Revisiting Stochastic Extragradient

Konstantin Mishchenko       Dmitry Kovalev       Egor Shulgin

Yura Malitsky               Peter Richtárik

Abstract

We consider a new extension of the extragradient method that is motivated by approximating implicit updates. Since in a recent work [11] it was shown that the existing stochastic extragradient algorithm (called mirror-prox) of [12] diverges on a simple bilinear problem, we prove guarantees for solving variational inequality that are more general than in [12]. Furthermore, we illustrate numerically that the proposed variant converges faster than many other methods on the example of [11]. We also discuss how extragradient can be applied to training Generative Adversarial Networks (GANs). 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. To further accelerate method’s convergence on problems such as bilinear minimax, we combine the extragradient step with negative momentum [8] and discuss the optimal momentum value.

1 Introduction

Variational inequality problem is a general framework which covers a variety of optimization problems such as constrained minimization and saddle-point problems. Roughly speaking, variational inequality is equivalent to the necessary first-order optimality condition for optimization problem (which is also sufficient in the convex case). The formulation has a lot of applications in machine learning, most prominent of which are empirical risk minimization and two-player games. In particular, recently invented generative adversarial neural networks [10] are often trained using schemes that resemble primal-dual and variational inequality methods, which we shall discuss in detail later.

The problem that we consider is that of finding a point $x^*$ satisfying

$$g(x) - g(x^*) + \langle F(x^*), x - x^* \rangle \geq 0, \text{ for all } x \in \mathcal{K},$$

where $\mathcal{K} \subset \mathbb{R}^d$ is a convex set, $g$ is a convex function and operator $F: \mathcal{K} \to \mathbb{R}^d$ is monotone. Having monotone operator is not directly related to training of neural networks, whose loss landscape has a lot of nonconvex regions, but, unfortunately, little is known about variational inequality and even minimax problems when convexity is missing. Thus, we stick to this assumption and rather try to model adversarial properties by considering particularly unstable bilinear minimax problems.

Of particular interest to us is the situation where $F(x)$ is the expectation with respect to random variable $\xi$ 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, $\xi$ can be a random variable sampled by one of the GAN networks, called generator.

A special case of (1) is constrained minimax optimization,

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

Preprint. Under review.
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 [9] even when $f$ is convex-concave. In particular, the objective $f(x, y) = x^\top y$ leads to geometrical divergence for any nontrivial initialization [3]. For more discussion on variational inequality and its relation to GANs see [7].

1.1 Related work

The extragradient method was first proposed in [14]. Since then there has been developed a number of its extensions, most famous of which is the mirror-prox method [19] 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 accurate enough, assuming the operator is smooth. We base our intuition upon this property and we shall discuss it in detail later in the paper.

While extragradient uses future information, convergence guarantees can be achieved even from past information. In particular, Optimistic mirror descent (OMD), first proposed by [21] for convex-concave zero-sum games, has been analyzed in a number of works [18, 4, 6] and it was applied to GAN training in [2]. 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.

There are other techniques that allow to improve stability and achieve convergence for monotone operators. While alternating gradient descent-ascent does not, in general, converge to a solution [8], the negative momentum trick proposed in [8] can fix this.

This work is not the first to consider a variant of stochastic extragradient. A stochastic version of the mirror-prox method [19] was analyzed in [12] under pretty restrictive assumptions. While deterministic extragradient approximates implicit update, the authors of [12] 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 [11] 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 [16] as a way to stabilize training of GANs. The main different of the method in [16] to what we consider is in applying extrasteps only on one of two neural networks. In addition, [16] 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.

1.2 Theoretical background

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

**Assumption 1.** Operator $F : \mathcal{K} \to \mathbb{R}^d$ is monotone, that is $(F(x) - F(y), x - y) \geq 0$ for all $x, y \in \mathcal{K}$. 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 most of the methods are analyzed under it. There are several versions of pseudo-monotonicity, but without it the variational inequality problem becomes extremely hard to solve.

**Assumption 2.** Operator $F : \mathcal{K} \to \mathbb{R}^d$ is $L$-Lipschitz, that is for all $x, y \in \mathcal{K}$

$$
\|F(x) - F(y)\| \leq L \|x - y\|. 
$$

**Assumption 3.** Function $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous and $\mu$-strongly convex for $\mu \geq 0$, i.e. for all $x, y \in \mathcal{K}$ 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.

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

$$
\mathbb{E}\|F(x^*; \xi) - F(x^*)\|^2 \leq \sigma^2.
$$
Let us show that extragradient efficiently approximates implicit update. It fails even on bilinear problems, as it was observed in [1]. When we use independent samples, it will rarely approximate the implicit update, so it is rather not surprising that stochastic extragradient, which was suggested in [12], does not make much sense. Since it uses two stepsizes, it is argued that the main goal is to approximate those [18]. From that perspective, the current work is almost surely monotonically convergent.

Theorem 1. Assume that $F$ is an $L$-Lipschitz operator and define $y \overset{\text{def}}{=} \text{prox}_{\eta g}(x - \eta F(x))$, $z \overset{\text{def}}{=} \text{prox}_{\eta g}(x - \eta F(y))$, and $w \overset{\text{def}}{=} \text{prox}_{\eta g}(x - \eta F(w))$, 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 and this is what makes it possible for the method to solve variational inequality. One can also mention that having extra variance of a set of points is almost surely monotonically convergent.

Below we show that extragradient efficiently approximates implicit update.

**Theorem 2.** Assume that $g$ is a $\mu$-strongly convex function, operator $F(\cdot ; \xi)$ is almost surely monotone and $L$-Lipschitz, and that its variance at the optimum is bounded by constant, $E\|F(x^*; \xi) - F(x^*; \xi')\|^2$.

2 Theory

It is argued that implicit updates are more stable when solving variational inequality and sometimes it is argued that the main goal is to approximate those [18]. From that perspective, the current stochastic extragradient, which was suggested in [12], 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 even on bilinear problems, as it was observed in [1].

Below we show that extragradient efficiently approximates implicit update.

**Theorem 1.** Let $F$ be an $L$-Lipschitz operator and define $y \overset{\text{def}}{=} \text{prox}_{\eta g}(x - \eta F(x))$, $z \overset{\text{def}}{=} \text{prox}_{\eta g}(x - \eta F(y))$, $w \overset{\text{def}}{=} \text{prox}_{\eta g}(x - \eta F(w))$, 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 and this is what makes it possible for the method to solve variational inequality. One can also mention that having extra variance of a set of points is almost surely monotonically convergent.

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 [12], 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, $\xi^t$, 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 precise rate that we obtained.

**Theorem 2.** Assume that $g$ is a $\mu$-strongly convex function, operator $F(\cdot ; \xi)$ is almost surely monotone and $L$-Lipschitz, and that its variance at the optimum is bounded by constant, $E\|F(x^*; \xi) - F(x^*; \xi')\|^2$.

**Algorithm 1** The Extragradient Method for Variational Inequality.

1: **Parameters:** $x^0 \in K$, stepsize $\eta > 0$
2: **for** $t = 0, 1, 2, \ldots$ **do**
3: Sample $\xi^t$
4: \hspace{1em} $y^t = \text{prox}_{\eta g}(x^t - \eta F(x^t; \xi^t))$
5: \hspace{1em} $x^{t+1} = \text{prox}_{\eta g}(x^t - \eta F(y^t; \xi^t))$
6: **end for**

**Algorithm 2** The extragradient method for min-max problems.

1: **Require:** Stepsizes $\eta_1, \eta_2$, initial vectors $x^0, y^0$
2: **for** $t = 0, 1, \ldots$ **do**
3: \hspace{1em} $u^t = x^t - \eta_1 \nabla_x f(x^t, y^t)$
4: \hspace{1em} $v^t = y^t + \eta_2 \nabla_y f(x^t, y^t)$
5: \hspace{1em} $x^{t+1} = x^t - \eta_2 \nabla_x f(u^t, v^t)$
6: **end for**

Depending on the assumptions, we will either work with the variance at the optimum or with a merit function, which involves variance of a set of points.
\[ F(x^*)^2 \leq \sigma^2. \] Then, for any \( \eta \leq 1/(2L) \)
\[ \mathbb{E}[\|x^t - x^*\|^2] \leq (1 - 2\eta \mu/3)^t \|x^0 - x^*\|^2 + 4\sigma^2/(3\mu). \]

In the case where at the optimum the noise is zero, we recover standard results for extragradient \([22]\).
This is also similar to the rate proved for optimistic mirror descent in \([7]\), 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 \([1]\) satisfy for any set \( X \)
\[ \sup_{x \in X} \{ g(\hat{x}^t) - g(x) + \langle F(x), \hat{x}^t - x \rangle \} \leq \frac{1}{\sqrt{t}L} \sup_{x \in X} \left\{ \frac{L^2}{2} \|x^0 - x\|^2 + \sigma_x^2 \right\}. \]
where \( \hat{x}^t = \frac{1}{t} \sum_{k=0}^{t} x^k \) and \( \sigma_x^2 \) is the variance of \( F \) at point \( x \).

This result is more general than the one obtained in \([7]\), where the authors require for the same rate bounded variance and even \( \mathbb{E}[\|F(x; \xi)\|^2] \leq M < \infty \) uniformly over \( x \). The left-hand side in the bound above is a merit function that has been used in variational inequality literature \([20]\).

### 2.2 Adversarial bilinear problems

The work \([8]\) argues that a good illustration of method’s stability can be obtained when considering minimax bilinear problems, which is given by
\[ \min_x \max_y f(x,y) = x^\top B y + a^\top x + b^\top y, \]
where \( B \) is a full rank square matrix. One can show that if there exists a Nash equilibrium point, then \( f(x,y) = (x-x^*)^\top B (y-y^*) + \text{const} \) for some pair \( (x^*, y^*) \). 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 \( B \) and apply Algorithm \([2]\) to it. Choose any \( \eta_1 \) and \( \eta_2 \) such that \( \eta_2 < 1/\sigma_{\max}(B) \) and \( \eta_1 \eta_2 < 2/\sigma_{\max}(B)^2 \), then the rate is
\[ \|x^t - x^*\|^2 + \|y^t - y^*\|^2 \leq \rho^t \left( \|x^0 - x^*\|^2 + \|y^0 - y^*\|^2 \right), \]
where \( \rho \) is the same assumption as in Theorem \([2]\), consider two choices of stepizes:

1. if \( \eta_1 = \eta_2 = \frac{1}{L} \) we get
\[ \|x^t - x^*\|^2 + \|y^t - y^*\|^2 \leq \left( 1 - \sigma_{\min}(B)/\sigma_{\max}(B) \right)^t \|x^0 - x^*\|^2 + \|y^0 - y^*\|^2, \]

2. if \( \sigma_{\min}(B) > 0, \) and \( \eta_1 = \frac{1}{\sqrt{2\sigma_{\max}(B)^2}}, \) \( \eta_2 = \frac{1}{\sqrt{2\sigma_{\max}(B)^2}} \) then the rate is
\[ \|x^t - x^*\|^2 + \|y^t - y^*\|^2 \leq \left( 1 - \sigma_{\min}(B)^2/4\sigma_{\max}(B)^2 \right)^t \|x^0 - x^*\|^2 + \|y^0 - y^*\|^2. \]

If we denote \( \kappa = \frac{\sigma_{\min}(B)}{\sigma_{\max}(B)} \) as in \([13]\), then the complexity in both cases is \( O(\kappa \log \frac{1}{\varepsilon}) \). However, we provide this result for potentially different stepizes to obtain new insights about how they should be chosen. One can see, in particular, that choosing a huge \( \eta_1 \) is possible if \( \eta_2 \) is chosen small, but not vice versa.

### 2.3 Negative momentum

The work \([8]\) suggests using negative momentum to improve game dynamics and achieve faster convergence of the iterates. We consider using two types of momentum together: \( \beta_1 \) in the first step and \( \beta_2 \) in the second. Detailed investigation on bilinear problems shows that \( \beta_1 \) can be chosen to be positive and \( \beta_2 \) should rather be negative. Intuitively, positive \( \beta_1 \) allows the method to look further ahead, while negative \( \beta_2 \) compensates for inaccuracy in the approximation of implicit update. In Appendix \([A.1]\), we discuss it in more details. See Algorithm \([3]\) for detailed description and our experiments section for some numerical investigations.

\(^1\)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 \( a = -B^\top y \) and \( b = -B^\top x^* \), we have \( (x - x^*)^\top B(y - y^*) = x^\top By + a^\top x + b^\top y + (x^*)^\top B y^* \).
Algorithm 3 The Extragradient Method for Variational Inequality with Momentum.

1: Parameters: $x^0 \in K$, stepsize $\eta > 0$, momentum parameters $\beta_1, \beta_2 \in (-1, 1)$
2: for $t = 0, 1, 2, \ldots$ do
3: Sample $\xi^t$
4: $y^t = \text{prox}_{\eta g}(x^t - \eta F(x^t; \xi^t)) + \beta_1(x^t - x^{t-1})$
5: $x^{t+1} = \text{prox}_{\eta g}(y^t - \eta F(y^t; \xi^t)) + \beta_2(x^t - x^{t-1})$
6: end for

3 Nonconvex extragradient

Since neural networks are not convex, it is desirable to see 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

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

(3)

where $f$ is a smooth but potentially nonconvex function. To show convergence, we need the following standard assumption.

**Assumption 5.** There exists a constant $\sigma > 0$ such that for all $x$$$
\mathbb{E}\|\nabla f(x; \xi) - \nabla f(x)\|^2 \leq \sigma^2.$$

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

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

where $\hat{x}^t$ is sampled uniformly from $\{x^0, \ldots, x^{t-1}\}$.

**Corollary 2.** If we choose $\eta = \Theta(1/(L\sqrt{t}))$, then the rate is $O((f(x^0) - f^*)/\sqrt{t} + \sigma^2/\sqrt{t})$, 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 [5]. 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 [11] on Cifar10 [15] in Appendix B.2, which gives a similar message.

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 2. The results are very encouraging and show the superiority of the proposed approach on this problem.

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 3, 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 optimal value of $\beta_1$ depends on $\eta\sigma_i$ and only for small values is significantly bigger 0. The dark area is where the method diverges.

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

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.

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 the update rule as in Adam [13].

In the first set of experiments, we compared the performance of ExtraAdam [6] and Adam in a Conditional GAN [17] setup on Fashion MNIST [23] dataset. The generator and discriminator were
Figure 3: 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.

| Generator | Discriminator |
|-----------|---------------|
|  
Input: \( z \in \mathbb{R}^{100} \sim \mathcal{N}(0, I) \)  
Embedding layer for the label  
Linear (110 → 256)  
LeakyReLU (negative slope: 0.2)  
Linear (256 → 512)  
LeakyReLU (negative slope: 0.2)  
Linear (512 → 1024)  
LeakyReLU (negative slope: 0.2)  
Linear (1024 → 784)  
\( \text{Tanh} (\cdot) \)  |
|  
Input: \( x \in \mathbb{R}^{1 \times 28 \times 28} \)  
Embedding layer for the label  
Linear (794 → 1024)  
LeakyReLU (negative slope: 0.2)  
Dropout (\( p=0.3 \))  
Linear (1024 → 512)  
LeakyReLU (negative slope: 0.2)  
Dropout (\( p=0.3 \))  
Linear (512 → 256)  
LeakyReLU (negative slope: 0.2)  
Dropout (\( p=0.3 \))  
Linear (1024 → 784)  
\( \text{Sigmoid} (\cdot) \)  |

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

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 are depicted in Figure 4. 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 [8], 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 [16], there was proposed a method that can be seen as a variant of extragradient, in which only one players makes several extra steps.

In the second experiment, following [1], we trained Self Attention GAN [24]. We note that the loss was generally a misleading metric of method comparison, so here we only provide the samples generated after training for two epochs. The results are provided in Figure 5.

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. Sometimes even established results such as effectiveness of negative momentum was not observed and positive momentum would perform better. 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. However, our work was not exclusively motivated by this application, but rather we wanted to fix a serious issue of the already popular extragradient method.
Figure 4: 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) \((5 \cdot 10^{-5}, 10^{-4})\). The first row shows the loss of the generator for different pairs. In the top row, we show the generator loss and in the bottom that of discriminator.

Figure 5: 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.
References

[1] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. *arXiv preprint arXiv:1904.08598*, 2019.

[2] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

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

[4] Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *arXiv preprint arXiv:1807.04252*, 2018.

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

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

[7] Gauthier Gidel, Hugo Berard, Gatan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019.

[8] Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Gabriel Huang, Remi Lepriol, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. *arXiv preprint arXiv:1807.04740*, 2018.

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

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

[11] 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.

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

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] GM Korpelevich. Extragradient method for finding saddle points and other problems. *Matekon*, 13(4):35–49, 1977.

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

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

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

[18] 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.

[19] 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.

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

[21] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. 2013.
[22] 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.

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

[24] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.
Appendix: “Revisiting Stochastic Extragradient”

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 extragradient 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 \( y_{m+1} \overset{def}{=} \text{prox}_{\eta g} (x - \eta F(y_m)) \) for \( m = 1, \ldots, k \) and let \( w \overset{def}{=} \text{prox}_{\eta g} (x - \eta F(w)) \) be the implicit update, where \( \eta > 0 \) is any stepsize. Then,
\[
\| w - y_k \| \leq \eta^k L \| w - x \| .
\]

**Proof.** We show the claim by induction. For \( k = 0 \) it holds simply because \( y_0 \overset{def}{=} x \). If it holds for \( k - 1 \), let us show it for \( k \). By non-expansiveness of the proximal operator we have
\[
\| w - y_k \| = \| \text{prox}_{\eta g} (x - \eta F(w)) - \text{prox}_{\eta g} (x - \eta F(y_{k-1})) \|
\leq \| x - \eta F(w) - (x - \eta F(y_{k-1})) \|
= \eta \| F(w) - F(y_{k-1}) \|
\leq \eta L \| w - y_{k-1} \|
\leq \eta^k L \| w - x \|. 
\]
\[\Box\]

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

**Lemma 1.** Let \( g \) be \( \mu \)-strongly convex and \( z = \text{prox}_{\eta g} (x) \). Then for all \( y \in \mathbb{R}^d \) the following inequality holds:
\[
\langle z - x, y - z \rangle \geq \eta \left( g(z) - g(y) + \frac{\mu}{2} \| z - y \|^2 \right).
\]

**Proof.** The lemma easily follows from the definitions. Indeed, since
\[
z \overset{def}{=} \arg \min \{ \eta g(u) + \frac{1}{2} \| u - x \|^2 \},
\]
we have necessary optimality condition \( 0 \in \eta \partial g(z) + (z - x) \). Thus,
\[
\langle z - x, y - z \rangle + \eta g(y) - \eta g(z) = \eta \langle g(y) - g(z) - \langle \partial g(z), y - z \rangle \rangle \geq \frac{\eta \mu}{2} \| z - y \|^2,
\]
where the last step follows from mere definition of strong convexity. \[\Box\]

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 \( \sigma_x^2 \overset{def}{=} \mathbb{E}[\| F(x; \xi) - F(x) \|^2] < +\infty \), i.e. the variance of \( F \) at \( x \) is bounded. Then,
\[
\mathbb{E} \langle F(x) - F(x; \xi^t), y^t - x \rangle \leq \eta \sigma_x^2 + \frac{1}{4\eta} \| y^t - x^t \|^2.
\]

**Proof.** As \( x^t \) and \( \xi^t \) are independent random variables and \( \mathbb{E} F(x; \xi^t) = F(x) \), we have
\[
\mathbb{E} \langle F(x) - F(x; \xi^t), y^t - x \rangle = \mathbb{E} \langle F(x) - F(x; \xi^t), x^t - x \rangle + \mathbb{E} \langle F(x) - F(x; \xi^t), y^t - x^t \rangle
= \mathbb{E} \langle F(x) - F(x; \xi^t), y^t - x^t \rangle.
\]

11
By Young’s inequality,
\[
\mathbb{E} \langle F(x) - F(x; \xi^t), y^t - x^t \rangle \leq \eta \mathbb{E} \|F(x) - F(x; \xi^t)\|^2 + \frac{1}{4\eta} \|y^t - x^t\|^2
\]
\[
= \eta \sigma_x^2 + \frac{1}{4\eta} \|y^t - x^t\|^2
\]
and the proof is complete.

Now we are ready to prove Theorem 2.

\textbf{Proof.} By Lemma 1 for points \(y^t = \text{prox}_{\eta g}(x^t - \eta F(x^t; \xi^t))\) and \(x^{t+1} = \text{prox}_{\eta g}(x^t - \eta F(y^t; \xi^t))\),
\[
\langle x^{t+1} - x^t + \eta F(y^t; \xi^t), x^* - x^{t+1} \rangle \geq \eta(g(x^{t+1}) - g(x^*) + \frac{\mu}{2} \|x^{t+1} - x^*\|^2)
\]
\[
\langle y^t - x^t + \eta F(x^t; \xi^t), x^{t+1} - y^t \rangle \geq \eta(g(y^t) - g(x^{t+1}) + \frac{\mu}{2} \|x^{t+1} - y^t\|^2).
\]
Summing these two inequalities together and rearranging, we get
\[
\langle x^{t+1} - x^t, x^* - x^{t+1} \rangle + \langle y^t - x^t, x^{t+1} - y^t \rangle + \eta(F(y^t; \xi^t) - F(x^t; \xi^t), y^t - x^{t+1}) + \eta(F(y^t; \xi^t), x^* - y^t) \geq \eta(g(y^t) - g(x^{t+1}) + \frac{\mu}{2} \|x^{t+1} - y^t\|^2).
\]
By substituting \(2(a, b) = \|a + b\|^2 - \|a\|^2 - \|b\|^2\) with \(a = x^t - x^*\) and \(b = x^* - x^{t+1}\), we deduce
\[
(1 + \eta\mu) \|x^{t+1} - x^*\|^2 \leq \|x^t - x^*\|^2 - \|x^t - y^t\|^2 - (1 + \eta\mu) \|x^{t+1} - y^t\|^2 \leq 2\eta(F(y^t; \xi^t) - F(x^t; \xi^t), y^t - x^{t+1}) - 2\eta(F(y^t; \xi^t), x^* - x^t) + g(y^t) - g(x^*)
\]
The first scalar product can be simplifies using Lipschitzness. Since \(F(\cdot; \xi^t)\) is almost surely \(L\)-Lipschitz, by Young’s inequality
\[
2\eta(F(y^t; \xi^t) - F(x^t; \xi^t), y^t - x^{t+1}) \leq \frac{\eta L}{2} \|F(y^t; \xi^t) - F(x^t; \xi^t)\|^2 + 2\eta \|y^t - x^{t+1}\|^2 \leq \frac{\eta L}{2} \|x^{t+1} - y^t\|^2 + \|y^t - x^{t+1}\|^2.
\]
To get rid of the other scalar product, we use monotonicity of \(F(\cdot; \xi^t)\), and then apply strong convexity of \(g\),
\[
(F(y^t; \xi^t), y^t - x^*) + g(y^t) - g(x^*) \geq (F(x^*; \xi^t), y^t - x^*) + g(y^t) - g(x^*)
\]
\[
= (F(x^*; \xi^t), y^t - x^*) + g(y^t) - g(x^*) + (F(x^*; \xi^t) - F(x^*), y^t - x^*) \geq \frac{\mu}{2} \|y^t - x^*\|^2 + (F(x^*; \xi^t) - F(x^*), y^t - x^*)
\]
So far, the proof has not involved any expectation, but now we shall use Lemma 2 to deduce from the produced bounds
\[
(1 + \eta\mu) \mathbb{E} \|x^{t+1} - x^*\|^2 \leq \mathbb{E} \left[\|x^t - x^*\|^2 - \eta\mu(\|y^t - x^*\|^2 + \|x^{t+1} - y^t\|^2)\right] + 2\eta^2 \sigma^2
\]
\[
- (1 - \frac{\eta L}{2}) \mathbb{E} \|y^t - x^t\|^2 \geq 0
\]
\[
\leq \mathbb{E} \left[\|x^t - x^*\|^2 - \eta\mu(\|y^t - x^*\|^2 + \|x^{t+1} - y^t\|^2)\right] + 2\eta^2 \sigma^2.
\]
Using inequality \(\|a\|^2 + \|b\|^2 \geq \frac{1}{2}\|a + b\|^2\), we arrive at
\[
(1 + \frac{3}{2}\eta\mu) \|x^{t+1} - x^*\|^2 \leq \|x^t - x^*\|^2 + 2\eta^2 \sigma^2.
\]
Note that \(\eta \mu \leq 1/2\) and, therefore, \(\frac{1}{1 + 3\eta^2/2} \leq (1 - 2\eta\mu/3)\). The statement of the theorem can be now easily obtained by induction. \qed
Proof of Theorem 3

Let us choose any $x$. Similarly to the proof of Theorem 2, we can obtain from Lemma 1 with $\mu = 0$ 
\[ |x_t^{i+1} - x|^2 \leq |x_t^i - x|^2 - |x_t^i - y|^2 + 2\eta L(|x_t^{i+1} - y|^2 + |y_t - x|^2) \]
\[ - 2\eta (F(y^i; \xi^i), y_t - x) + g(y^i) - g(x) \]
\[ \leq |x_t^i - x|^2 - \frac{1}{2} |x_t^i - y|^2 - |x_t^{i+1} - y|^2 \]
\[ - 2\eta (F(y^i; \xi^i), y_t - x) + g(y^i) - g(x) \].

By monotonicity of $F(\cdot; \xi^i)$ and Lemma 2, we deduce
\[ \mathbb{E} \langle F(y_t^i; \xi^i), x - y_t^i \rangle \leq \mathbb{E} \langle F(x; \xi^i), x - y_t^i \rangle \]
\[ \leq \eta \sigma^2 + \mathbb{E} \langle F(x), x - y_t^i \rangle + \frac{1}{4\eta} \mathbb{E} |y_t^i - x|^2. \]

Therefore,
\[ \mathbb{E} [g(y^i) - g(x) + \langle F(x), y_t^i - x \rangle] \leq \frac{1}{2\eta} \mathbb{E} [|x_t^i - x|^2 - |x_t^{i+1} - x|^2] + \eta \sigma^2. \]

Telescoping this inequality, we obtain
\[ \mathbb{E} \left[ \frac{1}{t + 1} \sum_{k=0}^t (g(y^k) - g(x) + \langle F(x), y_t^k - x \rangle) \right] \leq \frac{1}{2\eta} \mathbb{E} [x_0 - x]^2 + \eta \sigma^2 \leq \sup_{z \in x^*} \left\{ \frac{1}{2\eta} \|x_0 - z\|^2 + \eta \sigma^2 \right\}. \]

Choosing $\eta = \mathcal{O} \left( \frac{1}{\sqrt{t}} \right)$ and applying Jensen’s inequality to the left-hand side, we get the first claim.

Proof of Theorem 4

Proof. Since the function is bilinear, we can write
\[ \nabla_x f(x, y) = B(y - y^*), \quad \nabla_y f(x, y) = B^\top (x - x^*). \]

Then, we obtain the explicit update rules
\[ x_t^{i+1} = x_t^i - \eta_2 B(y_t^i - y^*) = x_t^i - \eta_2 B(y_t^i - y^* + \eta_1 B^\top (x_t^i - x^*)) \]
\[ y_t^{i+1} = y_t^i + \eta_2 B^\top (u_t^i - x^*) = y_t^i + \eta_2 B^\top (x_t^i - x^* - \eta_1 B(y_t^i - y^*)). \]

In matrix forms it is
\[ \begin{bmatrix} x_t^{i+1} - x^* \\ y_t^{i+1} - y^* \end{bmatrix} = \begin{bmatrix} I - \eta_1 \eta_2 B B^\top \\ \eta_2 B \end{bmatrix} \begin{bmatrix} -\eta_2 B \\ I - \eta_1 \eta_2 B^\top \end{bmatrix} \begin{bmatrix} x_t^i - x^* \\ y_t^i - y^* \end{bmatrix} \]

Apply SVD decomposition to $B$: $B = U \Sigma V^\top$, where $U$ and $V$ are orthogonal and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$. Then,
\[ \left\| \begin{bmatrix} x_t^{i+1} - x^* \\ y_t^{i+1} - y^* \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} I - \eta_1 \eta_2 B B^\top \\ \eta_2 B \end{bmatrix} \begin{bmatrix} -\eta_2 B \\ I - \eta_1 \eta_2 B^\top \end{bmatrix} \right\| \left\| \begin{bmatrix} x_t^i - x^* \\ y_t^i - y^* \end{bmatrix} \right\|. \]

Since $U$ and $V$ are orthogonal, we have
\[ BB^\top = U \Sigma^2 V^\top, \]
\[ B^\top B = V \Sigma^2 U^\top, \]

and
\[ \left\| \begin{bmatrix} I - \eta_1 \eta_2 B B^\top \\ \eta_2 B \end{bmatrix} \right\| = \left\| \begin{bmatrix} U \Sigma^2 \\ V \end{bmatrix} \begin{bmatrix} I - \eta_1 \eta_2 \Sigma^2 \\ \eta_2 \Sigma \end{bmatrix} \begin{bmatrix} U^\top \\ 0 \end{bmatrix} \right\| = \max_i \left\| \begin{bmatrix} 1 - \eta_1 \eta_2 \sigma_i^2 \\ \eta_2 \sigma_i \end{bmatrix} \right\| = \max_i \sqrt{(1 - \eta_1 \eta_2 \sigma_i^2)^2 + \eta_2^2 \sigma_i^2}. \]
A.1 Negative momentum

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

\[ \begin{bmatrix} x^t + \beta_2 y^t - x^* \\ y^t + \beta_1 y^t - y^* \\ x^t - y^* \\ y^t - y^* \end{bmatrix} = \begin{bmatrix} 1 + \beta_2 & -\eta_2 (1 + \beta_1) B & -\beta_2 I & -\eta_2 \beta_1 B \\ 0 & 1 + \beta_1 & -\eta_2 \beta_1 B & -\beta_2 I \\ 0 & 0 & 1 & -\eta_2 \beta_1 B \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x^{t+1} - x^* \\ y^{t+1} - y^* \\ x^t - x^* \\ y^t - y^* \end{bmatrix}. \]

Since up to reshuffling this a block-diagonal matrix, it can be simplified using SVD decomposition to the following 4 x 4 matrices

\[ T_i = \begin{bmatrix} 1 + \beta_2 - \eta_1 \eta_2 \sigma_i^2 & -\eta_2 (1 + \beta_1) \sigma_i & -\beta_2 & -\eta_2 \beta_1 \sigma_i \\ \eta_2 (1 + \beta_1) \sigma_i & 1 + \beta_2 - \eta_1 \eta_2 \sigma_i^2 & -\eta_2 \beta_1 & -\beta_2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \]

where \( \sigma_i \) is the \( i \)-th the singular value of \( \mathbf{B} \).

One can show that the spectral radius of this matrix improves with negative \( \beta_2 \), 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 \( \beta_2 \) when \( \beta = 1 = 0 \) and \( \eta_1 = \eta_2 = \eta \), see Figure 6. 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 \( \beta_1 \) is not very promising and \( \beta_2 \) leads only to a small improvement. Thus, it gives advantage mainly for large values of \( \eta \sigma \).
A.2 Proof of Theorem \[8\]

**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 \( x^t \):

\[
f(x^{t+1}) \leq f(x^t) + \langle \nabla f(x^t), x^{t+1} - x^t \rangle + \frac{L}{2} \|x^{t+1} - x^t\|^2
\]

\[
= f(x^t) - \eta \|\nabla f(x^t)\|^2 + \eta \langle \nabla f(x^t), \nabla f(x^t) - \nabla f(y^t; \xi^t) \rangle + \frac{L\eta^2}{2} \|\nabla f(y^t; \xi^t)\|^2.
\]

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 \( f(\cdot; \xi^t) \)

\[
\eta \langle \nabla f(x^t), \nabla f(x^t) - \nabla f(y^t; \xi^t) \rangle = \mathbb{E} \eta \langle \nabla f(x^t), \nabla f(x^t; \xi^t) - \nabla f(y^t; \xi^t) \rangle
\]

\[
\leq \eta^2 \frac{L}{2} \|\nabla f(x^t)\|^2 + \frac{1}{2L} \mathbb{E} \|\nabla f(x^t; \xi^t) - \nabla f(y^t; \xi^t)\|^2
\]

\[
\leq \eta^2 \frac{L}{2} \|\nabla f(x^t)\|^2 + \frac{L}{2} \mathbb{E} \|x^t - y^t\|^2
\]

\[
= \frac{\eta^2 L}{2} \|\nabla f(x^t)\|^2 + \frac{\eta^2 L}{2} \mathbb{E} \|\nabla f(y^t; \xi^t)\|^2.
\]

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

\[
\mathbb{E} \|\nabla f(y^t; \xi^t)\|^2 \leq 2 \mathbb{E} \|\nabla f(y^t; \xi^t) - \nabla f(x^t; \xi^t)\|^2 + 2 \mathbb{E} \|\nabla f(x^t; \xi^t)\|^2
\]

\[
\leq 2L^2 \mathbb{E} \|y^t - x^t\|^2 + 2 \mathbb{E} \|\nabla f(x^t; \xi^t)\|^2
\]

\[
= 2(1 + L^2 \eta^2) \mathbb{E} \|\nabla f(x^t; \xi^t)\|^2
\]

\[
\leq 2(1 + L^2 \eta^2)(\|\nabla f(x^t)\|^2 + \sigma^2).
\]

Thus,

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

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

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

Telescoping this inequality from 0 to \( t - 1 \), we get

\[
\frac{1}{t} \sum_{k=0}^{t-1} \|\nabla f(x^k)\|^2 \leq \frac{5}{\eta}(f(x^0) - f(x^t)) + 11\eta L \sigma^2
\]

\[
\leq \frac{5}{\eta}(f(x^0) - f^*) + 11\eta L \sigma^2.
\]

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

\( \square \)

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 \[8\].

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 \[8\] 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.

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.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 [11], Resnet-18, on Cifar10 [15] 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.