Last Iterate Convergence of the Extragradient Method for Variationally Coherent Min-Max Problems

Rumen Dangovski
rumenrd@mit.edu

Ileana Rugina
irugina@mit.edu

Kristian Georgiev
krisgrg@mit.edu

Abstract

In this work we analyze the behaviour of the last iterate of the Extragradient (EG) and Proximal Point (PP) algorithms on smooth variationally coherent problems[1]. In particular, we examine a relaxation of the conditions in Golowich et al. (2020), who show a rate of $\mathcal{O}\left(1/\sqrt{T}\right)$ for both algorithms when the problem is convex-concave. We show that best iterate convergence at a rate of $\mathcal{O}\left(1/\sqrt{T}\right)$ naturally carries over to to variationally coherent problems but key monotonicity properties are lost.

1 Introduction

Min-max problems of the type

$$\min_{\mathbf{x}\in\mathbb{R}^m}\max_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x},\mathbf{y}) \tag{MM}$$

have been extensively studied in the literature and have played a key role in many areas of math and science (Sion et al., 1958). Numerous first-order iterative algorithms have been explored as a means to approximate the solution of (MM) (Rockafellar, 1976; Popov, 1980; Korpelevich, 1976). Since the convex-concave setting was of primary interest, convergence analysis was developed for the ergodic (averaged) iterates (Nemirovski, 2004).

More recently, applications in machine learning, namely generative adversarial networks (GANs) (Goodfellow et al., 2014) and adversarially robust training (Madry et al., 2017) among others, have introduced important min-max problems that are not convex-concave. This has motivated exploring non-ergodic modes of convergence of the above algorithms. Their empirical utility has been recently demonstrated in GAN training (Daskalakis et al., 2017).

Mertikopoulos et al. (2018) show convergence of EG in variationally coherent problems (which they name coherent) without an explicit convergence rate. In a later work, Golowich et al. (2020) show that in convex-concave problems, PP enjoys last iterate convergence at a rate of $\mathcal{O}\left(1/\sqrt{T}\right)$ and the same is true for EG under sufficient smoothness conditions, which we discuss further in Section 2. While faster rates of last-iterate convergence are known for the strongly convex-concave setting, (Golowich et al., 2020) show that the $\mathcal{O}\left(1/\sqrt{T}\right)$ bound is tight without this further assumption.

In this paper we relax the convex-concave conditions to a natural generalization based on the variational inequality from convex analysis $\langle\nabla f(\mathbf{x}), \mathbf{x}-\mathbf{x}^*\rangle$ for a global optimum $\mathbf{x}^*$. In partcular,

[1] This paper is compiled for the MIT course 6.881

for (MM) we require that

$$\left\langle \begin{pmatrix} \nabla_x f(\mathbf{x}, \mathbf{y}) \\ \nabla_y f(\mathbf{x}, \mathbf{y}) \end{pmatrix}, (\mathbf{x}, \mathbf{y}) - (\mathbf{x}^*, \mathbf{y}^*) \right\rangle \geq 0, \tag{1}$$

for any $(\mathbf{x}^*, \mathbf{y}^*)$ such that $f(\mathbf{x}^*, \mathbf{y}) \leq f(\mathbf{x}^*, \mathbf{y}^*) \leq f(\mathbf{x}, \mathbf{y}^*)$ for all $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n$. While multiple names for this condition exist in the literature, we follow (Zhou et al., 2020) and refer to that condition as **variational coherence** (VC). Our main motivation, similarly to most recent body of works analyzing last-iterate convergence, is to make a step towards closing the gap between the empirical use of first-order iterative algorithms for general min-max problems and their theoretical understanding.

Summary of Contributions

- We show that the result of Mokhtari et al. (2019a) that all iterates $z^{(t)}$ of EG and PP lie in a compact set carries over naturally to the VC case.
- We show that $\nabla f(\mathbf{x}, \mathbf{y})$ is not sufficiently monotonic to have a bound of the type $\|\nabla f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)})\|^2 \leq (1 + c \cdot F) \cdot F$ for some $c \in \mathbb{R}$ and $F = \nabla f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$. This type of bound is key in translating best-iterate convergence to last-iterate convergence in (Golowich et al., 2020). As a result, considering the Hamiltonian of f as a measurement of convergence would likely not yield a last-iterate result.

Summary of Questions/ Directions for Further Steps

- Do you think there are any other convergence metrics that could be relevant to consider? Monotonicity of $\{||z^{(t)} - z^*||\}$ is available; does that mean anything in particular (in sufficiently well-conditioned problems)?
- Do you believe the arguments presented below are sufficient evidence that last-iterate results cannot be obtained for EG/ PP for VC problems?

The rest of the paper is organized as follows. In Section 2 we introduce the necessary notation and definitions. In Section 3 we briefly discuss the PP and EG algorithms. In Sections 5 and 6 we discuss our results for best-iterate, and last-iterate convergence, respectively.

2 Problem Statement

We borrow a lot of the notation from (Golowich et al., 2020) and (Mokhtari et al., 2019b). With that in mind, we optimize the following objective

$$\min_{\mathbf{x} \in \mathbb{R}^m} \max_{\mathbf{y} \in \mathbb{R}^n} f(\mathbf{x}, \mathbf{y}), \tag{2}$$

where f is smooth and $\langle F(\mathbf{z}), \mathbf{z} - \mathbf{z}^* \rangle \geq 0$ for any $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^m \times \mathbb{R}^n$ and any $\mathbf{z}^*$ satisfying (2) where $F(\mathbf{z}) \equiv F(\mathbf{x}, \mathbf{y}) = [\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})]^T$, and n is an even positive integer.

Remark 1. *Here we relaxed the convex-concave condition in (Golowich et al., 2020) to variational coherence, as discussed in (Zhou et al., 2020) and (Mertikopoulos et al., 2018).*

Remark 2. *Note that in general we do not have that F is a monotone operator. Hence, one cannot directly leverage the approach taken for Theorem 10 in (Golowich et al., 2020).*

Remark 3. *We highlight that $\mathcal{F}^{\text{convex}} \subset \mathcal{F}^{\textit{var. coh.}}$, hence the lower bound from Theorem 9 in Golowich et al. (2020) is still valid for EG on our objective.*

Finally, we note that due to Corollary 2.3 in (Zhou et al., 2017) similarly we have that the variationally coherent objective (2), admits an $\arg\min_{\mathbf{x} \in \mathbb{R}^n} \max_{\mathbf{y} \in \mathbb{R}^n}$ that is convex and compact.

2.1 Assumptions

We present a few assumptions that are standard within the literature.

Assumption 1. F is variationally coherent, i.e. F satisfies the condition $\langle F(\mathbf{z}), \mathbf{z} - \mathbf{z}^* \rangle \geq 0$.

Assumption 2. For some $L > 0$, the operator F is L**-Lipschitz**, i.e., for all $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$, we have that $\|F(\mathbf{z}) - F(\mathbf{z}')\| \leq L\|\mathbf{z} - \mathbf{z}'\|$.

Assumption 3. For some $\Lambda > 0$, the operator F has a **Λ-Lipschitz derivative**, i.e., for all $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$, we have that $\|\partial F(\mathbf{z}) - \partial F(\mathbf{z}')\|_\sigma \leq \Lambda\|\mathbf{z} - \mathbf{z}'\|$, where $\|\cdot\|_\sigma$ denotes the spectral norm and $\partial F \in \mathbb{R}^{n\times n}$ is the matrix of partial derivatives of F, that is $(\partial F)_{i,j} = \partial F_i(\mathbf{z})/\partial \mathbf{z}_j$.

2.2 Evaluating a solution

Golowich et al. (2020) discuss two standard measures of a solution's quality for saddle-point problems. The first one looks at the **Hamiltonian**:

$$\mathrm{Ham}_f(\mathbf{x}, \mathbf{y}) = \|F_f(\mathbf{z})\| = \|\nabla_\mathbf{x} f(\mathbf{x}, \mathbf{y})\|^2 + \|\nabla_\mathbf{y} f(\mathbf{x}, \mathbf{y})\|^2, \tag{3}$$

which achieves its minimal value 0 for global saddle points.

They also define the **primal-dual gap** w.r.t. a *convex region* $\mathcal{X}' \times \mathcal{Y}' \subseteq \mathcal{Z}$:

$$\mathrm{Gap}_f^{\mathcal{X}' \times \mathcal{Y}'}(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{y}' \in \mathcal{Y}'} f(\mathbf{x}, \mathbf{y}') - \min_{\mathbf{x}' \in \mathcal{X}'} f(\mathbf{x}', \mathbf{y}). \tag{4}$$

We follow (Golowich et al., 2020) and focus on the former evaluation of an iterate's quality.

3 Proximal Point and Extragradient Algorithms

Mokhtari et al. (2019) study gradient descent ascent (GDA) based methods for solving saddle point problems in convex-concave and bilinear settings. They exemplify scenarios in which simple GDA diverges while proximal point (PP) does not and then prove the latter's convergence.

The PP algorithm is an implicit optimization algorithm used here as a benchmark. For saddle point problems its iterates $\{\mathbf{x}_{k+1}, \mathbf{y}_{k+1}\}$are defined as the unique solution to:

$$\min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ f(\mathbf{x}, \mathbf{y}) + \frac{1}{2\eta}||\mathbf{x} - \mathbf{x}_k||_2^2 - \frac{1}{2\eta}||\mathbf{y} - \mathbf{y}_k||_2^2 \right\},$$

which can be rewritten as

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta\nabla_\mathbf{x} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}), \quad \mathbf{y}_{k+1} = \mathbf{y}_k + \eta\nabla_\mathbf{y} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$$

using optimality conditions.

They then look at two other methods, optimistic gradient descent ascent (OGDA) and extra-gradient (EG) algorithms, which converge in situations where simple GDA does not. They show both of these can be interpreted as approximations to PP and prove convergence results by analysing them as noisy approximations to PP. We focus here on EG, which calculates a midpoint and then uses the gradient evaluated at this midpoint to perform an update step:

$$\mathbf{x}_{k+1/2} = \mathbf{x}_k - \eta\nabla_\mathbf{x} f(\mathbf{x}_k, \mathbf{y}_k), \quad \mathbf{y}_{k+1/2} = \mathbf{y}_k + \eta\nabla_\mathbf{y} f(\mathbf{x}_k, \mathbf{y}_k)$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta\nabla_\mathbf{x} f(\mathbf{x}_{k+1/2}, \mathbf{y}_{k+1/2}), \quad \mathbf{y}_{k+1} = \mathbf{y}_k + \eta\nabla_\mathbf{y} f(\mathbf{x}_{k+1/2}, \mathbf{y}_{k+1/2}).$$

4 Related Work

In our work we revisit several results from (Golowich et al., 2020). Since our arguments rely heavily on these results we list them below.

Lemma 1 (Golowich et al. (2020))**.** *For all $\mathbf{z} \in \mathcal{Z}$, there are some matrices $\mathbf{A_z}, \mathbf{B_z}$ whose eigenvalues have non-negative real parts so that*

$$F(\mathbf{z} - \eta F(\mathbf{z} - \eta F(\mathbf{z}))) = F(\mathbf{z}) - \eta\mathbf{A_z}F(\mathbf{z}) + \eta^2 F(\mathbf{z}) + \eta^2\mathbf{A_z}\mathbf{B_z}F(\mathbf{z}) \tag{5}$$

and

$$\|\mathbf{A_z} - \mathbf{B_z}\|_\sigma \leq \frac{\eta\Lambda}{2}\|F(\mathbf{z}) - F(\mathbf{z} - \eta F(\mathbf{z}))\|, \quad \|\mathbf{A_z}\|_\sigma \leq L, \quad \|\mathbf{B_z}\|_\sigma \leq L. \tag{6}$$

Lemma 2 (Golowich et al. (2020)). *Suppose* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ *are matrices whose eigenvalues have non-negative real parts and* $\|\mathbf{A}\|_\sigma, \|\mathbf{B}\|_\sigma \leq 1/30$. *Then*

$$\|I - \mathbf{A} + \mathbf{AB}\|_\sigma \leq \sqrt{1 + 26\|\mathbf{A} - \mathbf{B}\|_\sigma^2}.$$

Lemma 3 (Golowich et al. (2020); Nesterov (2006)). *If* $\mathcal{Z} \subset \mathbb{R}^n$ *and* $F \colon \mathcal{Z} \to \mathbb{R}^n$ *is monotone, then for any* $\mathbf{z}, \mathbf{w} \in \mathbb{R}^n$, $\mathbf{z}^\top(\partial F(\mathbf{w})\mathbf{z} \geq 0$. *Equivalently, all eigenvalues of* $\partial F(\mathbf{w})$ *have positive real part.*

Lemma 4 (Golowich et al. (2020)). *Let* $\mathbf{S}, \mathbf{R} \in \mathbb{R}^{n \times n}$ *be (symmetric) PSD matrices. Then*

$$\mathbf{SR} + \mathbf{RS} \preceq 4\mathbf{S}^2 + 4\|\mathbf{S} - \mathbf{R}\|_\sigma^2 \cdot I. \tag{7}$$

5 Best Iterate Analysis

5.1 Extragradient

We first follow Golowich et al. (2020)'s theorem 10 proof to show that their results carries to the more general class of variationally coherent problems. First line of (60) in Lemma 5(b) in (Mokhtari et al., 2019a) follows due to variational coherence, which implies that the Lemma (which uses monotonicity in the above-mentioned line) follows also for variational coherence. Hence, we have that

$$\sum_{t=0}^{T-1} \eta^2 \|F(\mathbf{z}^{(t)})\|^2 = \sum_{t=0}^{T-1} \|\mathbf{z}^{(t)} - \mathbf{z}^{(t+1/2)}\|^2 \leq \frac{\|\mathbf{z}_0 - \mathbf{z}^*\|^2}{1 - \eta^2 L^2} \leq \frac{D^2}{1 - \eta^2 L^2}. \tag{8}$$

Therefore, there is a t^* between zero and $T-1$ for which

$$\|F(\mathbf{z}^{(t^*)})\|^2 \leq \frac{D^2}{\eta^2(1 - \eta^2 L^2)T}. \tag{9}$$

Hence, given the existence of the $O\left(1/\sqrt{T}\right)$ lower bound for variationally coherent problems, from (9) we obtain that *best iterate* converges as $O\left(1/\sqrt{T}\right)$, which parallels (22) in (Golowich et al., 2020).

5.2 Proximal Point

Similarly, for PP it naturally follows that

$$\|F(\mathbf{z}^{(t^*)})\| \leq \frac{D}{\eta} \cdot \frac{1}{\sqrt{T}}, \tag{10}$$

i.e. the best-iterate convergence matches the convex-concave setting. In Section 6.1 we provide a discussion on the last-iterate convergence of PP.

6 Last Iterate Analysis

To extend the analysis from the previous section, we need to show that $\|F(\mathbf{z}^{(t)})\|$ is not too far from $\|F(\mathbf{z}^{(t^*)})\|$ for all $t^* \leq t \leq T$. For that purpose, we are formulating an analogue to Theorem 10 for the upper bound as a conjecture.

Conjecture 1. *Suppose that* F *is a variationally coherent (Assumption 1) that is L-Lipschitz (Assumption 2) and has* Λ*-Lipschitz derivative (Assumption 3). Fix some* $\mathbf{z}^{(0)} \in \mathbb{R}^n$, *and suppose there is some* $\mathbf{z}^* \in \mathbb{R}^n$ *so that* $F(\mathbf{z}^*) = 0$ *and* $\|\mathbf{z}^* - \mathbf{z}\| \leq D$. *If the extragradient algorithm with step size* $\eta \leq \min\{5/(\Lambda D), 1/(30L)\}$ *is initialized at* $\mathbf{z}^{(0)}$, *then its iterates* $\mathbf{z}^{(T)}$ *satisfy*

$$\|F(\mathbf{z}^{(T)})\| \leq \frac{2D}{\eta\sqrt{T}}. \tag{11}$$

To prove the conjecture we need to obtain Lemma 1 for which we need to prove Lemma 3 in (Golowich et al., 2020) by replacing the property of monotonicity with variational coherence. This is not true unfortunately since the Lemma 3 is *if and only if*[2]. Then in Lemma 1 we cannot ensure that the eigenvalues of the matrices $\partial F(\mathbf{z} - (1-\alpha)\eta F(\mathbf{z} - \eta F(\mathbf{z})))$ and $\partial F(\mathbf{z} - (1-\alpha)\eta F(\mathbf{z}))$ have non-negative real parts. Hence, our works needs to start from seeing how breaking Lemma 1 would affect the rest of the conjecture.

In particular, the modified Lemma 1 now states

Lemma 5. *For all $z \in \mathcal{Z}$, there are some matrices $\mathbf{A}_z, \mathbf{B}_z$ s.t.*

$$F(\mathbf{z} - \eta F(\mathbf{z} - \eta F(\mathbf{z}))) = F(\mathbf{z}) - \eta \mathbf{A}_\mathbf{z} F(\mathbf{z}) + \eta^2 F(\mathbf{z}) + \eta^2 \mathbf{A}_\mathbf{z}\mathbf{B}_\mathbf{z} F(\mathbf{z}) \tag{12}$$

and

$$\|\mathbf{A}_\mathbf{z} - \mathbf{B}_\mathbf{z}\|_\sigma \leq \frac{\eta\Lambda}{2}\|F(\mathbf{z}) - F(\mathbf{z} - \eta F(\mathbf{z}))\|, \;\; \|\mathbf{A}_\mathbf{z}\|_\sigma \leq L, \;\; \|\mathbf{B}_\mathbf{z}\|_\sigma \leq L. \tag{13}$$

In other words, we lose the fact that $\mathbf{A}_\mathbf{z}$ and $\mathbf{B}_\mathbf{z}$ have positive real parts of their eigenvalues. This directly interferes with the proof of Lemma 2, which heavily relies on $\mathbf{A} + \mathbf{A}^\top$ and $\mathbf{B} + \mathbf{B}^\top$ being PSD. From L-Lipschitzness of F, however, we have $-L \leq \|\mathbf{A}_\mathbf{z}\|_\sigma \leq L$ and $-L \leq \|\mathbf{B}_\mathbf{z}\|_\sigma \leq L$. Hence we have $\mathbf{A}_\mathbf{z} + L \cdot I \succeq 0, \mathbf{B}_\mathbf{z} + L \cdot I \succeq 0$.

With this, we can attempt to show, instead of (37) from (Golowich et al., 2020), the following

$$(\mathbf{A} + \mathbf{A}^\top) - (\mathbf{A}\mathbf{B} + \mathbf{B}^\top\mathbf{A}^\top) - (1 + 2L + L^2)\mathbf{A}\mathbf{A}^\top \succeq -26\|\mathbf{A} - \mathbf{B}\|_\sigma^2 I. \tag{14}$$

Now, let us consider $\mathbf{S} := (\mathbf{A} + \mathbf{A}^\top)/2$ and $\mathbf{R} := (\mathbf{B} + \mathbf{B}^\top)/2$. The question is how to improve Lemma 4, since Lemma 4 is the only place that requires that $\mathbf{S}$ and $\mathbf{R}$ are PSD. Note that $\mathbf{S}$ and $\mathbf{R}$ are not PSD, but if we add L to each of them, then they both become PSD. Thus, for our non-PSD Lemma 4 now gives

$$\mathbf{S}\mathbf{R} + \mathbf{R}\mathbf{S} - 2L^2 \cdot I + 6L\mathbf{S} - 2L\mathbf{R} \preceq 4\mathbf{S}^2 + 4\|\mathbf{S} - \mathbf{R}\|_\sigma^2 \cdot I. \tag{15}$$

Using (15) we have from (38), (39) and (40) in Golowich et al. (2020) that

$$\begin{aligned}(1 + L^2)\mathbf{A}\mathbf{A}^\top + \mathbf{A}\mathbf{B} + \mathbf{B}^\top\mathbf{A}^\top &\preceq \\ 26\|\mathbf{A} - \mathbf{B}\|_\sigma^2 \cdot I + \mathbf{A} + \mathbf{A}^\top &- 2L^2 \cdot I + 3L(\mathbf{A} + \mathbf{A}^\top) - L(\mathbf{B} + \mathbf{B}^\top).\end{aligned} \tag{16}$$

With this in mind we have that (14) is satisfied if

$$2L^2 \cdot I - 3L(\mathbf{A} + \mathbf{A}^\top) + L(\mathbf{B} + \mathbf{B}^\top) - 2L\mathbf{A}^2 \succeq 0,$$

which simplifies to showing that

$$2L \cdot I - 3(\mathbf{A} + \mathbf{A}^\top) + (\mathbf{B} + \mathbf{B}^\top) - 2\mathbf{A}^2 \succeq 0, \tag{17}$$

which unfortunately is not true in general. To be precise, it would suffice to have

$$2L \cdot I - 3(\mathbf{A} + \mathbf{A}^\top) + (\mathbf{B} + \mathbf{B}^\top) - 2\mathbf{A}^2 \succeq C\|F(\mathbf{z}^{(t)})\|, \tag{18}$$

for some $C \in \mathbb{R}$, that we can later absorb into the bound for η.

Stepping back, (18) puts the additional rather unnatural constraint

$$2L \cdot I - 3(\mathbf{A}_\mathbf{z} + \mathbf{A}_\mathbf{z}^\top) + (\mathbf{B}_\mathbf{z} + \mathbf{B}_\mathbf{z}^\top) - 2\mathbf{A}_\mathbf{z}\mathbf{A}_\mathbf{z}^\top \succeq C\|F(\mathbf{z}^{(t)})\|^2, \tag{19}$$

for $\mathbf{A}_\mathbf{z} = \int_0^1 \partial F(\mathbf{z} - (1-\alpha)\eta F(\mathbf{z} - \eta F(\mathbf{z})))d\alpha$ and $\mathbf{B}_\mathbf{z} = \int_0^1 \partial F(\mathbf{z} - (1-\alpha)\eta F(\mathbf{z}))d\alpha$, and $\mathbf{z} = \mathbf{z}^{(t)}$.

In the next subsection, we provide an argument demonstrating the inability to control $\|F(\mathbf{z}^{(T)})\|$ for VC f.

[2] See for example, https://math.stackexchange.com/questions/1752317/show-that-monotonicity-implies-positive-definiteness-of-the-jacobian

6.1 Proximal Point

It is clear that we lose the monotonicity of F. Hence, even for PP, we would need to resort to bounding the deviation of the last iterate from the best iterate. Below we sketch out a simple example that illustrates that we might not be able to get useful bounds for the VC case. Intuitively, the setup below highlights that without violating the result that iterates do not leave a compact set, VC is not enough to control $\|F(z)\|$.

In particular, let $\mathbf{z}^{(t^*)}$ be the best iterate. We would like $\Delta^T \|F(\mathbf{z}^*)\| = \mathcal{O}\left(1/\sqrt{T}\right)$ for $\Delta := \|F(\mathbf{z}^{(t+1)})\| / \|F(\mathbf{z}^{(t)})\|$.

Now consider for simplicity $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^2 \cong \mathcal{C}$, there exists a unique $\mathbf{z}^*$ and we have $\mathbf{z}^{(t^*)}$ on the unit sphere $\partial \mathcal{B}_1(\mathbf{z}^*)$. Now let $F_{\mathbf{v}}$ denote the directional equivalent of F. Then suppose $F_{\mathbf{d}_\theta} e^{i\theta} = \epsilon > 0$ (with slight abuse of notation) where $\mathbf{d}_\theta := e^{i\theta} - \mathbf{z}^*$.

Now we can identify the restriction of f to S_1 with a 2π-periodic function $g(\theta)$. By making ϵ arbitrarily small, the smoothness of g is effectively controlled only by the smoothness of f. In particular, let WLOG $g := \sin(C\theta)$ and $\mathbf{z}^{(t^*)} = e^{i(\pi/2+\delta)}$ for some small $\delta > 0$. Then $|F_{|S_1}(\mathbf{z}^{(t^*)})| \approx C \cdot \delta$. By having a non-negative derivative in this direction, we can inflate $\|F(z^{(t^*+1)})\|$ by an arbitrary amount, controlled only by L, thus violating the desired conditions for Δ.

7 Conclusion

In this work we showed that the $\mathcal{O}\left(1/\sqrt{T}\right)$ best-iterate convergence guarantees of the Extragradient and Proximal Point algorithms for convex-concave problems can be extended for the considerably broader class of variationally coherent problems. In addition, we provide arguments suggesting the same cannot be said about last-iterate convergence.

8 Acknowledgements

The authors would like to express their gratitude towards the staff of 6.881 for their continual support throughout the semester. In particular, we would like to thank Sarath Pattathil for the insightful feedback.

References

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Golowich, N., Pattathil, S., Daskalakis, C., and Ozdaglar, A. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems, 2020.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Korpelevich, G. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mertikopoulos, P., Zenati, H., Lecouat, B., Foo, C., Chandrasekhar, V., and Piliouras, G. Mirror descent in saddle-point problems: Going the extra (gradient) mile. *CoRR*, abs/1807.02629, 2018. URL http://arxiv.org/abs/1807.02629.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach. *arXiv e-prints*, art. arXiv:1901.08511, January 2019.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems, 2019a.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach, 2019b.

Nemirovski, A. 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.

Nesterov, Y. Cubic regularization of newton's method for convex problems with constraints. *Universit catholique de Louvain, Center for Operations Research and Econometrics (CORE), CORE Discussion Papers*, 01 2006. doi: 10.2139/ssrn.921825.

Popov, L. D. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28:845–848, 1980.

Rockafellar, R. T. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

Sion, M. et al. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S., and Glynn, P. W. Stochastic mirror descent in variationally coherent optimization problems. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 7040–7049. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7279-stochastic-mirror-descent-in-variationally-coherent-optimization-problems.pdf.

Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S. P., and Glynn, P. W. On the convergence of mirror descent beyond stochastic convex programming. *SIAM Journal on Optimization*, 30(1):687–716, 2020.