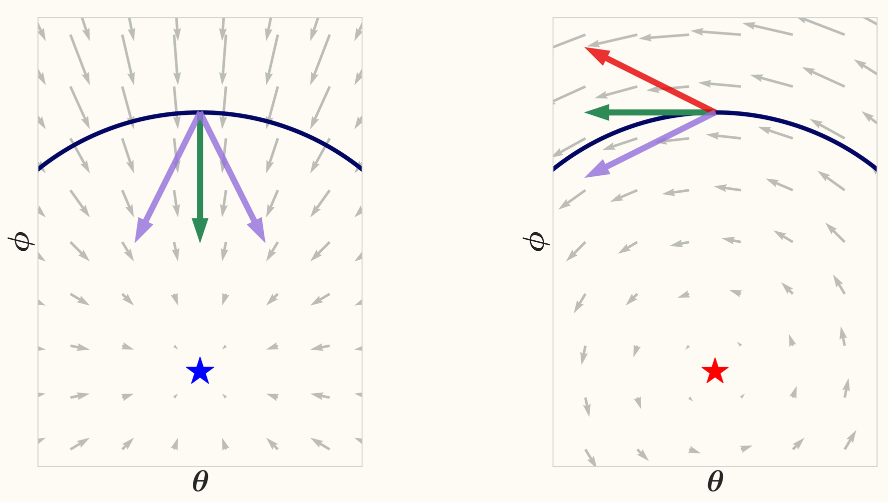
Overview

Takeaways

- Games harder to optimize → **Extragradient** damps the oscillations of the game.
- Not having the full gradient (e.g. stochastic gradient) breaks Extragradient.
- We propose "**SVRE**" that uses **variance reduction** to fix stochastic Extragradient
- Theoretically the fastest method (under some standard assumptions).
- Empirically much more stable than the baseline and yields improvements late in learning.

Motivation: VR for game optimization

- BigGAN: 8–fold increased batch size yields 46% relative improvement of Inception Score on ImageNet
- In practice (often) only the variance component of *Adam* is used: $\beta_1 = 0$ (tuned)
- Intuition on why noisy game vector field is more problematic than noisy minimization:



Minimization *Vs.* **Game**

Background

Two-player games Equilibrium

Generalizes mini-max formulation:

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

Stationary conditions

Objective: point with zero gradient.

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

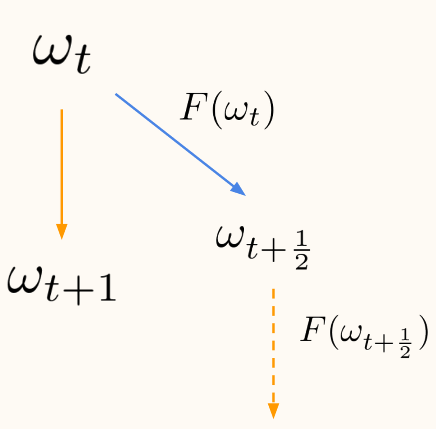
$\boldsymbol{\omega} \overset{\text{def}}{=} (\boldsymbol{\theta}, \boldsymbol{\varphi})$, $\boldsymbol{\omega}^* \overset{\text{def}}{=} (\boldsymbol{\theta}^*, \boldsymbol{\varphi}^*)$, $\Omega \overset{\text{def}}{=} \Theta \times \Phi$,
Can be reformulated as $F(\boldsymbol{\omega}^*) = 0$ where,

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

Extragradient

$$\begin{aligned} \boldsymbol{\omega}_{t+\frac{1}{2}} &= \boldsymbol{\omega}_t - \gamma_t F(\boldsymbol{\omega}_t) && \text{(extrapolation)} \\ \boldsymbol{\omega}_{t+1} &= \boldsymbol{\omega}_t - \gamma_t F(\boldsymbol{\omega}_{t+\frac{1}{2}}) && \text{(update)} \end{aligned}$$

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



Reducing Noise in GANs

Methods for solving bilinear games

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d}\max_{\boldsymbol{\varphi}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\theta}^\top \boldsymbol{A}_i \boldsymbol{\varphi}$$

Method	Gradient method	Extragradient
Batch	$\|\boldsymbol{\omega}_t - \boldsymbol{\omega}^*\| \to \infty$	$\|\boldsymbol{\omega}_t - \boldsymbol{\omega}^*\| \to 0$
Stochastic	No hope for convergence	$\|\boldsymbol{\omega}_t - \boldsymbol{\omega}^*\| \to \infty$

Stochasticity breaks extragradient !

Stochastic variance reduced gradient

Finite sum assumption:

$$\mathcal{L}^G(\boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i^G(\boldsymbol{\omega}), \quad \mathcal{L}^D(\boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i^D(\boldsymbol{\omega})$$

Unbiased estimates of the gradient are:

$$\begin{aligned} \boldsymbol{d}_i^G(\boldsymbol{\omega}) &:= \nabla\mathcal{L}_i^G(\boldsymbol{\omega}) - \nabla\mathcal{L}_i^G(\boldsymbol{\omega}^{\mathcal{S}}) + \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathcal{S}} \\ \boldsymbol{d}_i^D(\boldsymbol{\omega}) &:= \nabla\mathcal{L}_i^D(\boldsymbol{\omega}) - \nabla\mathcal{L}_i^D(\boldsymbol{\omega}^{\mathcal{S}}) + \boldsymbol{\mu}_{\boldsymbol{\varphi}}^{\mathcal{S}}. \end{aligned}$$

- $\boldsymbol{\mu}$: Full-batch gradient at the snapshot $\boldsymbol{\omega}^{\mathcal{S}}$.
- Index i sampled uniformly over $\{1, \dots, n\}$.
- $\mathbb{E}[\boldsymbol{d}_i^G(\boldsymbol{\omega})] = \frac{1}{n}\sum_{i=1}^{n}\nabla\mathcal{L}_i^G(\boldsymbol{\omega}) = \nabla\mathcal{L}^G(\boldsymbol{\omega})$.
- If $\boldsymbol{\omega}^{\mathcal{S}}$ is close to $\boldsymbol{\omega}$ → small variance.

SVRE: Stochastic Variance Reduced Extragradient

SVRE combines **Extragradient**:

$$\begin{aligned} \boldsymbol{\omega}_{t+\frac{1}{2}} &= \boldsymbol{\omega}_t - \gamma_t F_i(\boldsymbol{\omega}_t) && \text{(extrapolation)} \\ \boldsymbol{\omega}_{t+1} &= \boldsymbol{\omega}_t - \gamma_t F_i(\boldsymbol{\omega}_{t+\frac{1}{2}}) && \text{(update)} \end{aligned}$$

with **Variance Reduction**:

$$F_i(\boldsymbol{\omega}) := \begin{pmatrix} \nabla\mathcal{L}_i^G(\boldsymbol{\omega}) - \nabla\mathcal{L}_i^G(\boldsymbol{\omega}^{\mathcal{S}}) + \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\mathcal{S}} \\ \nabla\mathcal{L}_i^D(\boldsymbol{\omega}) - \nabla\mathcal{L}_i^D(\boldsymbol{\omega}^{\mathcal{S}}) + \boldsymbol{\mu}_{\boldsymbol{\varphi}}^{\mathcal{S}}. \end{pmatrix}$$

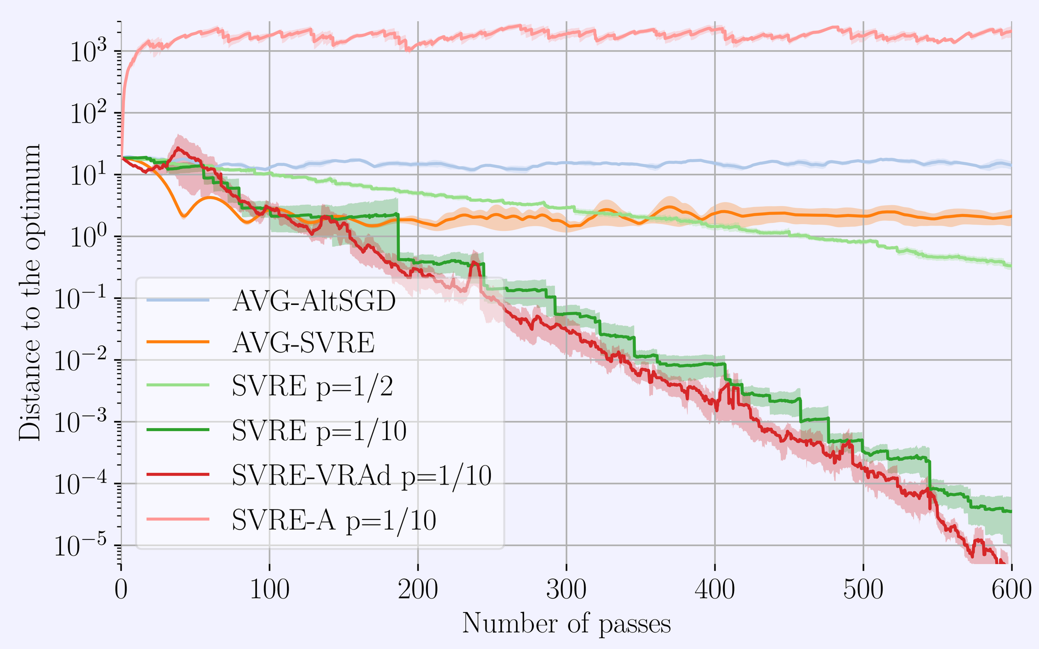
Recall $\boldsymbol{\omega} := (\boldsymbol{\theta}, \boldsymbol{\varphi})$ is the *joint* parameter.

Comparison of variance reduced methods for games

Method	Complexity	μ-adaptivity
SVRG	$\ln(1/\epsilon)\times(n + \frac{\bar{L}^2}{\mu^2})$	✗
A. SVRG	$\ln(1/\epsilon)\times(n + \sqrt{n}\frac{\bar{L}}{\mu})$	✗
SVRE	$\ln(1/\epsilon)\times(n + \frac{\bar{L}}{\mu})$	✓

- μ strong monotonicity.
- $\bar{L}$ average Lipschitz constant of the gradient.

Stochastic bilinear game



GAN Experiments

Optimization methods

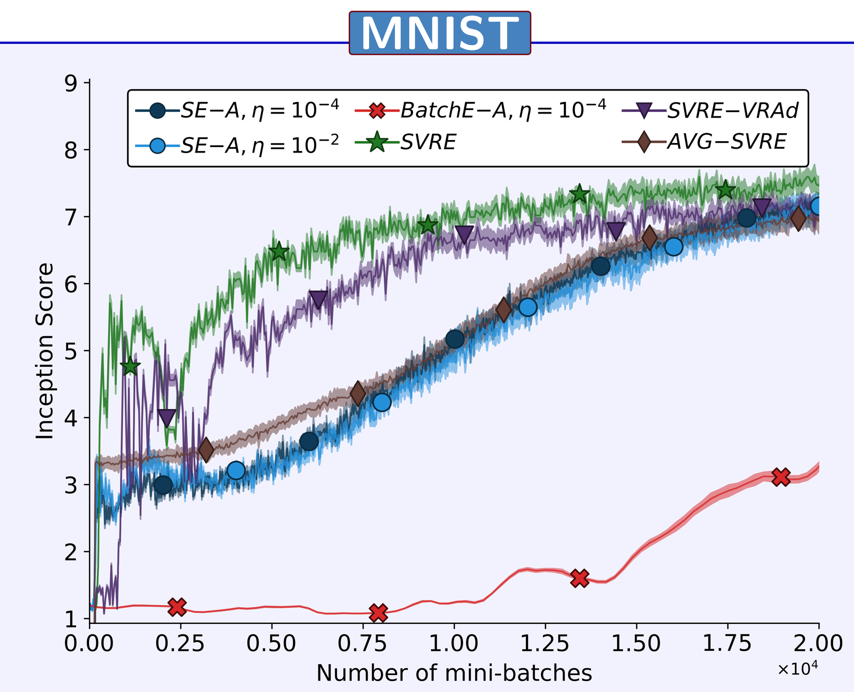
BatchE: full–batch extragradient
SG: stochastic gradient (alternating GAN)
SE: stochastic extragradient
WS–SVRE: warm-start SVRE
A: *Adam* (adaptive step size method)

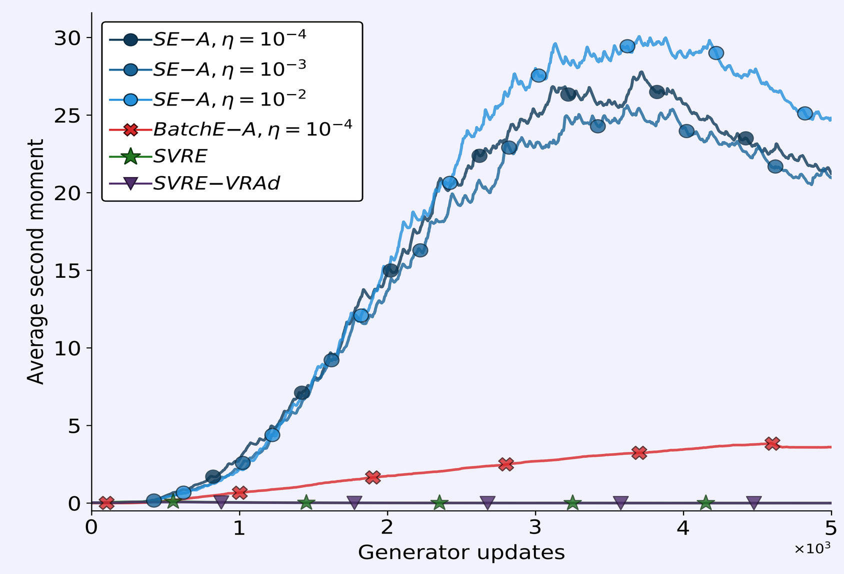
Summary

	SG–A	SE–A	SVRE	WS–SVRE
CIFAR-10	21.70	18.65	23.56	**16.77**
SVHN	5.66	5.14	**4.81**	4.88

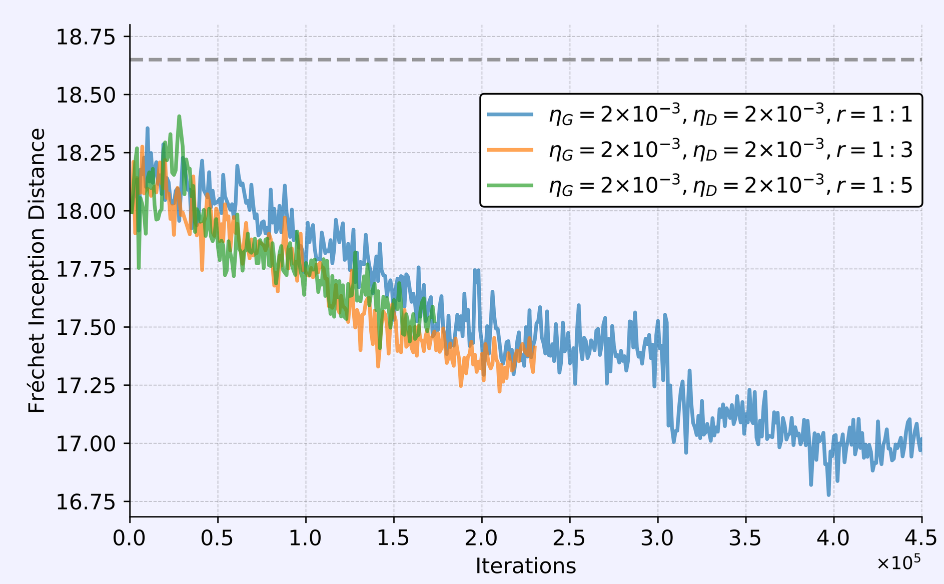
FID (lower is better)

MNIST





WS–SVRE on CIFAR10



SVRE is more robust and stable on SVHN