Efficient Methods for Structured
Nonconvex-Nonconcave Min-Max Optimization

Jelena Diakonikolas
University of Wisconsin-Madison
jelena@cs.wisc.edu

Constantinos Daskalakis
Massachusetts Institute of Technology
costis@csail.mit.edu

Michael I. Jordan
UC Berkeley
jordan@cs.berkeley.edu

Abstract

The use of min-max optimization in the adversarial training of deep neural network classifiers, and the training of generative adversarial networks has motivated the study of nonconvex-nonconcave optimization objectives, which frequently arise in these applications. Unfortunately, recent results have established that even approximate first-order stationary points of such objectives are intractable, even under smoothness conditions, motivating the study of min-max objectives with additional structure. We introduce a new class of structured nonconvex-nonconcave min-max optimization problems, proposing a generalization of the extragradient algorithm which provably converges to a stationary point. The algorithm applies not only to Euclidean spaces, but also to general ℓ_p-normed finite-dimensional real vector spaces. We also discuss its stability under stochastic oracles and provide bounds on its sample complexity. Our iteration complexity and sample complexity bounds either match or improve the best known bounds for the same or less general nonconvex-nonconcave settings, such as those that satisfy variational coherence or in which a weak solution to the associated variational inequality problem is assumed to exist.

1   Introduction

Min-max optimization and min-max duality theory lie at the foundations of game theory and mathematical programming, and have found far-reaching applications across a range of disciplines, including complexity theory, statistics, control theory, and online learning theory. Most recently, min-max optimization has played an important role in machine learning, notably in the adversarial training of deep neural network classifiers and the training of generative deep neural network models. These recent applications have heightened the importance of solving min-max optimization problems with nonconvex-nonconcave objectives, taking the following general form:

$$\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}), \tag{1.1}$$

where $\mathbf{x}$ and $\mathbf{y}$ are real-valued vectors and f is not (necessarily) convex in $\mathbf{x}$ for all $\mathbf{y}$ and/or not (necessarily) concave in $\mathbf{y}$ for all $\mathbf{x}$. There may also be constraints on $\mathbf{x}$ and $\mathbf{y}$, and in many applications $\mathbf{x}$ and $\mathbf{y}$ are high-dimensional vectors.

When the objective function is not convex-concave, von Neumann's celebrated min-max theorem fails to apply, and so do most standard optimization methods for solving (1.1). This has motivated several lines of investigation, which include extensions of the min-max theorem beyond convex-concave objectives (e.g. Sion's theorem for quasiconvex-quasiconcave objectives), and the pursuit of computational procedures that target solutions to (1.1) even in the absence of a min-max theorem; see Section 1.1 for a review of recent work. Of course, without strong assumptions on f, (1.1) is an intractable problem, at least as intractable as general nonconvex optimization. Thus, the literature has targeted locally optimal solutions, in the same spirit as the targeting of local optima in non-convex optimization. Naturally, there are various notions of local optimality that have been studied in the literature. Our focus here will be on the simplest such notion, namely first-order local optimality, for which, despite the apparent simplicity, many challenges arise (Daskalakis and Panageas, 2018, Mazumdar et al., 2020).

In contrast to classical optimization problems, where useful results can be obtained with very mild assumptions on the objective function, in min-max optimization it is necessary to impose non-trivial assumptions on f, even when the goal is only to compute locally optimal solutions. Indeed, Daskalakis et al. (2021) establish intractability results in the constrained setting of the problem, wherein first-order locally optimal solutions are guaranteed to exist whenever the objective is smooth. Moreover, they show that even the computation of *approximate* solutions is PPAD-complete and, if the objective function is accessible through value-queries and gradient-queries, exponentially many such queries are necessary (in particular, exponential in at least one of the following: the inverse approximation parameter, the smoothness constant of f, or the diameter of the constraint set).

We expect similar intractability results to hold in the unconstrained case, which is the case considered in this paper, even when restricting to smooth objectives that have a non-empty set of optimal solutions.[1] Indeed, fixed-point complexity-based intractability results for the constrained case are typically extendable to the unconstrained case, by embedding the hard instances within an unbounded domain.

Relatedly, we already know that the unconstrained Stampacchia variational inequality (SVI) problem for Lipschitz continuous operators $F : \mathbb{R}^d \to \mathbb{R}^d$—a problem which includes the unconstrained case of (1.1) by setting $F(\begin{bmatrix}\mathbf{x}\\\mathbf{y}\end{bmatrix}) = \begin{bmatrix}\nabla_\mathbf{x} f(\mathbf{x},\mathbf{y})\\-\nabla_\mathbf{y} f(\mathbf{x},\mathbf{y})\end{bmatrix}$—is computationally intractable, even when restricting to operators that have a non-empty set of SVI solutions.[2] This is because: (i) F is Lipschitz-continuous if and only if the operator $T(\mathbf{u}) = \mathbf{u} - F(\mathbf{u})$ is Lipschitz-continuous; (ii) for $\epsilon \geq 0$, points $\bar{\mathbf{u}} \in \mathbb{R}^d$ such that $\|F(\bar{\mathbf{u}})\|_2 \leq \epsilon$ satisfy $\|T(\bar{\mathbf{u}}) - \bar{\mathbf{u}}\|_2 \leq \epsilon$, i.e. they are ϵ-approximate fixed points of T, and vice versa; and (iii) it is known that finding approximate fixed points of Lipschitz operators over $\mathbb{R}^d$ is PPAD-hard, even when the operators are guaranteed to have fixed points (Papadimitriou, 1994). Moreover, if we restrict attention to algorithms that only make value queries to T (i.e. F, which corresponds to the type of access that all first-order algorithms have), the query complexity becomes exponential in the dimension (Hirsch et al., 1989). Finally, by the equivalence of norms, these results extend to arbitrary ℓ_p-normed finite dimensional real vector spaces. Of course, for these intractability results for SVI to apply to the nonconvex-nonconcave min-max problem (1.1), one would need to prove that these complexity results extend to operators F constructed from a smooth function f by setting $F(\begin{bmatrix}\mathbf{x}\\\mathbf{y}\end{bmatrix}) = \begin{bmatrix}\nabla_\mathbf{x} f(\mathbf{x},\mathbf{y})\\-\nabla_\mathbf{y} f(\mathbf{x},\mathbf{y})\end{bmatrix}$.

Our contributions. Given the aforedescribed intractability results, our goal is to identify structural properties that make it possible to solve min-max optimization problems with smooth objectives. Focusing on the unconstrained setting of (1.1), we view it as a special case (obtained by considering the operator $F(\begin{bmatrix}\mathbf{x}\\\mathbf{y}\end{bmatrix}) = \begin{bmatrix}\nabla_\mathbf{x} f(\mathbf{x},\mathbf{y})\\-\nabla_\mathbf{y} f(\mathbf{x},\mathbf{y})\end{bmatrix}$) of the unconstrained variational inequality problem (SVI), and consider instead this more general problem. We identify conditions for F under which a generalized version of the extragradient method of Korpelevich (1976), which we propose, converges to a solution of (SVI) (or, in the special case of (1.1), to a stationary point of f) at a rate of $1/\sqrt{k}$ in the number of iterations k. Our condition, presented as Assumption 1, postulates that there exists a solution to (SVI) that only violates the stronger (MVI) requirement in a controlled manner that we delineate. Our generalized extragradient method is based on an aggressive interpolation step, as specified by (EG+), and our main convergence result is Theorem 3.2. We additionally show, in Theorems 4.2 and 4.5, that the algorithm converges in non-Euclidean settings, under the stronger condition that an (MVI) solution exists, or when we only have stochastic oracle access to F (or, in the special case of (1.1), to the gradient of f).

The condition on F under which our main result applies is weaker than the assumption that a solution to (MVI) exists (Malitsky, 2019, Mertikopoulos et al., 2019, Song et al., 2020, Zhou et al., 2017), an assumption which is already satisfied by several interesting families of min-max objectives, including quasiconvex-concave families or starconvex-concave families. Our significantly weaker condition applies in particular to (min-max objectives f with corresponding) operators F that are negatively comonotone (Bauschke et al., 2020) or positively cohypomonotone (Combettes and Pennanen, 2004). These conditions have been studied in the literature for at least a couple of decades, but only asymptotic convergence results were available prior to our work for computing solutions to (SVI). In contrast, our rates are asymptotically identical to the rates that we would get under the stronger assumption that a solution to (MVI) exists, and sidestep the intractability results for (1.1) suggested by Daskalakis et al. (2021) for general smooth objectives.

Table 1: Comparison of iteration complexities required to find a point $\mathbf{x}$ with $\|F(\mathbf{x})\|_{p^*} \leq \epsilon$ using deterministic algorithms, where $\epsilon > 0$, $F : \mathbb{R}^d \to \mathbb{R}^d$ is a Lipschitz operator satisfying Assumption 1 (Section 2) with $\rho \geq 0$. Parameter p determines the ℓ_p setup, and $p^* = \frac{p}{p-1}$ is the exponent conjugate to p. Only the dependence on ϵ and possibly the dimension d is shown; the dependence on other problem parameters is comparable for all the results. $\widetilde{O}$ hides logarithmic factors. '—' indicates that the result does not exist/is not known.

Setup / Paper	$\rho \in (0, \frac{1}{4L})$, $p = 2$	$\rho = 0$, $p = 2$	$\rho = 0$, $p \in (1, 2)$	$\rho = 0$, $p > 2$
(Dang and Lan, 2015)	—	$O(\frac{1}{\epsilon^2})$	$O(\frac{\text{poly}(d^{1/p-1/2})}{\epsilon^2})$	$O(\frac{\text{poly}(d^{1/2-1/p})}{\epsilon^2})$
(Lin et al., 2018)	—	$\widetilde{O}(\frac{1}{\epsilon^2})$	—	—
(Song et al., 2020)	—	$O(\frac{1}{\epsilon^2})$	$O(\frac{1}{\epsilon^2})$	—
This Paper	$O(\frac{1}{\epsilon^2})$	$O(\frac{1}{\epsilon^2})$	$O(\frac{1}{\epsilon^2})$	$O(\frac{1}{\epsilon^p})$

1.1 Further Related Work

A large number of recent works target identifying practical first-order, low-order, or efficient online learning methods for solving min-max optimization problems in a variety of settings, ranging from the well-behaved setting of convex-concave objectives to the challenging setting of nonconvex-nonconcave objectives. There has been substantial work for convex-concave and nonconvex-concave objectives, targeting the computation of min-max solutions to (1.1) or, respectively, stationary points of f or $\Phi(\mathbf{x}) := \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$. This work has focused on attaining improved convergence rates (Alkousa et al., 2019, Azizian et al., 2020, Diakonikolas, 2020, Golowich et al., 2020, Kong and Monteiro, 2019, Lin et al., 2020a,b, Lu et al., 2020, Nouiehed et al., 2019, Thekumparampil et al., 2019, Zhao, 2019) and/or obtaining last-iterate convergence guarantees (Abernethy et al., 2019, Adolphs et al., 2019, Daskalakis and Panageas, 2018, 2019, Daskalakis et al., 2018, Gidel et al., 2019, Hamedani and Aybat, 2018, Liang and Stokes, 2019, Lin et al., 2018, Liu et al., 2020, Mazumdar et al., 2020, Mertikopoulos et al., 2018, Mokhtari et al., 2020).

In the nonconvex-nonconcave setting, research has focused on identifying different notions of local min-max solutions (Daskalakis and Panageas, 2018, Jin et al., 2020, Mangoubi and Vishnoi, 2021, Mazumdar et al., 2020) and studying the existence and (local) convergence properties of learning methods to these points (Mangoubi and Vishnoi, 2021, Mangoubi et al., 2020, Wang et al., 2019). As already discussed, recent work of Daskalakis et al. (2021) shows that, for general smooth objectives, the computation of even approximate first-order locally optimal min-max solutions is intractable, motivating the identification of structural assumptions on the objective function for which these intractability barriers can be bypassed.

An example such assumption, which is closely related to the one made in this work, is that an (MVI) solution exists for the operator $F(\begin{bmatrix}\mathbf{x}\\\mathbf{y}\end{bmatrix}) = \begin{bmatrix}\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\\-\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\end{bmatrix}$, as studied by Lin et al. (2018), Liu et al. (2020), Malitsky (2019), Mertikopoulos et al. (2019), Song et al. (2020), Zhou et al. (2017). As we have already discussed, the assumption we make for our main result in this work is weaker. Table 1 provides a comparison of our results to those of existing works, considering the deterministic setting (i.e. having exact value access to F).

In unconstrained Euclidean setups, the best known convergence rates are of the order $1/\sqrt{k}$ (Dang and Lan, 2015, Song et al., 2020), under the assumption that an (MVI) solution exists. We obtain the same rate under our weaker Assumption 1. Moreover, under our weaker assumption, we show that the accumulation points of the sequence of iterates of our algorithm are (SVI) solutions. This was previously established for alternative algorithms and under the stronger assumption that an (MVI) solution exists (Malitsky, 2019, Mertikopoulos et al., 2019).

When it comes to more general ℓ_p norms, Mertikopoulos et al. (2019) establish the asymptotic convergence of the iterates of an optimistic variant of the mirror descent algorithm, under the assumption that an (MVI) solution exists, but they do not provide any convergence rates. On the other hand, Dang and Lan (2015) prove a $1/\sqrt{k}$ rate of convergence for a variant of the mirror-prox algorithm in general normed spaces. This result, however, requires the regularizing (prox) function to be both smooth and strongly convex w.r.t. the same norm, and the constant in the convergence bound scales at least linearly with the condition number of the prox function. It is well-known that no function can be simultaneously smooth and strongly convex w.r.t. an ℓ_p norm with $p \neq 2$ and have a condition number independent of the dimension (Borwein et al., 2009). In fact, unless p is trivially close to 2, we only know of functions whose

---

[1]Note that these are stationary points of f in this case.

[2]We formally define the *Stampacchia variational inequality* problem, (SVI), in Section 2. We also define the harder *Minty variational inequality* problem, (MVI), in the same section.

condition number would scale polynomially with the dimension.

Very recent (and independent) work of Song et al. (2020) proposes an optimistic dual extrapolation method with linear convergence for a class of problems that have a "strong" (MVI) solution. (In particular, their assumption is that there exists $\mathbf{u}^* \in \mathbb{R}^d$ such that $\forall \mathbf{u} \in \mathbb{R}^d$: $\langle F(\mathbf{u}), \mathbf{u} - \mathbf{u}^* \rangle \geq m\|\mathbf{u} - \mathbf{u}^*\|^2$ for some constant $m \geq 0$; the case $m = 0$ recovers the existence of a standard (MVI) solution.) Their result only applies to norms that are strongly convex, which in the case of ℓ_p norms is true only for $p \in (1, 2]$. In that case, our results match those of Song et al. (2020). For the case of stochastic oracle access to F, our bounds also match those of Song et al. (2020) for $p \in (1, 2]$, and we also handle the case $p > 2$ which is not covered by Song et al. (2020).

Finally, it is worth noting that Malitsky (2019), Mertikopoulos et al. (2019), Song et al. (2020), Zhou et al. (2017) consider constrained optimization setups, which are not considered in our work. We believe that generalizing our results to constrained setups is possible, and defer such generalizations to future work.

2 Notation and Preliminaries

We consider real d-dimensional spaces $(\mathbb{R}^d, \|\cdot\|_p)$, where $\|\cdot\|_p$ is the standard ℓ_p norm for $p \geq 1$. In particular, $\|\cdot\|_2 = \sqrt{\langle \cdot, \cdot \rangle}$ is the ℓ_2 (Euclidean) norm and $\langle \cdot, \cdot \rangle$ denotes the inner product. When the context is clear, we omit the subscript 2 and just write $\|\cdot\|$ for the Euclidean norm $\|\cdot\|_2$. Moreover, we denote by $p^* = \frac{p}{p-1}$ the exponent conjugate to p.

We are interested in finding stationary points for min-max problems of the form:

$$\min_{\mathbf{x} \in \mathbb{R}^{d_1}} \max_{\mathbf{y} \in \mathbb{R}^{d_2}} f(\mathbf{x}, \mathbf{y}), \tag{P}$$

where f is a smooth (possibly nonconvex-nonconcave) function and $d_1 + d_2 = d$. In this case, stationary points can be defined as the points at which the gradient of f is the zero vector. As is standard, the ϵ-approximate variant of this problem for $\epsilon > 0$ is to find a point $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ such that $\|\nabla f(\mathbf{x}, \mathbf{y})\|_{p^*} \leq \epsilon$.

We will study Problem (P) through the lens of variational inequalities, described in Section 2.1. To do so, we consider the operator $F : \mathbb{R}^d \to \mathbb{R}^d$ defined via $F(\mathbf{u}) = \begin{bmatrix} \nabla_\mathbf{x} f(\mathbf{x}, \mathbf{y}) \\ -\nabla_\mathbf{y} f(\mathbf{x}, \mathbf{y}) \end{bmatrix}$, where $\mathbf{u} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$ and where $\nabla_\mathbf{x} f$ (respectively, $\nabla_\mathbf{y} f$) denotes the gradient of f w.r.t. $\mathbf{x}$ (respectively, $\mathbf{y}$). It is clear that F is Lipschitz-continuous whenever f is smooth and that $\|F(\mathbf{u})\|_{p^*} \leq \epsilon$ for $\mathbf{u} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$ holds if and only if $\|\nabla f(\mathbf{x}, \mathbf{y})\|_{p^*} \leq \epsilon$.

2.1 Variational Inequalities and Structured (Possibly Non-Monotone) Operators

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be an operator that is L-Lipschitz-continuous w.r.t. $\|\cdot\|_p$:

$$(\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^d): \quad \|F(\mathbf{u}) - F(\mathbf{v})\|_{p^*} \leq L\|\mathbf{u} - \mathbf{v}\|_p. \tag{A$_1$}$$

F is said to be *monotone* if:

$$(\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^d): \quad \langle F(\mathbf{u}) - F(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \geq 0. \tag{2.1}$$

Given a closed convex set $\mathcal{U} \subseteq \mathbb{R}^d$ and an operator F, the *Stampacchia Variational Inequality* problem consists in finding $\mathbf{u}^* \in \mathbb{R}^d$ such that:

$$(\forall \mathbf{u} \in \mathcal{U}): \quad \langle F(\mathbf{u}^*), \mathbf{u} - \mathbf{u}^* \rangle \geq 0. \tag{SVI}$$

In this case, $\mathbf{u}^*$ is referred to as the *strong solution* to the variational inequality corresponding to F and $\mathcal{U}$. When $\mathcal{U} \equiv \mathbb{R}^d$ (the case considered here), it must be the case that $\|F(\mathbf{u}^*)\|_{p^*} = 0$. We will assume that there exists at least one (SVI) solution, and will denote the set of all such solutions by $\mathcal{U}^*$.

The *Minty Variational Inequality* problem consists in finding $\mathbf{u}^*$ such that:

$$(\forall \mathbf{u} \in \mathcal{U}): \quad \langle F(\mathbf{u}), \mathbf{u}^* - \mathbf{u} \rangle \leq 0, \tag{MVI}$$

in which case $\mathbf{u}^*$ is referred to as the *weak solution* to the variational inequality corresponding to F and $\mathcal{U}$. If we assume that F is monotone, then (2.1) implies that every solution to (SVI) is also a solution to (MVI), and the two solution sets are equivalent. More generally, if F is *not* monotone, all that can be said is that the set of (MVI) solutions is a subset of the set of (SVI) solutions. In particular, (MVI) solutions may not exist even when (SVI) solutions exist. These facts follow from Minty's theorem (see, e.g., (Kinderlehrer and Stampacchia, 2000, Chapter 3)).

We will *not*, in general, be assuming that F is monotone. Note that the Lipschitzness of F on its own is not sufficient to guarantee that the problem is computationally tractable, as discussed in the introduction. Thus, additional structure is needed, which we introduce in the following.

Weak MVI solutions. We define the class of problems with weak (MVI) solutions as the class of problems in which F satisfies the following assumption.

Assumption 1 (Weak MVI). There exists $\mathbf{u}^* \in \mathcal{U}^*$ such that:

$$(\forall \mathbf{u} \in \mathbb{R}^d): \quad \langle F(\mathbf{u}), \mathbf{u} - \mathbf{u}^* \rangle \geq -\frac{\rho}{2} \| F(\mathbf{u}) \|_{p^*}^2, \tag{A_2}$$

for some parameter $\rho \in \left[0, \frac{1}{4L}\right)$.

We will only provide results for $\rho > 0$ in the case of the ℓ_2 norm. For $p \neq 2$, we will require a stronger assumption; namely, that an (MVI) solution exists, which holds when $\rho = 0$.

2.2 Example Settings Satisfying Assumption 1

The class of problems that have weak (MVI) solutions in the sense of Assumption 1 generalizes other structured non-monotone variational inequality problems, as we discuss in this section.

When $\rho = 0$, we recover the class of problems that have an (MVI) solution. This class contains all unconstrained variationally coherent problems studied in, e.g., Mertikopoulos et al. (2019), Zhou et al. (2017), which encompass all min-max problems with objectives that are bilinear, pseudo-convex-concave, quasi-convex-concave, and star-convex-concave.

When $\rho > 0$ and $p = 2$, Assumption 1 is implied by F being $-\frac{\rho}{2}$-comonotone (Bauschke et al., 2020) or $\frac{\rho}{2}$-cohypomonotone (Combettes and Pennanen, 2004), defined as follows:

$$(\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^d):$$
$$\langle F(\mathbf{u}) - F(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \geq -\frac{\rho}{2} \| F(\mathbf{u}) - F(\mathbf{v}) \|_2^2. \tag{2.2}$$

In particular, Assumption 1 is equivalent to requiring that (2.2) be satisfied for general $\mathbf{u}$ and $\mathbf{v} = \mathbf{u}^*$, where $\mathbf{u}^*$ is a solution to (SVI) (in which case $F(\mathbf{u}^*) = \mathbf{0}$). Note that Assumption 1 does *not* imply that a solution to (MVI) exists, unless $\rho = 0$. It is further important to note that cohypomonotone operators arise as inverses of operators that only need to be Lipschitz-continuous. (In fact, even a weaker property suffices; see Bauschke et al. (2020).) This is particularly interesting as, combined with our main result, it implies that we can efficiently find zeros of inverses of Lipschitz-continuous operators, as long as those inverses are sufficiently Lipschitz, even though finding zeros of Lipschitz-continuous operators is computationally intractable, in general, as we have discussed.

It is interesting to note that Assumption 1 does not imply that, in the min-max setting, f is convex-concave (or, more generally, that F is monotone), even in a neighborhood of an (SVI) solution $\mathbf{u}^* = \begin{bmatrix} \mathbf{x}^* \\ \mathbf{y}^* \end{bmatrix}$, i.e., a stationary point of f. To see this, fix $\mathbf{y} = \mathbf{y}^*$ and consider $f(\mathbf{x}, \mathbf{y}^*)$ for $\mathbf{x}$ in a small neighborhood of $\mathbf{x}^*$. Using the fact that a continuously-differentiable function is well-approximated by its linear approximation within small neighborhoods, all that we are able to deduce from Assumption 1 is that

$$f(\mathbf{x}^*, \mathbf{y}^*) - f(\mathbf{x}, \mathbf{y}^*) \approx \left\langle \begin{bmatrix} \nabla_\mathbf{x} f(\mathbf{x}, \mathbf{y}^*) \\ \nabla_\mathbf{y} f(\mathbf{x}, \mathbf{y}^*) \end{bmatrix}, \begin{bmatrix} \mathbf{x}^* - \mathbf{x} \\ \mathbf{y}^* - \mathbf{y}^* \end{bmatrix} \right\rangle \leq \frac{\rho}{2} \| \nabla f(\mathbf{x}, \mathbf{y}^*) \|_{p^*}^2.$$

In particular, Assumption 1 does not preclude that $f(\mathbf{x}^*, \mathbf{y}^*)$ is larger than $f(\mathbf{x}, \mathbf{y}^*)$; it only bounds how much larger it can be by a quantity proportional to $\| \nabla f(\mathbf{x}, \mathbf{y}^*) \|_{p^*}^2$. Compare this also to the Polyak-Łojasiewicz condition (see, e.g., Nouiehed et al. (2019), Yang et al. (2020)), which imposes the opposite inequality, namely, that $f(\mathbf{x}, \mathbf{y}^*) - f(\mathbf{x}^*, \mathbf{y}^*)$ is bounded *above* by a multiple of $\| \nabla f(\mathbf{x}, \mathbf{y}^*) \|_{p^*}^2$.

One way that a generic operator F may satisfy Assumption 1 is when there is a constant $\lambda > 0$ such that for some $\mathbf{u}^* \in \mathcal{U}^*$ we have

$$(\forall \mathbf{u} \in \mathbb{R}^d) \, \langle F(\mathbf{u}), \mathbf{u} - \mathbf{u}^* \rangle \geq -\frac{\lambda}{2} \| \mathbf{u} - \mathbf{u}^* \|_p^2, \tag{2.3}$$

and when the operator F does not plateau or become too close to a linear operator around $\mathbf{u}^*$; namely, $\| F(\mathbf{u}) - F(\mathbf{u}^*) \|_{p^*} \geq \mu \| \mathbf{u} - \mathbf{u}^* \|_p$. (Note that (2.3) is always satisfied with $\lambda = 2L$ for L-Lipschitz operators, but we may need λ to be smaller than $2L$). Then Assumption 1 would be satisfied with $\rho = \frac{\lambda}{\mu}$. For a min-max problem, assuming f is twice differentiable, this would mean that the lowest eigenvalue of the symmetric part of the Jacobian of $\begin{bmatrix} \nabla_\mathbf{x} f(\mathbf{x}, \mathbf{y}) \\ -\nabla_\mathbf{y} f(\mathbf{x}, \mathbf{y}) \end{bmatrix}$

is bounded below by $-\lambda/2$ in any direction $\mathbf{u} - \mathbf{u}^*$ and the function f is sufficiently "curved" (not close to a linear or a constant function) around $\mathbf{u}^* = \begin{bmatrix} \mathbf{x}^* \\ \mathbf{y}^* \end{bmatrix}$.

Finally, we discuss a concrete min-max application wherein there are no (MVI) solutions, but there do exist (SVI) solutions satisfying the weak (MVI) condition of Assumption 1. This application arises in the context of two-agent zero-sum reinforcement learning problems studied by many authors, including recently by Daskalakis et al. (2020). In Section 5.1 of that work, the authors consider a special case of the general two-agent zero-sum RL problem, called von Neumann's ratio game, for which they observe that, even on a random example, the (MVI) solution set is empty, yet the extragradient method still converges in practice (albeit at a slower rate). Interestingly, it is easy to construct examples of the von Neumann ratio game for which no (MVI) solution exists, but the weak (MVI) condition of Assumption 1 does hold, and yet the stronger cohypomonotonicity condition of (2.2) does not hold. Indeed, one such example is obtained for the game shown in Proposition 2 of their paper, setting $s = 1/2$ and $\epsilon = .49$. Here (MVI) fails, the weak (MVI) condition of Assumption 1 is satisfied, and cohypomonotonicity fails to hold, e.g., for $\mathbf{u} = (\mathbf{x}, \mathbf{y}) = (0.1, 0.3)$ and $\mathbf{v} = (\mathbf{x}', \mathbf{y}') = (0.8, 0.3)$. To be clear, the von Neumann ratio game gives rise to a constrained min-max problem while our algorithm targets the unconstrained setting. While extending our result to the constrained setting remains open, our example here demonstrates that there is value in further studying the weak (MVI) condition of Assumption 1 in the constrained setting as well.

2.3 Useful Definitions and Facts

We now list some useful definitions and facts that will subsequently be used in our analysis. We start with a presentation of uniformly convex functions, convex conjugates, and Bregman divergence, and then specialize these basic facts to the ℓ_p setups used in Section 4.

Definition 2.1 (Uniform convexity). Given $p \geq 2$, a differentiable function $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is said to be p-uniformly convex w.r.t. $\| \cdot \|$ and with constant m if $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\psi(\mathbf{y}) \geq \psi(\mathbf{x}) + \langle \nabla \psi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{m}{p} \|\mathbf{y} - \mathbf{x}\|^p.$$

Observe that when $p = 2$, we recover the standard definition of strong convexity. Thus, uniform convexity is a generalization of strong convexity.

Definition 2.2 (Convex conjugate). Given a convex function $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, its convex conjugate ψ^* is defined by:

$$(\forall \mathbf{z} \in \mathbb{R}^d) : \quad \psi^*(\mathbf{z}) = \sup_{\mathbf{x} \in \mathbb{R}^d} \{\langle \mathbf{z}, \mathbf{x} \rangle - \psi(\mathbf{x})\}.$$

The following standard fact can be derived using the Fenchel-Young inequality, which states that $\forall \mathbf{x}, \mathbf{z} \in \mathbb{R}^d$: $\psi(\mathbf{x}) + \psi^*(\mathbf{z}) \geq \langle \mathbf{z}, \mathbf{x} \rangle$, and it is a simple corollary of Danskin's theorem (see, e.g., (Bertsekas, 1971, Bertsekas et al., 2003)).

Fact 2.3. *Let $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a closed convex proper function and let ψ^* be its convex conjugate. Then, $\forall \mathbf{g} \in \partial \psi^*(\mathbf{z})$,*

$$\mathbf{g} \in \operatorname*{argsup}_{\mathbf{x} \in \mathbb{R}^d} \{\langle \mathbf{z}, \mathbf{x} \rangle - \psi(\mathbf{x})\},$$

where $\partial \psi^*(\mathbf{z})$ is the subdifferential set (the set of all subgradients) of ψ^* at point $\mathbf{z}$. In particular, if ψ^* is differentiable, then $\operatorname{argsup}_{\mathbf{x} \in \mathbb{R}^d} \{\langle \mathbf{z}, \mathbf{x} \rangle - \psi(\mathbf{x})\}$ is a singleton set and $\nabla \psi^*(\mathbf{z})$ is its only element.

Definition 2.4 (Bregman divergence). Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then its Bregman divergence between points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ is defined by

$$D_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla \psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

It is immediate that the Bregman divergence of a convex function is non-negative.

Useful facts for ℓ_p setups. We now outline some useful auxiliary results used specifically in Section 4, where we study the case that p is not necessarily equal to 2.

Proposition 2.5. *Given, $\mathbf{z}, \mathbf{u} \in \mathbb{R}^d$, $p \in (1, \infty)$ and $q \in \{p, 2\}$, let*

$$\mathbf{w} = \operatorname*{argmin}_{\mathbf{v} \in \mathbb{R}^d} \Big\{ \langle \mathbf{z}, \mathbf{v} \rangle + \frac{1}{q} \|\mathbf{u} - \mathbf{v}\|_p^q \Big\}.$$

Then, for $p^* = \frac{p}{p-1}$, $q^* = \frac{q}{q-1}$:

$$\mathbf{w} = \mathbf{u} - \nabla \Big(\frac{1}{q^*} \|\mathbf{z}\|_{p^*}^{q^*} \Big) \quad and \quad \frac{1}{q} \|\mathbf{w} - \mathbf{u}\|_p^q = \frac{1}{q} \|\mathbf{z}\|_{p^*}^{q^*}.$$

Proof. The statements in the proposition are simple corollaries of conjugacy of the functions $\psi(\mathbf{u}) = \frac{1}{q} \|\mathbf{u}\|_p^q$ and $\psi^*(\mathbf{z}) = \frac{1}{q^*} \|\mathbf{z}\|_{p^*}^{q^*}$. In particular, the first part follows from

$$\psi^*(\mathbf{z}) = \sup_{\mathbf{v} \in \mathbb{R}^d} \{ \langle \mathbf{z}, \mathbf{v} \rangle - \psi(\mathbf{v}) \},$$

by the definition of a convex conjugate and using that $\frac{1}{q} \|\mathbf{u}\|_p^q$ and $\frac{1}{q^*} \|\mathbf{z}\|_{p^*}^{q^*}$ are conjugates of each other, which are standard exercises in convex analysis for $q \in \{p, 2\}$ (see, e.g., (Borwein and Zhu, 2004, Exercise 4.4.2) and (Boyd et al., 2004, Example 3.27)).

The second part follows by $\nabla \psi^*(\mathbf{z}) = \arg \sup_{\mathbf{v} \in \mathbb{R}^d} \{ \langle \mathbf{z}, \mathbf{v} \rangle - \psi(\mathbf{v}) \}$, due to Fact 2.3 ($\psi$ and ψ^* are both continuously differentiable for $p \in (1, \infty)$). Lastly, $\frac{1}{q} \|\mathbf{w} - \mathbf{u}\|_p^q = \frac{1}{q} \|\mathbf{z}\|_{p^*}^{q^*}$ can be verified by setting $\mathbf{w} = \mathbf{u} - \nabla \big(\frac{1}{q^*} \|\mathbf{z}\|_{p^*}^{q^*} \big)$. $\qquad\square$

Another useful result is the following proposition, which will allow us to relate Lipschitzness of F to uniform convexity of the prox mapping $\frac{1}{q} \| \cdot \|_p^q$ in the definition of the algorithm. The ideas used in the proof can be found in the proofs of (d'Aspremont et al., 2018, Lemma 5.7), (Nesterov, 2015, Lemma 2), and in (Devolder et al., 2014, Section 2.3). The proof is provided for completeness.

Proposition 2.6. *For any $L > 0$, $\kappa > 0$, $q \geq \kappa$, $t \geq 0$, and $\delta > 0$,*

$$\frac{L}{\kappa} t^\kappa \leq \frac{\Lambda}{q} t^q + \frac{\delta}{2},$$

where

$$\Lambda = \Big(\frac{2(q - \kappa)}{\delta q \kappa} \Big)^{\frac{q-\kappa}{\kappa}} L^{q/\kappa}.$$

Proof. The proof is based on the Fenchel-Young inequality and the conjugacy of functions $\frac{|x|^r}{r}$ and $\frac{|y|^s}{s}$ for $r, s \geq 1$, $\frac{1}{r} + \frac{1}{s} = 1$, which implies $xy \leq \frac{x^r}{r} + \frac{y^s}{s}$, $\forall x, y \geq 0$. In particular, setting $r = q/\kappa$, $s = q/(q - \kappa)$, and $x = t^\kappa$, we have

$$\frac{L}{\kappa} t^\kappa \leq \frac{L t^q}{q y} + \frac{L(q - \kappa)}{q \kappa} y^{\frac{\kappa}{q-\kappa}}.$$

It remains to set $\frac{\delta}{2} = \frac{L(q-\kappa)}{q\kappa} y^{\frac{\kappa}{q-\kappa}}$, which, solving for y, gives $y = \big(\frac{\delta q \kappa}{2L(q-\kappa)} \big)^{q-\kappa}$, and verify that, under this choice, $\Lambda = \frac{L t^q}{q y}$. $\qquad\square$

3 Generalized Extragradient for Problems with Weak MVI Solutions

In this section, we consider the setup with the Euclidean norm $\| \cdot \| = \| \cdot \|_2$, i.e., $p = 2$. To address the class of problems with weak (MVI) solutions (see Assumption 1), we introduce the following generalization of the extragradient algorithm, to which we refer as Extragradient+ (EG+).

$$\bar{\mathbf{u}}_k = \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^d} \Big\{ \frac{a_k}{\beta} \langle F(\mathbf{u}_k), \mathbf{u} - \mathbf{u}_k \rangle + \frac{1}{2} \|\mathbf{u} - \mathbf{u}_k\|^2 \Big\} = \mathbf{u}_k - \frac{a_k}{\beta} F(\mathbf{u}_k),$$

$$\mathbf{u}_{k+1} = \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^d} \Big\{ a_k \langle F(\bar{\mathbf{u}}_k), \mathbf{u} - \mathbf{u}_k \rangle + \frac{1}{2} \|\mathbf{u} - \mathbf{u}_k\|^2 \Big\} = \mathbf{u}_k - a_k F(\bar{\mathbf{u}}_k),$$

$$(\text{EG+})$$

where $\beta \in (0, 1]$ is a parameter of the algorithm and $a_k > 0$ is the step size. When $\beta = 1$, we recover standard EG.

The analysis relies on the following merit (or gap) function:

$$h_k := a_k \Big(\langle F(\bar{\mathbf{u}}_k), \bar{\mathbf{u}}_k - \mathbf{u}^* \rangle + \frac{\rho}{2} \|F(\bar{\mathbf{u}}_k)\|^2 \Big), \tag{3.1}$$

for some $\mathbf{u}^* \in \mathcal{U}^*$ for which F satisfies Assumption 1. Then Assumption 1 implies that $h_k \geq 0, \forall k$.

The first (and main) step is to bound all h_k's above, as in the following lemma.

Lemma 3.1. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be an arbitrary L-Lipschitz operator that satisfies Assumption 1 for some $\mathbf{u}^* \in \mathcal{U}^*$. Given an arbitrary initial point $\mathbf{u}_0$, let the sequences of points $\{\mathbf{u}_i\}_{i \geq 1}$, $\{\bar{\mathbf{u}}_i\}_{i \geq 0}$ evolve according to* (EG+) *for some $\beta \in (0, 1]$ and positive step sizes $\{a_i\}_{i \geq 0}$. Then, for any $\gamma > 0$ and any $k \geq 0$, we have:*

$$h_k \leq \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|^2 - \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2 + \frac{a_k}{2}\big(\rho - a_k(1-\beta)\big)\|F(\bar{\mathbf{u}}_k)\|^2$$
$$+ \frac{a_k{}^2}{2\beta^2}\big(a_k L\gamma - \beta\big)\|F(\mathbf{u}_k)\|^2 + \frac{1}{2}\Big(\frac{a_k L}{\gamma} - \beta\Big)\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|^2, \tag{3.2}$$

where h_k is defined as in Eq. (3.1).

Proof. Fix any $k \geq 0$ and write h_k equivalently as

$$h_k = a_k \langle F(\bar{\mathbf{u}}_k), \mathbf{u}_{k+1} - \mathbf{u}^* \rangle + a_k \langle F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \rangle$$
$$+ a_k \langle F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \rangle + a_k \frac{\rho}{2}\|F(\bar{\mathbf{u}}_k)\|^2. \tag{3.3}$$

The proof proceeds by bounding from above individual terms on the right-hand side of Eq. (3.3). For the first term, the first-order optimality in the definition of $\mathbf{u}_{k+1}$ gives:

$$a_k F(\bar{\mathbf{u}}_k) + \mathbf{u}_{k+1} - \mathbf{u}_k = \mathbf{0}.$$

Thus, we have

$$a_k \langle F(\bar{\mathbf{u}}_k), \mathbf{u}_{k+1} - \mathbf{u}^* \rangle = - \langle \mathbf{u}_{k+1} - \mathbf{u}_k, \mathbf{u}_{k+1} - \mathbf{u}^* \rangle$$
$$= \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|^2 - \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2 - \frac{1}{2}\|\mathbf{u}_k - \mathbf{u}_{k+1}\|^2. \tag{3.4}$$

For the second term on the right-hand side of Eq. (3.3), the first-order optimality in the definition of $\bar{\mathbf{u}}_k$ implies:

$$\frac{a_k}{\beta} \langle F(\mathbf{u}_k) + \bar{\mathbf{u}}_k - \mathbf{u}_k, \mathbf{u}_{k+1} - \bar{\mathbf{u}}_k \rangle = 0,$$

which, similarly as for the first term, leads to:

$$a_k \langle F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \rangle = \frac{\beta}{2}\|\mathbf{u}_k - \mathbf{u}_{k+1}\|^2 - \frac{\beta}{2}\|\mathbf{u}_k - \bar{\mathbf{u}}_k\|^2 - \frac{\beta}{2}\|\mathbf{u}_{k+1} - \bar{\mathbf{u}}_k\|^2. \tag{3.5}$$

For the third term on the right-hand side of Eq. (3.3), applying the Cauchy-Schwarz inequality, L-Lipschitzness of F, and Young's inequality, respectively, we have:

$$a_k \langle F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \rangle \leq a_k \|F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k)\|\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|$$
$$\leq a_k L \|\bar{\mathbf{u}}_k - \mathbf{u}_k\|\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|$$
$$\leq \frac{a_k L\gamma}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|^2 + \frac{a_k L}{2\gamma}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|^2, \tag{3.6}$$

where the last inequality holds for any $\gamma > 0$.

Using the fact that $\bar{\mathbf{u}}_k - \mathbf{u}_k = -\frac{a_k}{\beta}F(\mathbf{u}_k)$, $\mathbf{u}_{k+1} - \mathbf{u}_k = -a_k F(\bar{\mathbf{u}}_k)$ and combining Eqs. (3.4)-(3.6) with Eq. (3.3), we have:

$$h_k \leq \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|^2 - \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2 + \frac{a_k}{2}\big(\rho - a_k(1-\beta)\big)\|F(\bar{\mathbf{u}}_k)\|^2$$
$$+ \frac{a_k{}^2}{2\beta^2}\big(a_k L\gamma - \beta\big)\|F(\mathbf{u}_k)\|^2 + \frac{1}{2}\Big(\frac{a_k L}{\gamma} - \beta\Big)\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|^2,$$

as claimed. $\qquad\square$

Using Lemma 3.1, we can now draw conclusions about the convergence of EG+ by choosing parameters β, γ and the step sizes a_k to guarantee that $h_k < \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|^2 - \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2$ as long as $\|F(\bar{\mathbf{u}}_k)\| \neq 0$.

Theorem 3.2. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be an arbitrary L-Lipschitz operator that satisfies Assumption 1 for some $\mathbf{u}^* \in \mathcal{U}^*$. Given an arbitrary initial point $\mathbf{u}_0 \in \mathbb{R}^d$, let the sequences of points $\{\mathbf{u}_i\}_{i\geq 1}$, $\{\bar{\mathbf{u}}_i\}_{i\geq 0}$ evolve according to (EG+) for $\beta = \frac{1}{2}$ and $a_k = \frac{1}{2L}$. Then:*

(i) all accumulation points of $\{\bar{\mathbf{u}}_k\}_{k\geq 0}$ are in $\mathcal{U}^*$.

(ii) for all $k \geq 1$:

$$\frac{1}{k+1}\sum_{i=0}^{k}\|F(\bar{\mathbf{u}}_i)\|^2 \leq \frac{2L\|\mathbf{u}_0 - \mathbf{u}^*\|^2}{(k+1)(1/(4L) - \rho)}.$$

In particular, we have that

$$\min_{0\leq i\leq k}\|F(\bar{\mathbf{u}}_i)\|^2 \leq \frac{2L\|\mathbf{u}_0 - \mathbf{u}^*\|^2}{(k+1)(1/(4L) - \rho)}$$

and

$$\mathbb{E}_{i\sim\mathrm{Unif}\{0,\ldots,k\}}\big[\|F(\bar{\mathbf{u}}_i)\|^2\big] \leq \frac{2L\|\mathbf{u}_0 - \mathbf{u}^*\|^2}{(k+1)(1/(4L) - \rho)},$$

where $i \sim \mathrm{Unif}\{0,\ldots,k\}$ denotes an index i chosen uniformly at random from the set $\{0,\ldots,k\}$.

Proof. Applying Lemma 3.1 with the choice of a_k and β from the theorem statement and with $\gamma = 1$, we get

$$h_k \leq \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|^2 - \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2 + \frac{1}{4L}\Big(\rho - \frac{1}{4L}\Big)\|F(\bar{\mathbf{u}}_k)\|^2.$$

By Assumption 1, $\rho < \frac{1}{4L}$, and, thus, the constant multiplying $\|F(\bar{\mathbf{u}}_k)\|^2$ is strictly negative.

As $h_k \geq 0$ (by Assumption 1), we can conclude that

$$\frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2 - \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|^2 \leq -\frac{1}{4L}\Big(\frac{1}{4L} - \rho\Big)\|F(\bar{\mathbf{u}}_k)\|^2 \leq 0. \tag{3.7}$$

As $\frac{1}{4L}\big(\frac{1}{4L} - \rho\big) > 0$, Eq. (3.7) implies that $\|F(\bar{\mathbf{u}}_k)\|$ converges to zero as $k \to \infty$. Further, as $\bar{\mathbf{u}}_k - \mathbf{u}_k = -\frac{a_k}{\beta}F(\mathbf{u}_k)$, using triangle inequality and $F(\mathbf{u}^*) = \mathbf{0}$:

$$\begin{aligned}\|\bar{\mathbf{u}}_k - \mathbf{u}^*\| &\leq \|\mathbf{u}_k - \mathbf{u}^*\| + \frac{a_k}{\beta}\|F(\mathbf{u}_k) - F(\mathbf{u}^*)\| \\ &\leq \Big(1 + L\frac{a_k}{\beta}\Big)\|\mathbf{u}_k - \mathbf{u}^*\| = 2\|\mathbf{u}_k - \mathbf{u}^*\|,\end{aligned} \tag{3.8}$$

where we have used that F is L-Lipschitz. Now, as $\|\mathbf{u}_k - \mathbf{u}^*\|$ is bounded (by $\|\mathbf{u}_0 - \mathbf{u}^*\|$, from Eq. (3.7)), it follows that the sequence $\{\bar{\mathbf{u}}_k\}$ is bounded as well, and thus has a converging subsequence. Let $\{\bar{\mathbf{u}}_{k_i}\}$ be any converging subsequence of $\{\bar{\mathbf{u}}_k\}$ and let $\bar{\mathbf{u}}^*$ be its corresponding accumulation point. Then, as $\|F(\bar{\mathbf{u}}_k)\|$ converges to zero as $k \to \infty$, it follows that $\|F(\bar{\mathbf{u}}_{k_i})\|$ converges to zero as $i \to \infty$, and so it must be $\bar{\mathbf{u}}^* \in \mathcal{U}^*$.

For Part (ii), telescoping Eq. (3.7), we get:

$$\begin{aligned}\frac{1}{4L}\Big(\frac{1}{4L} - \rho\Big)\sum_{i=0}^{k}\|F(\bar{\mathbf{u}}_i)\|^2 &\leq \frac{1}{2}\|\mathbf{u}_0 - \mathbf{u}^*\|^2 - \frac{1}{2}\|\mathbf{u}_{k+1} - \mathbf{u}^*\|^2 \\ &\leq \frac{1}{2}\|\mathbf{u}_0 - \mathbf{u}^*\|^2.\end{aligned}$$

Rearranging the last inequality:

$$\frac{1}{k+1}\sum_{i=0}^{k}\|F(\bar{\mathbf{u}}_i)\|^2 \leq \frac{2L\|\mathbf{u}_0 - \mathbf{u}^*\|^2}{(k+1)(1/(4L) - \rho)}.$$

9

It remains to observe that

$$\mathbb{E}_{i \sim \mathrm{Unif}\{0,\dots,k\}}[\|F(\bar{\mathbf{u}}_i)\|^2] = \frac{1}{k+1} \sum_{i=0}^{k} \|F(\bar{\mathbf{u}}_i)\|^2$$

and $\frac{1}{k+1} \sum_{i=0}^{k} \|F(\bar{\mathbf{u}}_i)\|^2 \geq \min_{0 \leq i \leq k} \|F(\bar{\mathbf{u}}_i)\|^2$. $\qquad\square$

Remark 3.3. Due to Eq. (3.8), we have that all the iterates of EG+ with the parameter setting as in Theorem 3.2 remain in the ball centered at $\mathbf{u}^*$ and of radius at most $2\|\mathbf{u}_0 - \mathbf{u}^*\|$. Thus, Assumption 1 does not need to hold globally for the result to apply; it suffices that it only applies locally to points from the ball around $\mathbf{u}^*$ with radius $2\|\mathbf{u}_0 - \mathbf{u}^*\|$.

Remark 3.4. It is possible to obtain similar convergence results as those of Theorem 3.2 under different parameter choices. In particular, for $\gamma \in (0, 1]$, it suffices that $a_k \leq \frac{\beta\gamma}{L}$ and $\rho < a_k(1 - \beta)$. We settled on the choice made in Theorem 3.2 as it is simple and requires tuning only one parameter, L.

Remark 3.5. Note that, in fact, we did not need to assume that $\mathbf{u}^*$ from Assumption 1 is from $\mathcal{U}^*$; it could have been any point from $\mathbb{R}^d$ for which Assumption 1 is satisfied. All that would change in the proof of Theorem 3.2 is that in Eq. (3.8), using $\|F(\mathbf{u}_k)\| \leq \|F(\mathbf{u}_k) - F(\mathbf{u}^*)\| + \|F(\mathbf{u}^*)\|$ (by triangle inequality) we would have $2\|\mathbf{u}_k - \mathbf{u}^*\| + \frac{1}{L}\|F(\mathbf{u}^*)\|$ on the right-hand side. Since $\mathbf{u}^* \in \mathbb{R}^d$ and F is Lipschitz-continuous, if F is bounded at any point $\mathbf{u} \in \mathbb{R}^d$, $\|F(\mathbf{u}^*)\|$ is bounded as well. Thus, we can still conclude that $\|\bar{\mathbf{u}}_k - \mathbf{u}^*\|$ is bounded and proceed with the rest of the proof. An interesting consequence of this observation and the proof of Theorem 3.2 is that Assumption 1 *guarantees* existence of an (SVI) solution.

4 Extensions: ℓ_p Norms and Stochastic Setups

In this section, we show how to extend the results of Section 3 to non-Euclidean, ℓ_p-normed setups (for $\rho = 0$) and stochastic evaluations of F. In particular, we let $\|\cdot\| = \|\cdot\|_p$ for $p \in (1, \infty)$[3] and $p^* = \frac{p}{p-1}$. Further, we let $\tilde{F}$ denote the stochastic estimate of F that at iteration k satisfies:

$$\mathbb{E}[\tilde{F}(\bar{\mathbf{u}}_k)|\bar{\mathcal{F}}_k] = F(\bar{\mathbf{u}}_k), \quad \mathbb{E}[\|\tilde{F}(\bar{\mathbf{u}}_k) - F(\bar{\mathbf{u}}_k)\|_*^2|\bar{\mathcal{F}}_k] \leq \bar{\sigma}_k^2$$
$$\mathbb{E}[\tilde{F}(\mathbf{u}_{k+1})|\mathcal{F}_{k+1}] = F(\mathbf{u}_{k+1}), \quad \mathbb{E}[\|\tilde{F}(\mathbf{u}_{k+1}) - F(\mathbf{u}_{k+1})\|_*^2|\mathcal{F}_{k+1}] \leq \sigma_{k+1}{}^2, \tag{4.1}$$

where $\mathcal{F}_k$ and $\bar{\mathcal{F}}_k$ denote the natural filtrations, including all the randomness up to the construction of points $\mathbf{u}_k$ and $\bar{\mathbf{u}}_k$, respectively, and $\bar{\sigma}_k^2, \sigma_{k+1}{}^2$ are the variance constants. Observe that $\mathcal{F}_k \subseteq \bar{\mathcal{F}}_k$ and $\bar{\mathcal{F}}_k \subseteq \mathcal{F}_{k+1}$. To simplify the notation, we denote:

$$\bar{\boldsymbol{\eta}}_k = \tilde{F}(\bar{\mathbf{u}}_k) - F(\bar{\mathbf{u}}_k), \ \boldsymbol{\eta}_{k+1} = \tilde{F}(\mathbf{u}_{k+1}) - F(\mathbf{u}_{k+1}). \tag{4.2}$$

The variant of the method we consider here is stated as follows:

$$\bar{\mathbf{u}}_k = \underset{\mathbf{u} \in \mathbb{R}^d}{\arg\min}\left\{ \frac{a_k}{\beta} \left\langle \tilde{F}(\mathbf{u}_k), \mathbf{u} - \mathbf{u}_k \right\rangle + \frac{1}{q}\|\mathbf{u} - \mathbf{u}_k\|_p^q \right\},$$
$$\mathbf{u}_{k+1} = \underset{\mathbf{u} \in \mathbb{R}^d}{\arg\min}\left\{ a_k \left\langle \tilde{F}(\bar{\mathbf{u}}_k), \mathbf{u} - \mathbf{u}_k \right\rangle + \phi_p(\mathbf{u}, \mathbf{u}_k) \right\}, \tag{EG_p+}$$

where

$$q = \begin{cases} 2, & \text{if } p \in (1, 2], \\ p^* = \frac{p}{p-1}, & \text{if } p \in (2, \infty) \end{cases} \tag{4.3}$$

and

$$\phi_p(\mathbf{u}, \mathbf{u}_k) = \begin{cases} D_{\frac{1}{2}\|\cdot - \mathbf{u}_0\|_p^2}(\mathbf{u}, \mathbf{u}_k), & \text{if } p \in (1, 2], \\ \frac{1}{p}\|\mathbf{u} - \mathbf{u}_k\|_p^p, & \text{if } p \in (2, \infty). \end{cases} \tag{4.4}$$

Notice that for $p = 2$, EG_p+ is equivalent to EG+. Thus, EG_p+ generalizes EG+ to arbitrary ℓ_p norms. However, EG_p+ is different from the standard Extragradient or Mirror-Prox, for two reasons. First is that, as is the case for

---

[3]Note that the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are within a constant factor of the ℓ_p-norm for $p = 1 + \frac{1}{\log(d)}$ and $p = \log(d)$, respectively, and so taking $p \in (1, \infty)$ is w.l.o.g.—for any $p < 1 + \frac{1}{\log(d)}$ or $p > \log(d)$, we can run the algorithm with $p = 1 + \frac{1}{\log d}$ or $p = \log d$, losing at most a constant factor in the convergence bound.

EG+, the step sizes that determine $\bar{\mathbf{u}}_k$ and $\mathbf{u}_{k+1}$ (i.e., a_k/β and a_k) are not the same in general, as we could (and will) choose $\beta \neq 1$. Second, unless $p = q = 2$, the function $\frac{1}{q}\|\mathbf{u} - \mathbf{u}_k\|_p^q$ in the definition of the algorithm is *not* a Bregman divergence between points $\mathbf{u}$ and $\mathbf{u}_k$ of any function ψ. Further, when $p > 2$, $\frac{1}{q}\|\mathbf{u} - \mathbf{u}_k\|_p^q$ is *not* strongly convex. Instead, it is *p-uniformly convex* with constant 1. Additionally, no function whose gap between the maximum and the minimum value is bounded by a constant on any ball of constant radius can have constant of strong convexity w.r.t. $\|\cdot\|_p$ that is larger than $O(\frac{1}{d^{1-2/p}})$ (d'Aspremont et al., 2018). When $p \in (1, 2]$, $\frac{1}{q}\|\mathbf{u} - \mathbf{u}_k\|_p^q$ is strongly convex with constant $p - 1$ (Nemirovski, 2004). We let m_p denote the constant of strong/uniform convexity of $\frac{1}{q}\|\mathbf{u} - \mathbf{u}_k\|_p^q$, that is:

$$m_p = \max\{p - 1, 1\}. \tag{4.5}$$

Observe that

$$\phi_p(\mathbf{u}, \mathbf{u}_k) \geq \frac{m_p}{q}\|\mathbf{u} - \mathbf{u}_k\|_p^q. \tag{4.6}$$

This is immediate for $p > 2$, by the definition of ϕ_p and using that $q = p$ and $m_p = 1$ when $p > 2$. For $p \in (1, 2]$, we have that $q = 2$, and Eq. (4.6) follows by strong convexity of $\frac{1}{2}\|\cdot\|_p^2$.

As in the case of Euclidean norms, the analysis relies on the following merit function:

$$h_k := a_k\left(\langle F(\bar{\mathbf{u}}_k), \bar{\mathbf{u}}_k - \mathbf{u}^* \rangle + \frac{\rho}{2}\|F(\bar{\mathbf{u}}_k)\|_{p^*}^2 \right). \tag{4.7}$$

Moreover, as before, Assumption 1 guarantees that $h_k \geq 0$, $\forall k$.

We start by first proving a lemma that holds for generic choices of algorithm parameters a_k and β. We will then use this lemma to deduce the convergence bounds for different choices of $p > 1$ and both deterministic and the stochastic oracle access to F.

Lemma 4.1. *Let $p > 1$ and let $F : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be an arbitrary L-Lipschitz operator w.r.t. $\|\cdot\|_p$ that satisfies Assumption 1 for some $\mathbf{u}^* \in \mathcal{U}^*$. Given an arbitrary initial point $\mathbf{u}_0$, let the sequences of points $\{\mathbf{u}_i\}_{i\geq 1}$, $\{\bar{\mathbf{u}}_i\}_{i\geq 0}$ evolve according to (EG_p+) for some $\beta \in (0, 1]$ and positive step sizes $\{a_i\}_{i\geq 0}$. Then, for any $\gamma > 0$ and any $k \geq 0$:*

$$
\begin{aligned}
h_k \leq &- a_k \langle \bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^* \rangle - a_k \langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \rangle + \frac{a_k \rho}{2}\|F(\bar{\mathbf{u}}_k)\|_{p^*}^2 \\
&+ \phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) + \frac{\beta - m_p}{q}\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^q \\
&+ \frac{a_k \Lambda_k \gamma - \beta}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^q + \frac{a_k \Lambda_k/\gamma - \beta m_p}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q + a_k \delta_k,
\end{aligned}
$$

where h_k is defined as in Eq. (4.7), δ_k is any positive number, and $\Lambda_k = \left(\frac{q-2}{\delta_k q}\right)^{\frac{q-2}{2}} L^{q/2}$. When $q = 2$, the statement also holds with $\delta_k = 0$ and $\Lambda_k = L$.

Proof. We begin the proof by writing h_k equivalently as:

$$
\begin{aligned}
h_k &= a_k \left\langle \tilde{F}(\bar{\mathbf{u}}_k), \bar{\mathbf{u}}_k - \mathbf{u}^* \right\rangle - a_k \langle \bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^* \rangle + \frac{a_k \rho}{2}\|F(\bar{\mathbf{u}}_k)\|_{p^*}^2 \\
&= a_k \left\langle \tilde{F}(\bar{\mathbf{u}}_k), \mathbf{u}_{k+1} - \mathbf{u}^* \right\rangle + a_k \left\langle \tilde{F}(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \right\rangle \\
&\quad + a_k \left\langle \tilde{F}(\bar{\mathbf{u}}_k) - \tilde{F}(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \right\rangle - a_k \langle \bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^* \rangle + \frac{a_k \rho}{2}\|F(\bar{\mathbf{u}}_k)\|_{p^*}^2.
\end{aligned} \tag{4.8}
$$

The proof now proceeds by bounding individual terms on the right-hand side of the last equality.

Let $M_{k+1}(\mathbf{u}) = a_k \left\langle \nabla \tilde{F}(\bar{\mathbf{u}}_k), \mathbf{u} - \mathbf{u}_k \right\rangle + \phi_p(\mathbf{u}, \mathbf{u}_k)$, so that $\mathbf{u}_{k+1} = \text{argmin}_{\mathbf{u} \in \mathbb{R}^d} M_{k+1}(\mathbf{u})$. By the definition of Bregman divergence of M_{k+1}:

$$M_{k+1}(\mathbf{u}^*) = M_{k+1}(\mathbf{u}_{k+1}) + \langle \nabla M_{k+1}(\mathbf{u}_{k+1}), \mathbf{u}^* - \mathbf{u}_{k+1} \rangle + D_{M_{k+1}}(\mathbf{u}^*, \mathbf{u}_{k+1}).$$

As $\mathbf{u}_{k+1} = \text{argmin}_{\mathbf{u} \in \mathbb{R}^d} M_{k+1}(\mathbf{u})$, we have $\nabla M_{k+1}(\mathbf{u}_{k+1}) = \mathbf{0}$. Further, $D_{M_{k+1}}(\mathbf{u}^*, \mathbf{u}_{k+1}) = D_{\phi_p(\cdot, \mathbf{u}_k)}(\mathbf{u}^*, \mathbf{u}_{k+1})$. When $p \leq 2$, ϕ_p itself is a Bregman divergence, and we have $D_{M_{k+1}}(\mathbf{u}^*, \mathbf{u}_{k+1}) = \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1})$. When $p > 2$,

$\phi_p(\mathbf{u}, \mathbf{u}_k) = \frac{1}{p}\|\mathbf{u} - \mathbf{u}_k\|_p^p$, and as ϕ_p is p-uniformly convex with constant 1, it follows that $D_{M_{k+1}}(\mathbf{u}^*, \mathbf{u}_{k+1}) \geq \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|_p^p = \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1})$. Thus:

$$M_{k+1}(\mathbf{u}^*) \geq M_{k+1}(\mathbf{u}_{k+1}) + \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}).$$

Equivalently, applying the definition of $M_{k+1}(\cdot)$ to the last inequality:

$$
\begin{aligned}
a_k \left\langle \nabla \tilde{F}(\bar{\mathbf{u}}_k), \mathbf{u}_{k+1} - \mathbf{u}^* \right\rangle &\leq \phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) - \phi_p(\mathbf{u}_{k+1,\mathbf{u}_k}) \\
&\leq \phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) - \frac{m_p}{q}\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^q,
\end{aligned}
\tag{4.9}
$$

where the last inequality follows from Eq. (4.6).

Now let $\bar{M}_k(\mathbf{u}) = \frac{a_k}{\beta}\left\langle \tilde{F}(\mathbf{u}_k), \mathbf{u} - \mathbf{u}_k \right\rangle + \frac{1}{q}\|\mathbf{u} - \mathbf{u}_k\|_p^q$ so that $\bar{\mathbf{u}}_k = \operatorname{argmin}_{\mathbf{u}\in\mathbb{R}^d} \bar{M}_k(\mathbf{u})$. By similar arguments as above,

$$
\begin{aligned}
\bar{M}_k(\mathbf{u}_{k+1}) &= \bar{M}_k(\bar{\mathbf{u}}_k) + \left\langle \nabla\bar{M}_k(\bar{\mathbf{u}}_k), \mathbf{u}_{k+1} - \bar{\mathbf{u}}_k \right\rangle + D_{M_k}(\mathbf{u}_{k+1}, \bar{\mathbf{u}}_k) \\
&\geq \bar{M}_k(\bar{\mathbf{u}}_k) + \frac{m_p}{q}\|\mathbf{u}_{k+1} - \bar{\mathbf{u}}_k\|_p^q,
\end{aligned}
$$

where the inequality is by $\nabla\bar{M}_k(\bar{\mathbf{u}}_k) = \mathbf{0}$ and the fact that $\frac{1}{q}\|\cdot\|_p^q$ is q-uniformly convex w.r.t. $\|\cdot\|_p$ with constant m_p, by the choice of q from Eq. (4.3). Applying the definition of $\bar{M}_k(\mathbf{u})$ to the last inequality:

$$
a_k \left\langle \tilde{F}(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \right\rangle \leq \frac{\beta}{q}\left(\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^q - \|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^q - m_p\|\mathbf{u}_{k+1} - \bar{\mathbf{u}}_k\|_p^q\right).
\tag{4.10}
$$

The remaining term that we need to bound is $\left\langle \tilde{F}(\bar{\mathbf{u}}_k) - \tilde{F}(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \right\rangle$. Using the definitions of $\bar{\boldsymbol{\eta}}_k, \boldsymbol{\eta}_k$, we have:

$$
\begin{aligned}
\left\langle \tilde{F}(\bar{\mathbf{u}}_k) - \tilde{F}(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \right\rangle &= \langle F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle - \langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle \\
&\overset{(i)}{\leq} -\langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle + \|F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k)\|_{p^*}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p \\
&\overset{(ii)}{\leq} -\langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle + L\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p \\
&\overset{(iii)}{\leq} -\langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle + \frac{L\gamma}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^2 + \frac{L}{2\gamma}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^2,
\end{aligned}
$$

where (i) is by Hölder's inequality, (ii) is by L-Lipschitzness of F, and (iii) is by Young's inequality, which holds for any $\gamma > 0$. Now, let $\delta_k > 0$ and $\Lambda_k = \left(\frac{2(q-\kappa)}{\delta_k q \kappa}\right)^{\frac{q-\kappa}{\kappa}} L^{q/\kappa}$. Then, applying Proposition 2.6 to the last two terms in the last inequality:

$$
\begin{aligned}
\left\langle \tilde{F}(\bar{\mathbf{u}}_k) - \tilde{F}(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \right\rangle \leq\ & -\langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle \\
& + \frac{\Lambda_k \gamma}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^q + \frac{\Lambda_k}{q\gamma}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q + \delta_k.
\end{aligned}
\tag{4.11}
$$

Observe that when $q = 2$, there is no need to apply Proposition 2.6, and the last inequality is satisfied with $\delta_k = 0$ and $\Lambda_k = L$.

Combining Eqs. (4.9)-(4.11) with Eq. (4.8), we have:

$$
\begin{aligned}
h_k \leq\ & -a_k\langle \bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^*\rangle - a_k\langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle + \frac{a_k\rho}{2}\|F(\bar{\mathbf{u}}_k)\|_{p^*}^2 \\
& + \phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) + \frac{\beta - m_p}{q}\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^q \\
& + \frac{a_k\Lambda_k\gamma - \beta}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^q + \frac{a_k\Lambda_k/\gamma - \beta m_p}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q + a_k\delta_k,
\end{aligned}
$$

as claimed. $\qquad\square$

12

We are now ready to state and prove the main convergence bounds. For simplicity, we start with the case of exact oracle access to F. We then show that we can build on this result by separately bounding the error terms due to the variance of the stochastic estimates $\tilde{F}$.

Deterministic oracle access. The main result is summarized in the following theorem.

Theorem 4.2. *Let $p > 1$ and let $F : \mathbb{R}^d \to \mathbb{R}^d$ be an arbitrary L-Lipschitz operator w.r.t. $\|\cdot\|_p$ that satisfies Assumption 1 with $\rho = 0$ for some $\mathbf{u}^* \in \mathcal{U}^*$. Assume that we are given oracle access to the exact evaluations of F, i.e., $\bar{\boldsymbol{\eta}}_i = \boldsymbol{\eta}_i = \mathbf{0}, \forall i$. Given an arbitrary initial point $\mathbf{u}_0 \in \mathbb{R}^d$, let the sequences of points $\{\mathbf{u}_i\}_{i \geq 1}$, $\{\bar{\mathbf{u}}_i\}_{i \geq 0}$ evolve according to (EG_p+) for $\beta \in (0, 1]$ and step sizes $\{a_i\}_{i \geq 0}$ specified below. Then, we have:*

(i) Let $p \in (1, 2]$. If $\beta = m_p = p - 1$, $a_k = \frac{m_p^{3/2}}{2L}$, then all accumulation points of $\{\mathbf{u}_k\}_{k \geq 0}$ are in $\mathcal{U}^*$, and, furthermore $\forall k \geq 0$:

$$\frac{1}{k+1} \sum_{i=0}^{k} \|F(\mathbf{u}_i)\|_{p^*}^2 \leq \frac{16L^2 \phi_p(\mathbf{u}^*, \mathbf{u}_0)}{m_p^2 (k+1)}$$

$$= O\Big(\frac{L^2 \|\mathbf{u}^* - \mathbf{u}_0\|_p^2}{(p-1)^2 (k+1)}\Big).$$

In particular, within $k = O\big(\frac{L^2 \|\mathbf{u}^* - \mathbf{u}_0\|_p^2}{(p-1)^2 \epsilon^2}\big)$ iterations EG_p+ can output a point $\mathbf{u}$ with $\|F(\mathbf{u})\|_{p^*} \leq \epsilon$.

(ii) Let $p \in (2, \infty)$. If $\beta = \frac{1}{2}$, $\delta_k = \delta > 0$, $\Lambda = \big(\frac{q-2}{\delta q}\big)^{\frac{q-2}{2}} L^{q/2}$, and $a_k = \frac{1}{2\Lambda} = a$, then, $\forall k \geq 0$:

$$\frac{1}{k+1} \sum_{i=0}^{k} \|F(\bar{\mathbf{u}}_i)\|_{p^*}^{p^*} \leq \frac{2\|\mathbf{u}^* - \mathbf{u}_0\|_p^p}{a^{p^*}(k+1)} + \frac{2p\delta}{a^{p^*-1}}.$$

In particular, for any $\epsilon > 0$, there is a choice of $\delta = \frac{\epsilon^2}{C_p L}$, where C_p is a constant that only depends on p, such that EG_p+ can output a point $\mathbf{u}$ with $\|F(\mathbf{u})\|_{p^*} \leq \epsilon$ in at most

$$k = O_p\Big(\Big(\frac{L\|\mathbf{u}^* - \mathbf{u}_0\|_p}{\epsilon}\Big)^p\Big)$$

iterations. Here, the O_p notation hides constants that only depend on p.

Proof. Observe that, as $\bar{\boldsymbol{\eta}}_i = \boldsymbol{\eta}_i = \mathbf{0}, \forall i \geq 0$ and $\rho = 0$, Lemma 4.1 and the definition of h_k give:

$$0 \leq h_k \leq \phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) + \frac{\beta - m_p}{q} \|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^q$$

$$+ \frac{a_k \Lambda_k \gamma - \beta}{q} \|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^q + \frac{a_k \Lambda_k / \gamma - \beta m_p}{q} \|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q + a_k \delta_k. \tag{4.12}$$

Proof of Part (i). In this case, we can set $\delta_k = 0$ (see Lemma 4.1), $\Lambda_k = L$, and $q = 2$. Therefore, setting $\beta = m_p$, $a_k = \frac{m_p^{3/2}}{2L}$, and $\gamma = \frac{1}{\sqrt{m_p}}$ we get from Eq. (4.12) that

$$\phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) \leq \phi_p(\mathbf{u}^*, \mathbf{u}_k) - \frac{m_p}{4} \|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^2. \tag{4.13}$$

It follows that $\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^2$ converges to zero as $k \to \infty$. By the definition of $\bar{\mathbf{u}}_k$ and Proposition 2.5, $\frac{1}{2} \|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^2 = \frac{a_k^2}{2\beta^2} \|F(\mathbf{u}_k)\|_{p^*}^2$, and so $\|F(\mathbf{u}_k)\|_{p^*}$ converges to zero as $k \to \infty$. Further, as $\phi_p(\mathbf{u}^*, \mathbf{u}_k) \leq \phi_p(\mathbf{u}^*, \mathbf{u}_0) < \infty$ and $\phi_p(\mathbf{u}^*, \mathbf{u}_k) \geq \frac{m_p}{2} \|\mathbf{u}^* - \mathbf{u}_k\|_p^2$, $m_p > 0$, it follows that $\|\mathbf{u}^* - \mathbf{u}_k\|_p$ is bounded, and, thus, $\{\mathbf{u}_k\}_{k \geq 0}$ is a bounded sequence. The proof that all accumulation points of $\{\mathbf{u}_k\}_{k \geq 0}$ are in $\mathcal{U}^*$ is standard and omitted (see the proof of Theorem 3.2 for a similar argument).

To bound $\frac{1}{k+1}\sum_{i=0}^{k}\|F(\mathbf{u}_i)\|_{p^*}^2$, we telescope the inequality from Eq. (4.13) to get:

$$m_p \sum_{i=0}^{k}\|\bar{\mathbf{u}}_i - \mathbf{u}_i\|_p^2 \leq 4(\phi_p(\mathbf{u}^*, \mathbf{u}_0) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1})) \leq 4\phi_p(\mathbf{u}^*, \mathbf{u}_0).$$

To complete the proof of this part, it remains to use the fact that $\|\bar{\mathbf{u}}_i - \mathbf{u}_i\|_p^2 = \frac{a_k^2}{\beta^2}\|F(\mathbf{u}_i)\|_{p^*}^2$ (already argued above), the definitions of a_k and β, and $m_p = p - 1$. The bound on $\phi_p(\mathbf{u}^*, \mathbf{u}_0)$ follows from the definition of ϕ_p in this case. In particular, if we denote $\psi(\mathbf{u}) = \frac{1}{2}\|\mathbf{u} - \mathbf{u}_0\|_p^2$, then $\phi_p(\mathbf{u}^*, \mathbf{u}_0) = D_\psi(\mathbf{u}^*, \mathbf{u}_0)$. Using the definition of Bregman divergence and the fact that, for this choice of ψ, we have $\|\nabla\psi(\mathbf{u})\|_{p^*} = \|\mathbf{u} - \mathbf{u}_0\|_p, \forall \mathbf{u} \in \mathbb{R}^d$, (see the last part of Proposition 2.5) it follows that:

$$\phi_p(\mathbf{u}^*, \mathbf{u}_0) = \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_0\|_p^2 - \frac{1}{2}\|\mathbf{u}_0 - \mathbf{u}_0\|_p^2 - \left\langle \nabla_{\mathbf{u}}\left(\frac{1}{2}\|\mathbf{u} - \mathbf{u}_0\|_p^2\right)\Big|_{\mathbf{u}=\mathbf{u}_0}, \mathbf{u}^* - \mathbf{u}_0 \right\rangle$$

$$= \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_0\|_p^2.$$

Proof of Part (ii). In this case, $q = p$, $\phi_p(\mathbf{u}, \mathbf{v}) = \frac{1}{p}\|\mathbf{u} - \mathbf{v}\|_p^p$, and $m_p = 1$. Using Proposition 2.5, $\|\mathbf{u}_k - \bar{\mathbf{u}}_k\|_p^p = \frac{a_k^{p^*}}{\beta^{p^*}}\|F(\mathbf{u}_k)\|_{p^*}^{p^*}$ and $\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^p = a_k^{p^*}\|F(\bar{\mathbf{u}}_k)\|_{p^*}^{p^*}$. Combining with Eq. (4.12), we have:

$$\begin{aligned}
0 \leq \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_k\|_p^p - \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|_p^p &+ \frac{(\beta - 1)a_k^{p^*}}{p}\|F(\bar{\mathbf{u}}_k)\|_{p^*}^{p^*} \\
+ \frac{(a_k\Lambda_k\gamma - \beta)a_k^{p^*}}{p\beta^{p^*}}\|F(\mathbf{u}_k)\|_{p^*}^{p^*} &+ \frac{a_k\Lambda_k/\gamma - \beta}{p}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^p + a_k\delta_k.
\end{aligned} \tag{4.14}$$

Now let $\gamma = 1$, $\beta = \frac{1}{2}$, $\delta_k = \delta > 0$, and $a_k = \frac{1}{2\Lambda_k} = \frac{1}{2\Lambda} = a$. Then $a_k\Lambda_k\gamma - \beta = a_k\Lambda_k/\gamma - \beta = 0$ and Eq. (4.14) simplifies to:

$$\frac{a^{p^*}}{2p}\|F(\bar{\mathbf{u}}_k)\|_{p^*}^{p^*} \leq \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_k\|_p^p - \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|_p^p + a\delta.$$

Telescoping the last inequality and then dividing by $\frac{a_k^{p^*}(k+1)}{2p}$, we have:

$$\frac{1}{k+1}\sum_{i=0}^{k}\|F(\bar{\mathbf{u}}_i)\|_{p^*}^{p^*} \leq \frac{2\|\mathbf{u}^* - \mathbf{u}_0\|_p^p}{a^{p^*}(k+1)} + \frac{2p\delta}{a^{p^*-1}}. \tag{4.15}$$

Now, for EG$_p$+ to be able to output a point $\mathbf{u}$ with $\|F(\mathbf{u})\|_{p^*} \leq \epsilon$, it suffices to show that for some choice of δ and k we can make the right-hand side of Eq. (4.15) at most ϵ^{p^*}. This is true because then EG$_p$+ can output the point $\bar{\mathbf{u}}_i = \arg\min_{0 \leq i \leq k}\|F(\bar{\mathbf{u}}_i)\|_{p^*}$. For stochastic setups, the guarantee would be in expectation, and EG$_p$+ could output a point $\bar{\mathbf{u}}_i$ with i chosen uniformly at random from $\{0, \ldots, k\}$, as discussed in the proof of Theorem 3.2.

Observe first that, as $\Lambda = \left(\frac{p-2}{p\delta}\right)^{\frac{p-2}{2}}L^{p/2}$ and $p^* = \frac{p}{p-1}$, we have that:

$$\begin{aligned}
\frac{\delta}{a^{p^*-1}} &= \delta(2\Lambda)^{p^*-1} = \delta 2^{\frac{1}{p-1}}\Lambda^{\frac{1}{p-1}} \\
&= 2^{\frac{1}{p-1}}\delta^{\frac{p}{2(p-1)}}\left(\frac{p-2}{p}\right)^{\frac{p-2}{2(p-1)}}L^{\frac{p}{2(p-1)}}.
\end{aligned}$$

Setting $\frac{2p\delta}{a^{p^*-1}} \leq \frac{\epsilon^{p^*}}{2}$, recalling that $p^* = \frac{p}{p-1}$, and rearranging, we have:

$$\delta^{\frac{p^*}{2}} \leq \frac{\epsilon^{p^*}}{2^{\frac{2p-1}{p}}p}\left(\frac{p}{p-2}\right)^{\frac{p-2}{2p}p^*}L^{-p^*/2}.$$

Equivalently:

$$\delta \leq \frac{\epsilon^2}{L \cdot 2^{\frac{2(2p-1)}{p}}p^{\frac{2(p-1)}{p}}\left(\frac{p-2}{p}\right)^{\frac{p-2}{p}}}.$$

It can be verified numerically that $\left(\frac{p-2}{p}\right)^{\frac{p-2}{p}}$ is a constant between $\frac{1}{e}$ and 1, while it is clear that $2^{\frac{2(2p-1)}{p}}p^{\frac{2(p-1)}{p}} = O(p^2)$ is a constant that only depends on p. Hence, it suffices to set $\delta = \frac{\epsilon^2}{C_p L}$, where $C_p = 2^{\frac{2(2p-1)}{p}}p^{\frac{2(p-1)}{p}}$.

It remains to bound the number of iterations k so that $\frac{2\|\mathbf{u}^* - \mathbf{u}_0\|_p^p}{a^{p^*}(k+1)} \leq \frac{\epsilon^{p^*}}{2}$. Equivalently, we need $k+1 \geq \frac{4\|\mathbf{u}^* - \mathbf{u}_0\|_p^p}{a^{p^*}\epsilon^{p^*}}$. Plugging $\delta = \frac{\epsilon^2}{C_p L}$ into the definition of Λ, using that $p^* = \frac{p}{p-1}$, and simplifying, we have:

$$a^{p^*} = (2\Lambda)^{p^*} = 2^{\frac{p}{p-1}}\left(\frac{p-2}{p\delta}\right)^{\frac{p-2}{2}\cdot\frac{p}{p-1}}L^{\frac{p}{2}\cdot\frac{p}{p-1}}$$

$$= O_p\left(\left(\frac{1}{\epsilon}\right)^{\frac{p(p-2)}{p-1}}L^p\right).$$

Thus,

$$k = O_p\left(\left(\frac{1}{\epsilon}\right)^{\frac{p(p-2)}{p-1}+\frac{p}{p-1}}L^p\|\mathbf{u}^* - \mathbf{u}_0\|_p^p\right) = O_p\left(\left(\frac{L\|\mathbf{u}^* - \mathbf{u}_0\|_p}{\epsilon}\right)^p\right),$$

as claimed. □

Remark 4.3. There are significant technical obstacles in generalizing the results from Theorem 4.2 to settings with $\rho > 0$. In particular, when $p \in (1, 2)$, the proof fails because we take $\phi_p(\mathbf{u}^*, \mathbf{u})$ to be the Bregman divergence of $\|\cdot - \mathbf{u}_0\|_p^2$, and relating $\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p$ to $\|F(\mathbf{u}_k)\|_{p^*}$ would require $\|\cdot\|_p^2$ to be smooth, which is not true. If we had, instead, used $\|\mathbf{u}^* - \mathbf{u}\|_p^2$ in place of $\phi_p(\mathbf{u}^*, \mathbf{u})$, we would have incurred $\frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|_p^2 - \frac{m_p}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|_p^2$ in the upper bound on h_k, which would not telescope, as in this case $m_p < 1$. In the case of $p > 2$, the challenges come from a delicate relationship between the step sizes a_k and error terms δ_k. It turns out that it is possible to guarantee local convergence (in the region where $\|F(\bar{\mathbf{u}}_k)\|_2$ is bounded by a constant less than 1) with $\rho > 0$, but ρ would need to scale with $\text{poly}(\epsilon)$ in this case. As this is a weak result whose usefulness is unclear, we have omitted it.

Stochastic oracle access. To obtain results for stochastic oracle access to F, we only need to bound the terms $\mathcal{E}^s \overset{\text{def}}{=} -a_k\langle\bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^*\rangle - a_k\langle\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle$ from Lemma 4.1 corresponding to the stochastic error in expectation, while for the rest of the analysis we can appeal to the results for the deterministic oracle access to F. In the case of $p = 2$, there is one additional term that appears in h_k due to replacing $F(\bar{\mathbf{u}}_k)$ with $\tilde{F}(\bar{\mathbf{u}}_k)$. This term is simply equal to:

$$\frac{a_k\rho}{2}\mathbb{E}[\|\tilde{F}(\bar{\mathbf{u}}_k)\|_2^2 - \|F(\bar{\mathbf{u}}_k)\|_2^2|\bar{\mathcal{F}}_k] = \frac{a_k\rho}{2}\mathbb{E}[\|F(\bar{\mathbf{u}}_k) + \bar{\boldsymbol{\eta}}_k\|_2^2 - \|F(\bar{\mathbf{u}}_k)\|_2^2|\bar{\mathcal{F}}_k] = \frac{a_k\rho}{2}\bar{\sigma}_k^2. \quad (4.16)$$

We start by bounding the stochastic error $\mathcal{E}^s$ in expectation.

Lemma 4.4. *Let $\mathcal{E}^s = -a_k\langle\bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^*\rangle - a_k\langle\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle$, where $\bar{\boldsymbol{\eta}}_k$ and $\boldsymbol{\eta}_k$ are defined as in Eq. (4.2) and all the assumptions of Theorem 4.5 below apply. Then, for q defined by Eq. (4.3) and any $\tau > 0$:*

$$\mathbb{E}[\mathcal{E}^s] \leq \frac{2^{q^*/2}a_k^{q^*}(\sigma_k^2 + \bar{\sigma}_k^2)^{q^*/2}}{q^*\tau^{q^*}} + \mathbb{E}\left[\frac{\tau^q}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q\right],$$

where the expectation is w.r.t. all the randomness in the algorithm.

Proof. Let us start by bounding $-a_k\langle\bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^*\rangle$ first. Conditioning on $\bar{\mathcal{F}}_k$, $\bar{\boldsymbol{\eta}}_k$ is independent of $\bar{\mathbf{u}}_k$ and $\mathbf{u}^*$, and, thus:

$$\mathbb{E}[-a_k\langle\bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^*\rangle] = \mathbb{E}[\mathbb{E}[-a_k\langle\bar{\boldsymbol{\eta}}_k, \bar{\mathbf{u}}_k - \mathbf{u}^*\rangle|\bar{\mathcal{F}}_k]] = 0.$$

The second term, $-a_k\langle\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle$, can be bounded using Hölder's inequality and Young's inequality as follows:

$$\mathbb{E}\left[-a_k\langle\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle\right] \leq \mathbb{E}\left[a_k\|\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k\|_{p^*}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p\right]$$

$$\leq \mathbb{E}\left[\frac{a_k^{q^*}\|\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k\|_{p^*}^{q^*}}{q^*\tau^{q^*}}\right] + \mathbb{E}\left[\frac{\tau^q}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q\right].$$

15

It remains to bound $\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k\|_{p^*}^{q^*}\big]$. Using the triangle inequality,

$$\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k\|_{p^*}^{q^*}\big] \leq \mathbb{E}\Big[\big(\|\bar{\boldsymbol{\eta}}_k\|_{p^*} + \|\boldsymbol{\eta}_k\|_{p^*}\big)^{q^*}\Big]$$

$$= \mathbb{E}\Big[\big((\|\bar{\boldsymbol{\eta}}_k\|_{p^*} + \|\boldsymbol{\eta}_k\|_{p^*})^2\big)^{q^*/2}\Big]$$

$$\leq \Big(\mathbb{E}\big[(\|\bar{\boldsymbol{\eta}}_k\|_{p^*} + \|\boldsymbol{\eta}_k\|_{p^*})^2\big]\Big)^{q^*/2},$$

where the last line is by Jensen's inequality, as $q^* \in (1,2]$, and so $(\cdot)^{q^*/2}$ is concave. Using Young's inequality and linearity of expectation:

$$\mathbb{E}\big[(\|\bar{\boldsymbol{\eta}}_k\|_{p^*} + \|\boldsymbol{\eta}_k\|_{p^*})^2\big] \leq 2\Big(\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_k\|_{p^*}^2\big] + \mathbb{E}\big[\|\boldsymbol{\eta}_k\|_{p^*}^2\big]\Big)$$

$$\leq 2(\sigma_k{}^2 + \bar{\sigma}_k^2).$$

Putting everything together:

$$\mathbb{E}\big[\|\bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k\|_{p^*}^{q^*}\big] \leq 2^{q^*/2}(\sigma_k{}^2 + \bar{\sigma}_k^2)^{q^*/2}$$

and

$$\mathbb{E}[\mathcal{E}^s] = \mathbb{E}\big[-a_k \langle \bar{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k, \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle\big]$$

$$\leq \frac{2^{q^*/2} a_k{}^{q^*}(\sigma_k{}^2 + \bar{\sigma}_k^2)^{q^*/2}}{q^* \tau^{q^*}} + \mathbb{E}\Big[\frac{\tau^q}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q\Big],$$

as claimed. $\qquad\square$

We are now ready to bound the total oracle complexity of EG_p+ (and its special case $\mathrm{EG}+$), as follows.

Theorem 4.5. *Let $p > 1$ and let $F : \mathbb{R}^d \to \mathbb{R}^d$ be an arbitrary L-Lipschitz operator w.r.t. $\|\cdot\|_p$ that satisfies Assumption 1 for some $\mathbf{u}^* \in \mathcal{U}^*$. Given an arbitrary initial point $\mathbf{u}_0 \in \mathbb{R}^d$, let the sequences of points $\{\mathbf{u}_i\}_{i\geq 1}$, $\{\bar{\mathbf{u}}_i\}_{i\geq 0}$ evolve according to (EG_p+) for some $\beta \in (0,1]$ and positive step sizes $\{a_i\}_{i\geq 0}$. Let the variance of a single query to the stochastic oracle $\tilde{F}$ be bounded by some $\sigma^2 < \infty$.*

(i) Let $p = 2$ and $\rho \in \big[0, \bar{\rho}\big)$, where $\bar{\rho} = \frac{1}{4\sqrt{2}L}$. If $\beta = \frac{1}{2}$ and $a_k = \frac{1}{2\sqrt{2}L}$, then EG_p+ can output a point $\mathbf{u}$ with $\mathbb{E}[\|\tilde{F}(\mathbf{u})\|_2] \leq \epsilon$ with at most

$$O\Big(\frac{L\|\mathbf{u}^* - \mathbf{u}_0\|_2^2}{\epsilon^2(\bar{\rho} - \rho)}\Big(1 + \frac{\sigma^2}{L\epsilon^2(\bar{\rho} - \rho)}\Big)\Big)$$

oracle queries to $\tilde{F}$.

(ii) Let $p \in (1,2]$ and $\rho = 0$. If $a_k = \frac{m_p{}^{3/2}}{2L}$ and $\beta = m_p$, then EG_p+ can output a point $\mathbf{u}$ with $\mathbb{E}[\|\tilde{F}(\mathbf{u})\|_{p^*}] \leq \epsilon$ with at most

$$O\Big(\frac{L^2\|\mathbf{u}^* - \mathbf{u}_0\|_p^2}{m_p{}^2\epsilon^2}\Big(1 + \frac{\sigma^2}{m_p\epsilon^2}\Big)\Big)$$

oracle queries to $\tilde{F}$, where $m_p = p - 1$.

(iii) Let $p > 2$ and $\rho = 0$. If $\beta = \frac{1}{2}$ and $a_k = a = \frac{1}{4\Lambda}$, then EG_p+ can output a point $\mathbf{u}$ with $\mathbb{E}[\|\tilde{F}(\mathbf{u})\|_{p^*}] \leq \epsilon$ with at most

$$O_p\Big(\Big(\frac{L\|\mathbf{u}^* - \mathbf{u}_0\|_p}{\epsilon}\Big)^p\Big(1 + \Big(\frac{\sigma}{\epsilon}\Big)^{p^*}\Big)\Big)$$

oracle queries to $\tilde{F}$, where $p^* = \frac{p}{p-1}$.

Proof. Combining Lemmas 4.1 and 4.4, we have, $\forall k \geq 0$:

$$0 \leq \mathbb{E}[h_k] \leq \frac{2^{q^*/2} a_k{}^{q^*} (\sigma_k{}^2 + \bar{\sigma}_k^2)^{q^*/2}}{q^* \tau^{q^*}} + \mathbb{E}\Big[\frac{\tau^q}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q\Big] + \frac{a_k \rho \bar{\sigma}_k{}^2}{2} + \mathbb{E}\Big[\frac{a_k \rho}{2}\|\tilde{F}(\bar{\mathbf{u}}_k)\|_{p^*}^2\Big]$$

$$+ \mathbb{E}\Big[\phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) + \frac{\beta - m_p}{q}\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^q\Big] \qquad (4.17)$$

$$+ \mathbb{E}\Big[\frac{a_k \Lambda_k \gamma - \beta}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^q + \frac{a_k \Lambda_k / \gamma - \beta m_p}{q}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^q + a_k \delta_k\Big],$$

Proof of Part (i). In this case, $q = 2$, $m_p = 1$, $\delta = 0$, $\Lambda_k = L$, and $\phi_p(\mathbf{u}^*, \mathbf{u}) = \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}\|_2^2$, and, further, $\mathbf{u}_{k+1} - \mathbf{u}_k = -a_k F(\bar{\mathbf{u}}_k)$, so Eq. (4.17) simplifies to

$$0 \leq \mathbb{E}[h_k] \leq \frac{2 a_k{}^2 (\bar{\sigma}_k{}^2 + \bar{\sigma}_k^2)}{2\tau^2} + \frac{a_k \rho \sigma_k{}^2}{2}$$

$$+ \mathbb{E}\Big[\frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|_2^2 - \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|_2^2 + \frac{a_k{}^2(\beta - 1) + a_k \rho}{2}\|\tilde{F}(\bar{\mathbf{u}}_k)\|_2^2\Big]$$

$$+ \mathbb{E}\Big[\frac{a_k L \gamma - \beta}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_2^2 + \frac{a_k L/\gamma - \beta + \tau^2}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_2^2\Big].$$

Taking $\beta = \frac{1}{2}$, $\tau^2 = \frac{1}{4}$, $\gamma = \sqrt{2}$, and $a_k = \frac{1}{2\sqrt{2}L}$, and recalling that $\bar{\rho} = \frac{1}{4\sqrt{2}L}$, we have:

$$a_k(\bar{\rho} - \rho)\mathbb{E}\big[\|\tilde{F}(\bar{\mathbf{u}}_k)\|_2^2\big] \leq \mathbb{E}\big[\|\mathbf{u}^* - \mathbf{u}_k\|_2^2 - \|\mathbf{u}^* - \mathbf{u}_{k+1}\|_2^2\big] + 4 a_k{}^2(\sigma_k{}^2 + \bar{\sigma}_k)^2 + \frac{a_k \rho \bar{\sigma}_k{}^2}{2}.$$

Telescoping the last inequality and dividing both sides by $a_k(\bar{\rho} - \rho)(k + 1)$, we get:

$$\frac{1}{k+1}\sum_{i=0}^k \mathbb{E}\big[\|\tilde{F}(\bar{\mathbf{u}}_i)\|_2^2\big] \leq \frac{2\sqrt{2}L\|\mathbf{u}^* - \mathbf{u}_0\|_2^2}{(k+1)(\bar{\rho} - \rho)} + \frac{\sqrt{2}\sum_{i=0}^k (\sigma_i{}^2 + \bar{\sigma}_i^2)}{L(\bar{\rho} - \rho)(k+1)} + \frac{\rho \sum_{i=0}^k \bar{\sigma}_i^2}{2(k+1)(\bar{\rho} - \rho)}.$$

In particular, if the variance of a single sample of $\tilde{F}$ evaluated at an arbitrary point is σ^2 and we take n samples of $\tilde{F}$ in each iteration, then:

$$\frac{1}{k+1}\sum_{i=0}^k \mathbb{E}\big[\|\tilde{F}(\bar{\mathbf{u}}_i)\|_2^2\big] \leq \frac{2\sqrt{2}L\|\mathbf{u}^* - \mathbf{u}_0\|_2^2}{(k+1)(\bar{\rho} - \rho)} + \frac{\sigma^2(4\sqrt{2}/L + \rho)}{2n(\bar{\rho} - \rho)}.$$

To finish the proof of this part, we require that both terms on the right-hand side of the last inequality are bounded by $\frac{\epsilon^2}{2}$. For the first term, this leads to:

$$k = \Big\lceil \frac{4\sqrt{2}L\|\mathbf{u}^* - \mathbf{u}_0\|_2^2}{\epsilon^2(\bar{\rho} - \rho)} - 1\Big\rceil = O\Big(\frac{L\|\mathbf{u}^* - \mathbf{u}_0\|_2^2}{\epsilon^2(\bar{\rho} - \rho)}\Big).$$

For the second term, the bound is:

$$n = \Big\lceil \frac{2\sigma^2(4\sqrt{2}/L + \rho)}{\epsilon^2(\bar{\rho} - \rho)}\Big\rceil = O\Big(\frac{\sigma^2}{L\epsilon^2(\bar{\rho} - \rho)}\Big).$$

Thus, the total number of required oracle queries to $\tilde{F}$ is bounded by:

$$k(1 + n) = O\Big(\frac{L\|\mathbf{u}^* - \mathbf{u}_0\|_2^2}{\epsilon^2(\bar{\rho} - \rho)}\Big(1 + \frac{\sigma^2}{L\epsilon^2(\bar{\rho} - \rho)}\Big)\Big).$$

As discussed before, $\bar{\mathbf{u}}_i$ with i chosen uniformly at random from $\{0, \ldots, k\}$ will satisfy $\|\tilde{F}(\bar{\mathbf{u}}_i)\|_2 \leq \epsilon$ in expectation.

Proof of Part (ii). In this case, $q = 2$, $m_p = p - 1$, $\delta = 0$, $\Lambda_k = L$, and $\rho = 0$. Thus, Eq. (4.17) simplifies to:

$$0 \le \mathbb{E}[h_k] \le \frac{2a_k{}^2(\sigma_k{}^2 + \bar{\sigma}_k^2)}{2\tau^2}$$
$$+ \mathbb{E}\Big[\phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1}) + \frac{\beta - m_p}{2}\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^2\Big]$$
$$+ \mathbb{E}\Big[\frac{a_k L\gamma - \beta}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^2 + \frac{a_k L/\gamma - \beta m_p + \tau^2}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^2\Big].$$

In this case, the same choices for a_k and β as in the deterministic case suffice. In particular, let $a_k = \frac{m_p{}^{3/2}}{2L}$, $\beta = m_p$, $\gamma = \frac{1}{\sqrt{m_p}}$, and $\tau^2 = \frac{m_p{}^2}{2}$. Then, using the fact that $\frac{1}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^2 = \frac{a_k{}^2}{2\beta^2}\|\tilde{F}(\mathbf{u}_k)\|_{p^*}^2$, from Proposition 2.5, we have

$$\frac{a_k{}^2 m_p}{4\beta^2}\mathbb{E}\big[\|\tilde{F}(\mathbf{u}_k)\|_{p^*}^2\big] \le \mathbb{E}\Big[\phi_p(\mathbf{u}^*, \mathbf{u}_k) - \phi_p(\mathbf{u}^*, \mathbf{u}_{k+1})\Big] + \frac{a_k{}^2(\sigma_k{}^2 + \bar{\sigma}_k^2)}{\tau}.$$

Telescoping the last inequality and dividing both sides by $(k+1)\frac{a_k{}^2 m_p}{4\beta^2}$, we have:

$$\frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\big[\|\tilde{F}(\mathbf{u}_i)\|_{p^*}^2\big] \le \frac{16L^2\phi_p(\mathbf{u}^*, \mathbf{u}_0)}{(k+1)m_p{}^2} + \frac{8\sum_{i=0}^{k}(\sigma_i{}^2 + \bar{\sigma}_i{}^2)}{(k+1)m_p}. \tag{4.18}$$

Now let $\sigma_i{}^2 = \bar{\sigma}_i^2 = \sigma^2/n$, where σ^2 is the variance of a single sample of $\tilde{F}$ and n is the number of samples taken per iteration. Then, similarly as in Part (i), to bound the total number of samples it suffices to bound each term on the right-hand side of Eq. (4.18) by $\frac{\epsilon^2}{2}$. The first term was already bounded in Theorem 4.2, and thus we obtain:

$$k = O\Big(\frac{L^2\|\mathbf{u}^* - \mathbf{u}_0\|_p^2}{m_p{}^2\epsilon^2}\Big).$$

For the second term, it suffices that:

$$n = O\Big(\frac{\sigma^2}{m_p\epsilon^2}\Big),$$

and the bound on the total number of samples follows.

Proof of Part (iii). In this case, $q = p$, $m_p = 1$, $\rho = 0$, $\phi_p(\mathbf{u}^*, \mathbf{u}) = \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}\|_p^p$, and we take $\delta_k = \delta > 0$, $\Lambda_k = \Lambda = \big(\frac{p-2}{p\delta}\big)^{\frac{p-2}{2}}L^{\frac{p}{2}}$. Eq. (4.17) now simplifies to:

$$0 \le \mathbb{E}[h_k] \le \frac{2^{p^*/2}a_k{}^{p^*}(\sigma_k{}^2 + \bar{\sigma}_k^2)^{p^*/2}}{p^*\tau^{p^*}}$$
$$+ \mathbb{E}\Big[\frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_k\|_p^p - \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|_p^p + \frac{\beta - 1}{p}\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^p\Big] \tag{4.19}$$
$$+ \mathbb{E}\Big[\frac{a_k\Lambda\gamma - \beta}{p}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|_p^p + \frac{a_k\Lambda/\gamma + \tau^p - \beta}{p}\|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|_p^p + a_k\delta\Big].$$

Recall that, by Proposition 2.5, $\frac{1}{p}\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_p^p = \frac{a^{p^*}}{p}\|\tilde{F}(\bar{\mathbf{u}}_k)\|_{p^*}^{p^*}$. Let $\beta = \frac{1}{2}$, $a_k = a = \frac{1}{4\Lambda}$, $\tau^p = \frac{1}{4}$, and $\gamma = 1$. Then $\beta - 1 = -\frac{1}{2}$, $a_k\Lambda\gamma - \beta = -\frac{1}{4} < 0$, and $a_k\Lambda/\gamma + \tau^p - \beta = 0$, and Eq. (4.19) leads to:

$$\frac{a^{p^*}}{2p}\mathbb{E}\big[\|\tilde{F}(\bar{\mathbf{u}}_k)\|_{p^*}^{p^*}\big] \le \mathbb{E}\Big[\frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_k\|_p^p - \frac{1}{p}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|_p^p\Big] + \frac{2^{\frac{4+p}{2(p-1)}}a^{p^*}(\sigma_k{}^2 + \bar{\sigma}_k^2)^{p^*/2}}{p^*} + a\delta.$$

Telescoping the last inequality and then dividing both sides by $\frac{a^{p^*}}{2p}(k+1)$, we have:

$$\frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\big[\|\tilde{F}(\bar{\mathbf{u}}_i)\|_{p^*}^{p^*}\big] \le \frac{2\|\mathbf{u}^* - \mathbf{u}_0\|_p^p}{a^{p^*}(k+1)} + \frac{2^{\frac{3p+2}{2(p-1)}}p\sum_{i=0}^{k}(\sigma_i{}^2 + \bar{\sigma}_i^2)^{p^*/2}}{p^*(k+1)} + \frac{2p\delta}{a^{p^*} - 1}.$$

18

Now let σ^2 be the variance of a single sample of $\tilde{F}$ and suppose that in each iteration we take n samples to estimate $F(\bar{\mathbf{u}}_i)$ and $F(\mathbf{u}_i)$. Then $\sigma_i^2 = \bar{\sigma}_i^2 = \frac{\sigma^2}{n}$, and the last equation simplifies to

$$\frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\big[\|\tilde{F}(\bar{\mathbf{u}}_i)\|_{p^*}^{p^*}\big] \leq \frac{2\|\mathbf{u}^* - \mathbf{u}_0\|_p^p}{a^{p^*}(k+1)} + \frac{2^{\frac{p+2}{p-1}}p\sigma^{p^*}}{p^*n} + \frac{2p\delta}{a^{p^*}-1}.$$

To complete the proof, as before it suffices to show that we can choose k and n so that $\frac{2p\|\mathbf{u}^*-\mathbf{u}_0\|_p^p}{a^{p^*}(k+1)} + \frac{2p\delta}{a^{p^*}-1} \leq \frac{\epsilon^{p^*}}{2}$ and $\frac{2^{\frac{p+2}{p-1}}p\sigma^{p^*}}{p^*n} \leq \frac{\epsilon^{p^*}}{2}$. For the former, following the same argument as in the proof of Theorem 4.2, Part (ii), it suffices to choose $\delta = O_p(\frac{\epsilon^2}{L})$, which leads to:

$$k = O_p\left(\left(\frac{L\|\mathbf{u}^* - \mathbf{u}_0\|_p}{\epsilon}\right)^p\right).$$

For the latter, it suffices to choose:

$$n = \frac{2^{\frac{p+2}{p-1}+1}p\sigma^{p^*}}{p^*\epsilon^{p^*}} = O\left(\frac{p\sigma^{p^*}}{\epsilon^{p^*}}\right).$$

The total number of queries to the stochastic oracle is then bounded by $k(1+n)$. $\qquad\square$

5 Discussion

We introduced a new class of structured nonconvex-nonconcave min-max optimization problems and proposed a new generalization of the extragradient method that provably converges to a stationary point in Euclidean setups. Our algorithmic results guarantee that problems in this class contain at least one stationary point (an (SVI) solution, see Remark 3.5). The class we introduced generalizes other important classes of structured nonconvex-nonconcave problems, such as those in which an (MVI) solution exists. We further generalized our results to stochastic setups and ℓ_p-normed setups in which an (MVI) solution exists. An interesting direction for future research is to understand to what extent we can further relax the assumptions about the structure of nonconvex-nonconcave problems, while maintaining computational feasibility of algorithms that can address them.

Acknowledgements

References

Jacob Abernethy, Kevin A Lai, and Andre Wibisono. Last-iterate convergence rates for min-max optimization. *arXiv preprint arXiv:1906.02027*, 2019.

Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach. In *Proc. AISTATS'19*, 2019.

Mohammad Alkousa, Darina Dvinskikh, Fedor Stonyakin, and Alexander Gasnikov. Accelerated methods for composite non-bilinear saddle point problem. *arXiv preprint arXiv:1906.03620*, 2019.

Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of extragradient for a whole spectrum of differentiable games. In *Proc. AISTATS'20*, 2020.

Heinz H Bauschke, Walaa M Moursi, and Xianfu Wang. Generalized monotone operators and their averaged resolvents. *Mathematical Programming*, pages 1–20, 2020.

Dimitri P Bertsekas. *Control of uncertain systems with a set-membership description of the uncertainty*. PhD thesis, MIT, 1971.

Dimitri P Bertsekas, Angelia Nedic, and Asuman E Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, 2003.

J Borwein, A Guirao, Petr Hájek, and J Vanderwerff. Uniformly convex functions on Banach spaces. *Proceedings of the American Mathematical Society*, 137(3):1081–1091, 2009.

Jonathan M Borwein and Qiji J Zhu. *Techniques of Variational Analysis*. Springer, 2004.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Patrick L Combettes and Teemu Pennanen. Proximal methods for cohypomonotone operators. *SIAM journal on control and optimization*, 43(2):731–742, 2004.

Cong D Dang and Guanghui Lan. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and applications*, 60(2): 277–310, 2015.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Proc. NeurIPS'18*, pages 9236–9246, 2018.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *Proc. ITCS'19*, 2019.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *Proc. ICLR'18*, 2018.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Proc. NeurIPS'20*, 2020.

Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. In *Proc. STOC'21*, 2021.

Alexandre d'Aspremont, Cristobal Guzman, and Martin Jaggi. Optimal affine-invariant smooth minimization algorithms. *SIAM Journal on Optimization*, 28(3):2384–2405, 2018.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.

Jelena Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *Proc. COLT'20*, 2020.

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 *Proc. AISTATS'19*, 2019.

Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman E. Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Proc. COLT'2020*, 2020.

Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401*, 2018.

Michael D Hirsch, Christos H Papadimitriou, and Stephen A Vavasis. Exponential lower bounds for finding brouwer fix points. *Journal of Complexity*, 5(4):379–416, 1989.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *Proc. ICML'20*, 2020.

David Kinderlehrer and Guido Stampacchia. *An introduction to variational inequalities and their applications*. SIAM, 2000.

Weiwei Kong and Renato DC Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *arXiv preprint arXiv:1905.13433*, 2019.

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

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *Proc. AISTATS'19*, 2019.

Qihang Lin, Mingrui Liu, Hassan Rafique, and Tianbao Yang. Solving weakly-convex-weakly-concave saddle-point problems as successive strongly monotone variational inequalities. *arXiv preprint arXiv:1810.10207*, 2018.

Tianyi Lin, Chi Jin, and Michael Jordan. Near-optimal algorithms for minimax optimization. In *Proc. COLT'20*, 2020a.

Tianyi Lin, Chi Jin, and Michael I Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *Proc. ICML'20*, 2020b.

Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. In *Proc. ICLR'20*, 2020.

Songtao Lu, Ioannis Tsaknakis, Mingyi Hong, and Yongxin Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications. *IEEE Transactions on Signal Processing*, 68: 3676–3691, 2020.

Yura Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, Jul 2019.

Oren Mangoubi and Nisheeth K Vishnoi. Greedy adversarial equilibrium: An efficient alternative to nonconvex-nonconcave min-max optimization. In *Proc. ICLR'21*, 2021.

Oren Mangoubi, Sushant Sachdeva, and Nisheeth K Vishnoi. A provably convergent and practical algorithm for min-max optimization with applications to GANs. *arXiv preprint arXiv:2006.12376*, 2020.

Eric Mazumdar, Lillian J Ratliff, and S Shankar Sastry. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020.

Panayotis Mertikopoulos, Christos H. Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proc. ACM-SIAM SODA'18*, 2018.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *Proc. ICLR'19*, 2019.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *Proc. AISTATS'20*, 2020.

Arkadi Nemirovski. Regular Banach spaces and large deviations of random sums. *Unpublished, E-print: https://www2.isye.gatech.edu/~nemirovs/LargeDev2004.pdf*, 2004.

Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2): 381–404, 2015.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Proc. NeurIPS'19*, 2019.

Christos H Papadimitriou. On the complexity of the parity argument and other inefficient proofs of existence. *Journal of Computer and system Sciences*, 48(3):498–532, 1994.

Chaobing Song, Yichao Zhou, Zhengyuan Zhou, Yong Jiang, and Yi Ma. Optimistic dual extrapolation for coherent non-monotone variational inequalities. In *Proc. NeurIPS'20*, 2020.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth mini-max optimization. In *Proc. NeurIPS'19*, 2019.

Yuanhao Wang, Guodong Zhang, and Jimmy Ba. On solving minimax optimization locally: A follow-the-ridge approach. In *Proc. ICLR'19*, 2019.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. In *Proc. NeurIPS'20*, 2020.

Renbo Zhao. Optimal algorithms for stochastic three-composite convex-concave saddle point problems. *arXiv preprint arXiv:1903.01687*, 2019.

Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen Boyd, and Peter W Glynn. Stochastic mirror descent in variationally coherent optimization problems. In *Proc. NIPS'17*, 2017.