Московский физико-технический институт
(Государственный университет)

Физтех-школа Прикладной Математики и Информатики (ФПМИ)
Кафедра интеллектуальных систем

Ковалев Дмитрий Александрович

О стохастическом экстраградиентном методе для вариационных неравенств

Магистерская диссертация

Направление подготовки 03.04.01 «Прикладные математика и физика»
Магистерская программа «Математическая физика, компьютерные технологии и математическое моделирование в экономике»

Студент:
Ковалев Дмитрий Александрович
группа М05-904б

Научный руководитель:
Гасников Александр Владимирович
д-р физ.-мат. наук, доц

Москва, 2021

Аннотация

В данной работе исправлена фундаментальная проблема стохастического экстраградиентного метода с помощью новой стратегии семплирования, мотивированной аппроксимацией неявного градиентного метода. Так как существующий стохастический экстраградиентный метод Mirror-Prox [Juditsky et al., 2011] расходится на простой билинейной задаче, когда область определения неограничена, в данной работе доказываются гарантии сходимости нового метода для более общих постановок, чем в существующих результатах. Численные эксперименты в данной работе показывают, что предложенный вариант экстраградиентного метода сходится на билинейных седловых задачах быстрее, чем многие другие методы. Также в работе рассматривается применение экстраградиентного метода для обучения генеративно-состязательных нейронных сетей и показывается с помощью численных экспериментов, что предложенный подход имеет преимущество по количеству проходов по обучающей выборке, в то время как более высокая стоимость итераций метода уменьшает это преимущество.

Данная работа основана на статье «Revisiting Stochastic Extragradient» [Mishchenko et al., 2020], написанной в соавторстве с Константином Мищенко, Егором Шульгиным, Питером Рихтариком и Юрой Малицким.

Abstract

We fix a fundamental issue in the stochastic extragradient method by providing a new sampling strategy that is motivated by approximating implicit updates. Since the existing stochastic extragradient algorithm, called Mirror-Prox, of [Juditsky et al., 2011] diverges on a simple bilinear problem when the domain is not bounded, we prove guarantees for solving variational inequality that go beyond existing settings. Furthermore, we illustrate numerically that the proposed variant converges faster than many other methods on bilinear saddle-point problems. We also discuss how extragradient can be applied to training Generative Adversarial Networks (GANs) and how it compares to other methods. Our experiments on GANs demonstrate that the introduced approach may make the training faster in terms of data passes, while its higher iteration complexity makes the advantage smaller.

This work is based on a paper «Revisiting Stochastic Extragradient» [Mishchenko et al., 2020] written in collaboration with Konstantin Mishchenko, Egor Shulgin, Peter Richtárik, and Yura Malitsky.

Contents

1 Introduction

Algorithmic machine learning has for a long time been centered around minimization of a single function. A lot of works are still targeting solving empirical risk minimization and new results touch upon methods as old as gradient descent itself.

However, as the gap between lower bounds and available minimization algorithms is shrinking, the focus is shifting towards more challenging problems such as variational inequality, where a significant number of long-unresolved questions is remaining. This problem has a rich history with applications in economics and computer science, but the arising applications provide new desiderata on algorithm properties. In particular, due to high dimensionality and large scale of the corresponding problems, we shall consider the impact of having a *stochastic* objective. In particular, recently invented generative adversarial neural networks [Goodfellow et al., 2014] are often trained using schemes that resemble primal-dual and variational inequality methods, which we shall discuss in detail later.

Variational inequality can be seen as an extension of the necessary first-order optimality condition for minimization problem, which is also sufficient in the convex case. When the operator involved in its formulation is monotone and is equal to the gradient of a function, this corresponds to convex minimization.

Formally, the problem that we consider is that of finding a point x^* satisfying

$$g(x) - g(x^*) + \langle F(x^*), x - x^* \rangle \geq 0, \text{ for all } x \in \mathbb{R}^d, \tag{1}$$

where $g\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper lower semi-continuous convex function and $F\colon \mathbb{R}^d \to \mathbb{R}^d$ is a monotone operator. Some application of interest are not covered by the monotonicity framework, but, unfortunately, little is known about variational inequality and even saddle point problems when monotonicity is missing. Thus, we stick to this assumption and rather try to model oscillations arising in some problems by considering particularly unstable [Gidel et al., 2019b, Chavdarova et al., 2019] bilinear minimax problems.

Of particular interest to us is the situation where $F(x)$ is the expectation with respect to random variable ξ of the random operator $F(x;\xi)$. This formulation has two aspects. First, one can model data distribution, especially when a large dataset is available and the problem is that of minimizing empirical loss. Second, ξ can be a random variable sampled by one of the GAN networks, called *generator*. In any case, throughout the work we assume that we sample unbiased estimates $F(\cdot;\xi)$ of $F(\cdot)$ such that $\mathbb{E}_\xi F(\cdot;\xi) = F(\cdot)$.

Let us explicitly mention that a special case of (1) is constrained saddle point optimization,

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y),$$

where $\mathcal{X}$ and $\mathcal{Y}$ are some convex sets and f is a smooth function. While this example looks deceptively simple, simultaneous gradient descent-ascent is known to diverge on this problem [Goodfellow, 2016] even when f is convex-concave. In particular, the objective $f(x,y) = x^\top y$ leads to geometrical divergence for any nontrivial initialization [Daskalakis et al., 2018]. See [Mishchenko and Richtárik, 2019] for more applications of the convex-concave saddle point problem in machine learning and [Gidel et al., 2019a] for extra discussion on variational inequality and its relation to GANs.

1.1 Related work

The extragradient method was first proposed by [Korpelevich, 1977]. Since then there have been developed a number of its extensions, most famous of which is the Mirror-Prox method [Nemirovski, 2004] that uses mirror descent update. At each iteration, the standard extragradient

method is trying to approximate the implicit update, which is known to be much more stable. Assuming the operator is Lipschitz, it is enough to compute the operator twice to do the approximation accurately enough. We base our intuition upon this property and we shall discuss it in detail later in the paper.

While extragradient uses future information, i.e., information from one gradient step ahead, past information can also help to stabilize convergence. In particular, *Optimistic mirror descent* (OMD), first proposed by [Rakhlin and Sridharan, 2013] for convex-concave zero-sum games, has been analyzed in a number of works [Mokhtari et al., 2019, Daskalakis and Panageas, 2019, Gidel et al., 2019a] and it was applied to GAN training in [Daskalakis et al., 2018]. The rates that we prove in this work for stochastic extragradient match the best known results for OMD, but are given under more general assumptions. Moreover, the method of [Gidel et al., 2019a] diverges on bilinear problems.

Many other techniques also allow to improve stability and achieve convergence for monotone operators in the particular case of saddle point problems. For instance, *alternating* gradient descent-ascent does not, in general, converge to a solution [Gidel et al., 2019b], the negative momentum trick proposed in [Gidel et al., 2019b] can fix this.

We note that our work is not the first to consider a variant of stochastic extragradient. A stochastic version of the Mirror-Prox method [Nemirovski, 2004] was analyzed in [Juditsky et al., 2011] under pretty restrictive assumptions. While deterministic extragradient approximates implicit update, the authors of [Juditsky et al., 2011] chose to sample two different instances of the stochastic operator, which leads to a poor approximation of stochastic implicit update unless the variance is tiny. It was observed in [Chavdarova et al., 2019] that this approach leads to terrible practical performance, dubious convergence guarantees and divergence on bilinear problems. All later variants of stochastic extragradient, that we are aware of, consider the same update model.

Surprisingly, a variant of extragradient was also rediscovered by practitioners [Metz et al., 2016] as a way to stabilize training of GANs. The main difference of the method of [Metz et al., 2016] to what we consider is in applying extra steps only on one of two neural networks. In addition, [Metz et al., 2016] proposed to use more than one extra step and claim that in on specific problems 5 steps is a good trade-off between results quality and computation.

[Chavdarova et al., 2019] showed that the methods of [Juditsky et al., 2011] and [Gidel et al., 2019a] diverge on stochastic bilinear saddle point problem. As a fix, they proposed a stochastic extragradient method with variance reduction (SVRE), which achieves a linear rate $\mathcal{O}((n + \frac{L}{\mu}) \log \frac{1}{\varepsilon})$. However, their theory works only for saddle point problems and it does not cover the case without strong monotonicity, so it is less general than ours.

1.2 Theoretical background

Here we provide several technical assumptions that are standard for variational inequality.

Assumption 1. *Operator $F\colon \mathbb{R}^d \to \mathbb{R}^d$ is monotone, that is $\langle F(x) - F(y), x - y\rangle \geq 0$ for all $x, y \in \mathbb{R}^d$. In stochastic case, we assume that $F(x;\xi)$ is monotone almost surely.*

The monotonicity assumption is an extension of the notion of convexity and is quite standard in the literature. There are several versions of pseudo-monotonicity, but without it the variational inequality problem becomes extremely hard to solve.

Assumption 2. *Operator $F(\cdot;\xi)$ is almost-surely L-Lipschitz, that is for all $x, y \in \mathbb{R}^d$*

$$\|F(x;\xi) - F(y;\xi)\| \leq L \|x - y\| . \tag{2}$$

In addition to operator monotonicity, we ask for convexity and some regularity properties of $g(\cdot)$ as given below.

Algorithm 1 Same-Sample Stochastic Extragradient Method for Variational Inequality.

Parameters: $x^0 \in \mathcal{K}$, stepsize $\eta > 0$
for $t = 0, 1, 2, \ldots$ **do**
  Sample ξ^t
  $y^t = \text{prox}_{\eta g}\left(x^t - \eta F(x^t; \xi^t)\right)$
  $x^{t+1} = \text{prox}_{\eta g}\left(x^t - \eta F(y^t; \xi^t)\right)$
end for

Algorithm 2 The extragradient method for min-max problems.

Require: Stepsizes η_1, η_2, initial vectors x^0, y^0
for $t = 0, 1, \ldots$ **do**
  $u^t = x^t - \eta_1 \nabla_x f(x^t, y^t)$
  $v^t = y^t + \eta_1 \nabla_y f(x^t, y^t)$
  $x^{t+1} = x^t - \eta_2 \nabla_x f(u^t, v^t)$
  $y^{t+1} = y^t + \eta_2 \nabla_y f(u^t, v^t)$
end for

Assumption 3. *Function $g\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous and μ-strongly convex for $\mu \geq 0$, i.e., for all $x, y \in \mathbb{R}^d$ and any $h \in \partial g(y)$*

$$g(x) - g(y) - \langle h, x - y\rangle \geq \frac{\mu}{2}\|x - y\|^2.$$

If $\mu = 0$, then g is just convex.

Even in simple minimization problems, the classical theoretical analysis of stochastic methods ask for uniformly bounded variance, an assumption rarely satisfied in practice. Recent developments of the theory for SGD have removed this assumption, but we are not aware of any results in more general settings. Thus, it is one of our contributions is to relax the uniform variance bound the one below.

Assumption 4. *In the strongly convex case, we assume that F has bounded variance at the optimum, i.e.,*

$$\mathbb{E}\|F(x^*; \xi) - F(x^*)\|^2 \leq \sigma^2.$$

Depending on the assumptions, we will either work with the variance at the optimum or with a merit function, which involves the variance of a bounded set.

2 Theory

It is known that implicit updates are more stable when solving variational inequality and sometimes it is argued that the main goal of algorithmic design is to approximate those [Mokhtari et al., 2019]. From that perspective, the current stochastic extragradient, which was suggested in [Juditsky et al., 2011], does not make much sense. Since it uses two independent samples, it will rarely approximate the implicit update, so it is rather not surprisingly that it fails on bilinear problems.

To better explain this phenomenon, below we show that extragradient efficiently approximates implicit update.

Theorem 1. *Let F be an L-Lipschitz operator and define $y \overset{def}{=} \text{prox}_{\eta g}\left(x - \eta F(x)\right)$, $z \overset{def}{=} \text{prox}_{\eta g}\left(x - \eta F(y)\right)$, $w \overset{def}{=} \text{prox}_{\eta g}\left(x - \eta F(w)\right)$, where $\eta > 0$ is any stepsize. Then,*

$$\|w - z\| \leq \eta^2 L^2 \|w - x\|.$$

The right-hand side in Theorem 1 serves as a measure of stationarity and decreases as x gets closer to the problem's solution. The essential part of the bound is that the error is of order $O(\eta^2)$ rather than $O(\eta)$. This allows the approximation to be better than simple gradient update making it possible for the method to solve variational inequality. One can also mention that having extra factor of ηL is beneficial only when $\eta < 1/L$, which provides a good intuition on why extragradient uses smaller stepsizes than gradient.

However, when the stochastic update is used, this result is not applicable directly. If two different samples of the operator are used, $F(\cdot;\xi^t)$ and $F(\cdot;\xi^{t+1/2})$, as is done in stochastic Mirror-Prox [Juditsky et al., 2011], then the update does not seem to approximate implicit update of any operator. This is why we propose in this work to use the same sample, ξ^t, when computing y^t and x^{t+1}, see Algorithm 1. Equipped with our update, we are always approximating the implicit update of stochastic operator $F(\cdot;\xi^t)$ and our theoretical results suggest that this is the right approach.

2.1 Stochastic variational inequality

Our first goal is to show that our stochastic version of the extragradient method converges for strongly monotone variational inequality. The next theorem provides the rate that we obtained.

Theorem 2. *Assume that g is a μ-strongly convex function, operator $F(\cdot;\xi)$ is almost surely monotone and L-Lipschitz, and that its variance at the optimum x^* is bounded by constant, $\mathbb{E}\|F(x^*;\xi) - F(x^*)\|^2 \leq \sigma^2$. Then, for any $\eta \leq 1/(2L)$*

$$\mathbb{E}\|x^t - x^*\|^2 \leq (1 - 2\eta\mu/3)^t \|x^0 - x^*\|^2 + 3\eta\sigma^2/\mu.$$

In the case where at the optimum the noise is zero, we recover a slight generalization of linear convergence of extragradient [Tseng, 1995]. This is also similar to the rate proved for optimistic mirror descent in [Gidel et al., 2019a], however we do not ask for uniform bounds on the variance. Therefore, we believe that this result is significantly more general.

Theorem 3. *Let g be a convex function, $F(\cdot;\xi)$ be monotone and L-Lipschitz almost surely. Then, the iterates of Algorithm 1 with stepsize $\eta = \mathcal{O}(1/(\sqrt{t}L))$ satisfy for any set $\mathcal{X}$ and $x \in \mathcal{X}$*

$$\begin{aligned}&\mathbb{E}\left[g(\hat{x}^t) - g(x) + \left\langle F(x), \hat{x}^t - x\right\rangle\right]\\ &\leq \frac{1}{\sqrt{t}L}\sup_{x\in\mathcal{X}}\left\{\frac{L^2}{2}\|x^0 - x\|^2 + \sigma_x^2\right\}.\end{aligned}$$

where $\hat{x}^t = \frac{1}{t}\sum_{k=0}^t y^k$ and $\sigma_x^2 \stackrel{def}{=} \mathbb{E}\|F(x) - F(x;\xi)\|^2$, i.e., σ_x^2 is the variance of F at point x.

The left-hand side in the bound above is a merit function that has been used in variational inequality literature [Nesterov, 2007]. This result is more general than the one obtained in [Gidel et al., 2019a], where the authors require for the same rate bounded variance and even $\mathbb{E}\|F(x;\xi)\|^2 \leq M < \infty$ uniformly over x.

In fact, the claim that we prove in the appendix is a bit more general than the one presented in the previous theorem. If we know that σ_x is sufficiently small on a bounded set $\mathcal{X}$, then we can get a $\mathcal{O}(1/t + \sup_{x\in\mathcal{X}}\sigma_x)$ rate, i.e., fast convergence to a neighborhood.

2.2 Adversarial bilinear problems

The work [Gidel et al., 2019b] argues that a good illustration of method's stability can be obtained when considering minimax bilinear problems, which is given by

$$\min_x \max_y f(x,y) = x^\top \mathbf{B} y + a^\top x + b^\top y,$$

where $\mathbf{B}$ is a full rank square matrix. One can show that if there exists a Nash equilibrium point, then $f(x,y) = (x-x^*)^\top \mathbf{B}(y-y^*) + \text{const}$ for some pair (x^*, y^*)[1]. This problem is particularly interesting because simple gradient descent-ascent diverges geometrically when solving it,

Theorem 4. *Let f be bilinear with a full-rank matrix $\mathbf{B}$ and apply Algorithm 2 to it. Choose any η_1 and η_2 such that $\eta_2 < 1/\sigma_{\max}(\mathbf{B})$ and $\eta_1\eta_2 < 2/\sigma_{\max}(\mathbf{B})^2$, then the rate is*

$$\|x^t - x^*\|^2 + \|y^t - y^*\|^2 \leq \rho^{2t}(\|x^0 - x^*\|^2 + \|y^0 - y^*\|^2),$$

where $\rho \overset{def}{=} \max\left\{(1-\eta_1\eta_2\sigma_{\max}(\mathbf{B})^2)^2 + \eta_2^2\sigma_{\max}(\mathbf{B})^2,\ (1-\eta_1\eta_2\sigma_{\min}(\mathbf{B})^2)^2 + \eta_2^2\sigma_{\min}(\mathbf{B})^2\right\}$.

The conditions for η_1 and η_2 in Theorem 4 are necessary, but not sufficient. To guarantee convergence, one needs to have $\rho < 1$ and below we provide two such examples.

Corollary 1. *Under the same assumption as in Theorem 4, consider two choices of stepsizes:*

1. *if $\eta_1 = \eta_2 = 1/(\sqrt{2}\sigma_{\max}(\mathbf{B}))$ we get*

$$\begin{aligned}&\|x^t - x^*\|^2 + \|y^t - y^*\|^2\\ &\leq \left(1 - \frac{\sigma_{\min}(\mathbf{B})^2}{6\sigma_{\max}(\mathbf{B})^2}\right)^{2t}(\|x^0 - x^*\|^2 + \|y^0 - y^*\|^2),\end{aligned}$$

2. *if $\sigma_{\min}(\mathbf{B}) > 0$, and $\eta_1 = \kappa/(\sqrt{2}\sigma_{\max}(\mathbf{B})^2)$, $\eta_2 = 1/(\sqrt{2}\kappa\sigma_{\max}(\mathbf{B})^2)$ with $\kappa \overset{def}{=} \sigma^2_{\min(\mathbf{B})}/\sigma^2_{\max(\mathbf{B})}$, then the rate is*

$$\begin{aligned}&\|x^t - x^*\|^2 + \|y^t - y^*\|^2\\ &\leq \left(1 - \frac{\sigma_{\min}(\mathbf{B})^2}{4\sigma_{\max}(\mathbf{B})^2}\right)^{2t}(\|x^0 - x^*\|^2 + \|y^0 - y^*\|^2).\end{aligned}$$

If we denote $\kappa \overset{\text{def}}{=} \frac{\sigma^2_{\min(\mathbf{B})}}{\sigma^2_{\max(\mathbf{B})}}$ as in [Mokhtari et al., 2019], then the complexity in both cases is $O(\kappa \log \frac{1}{\varepsilon})$. However, we provide this result for potentially different stepsizes to obtain new insights about how they should be chosen. One can see, in particular, that choosing a huge η_1 is possible if η_2 is chosen small, but not vice versa.

3 Nonconvex extragradient

Since the objective of neural networks is not convex, it is desirable to have a guarantee for convergence that would not assume operator monotonicity. Alas, there is almost no theory even for nonconvex minimax problems and full gradient updates as even the notion of stationarity becomes tricky. Therefore, in this section we only discuss the method performance when minimizing loss function.

Formally, the problem that we consider here is

$$\min_x \mathbb{E}_\xi f(x;\xi), \tag{3}$$

where f is a smooth bounded from below and potentially nonconvex function. To show convergence, we need the following standard assumption.

[1] If a does not belong to the column space of B or b does not belong to the column space of $B^\top$, the unconstrained minimax problem admits no equilibrium. Otherwise, if we introduce $\tilde{a}, \tilde{b}$ such that $a = -By^*$ and $b = -B^\top x^*$, we have $(x-x^*)^\top B(y-y^*) = x^\top By + a^\top x + b^\top y + (x^*)^\top By^*$.

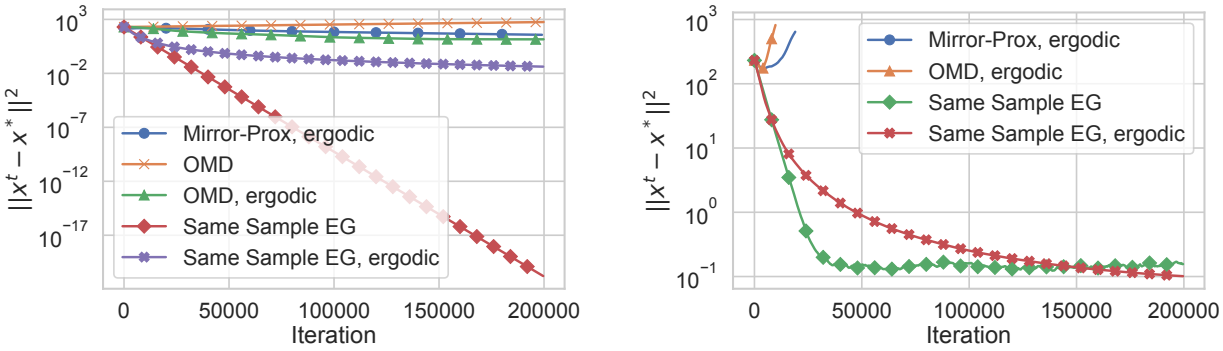


Figure 1: Left: comparison of using independent samples and averaging as suggested by [Juditsky et al., 2011] and the same sample as proposed in this work. The problem here is the sum of randomly sampled matrices $\min_x \max_y \sum_{i=1}^n x^\top \mathbf{B}_i y$. Since at point (x^*, y^*) the noise is equal 0, the convergence of Algorithm 1 is linear unlike the slow rates of [Juditsky et al., 2011] and [Gidel et al., 2019a]. 'EGm' is the version with negative momentum [Gidel et al., 2019b] equal $\beta = -0.3$. Right: bilinear example with linear terms.

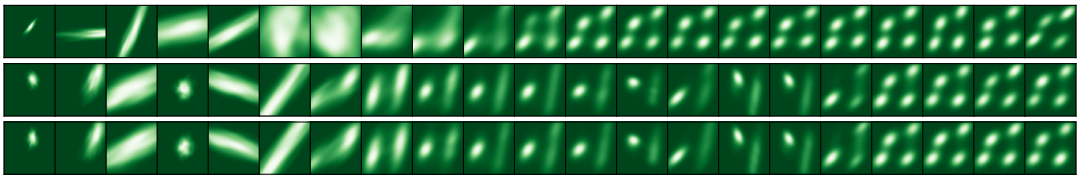


Figure 2: Top line: extragradient with the same sample. Middle line: gradient descent-ascent. Bottom line: extragradient with different samples. Since the same seed was used for all methods, the former two methods performed extremely similarly, although when zooming it should be clear that their results are slightly different.

Assumption 5. *There exists a constant $\sigma > 0$ such that for all x it holds*

$$\mathbb{E}\|\nabla f(x;\xi) - \nabla f(x)\|^2 \leq \sigma^2.$$

Then, we are able to show that the method converges to a local minimum.

Theorem 5. *Choose $\eta \leq \frac{1}{4L}$ and apply extragradient to (3). Then, its iterates satisfy*

$$\mathbb{E}\|\nabla f(\hat{x}^t)\|^2 \leq \frac{5}{\eta t}(f(x^0) - f^*) + 11\eta L\sigma^2,$$

where $\hat{x}^t$ is sampled uniformly from $\{x^0, \ldots, x^{t-1}\}$ and $f^* = \inf_x f(x)$.

Corollary 2. *If we choose $\eta = \Theta\left(1/(L\sqrt{t})\right)$, then the rate is $O\left((f(x^0)-f^*)/\sqrt{t} + \sigma^2/\sqrt{t}\right)$, which is the same as the rate of SGD under our assumptions.*

The statement of the theorem almost coincides with that of SGD, see for instance [Ghadimi and Lan, 2013]. This suggests that extragradient in most cases should not be seen as an alternative to SGD. We also provide a simple experiment with training Resnet-18 [He et al., 2016] on Cifar10 [Krizhevsky and Hinton, 2009] in Appendix B.2, which gives a similar message.

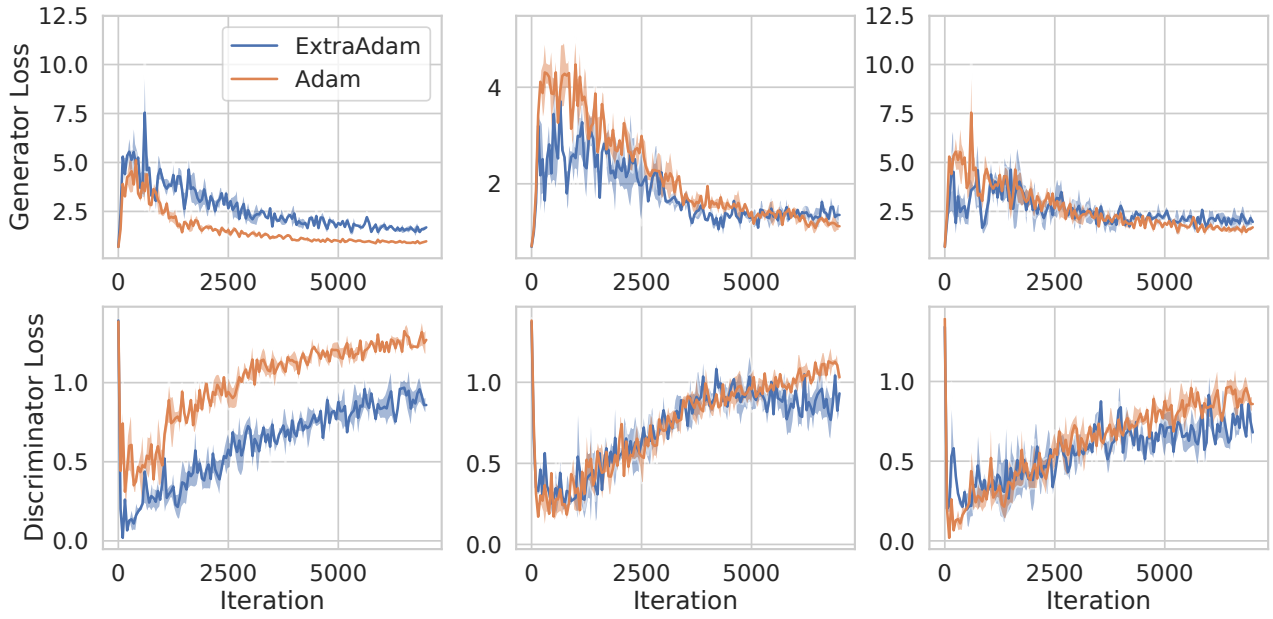


Figure 3: The columns differ in step sizes for generator and discriminator: 1) $(10^{-4}, 10^{-4})$, 2) $(5 \cdot 10^{-5}, 5 \cdot 10^{-5})$, 3) $(10^{-4}, 5 \cdot 10^{-5})$. In the top row, we show the generator loss and in the bottom row that of discriminator.

4 Experiments

4.1 Bilinear minimax

In this experiment, we generated a matrix with entries from standard normal distribution and dimensions 200. Since we did not observe much difference when changing the matrix size, we provide only one run in Figure 1. The results are very encouraging and show the superiority of the proposed approach on this problem. We provide two cases, with zero noise at the optimum and non-zero noise. In the latter case, only our method did not diverge.

When the noise at the optimum is zero, this is mostly like deterministic case for our method, but for the rest it is a difficult problem. On the other hand, when the noise is not equal 0 at the solution, the ergodic convergence of our method is faster, just as predicted by Theorem 3.

4.2 Generating mixture of Gaussians

Here we compare gradient descent-ascent as well as Mirror-Prox to our method on the task of learning mixture of 4 Gaussians. We provide the evolution of the process in Figure 2, although we note that the process is rather unstable and all results should be taken with a grain of salt.

To our surprise, negative momentum was rarely helpful and even positive momentum sometimes was giving significant improvement. We suspect that this is due to the different roles of generator and discriminator, but leave further exploration for future work.

The details of the experiment are as follows. For generator we use neural net with 2 hidden layers of size 16 and tanh activation function and output layer with size 2 and no activation function, which represents coordinates in 2D. Generator uses standard Gaussian vector of size 16 as an input. For discriminator we use neural net with input layer of size 2, which takes a point from 2D, 2 hidden layers of size 16 and tanh activation function and output layer with size 1 and sigmoid activation function, which represents probability of input point to be sampled from data distribution. We choose the same stepsize $5 \cdot 10^{-3}$ for all methods, which is close to maximal possible stepsize under which the methods rarely diverge.

Generator
Input: $z \in \mathbb{R}^{100} \sim \mathcal{N}(0, I)$
Embedding layer for the label
Linear ($110 \to 256$)
LeakyReLU (negative slope: 0.2)
Linear ($256 \to 512$)
LeakyReLU (negative slope: 0.2)
Linear ($512 \to 1024$)
LeakyReLU (negative slope: 0.2)
Linear ($1024 \to 784$)
$Tanh(\cdot)$

Discriminator
Input: $x \in \mathbb{R}^{1\times28\times28}$
Embedding layer for the label
Linear ($794 \to 1024$)
LeakyReLU (negative slope: 0.2)
Dropout (p=0.3)
Linear ($1024 \to 512$)
LeakyReLU (negative slope: 0.2)
Dropout (p=0.3)
Linear ($512 \to 256$)
LeakyReLU (negative slope: 0.2)
Dropout (p=0.3)
Linear ($1024 \to 784$)
$Sigmoid(\cdot)$

Table 1: Architectures used for our experiments on *Fashion MNIST*.

4.3 Comparison of Adam and ExtraAdam

Unfortunately, pure extragradient did not perform extremely well on big datasets, so for the Fashion MNIST and Celeba experiments we used adaptive stepsizes as in Adam [Kingma and Ba, 2014].

In the first set of experiments, we compared the performance of ExtraAdam [Gidel et al., 2019a] and Adam in a Conditional GAN [Mirza and Osindero, 2014] setup on Fashion MNIST [Xiao et al., 2017] dataset. The generator and discriminator were simple feedforward networks (detailed architectures description in Table 1). Optimizers were run with mini-batch size of 64 samples, no weight decay and $\beta_1 = 0.5, \beta_2 = 0.999$. One iteration of ExtraAdam was counted as two due to a double gradient calculation. The results (mean and variance) are depicted in Figure 3 and were obtained using 3 runs with different seeds. One can see that extragradient is slower because of the need to compute twice more gradients.

We suspect that Adam is faster partially due to that the problem's structure is something more specific than just a variational inequality. One validation of this guess is that in [Gidel et al., 2019b], the networks were trained with negative momentum only on discriminator, while generator was trained with constant momentum +0.5. Another reason we make this conjecture is that in [Metz et al., 2016] there was proposed a method that can be seen as a variant of extragradient, in which parameters of only one network requires extra steps.

In the second experiment, following [Chavdarova et al., 2019], we trained Self Attention GAN [Zhang et al., 2018]. We note that the loss was generally an ambiguous metric of method comparison, so we provide the Inception score [Salimans et al., 2016][1] in Figure 4 as performance measure for image synthesis. Besides, samples generated after training for two epochs are provided in Figure 9 in the Appendix.

The work [Gidel et al., 2019b] suggests using negative momentum to improve game dynamics and achieve faster convergence of the iterates. We consider using two types of momentum together: β_1 in the first step and β_2 in the second, i.e., we use $y^t = x^t - \eta_1 F(x^t; \xi^t) + \beta_1(x^t - x^{t-1})$ and $x^{t+1} = x^t - \eta_2 F(y^t; \xi^t) + \beta_2(x^t - x^{t-1})$. Detailed investigation on bilinear problems shows that β_1 can be chosen to be positive and β_2 should rather be negative. Intuitively, positive β_1 allows the method to look further ahead, while negative β_2 compensates for inaccuracy in the approximation of implicit update. In Appendix A.1, we discuss it in more details.

The results (mean and variance) are depicted in Figure 3 and were obtained using 3 runs

[1] We used implementation from this GitHub repository.

with different seeds.

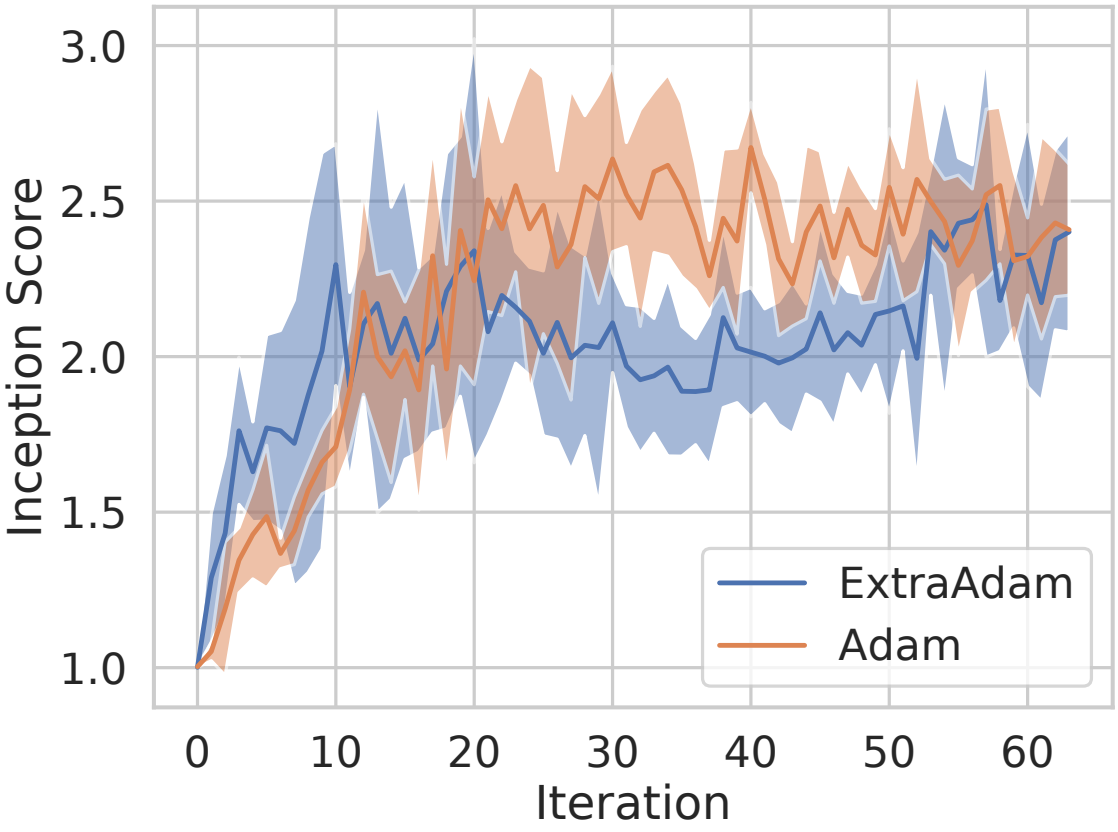
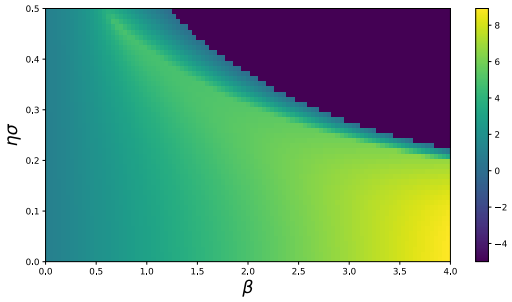


Figure 4: Inception score (mean and variance obtained by 5 runs) computed every 50 iterations during the training process on *CelebA* dataset for 2 epochs.

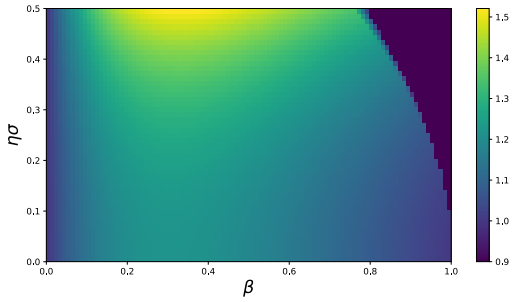
4.4 Discussion

The bilinear example is very clear and the results that we obtained showed enough stability. However, the message from training GANs is very vague due to their well-known instability. We did not observe a significant impact of negative momentum on convergence speed or stability, but at the same time we mentioned that setting first momentum to 0 in Adam is important for the extra update to have impact. We believe that the bilinear problem in this situation is the best way to make conclusion, but we still aim to obtain new methods for GANs in future.

It is also worth mentioning that the actual loss functions used in GANs are typically nonsmooth due to the choice of loss functions. For instance, the popular WGAN formulation [Arjovsky et al., 2017] includes hinge loss. On top of that, neural networks themselves have nonsmooth activations such as ReLU and its variants. Therefore, it is an interesting direction to understand what happens when the assumptions typical to variational inequalities are violated.



(a) $\eta_1 = \eta_2$, $\beta_2 = 0$, $\beta = \beta_1$ is the x-axis, $\eta\sigma_i$ is the y-axis. The optimal value of β_1 depends on $\eta\sigma_i$ and only for small values is significantly bigger 0. The dark area is where the method diverges.



(b) $\eta_1 = \eta_2$, $\beta_1 = 0$, $\beta = -\beta_2$ (negative momentum) is the x-axis, $\eta\sigma_i$ is the y-axis. The optimal value of β_2 is always very close to -0.3. The dark area is where the method diverges.

Figure 5: Values of the spectral radius of the extragradient momentum matrix (5) for bilinear problems for different values of $\eta\sigma$ and β. The heat values is the multiplicative speed up from using $\beta > 0$ compared to $\beta = 0$, which we define as the ratio $\frac{\rho(\mathbf{T}(\eta\sigma,\beta))}{\rho(\mathbf{T}(\eta\sigma,0))}$, where $\rho(\mathbf{A})$ is the spectral radius of a matrix $\mathbf{A}$ for any $\mathbf{A}$ and $\mathbf{T}(\eta\sigma,\beta)$ is the value of matrix in the update under given $\eta\sigma$ and β, see (5) in Appendix A.1.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. In *Advances in Neural Information Processing Systems 32*, pages 391–401. Curran Associates, Inc., 2019.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *Innovations in Theoretical Computer Science*, 2019.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations (ICLR 2018)*, 2018.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019a. URL `https://openreview.net/forum?id=r1laEnA5Ym`.

Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1802–1811. PMLR, 16–18 Apr 2019b.

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

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

G. M. Korpelevich. Extragradient method for finding saddle points and other problems. *Matekon*, 13(4):35–49, 1977.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

Konstantin Mishchenko and Peter Richtárik. A stochastic decoupling method for minimizing the sum of smooth and non-smooth functions. *arXiv preprint arXiv:1905.11535*, 2019.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pages 4573–4582. PMLR, 2020.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *arXiv preprint arXiv:1901.08511*, 2019.

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.

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. 2013.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2234–2242. Curran Associates, Inc., 2016. URL `http://papers.nips.cc/paper/6125-improved-techniques-for-training-gans.pdf`.

Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

Appendix

A Proofs

Proof of Theorem 1
We prove a more general version of the claim made in the main part, in particular we provide $O(\eta^k)$ bound for extragraident with k steps. The precise claim is given below.

Theorem 6. *Let F be an L-Lipschitz operator and define recursively $y_0 = x$ and $y_{m+1} \stackrel{def}{=} \mathrm{prox}_{\eta g}(x - \eta F(y_m))$ for $m = 1, \dots, k$ and let $w \stackrel{def}{=} \mathrm{prox}_{\eta g}(x - \eta F(w))$ be the implicit update, where $\eta > 0$ is any stepsize. Then,*

$$\|w - y_k\| \leq \eta^k L^k \|w - x\|.$$

Proof. We show the claim by induction. For $k = 0$ it holds simply because $y_0 \stackrel{\text{def}}{=} x$. If it holds for $k-1$, let us show it for k. By non-expansiveness of the proximal operator we have

$$\begin{aligned}
\|w - y_k\| &= \|\mathrm{prox}_{\eta g}(x - \eta F(w)) - \mathrm{prox}_{\eta g}(x - \eta F(y_{k-1}))\| \\
&\leq \|x - \eta F(w) - (x - \eta F(y_{k-1}))\| \\
&= \eta \|F(w) - F(y_{k-1})\| \\
&\leq \eta L \|w - y_{k-1}\| \\
&\leq \eta^k L^k \|w - x\|.
\end{aligned}$$

□

Proof of Theorem 2
First, let us introduce the following lemma that will be very useful in our analysis.

Lemma 1. *Let g be μ–strongly convex and $z = \mathrm{prox}_{\eta g}(x)$. Then for all $y \in \mathbb{R}^d$ the following inequality holds:*

$$\langle z - x, y - z \rangle \geq \eta\big(g(z) - g(y) + \frac{\mu}{2}\|z - y\|^2\big). \tag{4}$$

Proof. The lemma easily follows from the definitions. Indeed, since

$$z \stackrel{\text{def}}{=} \arg\min_u \{\eta g(u) + \frac{1}{2}\|u - x\|^2\},$$

we have necessary optimality condition $0 \in \eta \partial g(z) + (z - x)$. Thus, by the definition of a subdifferential and by strong convexity,

$$\eta(g(y) - g(z)) \geq \langle x - z, y - z \rangle + \frac{\eta\mu}{2}\|z - y\|^2$$

and the proof is complete. □

In addition, let us also separately state how we are going to deal with the update variance.

Lemma 2. *Let $F(\cdot; \xi)$ be almost surely monotone and assume that point x is such that $\sigma_x^2 \stackrel{def}{=} \mathbb{E}\|F(x; \xi) - F(x)\|^2 < +\infty$, i.e., the variance of F at x is bounded. Then,*

$$\mathbb{E}\left\langle F(x) - F(x; \xi^t), y^t - x \right\rangle \leq \eta\sigma_x^2 + \frac{1}{4\eta}\mathbb{E}\|y^t - x^t\|^2.$$

Proof. As x^t and ξ^t are independent random variables and $\mathbb{E}F(x;\xi^t) = F(x)$, we have

$$\begin{aligned}\mathbb{E}\left\langle F(x) - F(x;\xi^t), y^t - x\right\rangle &= \mathbb{E}\left\langle F(x) - F(x;\xi^t), x^t - x\right\rangle + \mathbb{E}\left\langle F(x) - F(x;\xi^t), y^t - x^t\right\rangle \\ &= \mathbb{E}\left\langle F(x) - F(x;\xi^t), y^t - x^t\right\rangle.\end{aligned}$$

By Young's inequality,

$$\begin{aligned}\mathbb{E}\left\langle F(x) - F(x;\xi^t), y^t - x^t\right\rangle &\leq \eta\mathbb{E}\|F(x) - F(x;\xi^t)\|^2 + \frac{1}{4\eta}\mathbb{E}\|y^t - x^t\|^2 \\ &= \eta\sigma_x^2 + \frac{1}{4\eta}\mathbb{E}\|y^t - x^t\|^2\end{aligned}$$

and the proof is complete. □

Now we are ready to prove Theorem 2.

Proof. By Lemma 1 for points $y^t = \operatorname{prox}_{\eta g}\left(x^t - \eta F(x^t;\xi^t)\right)$ and $x^{t+1} = \operatorname{prox}_{\eta g}\left(x^t - \eta F(y^t;\xi^t)\right)$,

$$\begin{aligned}\left\langle x^{t+1} - x^t + \eta F(y^t;\xi^t), x^* - x^{t+1}\right\rangle &\geq \eta\big(g(x^{t+1}) - g(x^*) + \frac{\mu}{2}\|x^{t+1} - x^*\|^2\big) \\ \left\langle y^t - x^t + \eta F(x^t;\xi^t), x^{t+1} - y^t\right\rangle &\geq \eta\big(g(y^t) - g(x^{t+1}) + \frac{\mu}{2}\|x^{t+1} - y^t\|^2\big).\end{aligned}$$

Summing these two inequalities together and rearranging, we get

$$\begin{aligned}\langle x^{t+1} - x^t, x^* - x^{t+1}\rangle + \langle y^t - x^t, x^{t+1} - y^t\rangle + \eta\langle F(y^t;\xi^t) - F(x^t;\xi^t), y^t - x^{t+1}\rangle + \eta\langle F(y^t;\xi^t), x^* - y^t\rangle \\ \geq \eta\big(g(y^t) - g(x^*) + \frac{\mu}{2}\|x^{t+1} - x^*\|^2 + \frac{\mu}{2}\|x^{t+1} - y^t\|^2\big).\end{aligned}$$

Using identity $2\langle a, b\rangle = \|a+b\|^2 - \|a\|^2 - \|b\|^2$ for the first two scalar products, we deduce

$$\begin{aligned}(1+\eta\mu)\|x^{t+1} - x^*\|^2 \leq \|x^t - x^*\|^2 - \|x^t - y^t\|^2 - (1+\eta\mu)\|x^{t+1} - y^t\|^2 \\ + 2\eta\langle F(y^t;\xi^t) - F(x^t;\xi^t), y^t - x^{t+1}\rangle - 2\eta\big(\langle F(y^t;\xi^t), y^t - x^*\rangle + g(y^t) - g(x^*)\big).\end{aligned}$$

The first scalar product can be simplified using Lipschitzness. Since $F(\cdot;\xi^t)$ is almost surely L–Lipschitz, by Young's inequality

$$\begin{aligned}2\eta\langle F(y^t;\xi^t) - F(x^t;\xi^t), y^t - x^{t+1}\rangle &\leq \frac{\eta}{L}\|F(y^t;\xi^t) - F(x^t;\xi^t)\|^2 + \eta L\|y^t - x^{t+1}\|^2 \\ &\leq \eta L\big(\|x^{t+1} - y^t\|^2 + \|y^t - x^t\|^2\big).\end{aligned}$$

To get rid of the other scalar product, we use monotonicity of $F(\cdot;\xi^t)$, and then apply strong convexity of g,

$$\begin{aligned}\langle F(y^t;\xi^t), y^t - x^*\rangle + g(y^t) - g(x^*) &\geq \langle F(x^*;\xi^t), y^t - x^*\rangle + g(y^t) - g(x^*) \\ &= \langle F(x^*), y^t - x^*\rangle + g(y^t) - g(x^*) + \left\langle F(x^*;\xi^t) - F(x^*), y^t - x^*\right\rangle \\ &\geq \frac{\mu}{2}\|y^t - x^*\|^2 + \left\langle F(x^*;\xi^t) - F(x^*), y^t - x^*\right\rangle.\end{aligned}$$

So far, the proof has not involved any expectation, but now we shall use Lemma 2 to deduce from the produced bounds

$$\begin{aligned}(1+\eta\mu)\mathbb{E}\|x^{t+1} - x^*\|^2 &\leq \mathbb{E}\Big[\|x^t - x^*\|^2 - \eta\mu\big(\|y^t - x^*\|^2 + \|x^{t+1} - y^t\|^2\big)\Big] + 2\eta^2\sigma^2 \\ &\quad - \underbrace{(1 - \eta L - \tfrac{1}{2})}_{\geq 0}\mathbb{E}\|y^t - x^t\|^2 \\ &\leq \mathbb{E}\Big[\|x^t - x^*\|^2 - \eta\mu\big(\|y^t - x^*\|^2 + \|x^{t+1} - y^t\|^2\big)\Big] + 2\eta^2\sigma^2.\end{aligned}$$

Using inequality $\|a\|^2 + \|b\|^2 \geq \frac{1}{2}\|a+b\|^2$, we arrive at

$$\Big(1 + \frac{3}{2}\eta\mu\Big)\|x^{t+1} - x^*\|^2 \leq \|x^t - x^*\|^2 + 2\eta^2\sigma^2.$$

Note that $\eta\mu \leq 1/2$ and, therefore, $\frac{1}{1+3\eta\mu/2} \leq (1 - 2\eta\mu/3)$. The statement of the theorem can be now easily obtained by induction. □

Proof of Theorem 3

Let $x \in \mathcal{X}$. Similarly to the proof of Theorem 2, we can obtain from Lemma 1 with $\mu = 0$

$$\begin{aligned}\|x^{t+1} - x\|^2 &\leq \|x^t - x\|^2 - \|x^t - y^t\|^2 - \|x^{t+1} - y^t\|^2 + 2\eta L(\|x^{t+1} - y^t\|^2 + \|y^t - x^t\|^2\|) \\ &\quad - 2\eta\big(\langle F(y^t;\xi^t), y^t - x\rangle + g(y^t) - g(x)\big) \\ &\leq \|x^t - x\|^2 - \frac{1}{2}\|x^t - y^t\|^2 - \|x^{t+1} - y^t\|^2 \\ &\quad - 2\eta\big(\langle F(y^t;\xi^t), y^t - x\rangle + g(y^t) - g(x)\big).\end{aligned}$$

By monotonicity of $F(\cdot;\xi^t)$ and Lemma 2 we deduce

$$\begin{aligned}\mathbb{E}\left\langle F(y^t;\xi^t), x - y^t\right\rangle &\leq \mathbb{E}\left\langle F(x;\xi^t), x - y^t\right\rangle \\ &\leq \eta\sigma_x^2 + \mathbb{E}\left\langle F(x), x - y^t\right\rangle + \frac{1}{4\eta}\mathbb{E}\|y^t - x^t\|^2.\end{aligned}$$

Therefore,

$$\mathbb{E}\left[g(y^t) - g(x) + \left\langle F(x), y^t - x\right\rangle\right] \leq \frac{1}{2\eta}\mathbb{E}\left[\|x^t - x\|^2 - \|x^{t+1} - x\|^2\right] + \eta\sigma_x^2.$$

Telescoping this inequality, we obtain

$$\mathbb{E}\frac{1}{t+1}\sum_{k=0}^{t}(g(y^k) - g(x) + \left\langle F(x), y^k - x\right\rangle) \leq \frac{1}{2\eta t}\|x^0 - x\|^2 + \eta\sigma_x^2 \leq \sup_{z\in\mathcal{X}}\left\{\frac{1}{2\eta t}\|x^0 - z\|^2 + \eta\sigma_z^2\right\}.$$

The left-hand side is a convex function y^k. Therefore, choosing $\eta = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$ and applying Jensen's inequality to the left-hand side, we get the claim.

Proof of Theorem 4

Proof. Since the function is bilinear, we can write

$$\nabla_x f(x,y) = \mathbf{B}(y - y^*), \qquad \nabla_y f(x,y) = \mathbf{B}^\top(x - x^*).$$

Then, we obtain the explicit update rules

$$\begin{aligned}x^{t+1} &= x^t - \eta_2\mathbf{B}(v^t - y^*) = x^t - \eta_2\mathbf{B}(y^t - y^* + \eta_1\mathbf{B}^\top(x^t - x^*)) \\ y^{t+1} &= y^t + \eta_2\mathbf{B}^\top(u^t - x^*) = y^t + \eta_2\mathbf{B}^\top(x^t - x^* - \eta_1\mathbf{B}(y^t - y^*)).\end{aligned}$$

In matrix forms it is

$$\begin{bmatrix}x^{t+1} - x^* \\ y^{t+1} - y^*\end{bmatrix} = \begin{pmatrix}\mathbf{I} - \eta_1\eta_2\mathbf{B}\mathbf{B}^\top & -\eta_2\mathbf{B} \\ \eta_2\mathbf{B}^\top & \mathbf{I} - \eta_1\eta\mathbf{B}^\top\mathbf{B}\end{pmatrix}\begin{bmatrix}x^t - x^* \\ y^t - y^*\end{bmatrix}$$

Apply SVD decomposition to $\mathbf{B}$: $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$. Then,

$$\left\|\begin{bmatrix}x^{t+1} - x^* \\ y^{t+1} - y^*\end{bmatrix}\right\| \leq \left\|\begin{pmatrix}\mathbf{I} - \eta_1\eta_2\mathbf{B}\mathbf{B}^\top & -\eta_2\mathbf{B} \\ \eta_2\mathbf{B}^\top & \mathbf{I} - \eta_1\eta_2\mathbf{B}^\top\mathbf{B}\end{pmatrix}\right\|\left\|\begin{bmatrix}x^t - x^* \\ y^t - y^*\end{bmatrix}\right\|.$$

Since $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, we have

$$\mathbf{B}\mathbf{B}^\top = \mathbf{U}\mathbf{\Sigma}^2\mathbf{V}^\top,$$
$$\mathbf{B}^\top\mathbf{B} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{U}^\top,$$

and

$$\begin{aligned}
\left\|\begin{pmatrix}\mathbf{I}-\eta_1\eta\mathbf{B}\mathbf{B}^\top & -\eta_2\mathbf{B}\\ \eta_2\mathbf{B}^\top & \mathbf{I}-\eta_1\eta\mathbf{B}^\top\mathbf{B}\end{pmatrix}\right\| &= \left\|\begin{pmatrix}\mathbf{U} & 0\\ 0 & \mathbf{V}\end{pmatrix}\begin{pmatrix}\mathbf{I}-\eta_1\eta\mathbf{\Sigma}^2 & -\eta_2\mathbf{\Sigma}\\ \eta_2\mathbf{\Sigma} & \mathbf{I}-\eta_1\eta\mathbf{\Sigma}^2\end{pmatrix}\begin{pmatrix}\mathbf{U}^\top & 0\\ 0 & \mathbf{V}^\top\end{pmatrix}\right\|\\
&= \left\|\begin{pmatrix}\mathbf{I}-\eta_1\eta_2\mathbf{\Sigma}^2 & -\eta_2\mathbf{\Sigma}\\ \eta_2\mathbf{\Sigma} & \mathbf{I}-\eta_1\eta_2\mathbf{\Sigma}^2\end{pmatrix}\right\|\\
&= \max_i \left\|\begin{pmatrix}1-\eta_1\eta_2\sigma_i^2 & -\eta_2\sigma_i\\ \eta_2\sigma_i & 1-\eta_1\eta_2\sigma_i^2\end{pmatrix}\right\|\\
&= \max_i \sqrt{(1-\eta_1\eta_2\sigma_i^2)^2+\eta_2^2\sigma_i^2}.
\end{aligned}$$

Assume without loss of generality that $\sigma_1 \geq \cdots \geq \sigma_n$. Note that function $x \mapsto \left(1-\frac{\eta_1}{\eta_2}x^2\right)^2 + x^2$ is monotonically decreasing on $(0, c)$ and monotonically increasing on $(c, +\infty)$, where c is $+\infty$ if $\eta_2 \geq 2\eta_1$ and $\frac{\eta_2}{\sqrt{2}\eta_1}\sqrt{2\frac{\eta_1}{\eta_2}-1}$ otherwise. Consequently, it holds

$$\max_i\{(1-\eta_1\eta_2\sigma_i^2)^2+\eta_2^2\sigma_i^2\} = \max\{(1-\eta_1\eta_2\sigma_1^2)^2+\eta_2^2\sigma_1^2, (1-\eta_1\eta_2\sigma_n^2)^2+\eta_2^2\sigma_n^2\}.$$

□

Proof of Corollary 1

Proof. These statements follow from the bound obtained in Theorem 4. Since function $(1-x^2)^2+x^2$ monotonically decreases when $x \in \left(0, \frac{1}{\sqrt{2}}\right)$, we have $\rho = (1-\eta_1\eta_2\sigma_{\min}(\mathbf{B})^2)^2 + \eta_2^2\sigma_{\min}(\mathbf{B})^2 = \left(1-\frac{\sigma_{\min}(\mathbf{B})^2}{2\sigma_{\max}(\mathbf{B})^2}\right)^2 + \frac{\sigma_{\min}(\mathbf{B})^2}{2\sigma_{\max}(\mathbf{B})^2}$. The second case follows similarly. □

A.1 Negative momentum

For bilinear problems with two types of momentum the update recurrence is

$$\begin{bmatrix}x^{t+1}-x^*\\ y^{t+1}-y^*\\ x^t-x^*\\ y^t-y^*\end{bmatrix} = \begin{pmatrix}(1+\beta_2)\mathbf{I}-\eta_1\eta_2\mathbf{B}\mathbf{B}^\top & -\eta_2(1+\beta_1)\mathbf{B} & -\beta_2\mathbf{I} & \eta_2\beta_1\mathbf{B}\\ \eta_2(1+\beta_1)\mathbf{B}^\top & (1+\beta_2)\mathbf{I}-\eta_1\eta\mathbf{B}^\top\mathbf{B} & -\eta_2\beta_1\mathbf{I} & -\beta_2\mathbf{I}\\ \mathbf{I} & 0 & 0 & 0\\ 0 & \mathbf{I} & 0 & 0\end{pmatrix}\begin{bmatrix}x^t-x^*\\ y^t-y^*\\ x^{t-1}-x^*\\ y^{t-1}-y^*\end{bmatrix}.$$

Using SVD decomposition, we can represent the above matrix as block-diagonal with blocks $\mathbf{T}_i$

$$\mathbf{T}_i = \begin{pmatrix}1+\beta_2-\eta_1\eta_2\sigma_i^2 & -\eta_2(1+\beta_1)\sigma_i & -\beta_2 & \eta_2\beta_1\sigma_i\\ \eta_2(1+\beta_1)\sigma_i & 1+\beta_2-\eta_1\eta\sigma_i^2 & -\eta_2\beta_1 & -\beta_2\\ 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{pmatrix}, \tag{5}$$

where σ_i is the i-th the singular value of $\mathbf{B}$.

One can show that the spectral radius of this matrix improves with negative β_2, however this is not true for its second norm. Since this is a very technical property that can be easily illustrated numerically, we simply provided a plot of how spectral radius changes depending on values of $\eta\sigma$ and β_2 when $\beta = 1 = 0$ and $\eta_1 = \eta_2 = \eta$, see Figure 5. In addition, here we provide the heatmap for $\eta_1 = \eta_2$ and product $\eta\sigma = 0.01$. As can be seen from Figure 6, nonzero β_1 is not very promising and β_2 leads only to a small improvement. Thus, it gives advantage mainly for large values of $\eta\sigma$.

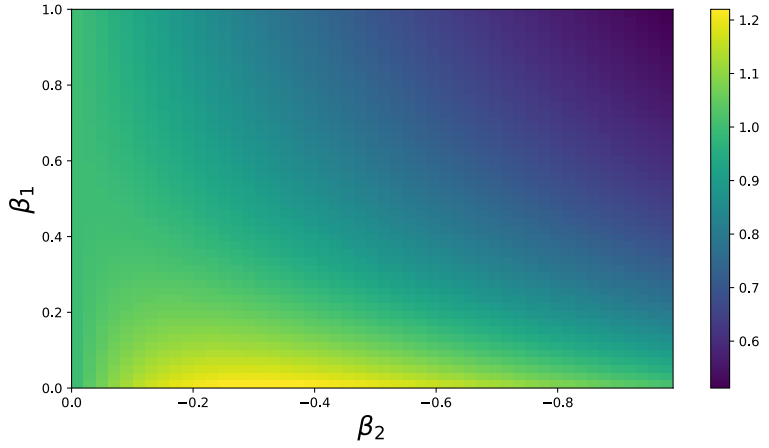


Figure 6: Ratio of spectral radii as in Figure 5 but with fixed $\eta\sigma = 0.01$ and different values of β_1 and β_2.

A.2 Proof of Theorem 5

Let us introduce a notation that simplifies the proof. We will denote by $\mathbb{E}_t$ the expectation conditioned on x^t, i.e., $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid x^t]$.

Proof. Recall that $y^t = x^t - \eta\nabla f(x^t;\xi^t)$, $x^{t+1} = x^t - \eta\nabla f(y^t;\xi^t)$, and apply smoothness of f to x^{t+1} and x^t:

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \left\langle \nabla f(x^t), x^{t+1} - x^t \right\rangle + \frac{L}{2}\|x^{t+1} - x^t\|^2 \\ &= f(x^t) - \eta\|\nabla f(x^t)\|^2 + \eta\left\langle \nabla f(x^t), \nabla f(x^t) - \nabla f(y^t;\xi^t)\right\rangle + \frac{L\eta^2}{2}\|\nabla f(y^t;\xi^t)\|^2. \end{aligned}$$

Since $\nabla f(x^t;\xi^t)$ is an unbiased estimate of $\nabla f(x^t)$, it follows by Young's inequality and smoothness of $f(\cdot;\xi^t)$

$$\begin{aligned} \eta\left\langle \nabla f(x^t), \nabla f(x^t) - \nabla f(y^t;\xi^t)\right\rangle &= \mathbb{E}_t\eta\left\langle \nabla f(x^t), \nabla f(x^t;\xi^t) - \nabla f(y^t;\xi^t)\right\rangle \\ &\leq \frac{\eta^2 L}{2}\|\nabla f(x^t)\|^2 + \frac{1}{2L}\mathbb{E}_t\|\nabla f(x^t;\xi^t) - \nabla f(y^t;\xi^t)\|^2 \\ &\leq \frac{\eta^2 L}{2}\|\nabla f(x^t)\|^2 + \frac{L}{2}\mathbb{E}_t\|x^t - y^t\|^2 \\ &= \frac{\eta^2 L}{2}\|\nabla f(x^t)\|^2 + \frac{\eta^2 L}{2}\mathbb{E}_t\|\nabla f(y^t;\xi^t)\|^2. \end{aligned}$$

Moreover, similar arguments show how to bound the expectation of the squared gradient norm:

$$\begin{aligned} \mathbb{E}_t\|\nabla f(y^t;\xi^t)\|^2 &\leq 2\mathbb{E}_t\|\nabla f(y^t;\xi^t) - \nabla f(x^t;\xi^t)\|^2 + 2\mathbb{E}_t\|\nabla f(x^t;\xi^t)\|^2 \\ &\leq 2L^2\mathbb{E}_t\|y^t - x^t\|^2 + 2\mathbb{E}_t\|\nabla f(x^t;\xi^t)\|^2 \\ &= 2(1 + L^2\eta^2)\mathbb{E}_t\|\nabla f(x^t;\xi^t)\|^2 \\ &\leq 2(1 + L^2\eta^2)(\|\nabla f(x^t)\|^2 + \sigma^2). \end{aligned}$$

Thus,

$$\mathbb{E}_t f(x^{t+1}) \leq f(x^t) - \eta \left[1 - \eta L - 2\eta L(1+\eta^2 L^2)\right] \|\nabla f(x^t)\|^2 + 2\eta^2 L(1+\eta^2 L^2)\sigma^2.$$

If $\eta L \leq \frac{1}{4}$, we have $1 - \eta L - 2\eta L(1+\eta^2 L^2) > \frac{1}{5}$, so this bound can be simplified to

$$\|\nabla f(x^t)\|^2 \leq \frac{5}{\eta}\mathbb{E}_t[f(x^t) - f(x^{t+1})] + 11\eta L\sigma^2.$$

Telescoping this inequality from 0 to $t-1$ and taking full expectation with respect to all randomness, we get

$$\begin{aligned}\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\|\nabla f(x^k)\|^2 &\leq \frac{5}{\eta t}(f(x^0) - f(x^t)) + 11\eta L\sigma^2 \\ &\leq \frac{5}{\eta t}(f(x^0) - f^*) + 11\eta L\sigma^2.\end{aligned}$$

It remains to mention that the left-hand side is exactly the expectation of $\mathbb{E}\|\nabla f(\hat{x}^t)\|^2$. □

B Additional experiments

B.1 Reproducing mixture of eight Gaussians

We also double check that extragradient converges on the mixture of 8 Gaussians. This experiment is a sanity that allows us to show that the method can do at least as well as alternating gradient [Gidel et al., 2019b]. To directly relate to their experiments, we ran extragradient on the same type of network, although we changed activation from ReLU to tanh, which was more stable in our experiments. Note that [Gidel et al., 2019b] ran alternating method for 100,000 iterations, while we required only 20,000, which corresponds to 40,000 generator updates. The result is presented in Figure 7.

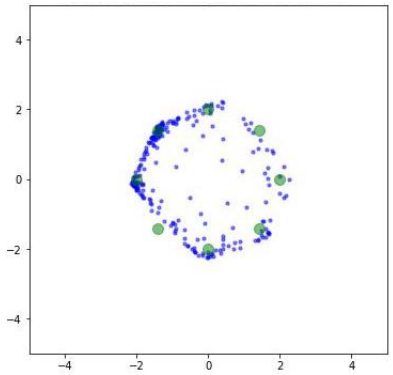


Figure 7: Samples from generator after training for 20,000 iterations of minibatch 512 with extragradient. Both generator and discriminator are 4-layers neural networks with tanh activation and the dimension of the noise distribution is 256.

B.2 Empirical risk minimization

As our theory suggests, stochastic extragradient might not be better than SGD when solving a simple task such as function minimization. To see how it works in practice, we trained Residual Network [He et al., 2016], Resnet-18, on Cifar10 [Krizhevsky and Hinton, 2009] dataset with

cross-entropy loss and different stepsizes, and compared the results to SGD. In order to see the effect of the update rule, we do not use any type of momentum in this experiment and keep the learning rate constant. Our observation in this situation is that extragradient is indeed slower, both because of the need to compute two gradients per iterations and because of worse final accuracy.

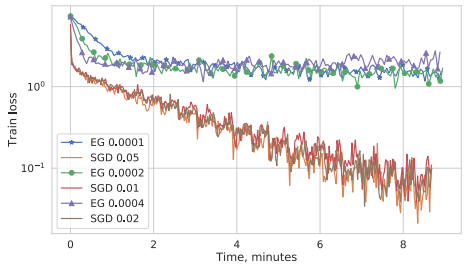


Figure 8: Comparison of the proposed stochastic extragradient and stochastic gradient descent when optimizing Residual Network with 18 hidden layers on Cifar10 dataset. We report only the train loss as this is the most relevant metric for an optimization method, and test accuracy in this experiment behaved similarly.

B.3 Samples of generated images

Figure 9: Adam (top) and ExtraAdam (bottom) results of training self attention GAN for two epochs. The results of training with the three best performing stepsizes, $10^{-3}, 2 \cdot 10^{-3}, 4 \cdot 10^{-3}$, are provided for each method (from the left to the right). Best seen in color by zooming on a computer screen.