Efficient Algorithms for Smooth Minimax Optimization

Kiran Koshy Thekumprampil
University of Illinois at Urbana-Champaign
thekump2@illinois.edu

Prateek Jain
Microsoft Research, India
prajain@microsoft.com

Praneeth Netrapalli
Microsoft Research, India
praneeth@microsoft.com

Sewoong Oh
University of Washington, Seattle
sewoong@cs.washington.edu

Abstract

This paper studies first order methods for solving smooth minimax optimization problems $\min_x \max_y g(x, y)$ where $g(\cdot, \cdot)$ is smooth and $g(x, \cdot)$ is concave for each x. In terms of $g(\cdot, y)$, we consider two settings – strongly convex and nonconvex – and improve upon the best known rates in both. For strongly-convex $g(\cdot, y),\ \forall y$, we propose a new direct optimal algorithm combining Mirror-Prox and Nesterov's AGD, and show that it can find global optimum in $\widetilde{O}\left(1/k^2\right)$ iterations, improving over current state-of-the-art rate of $O(1/k)$. We use this result along with an inexact proximal point method to provide $\widetilde{O}\left(1/k^{1/3}\right)$ rate for finding stationary points in the nonconvex setting where $g(\cdot, y)$ can be nonconvex. This improves over current best-known rate of $O(1/k^{1/5})$. Finally, we instantiate our result for finite nonconvex minimax problems, i.e., $\min_x \max_{1\leq i\leq m} f_i(x)$, with nonconvex $f_i(\cdot)$, to obtain convergence rate of $O(m^{1/3}\sqrt{\log m}/k^{1/3})$.

1 Introduction

In this paper we study smooth minimax problems of the form:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} g(x, y) \ , \ \ g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R},\ g \text{ is smooth i.e., gradient Lipschitz.} \tag{1}$$

The problem has applications in several domains such as machine learning [15, 29], optimization [5], statistics [3], mathematics [23], and game theory [31]. Given the importance of these problems, there is an extensive body of work that studies various algorithms and their convergence properties. The vast majority of existing results for this problem focus on the convex-concave setting, where $g(\cdot, y)$ is convex for every y and $g(x, \cdot)$ is concave for every x. The best known convergence rate in this setting is $O(1/k)$ for the primal-dual gap, achieved for example by Mirror-Prox [34]. This rate is also known to be optimal for the class of smooth convex-concave problems [41]. A natural question is whether we can achieve a faster convergence if we have strong convexity (as opposed to just convexity) of $g(\cdot, y)$. We answer this in the affirmative, by introducing an algorithm that achieves a convergence rate of $\widetilde{O}\left(1/k^2\right)$ for the general smooth, strongly-convex–concave minimax problem. The algorithm we propose is a novel combination of Mirror-Prox and Nesterov's accelerated gradient descent. This matches the known lower bound of $\Omega(1/k^2)$ from [41], closing the gap up to a poly-logarithmic factor. There also exists a conceptually simple smoothing technique based indirect algorithm, which prefixes the tolerance of ε. However, our goal is to find a direct algorithm which does not prefix the tolerance. Other known methods that obtain a rate of $O(1/k^2)$ in this context are for very special cases, where x and y are connected through a bi-linear term or $g(x, \cdot)$ is linear in y [35, 20, 14, 8, 49, 16, 48].

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Setting	Optimality notion	Previous state-of-the-art	Our results	Smoothing schemes	Lower bound
Convex	Primal-dual gap	$O\left(k^{-1}\right)$ [34]	-	-	$\Omega(k^{-1})$ [41]
Strongly convex	Primal-dual gap	$O\left(k^{-1}\right)$ [34]	$\widetilde{O}\left(k^{-2}\right)$	$\widetilde{O}\left(k^{-2}\right)$	$\Omega(k^{-2})$ [41]
Nonconvex	Approx. stat. point	$O\left(k^{-1/5}\right)$ [18]	$\widetilde{O}\left(k^{-1/3}\right)$	$\widetilde{O}\left(k^{-1/3}\right)$ [26]	-

Table 1: Comparison of our results with previous state-of-the-art. We assume that $g(\cdot,\cdot)$ is smooth (i.e., has Lipschitz gradients) and $g(x,\cdot)$ is concave $\forall x \in \mathcal{X}$. Convexity, strong convexity and nonconvexity in the first column refers to $g(\cdot,y)$ for fixed y. Smoothing schemes are indirect methods using the smoothing technique [36].

While most theoretical results focus on the convex-concave setting, several real world problems fall outside this class. A slightly larger class, which captures several more applications, is the class of smooth nonconvex–concave minimax problems, where $g(x,\cdot)$ is *concave* for every x but $g(\cdot,y)$ can be nonconvex. For example, finite minimax problems, i.e., $\min_x \max_{i=1}^m f_i(x) = \min_x \max_{0\preceq y\preceq 1, \sum_{i=1}^m y_i=1} \sum_i y_i \cdot f_i(x) := g(x,y)$ belong to this class, and so do smooth nonconvex constrained optimization problems [25]. In addition, several machine learning problems with non-decomposable loss functions [22] also belong to this class.

In this general nonconvex concave setting however, we cannot hope to find global optimum efficiently as even the special case of nonconvex optimization is NP-hard. Similar to nonconvex optimization, we might hope to find an approximate stationary point [37].

Our second contribution is a new algorithm and a faster rate for the *general smooth nonconvex–concave* minimax problem. Our algorithm is an inexact proximal point method for the nonconvex function $f(x) := \max_{y\in\mathcal{Y}} g(x,y)$. The key insight is that the proximal point problem in each iteration results in a strongly-convex concave minimax problem, for which we use our improved algorithm to obtain the overall computation/iteration complexity of $\widetilde{O}\left(1/k^{1/3}\right)$ thus improving over the previous best known rate of $O(1/k^{1/5})$ [18][1]. More recently, independent to our work, a smoothing based algorithm has also been proposed to achieve the same $O\left(k^{-1/3}\right)$ rate [26].

Finally, we specialize our result to finite minimax problems, i.e., $\min_x \max_{1\le i\le m} f_i(x)$ where $f_i(x)$ can be nonconvex function but each f_i is a smooth function; nonconvex constrained optimization problems can be reduced to such finite minimax problems. For these, we obtain a rate of $\widetilde{O}\left(m^{1/3}\sqrt{\log m}/k^{1/3}\right)$ total gradient computations which improves upon the state-of-the-art rate $O(m^{1/4}/k^{1/4})$ [11] in this setting as well.

Summary of contributions: See also Table 1.
1. Optimal $\widetilde{O}\left(1/k^2\right)$ convergence rate for smooth, strongly-convex – concave problems, improving upon the previous best known rate of $O\left(1/k\right)$ for a direct algorithm and,
2. $\widetilde{O}\left(1/k^{1/3}\right)$ convergence rate for smooth, nonconvex – concave problems, improving upon the previous best known rate of $O\left(1/k^{1/5}\right)$.

Related works: For strongly-convex-concave minimax problems with special structures, several algorithms have been proposed. In an increasing order of generality, [14, 49, 50] study optimizing a strongly convex function with linear constraints, which can be posed as a special case of minimax optimization, [35, 8] study a minimax problem where x and y are connected only through a bi-linear term $y^T Ax$, and [16, 20] study a case where $g(x,\cdot)$ is linear in y. In all these cases, it is shown that $O(1/k^2)$ convergence rate is achievable if $g(\cdot,y)$ is strongly-convex $\forall$ y. Recently, [12] showed linear convergence of gradient descent ascent for strongly-convex–concave problem with bilinear coupling when A has full row rank. However, it has remained an open question if the fast rate of $O(1/k^2)$ can be achieved for general strongly-convex-concave minimax problems. See [32, 9, 7, 17, 51, 1]

[1]While [18] gives a rate of $O\left(1/k^{1/4}\right)$ with an approximate maximization oracle for $\max_{y\in\mathcal{Y}} g(x,y)$, taking into account the cost of implementing such a maximization oracle gives a rate of $O\left(1/k^{1/5}\right)$.

for detailed surveys on the results for the convex-concave minimax problems. For examples and application of saddle point problems refer [36, 19, 20, 7, 43].

For nonconvex-concave minimax problems, [42] considers both deterministic and stochastic settings, and proposes inexact proximal point methods for solving smooth nonconvex–concave problems. In the deterministic setting, their result guarantees an error of $O(1/k^{1/6})$. We note that there have also been other notions of stationarity proposed in literature for nonconvex-concave minimax problems [28, 40]. These notions however are weaker than the one considered in this paper, in the sense that, our notion of stationarity implies these other notions (without loss in parameters). For one such weaker notion, [40] proposes an algorithm with a convergence rate of $O\left(k^{-1/3.5}\right)$. Since the notion they consider is weaker, it does not imply the same convergence rate in our setting.

We would also like to highlight the works [6, 13, 33, 46, 34, 47, 10] designing efficient algorithms for solving monotone variational inequalities which generalizes the convex-concave minimax problems.

Notations: $\mathbb{R}$ is the real line and for any natural number p, $\mathbb{R}^p$ is the real vector space of dimension p. $\|\cdot\|$ is a norm on some metric space which would be evident from the context. For a convex set $\mathcal{X} \subseteq \mathbb{R}^p$ and $x \in \mathbb{R}^p$, $\mathcal{P}_{\mathcal{X}}(x) = \arg\min_{x' \in \mathcal{X}} \|x - x'\|$ is the projection of x on to $\mathcal{X}$. For a differentiable function $g(x, y)$, $\nabla_x g(x, y)$ is its gradient with respect to x at (x, y). We use the standard big-O notations. For functions $T, S : \mathbb{R} \to \mathbb{R}$ such that $0 < \liminf_{x\to\infty} T(x)$, $0 < \liminf_{x\to\infty} S(x)$, (a) $T(x) = O(S(x))$ means $\limsup_{x\to\infty} T(x)/S(x) < \infty$; (b) $T(x) = \Theta(S(x))$ means $T(x) = O(S(x))$ and $S(x) = O(T(x))$; and (c) $T(x) = \widetilde{O}\left(S(x)\right)$ means that $T(x) = O(S(x)R(x))$ for some poly-logarithmic function $R : \mathbb{R} \to \mathbb{R}$.

Paper organization: In Section 2, we present preliminaries and all relevant background. In Section 3, we present our results for strongly-convex–concave setting and in section 4, results for nonconvex–concave setting. In Section 5, we present empirical evaluation of our algorithm for nonconvex-concave setting and compare it to a state-of-the-art algorithm. We conclude in Section 6. Several technical details are presented in the appendix.

2 Preliminaries and background material

In this section, we will present some preliminaries, describing the setup and reviewing some background material that will be useful in the sequel.

2.1 Minimax problems

We are interested in the minimax problems of the form (1) where $g(x, y)$ is a smooth function.

Definition 1. *A function $g(x, y)$ is said to be L-smooth if:*

$$\max\left\{\|\nabla_x g(x, y) - \nabla_x g(x', y')\|, \|\nabla_y g(x, y) - \nabla_y g(x', y')\|\right\} \leq L\left(\|x - x'\| + \|y - y'\|\right).$$

Throughout, we assume that $g(x, .)$ is *concave* for every $x \in \mathcal{X}$. For $g(\cdot, y)$ behavior in terms of x, there are broadly two settings:

2.1.1 Convex-concave setting

In this setting, $g(\cdot, y)$ is convex $\forall\, y \in \mathcal{Y}$. Given any g and $\forall(\widehat{x}, \widehat{y})$, the following holds trivially:

$$\min_{x\in\mathcal{X}} g(x, \widehat{y}) \leq g(\widehat{x}, \widehat{y}) \leq \max_{y\in\mathcal{Y}} g(\widehat{x}, y),$$

which then implies that $\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} g(x, y) \leq \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} g(x, y)$. The celebrated minimax theorem for the convex-concave setting [44] says that if $\mathcal{Y}$ is a compact set then the above inequality is in fact an equality, i.e., $\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} g(x, y) = \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} g(x, y)$. Furthermore, any point (x^*, y^*) is an optimal solution to (1) if and only if:

$$\min_{x\in\mathcal{X}} g(x, y^*) = g(x^*, y^*) = \max_{y\in\mathcal{Y}} g(x^*, y). \tag{2}$$

Hence, our goal is to find ε-primal-dual pair $(\widehat{x}, \widehat{y})$ with small primal-dual gap: $\max_{y\in\mathcal{Y}} g(\widehat{x}, y) - \min_{x\in\mathcal{X}} g(x, \widehat{y})$.

Definition 2. *For a convex-concave function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, $(\hat{x}, \hat{y})$ is an ε-primal-dual-pair of g if the primal-dual gap is less than ε:* $\max_{y\in\mathcal{Y}} g(\widehat{x}, y) - \min_{x\in\mathcal{X}} g(x, \widehat{y}) \leq \varepsilon$.

2.1.2 Nonconvex-concave setting

In this setting the function $g(\cdot, y)$ need not be convex. One cannot hope to solve such problems in general, since the special case of nonconvex optimization is already NP-hard [39]. Furthermore, the minimax theorem no longer holds, i.e., $\max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} g(x,y)$ can be strictly smaller than $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} g(x,y)$, and therefore the order of min and max might be important for a given application i.e., we might be interested only in minimax but not maximin (or vice versa). So, the primal-dual gap may not be a meaningful quantity to measure convergence. In this paper we will focus on the minimax problem: $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} g(x,y)$. One approach, inspired by nonconvex optimization, to measure convergence is to consider the function $f(x) = \max_{y\in\mathcal{Y}} g(x,y)$ and consider the convergence rate to approximate first order stationary points (i.e., $\nabla f(x)$ is small)[42, 18]. But as $f(x)$ could be non-smooth, $\nabla f(x)$ might not even be defined. It turns out that whenever $g(x,y)$ is smooth, $f(x)$ is weakly convex (Definition 4) for which first order stationarity notions are well-studied and are discussed below.

Approximate first-order stationary point for weakly convex functions: We first need to generalize the notion of gradient for a non-smooth function.

Definition 3. *The Fréchet sub-differential of a function $f(\cdot)$ at x is defined as the set,* $\partial f(x) = \{u \,|\, \liminf_{x'\to x} f(x') - f(x) - \langle u, x'-x\rangle/\|x'-x\| \geq 0\}$.

In order to define approximate stationary points, we also need the notion of weakly convex function and Moreau envelope.

Definition 4. *A function $f : \mathcal{X} \to \mathbb{R}\cup\{\infty\}$ is L-*weakly convex *if,*

$$f(x) + \langle u_x, x'-x\rangle - \frac{L}{2}\|x'-x\|^2 \quad \leq \quad f(x')\,, \tag{3}$$

for all Fréchet subgradients $u_x \in \partial f(x)$, for all $x, x' \in \mathcal{X}$.

Definition 5. *For a proper lower semi-continuous (l.s.c.) function $f : \mathcal{X} \to \mathbb{R}\cup\{\infty\}$ and $\lambda > 0$ ($\mathcal{X} \subseteq \mathbb{R}^p$), the* Moreau envelope *function is given by*

$$f_\lambda(x) \quad = \quad \min_{x'\in\mathcal{X}} f(x') + \frac{1}{2\lambda}\|x-x'\|^2\,. \tag{4}$$

Lemma 4 (in Appendix B.2) provides some useful properties of the Moreau envelope for weakly convex functions. Now, first order stationary point of a non-smooth nonconvex function is well-defined, i.e., x^* is a *first order stationary point (FOSP)* of a function $f(x)$ if, $0 \in \partial f(x^*)$ (see Definition 3). However, unlike smooth functions, it is nontrivial to define an *approximate* FOSP. For example, if we define an ε-FOSP as the point x with $\min_{u\in\partial f(x)}\|u\| \leq \varepsilon$, there may never exist such a point for sufficiently small ε, unless x is exactly a FOSP. In contrast, by using above properties of the Moreau envelope of a weakly convex function, it's approximate FOSP can be defined as [11]:

Definition 6. *Given an L-weakly convex function f, we say that x^* is an ε-first order stationary point (ε-FOSP) if, $\|\nabla f_{\frac{1}{2L}}(x^*)\| \leq \varepsilon$, where $f_{\frac{1}{2L}}$ is the Moreau envelope with parameter $1/2L$.*

Using Lemma 4, we can show that for any ε-FOSP x^*, there exists $\hat{x}$ such that $\|\hat{x}-x^*\| \leq \varepsilon/2L$ and $\min_{u\in\partial f(\hat{x})}\|u\| \leq \varepsilon$. In other words, an ε-FOSP is $O(\varepsilon)$ close to a point $\hat{x}$ which has a subgradient smaller than ε. We note that other notions of FOSP have also been proposed recently such as in [40]. However, it can be shown that an ε-FOSP according to the above definition is also an ϵ-FOSP with [40]'s definition as well, but the reverse is not necessarily true.

2.2 Mirror-Prox

Mirror-Prox [34] is a popular algorithm proposed for solving convex-concave minimax problems (1). It achieves a convergence rate of $O(1/k)$ for the primal dual gap. The original Mirror-Prox paper [34] motivates the algorithm through a *conceptual* Mirror-Prox (CMP) method, which brings out the main idea behind its convergence rate of $O(1/k)$. CMP for Euclidean norm (after ignoring projections to $\mathcal{X}$ and $\mathcal{Y}$) does the following update:

$$(x_{k+1}, y_{k+1}) = (x_k, y_k) + \frac{1}{\beta}\left(-\nabla_x g\left(x_{k+1}, y_{k+1}\right), \nabla_y g\left(x_{k+1}, y_{k+1}\right)\right). \tag{5}$$

The main difference between CMP and standard gradient descent ascent (GDA) is that in the k^{th} step, while GDA uses gradients at (x_k, y_k), CMP uses gradients at (x_{k+1}, y_{k+1}). The key observation of [34] is that if $g(\cdot,\cdot)$ is smooth, it can be implemented efficiently. CMP is analyzed as follows:

Implementability of CMP: Let $(x_k^{(0)}, y_k^{(0)}) = (x_k, y_k)$. For $\beta < \frac{1}{L}$, the iteration

$$(x_k^{(i+1)}, y_k^{(i+1)}) = (x_k, y_k) + \frac{1}{\beta}\left(-\nabla_x g\left(x_k^{(i)}, y_k^{(i)}\right), \nabla_y g\left(x_k^{(i)}, y_k^{(i)}\right)\right). \tag{6}$$

can be shown to be $\frac{1}{\sqrt{2}}$-contraction (when $g(\cdot,\cdot)$ is smooth) and that its fixed point is (x_{k+1}, y_{k+1}). So, in $\log\frac{1}{\epsilon}$ iterations of (6), we can obtain an accurate version of the update required by CMP. In fact, [34] showed that just *two* iterations of (6) suffice [30].

Convergence rate of CMP: Using CMP update with simple manipulations leads to the following:

$$g(x_{k+1}, y) - g(x, y_{k+1}) \le \frac{\beta}{2}\left(\|x - x_k\|^2 - \|x - x_{k+1}\|^2 + \|y - y_k\|^2 - \|y - y_{k+1}\|^2\right), \forall x \in \mathcal{X},\ y \in \mathcal{Y}.$$

$O(1/k)$ convergence rate follows easily using the above result.

Finally, our method and analysis also requires Nesterov's accelerated gradient descent method (see Algorithm 3 in Appendix A)and it's per-step analysis by [2] (Lemma 2 in Appendix A).

3 Strongly-convex concave saddle point problem

We first study the minimax problem of the form:

$$\min_{x\in\mathcal{X}}\ [\ f(x) = \max_{y\in\mathcal{Y}} g(x,y)\ ], \tag{P1}$$

where $g(x,\cdot)$ is concave, $g(\cdot,y)$ is σ-*strongly-convex*, $g(\cdot,\cdot)$ is L-smooth, i.e., $0 < \sigma \le L$. $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} \subset \mathbb{R}^q$ is a convex compact sub-set of $\mathbb{R}^q$ and let the function f take a minimum value $f^*(> -\infty)$. Let $D_{\mathcal{Y}} = \max_{y,y'\in\mathcal{Y}} \|y - y'\|$ be the diameter of $\mathcal{Y}$.

Our objective here is to find an ϵ-primal-dual pair $(\widehat{x}, \widehat{y})$ (see Definition 2). Now the fact that $f(\hat{x}) - f^* \le \max_{y\in\mathcal{Y}} g(\widehat{x}, y) - \min_{x\in\mathcal{X}} g(x, \widehat{y})$ implies that if $(\hat{x}, \hat{y})$ is an ε-primal-dual-pair, then $\hat{x}$ is also an ε-approximate minima of f. Furthermore, by Sion's minimax theorem [24], strong-convexity–concavity of $g(\cdot,\cdot)$ ensures that: $\min_x[f(x) := \max_y g(x,y)] = \max_y[h(y) := \min_x g(x,y)]$. Hence, one approach to efficiently solving the problem is by optimizing the dual problem $\max_y h(y)$. By Lemma 6 (in Appendix B.6), $h(y)$ is an $(L + \frac{L^2}{\sigma})$-smooth function. So we can use AGD to ensure that $h(y_k) - h(y^*) = O(1/k^2)$. Now, each step of AGD requires computing $\arg\min_x g(x, y_k)$ which can be done efficiently (i.e., logarithmic number of steps) as $g(\cdot, y_k)$ is strongly-convex and smooth. So, the overall first-order oracle complexity is $h(y_k) - h(y^*) = \widetilde{O}\left(1/k^2\right)$.

So does this simple approach give us our desired result? Unfortunately that is not the case, as the above bound on the dual function h does not translate to the same error rate for primal function f, i.e., the solution need not be $\widetilde{O}\left(1/k^2\right)$-primal-dual pair. E.g., consider $\min_{x\in\mathbb{R}}\max_{y\in[-1,1]}[g(x,y) = xy + x^2/2]$, where $\min_x \max_y g(x,y) = 0$, $f(x) = x^2/2 + |x|$ and $h(y) = -y^2/2$. If $h(y_k) = \Theta(k^{-2})$, then $x_k \in \operatorname{argmin}_x g(x, y_k) = \Theta(1/k)$ and so $f(x_k)$ is $\Theta(k^{-1})$. This is due to the non-smoothness of $\arg\max_{y\in\mathcal{Y}} g(x,y)$ w.r.t. x.

Instead of using AGD, we introduce a new method to solve the dual problem that we refer to as DIAG, which stands for Dual Implicit Accelerated Gradient. DIAG combines ideas from AGD [38] and Nemirovski's original derivation of the Mirror-Prox algorithm [34], and can ensure a fast convergence rate of $\tilde{O}(k^{-2})$ for the primal-dual gap. We note that there also exists a conceptually simpler smoothing technique based indirect algorithm, which prefixes the tolerance of ε (Appendix D). However, our goal is to find a direct algorithm which does note require prefixing the tolerance at ε. For better exposition, we first present a conceptual version of DIAG (C-DIAG), which is not implementable *exactly*, but brings out the main new ideas in our algorithm. We then present a detailed error analysis for the *inexact* version of this algorithm, which is implementable.

3.1 Conceptual version: C-DIAG

Consider the following updates which is a modified version of AGD (see Algorithm 3 in Appendix A):

(a) $w_k = (1 - \tau_k)y_k + \tau_k z_k$

(b) Choose x_{k+1}, y_{k+1} ensuring:
$x_{k+1} \in \arg\min_x g(x, y_{k+1})$, and $y_{k+1} = \mathcal{P}_{\mathcal{Y}}(w_k + \frac{1}{\beta}\nabla_y g(x_{k+1}, w_k))$

(c) $z_{k+1} = \mathcal{P}_{\mathcal{Y}}(z_k + \eta_k \nabla_y g(x_{k+1}, w_k))$

Complete pseudocode for C-DIAG algorithm is presented in Algorithm 4 in Appendix B.4. The main idea of the algorithm is in Step (b) above (i.e., Step 4 of Algorithm 4 in Appendix B.4), where we simultaneously find x_{k+1} and y_{k+1} satisfying the following requirements:

- x_{k+1} is the minimizer of $g(\cdot, y_{k+1})$, and
- y_{k+1} corresponds to an AGD step (see Algorithm 3 in Appendix A) for $g(x_{k+1}, \cdot)$

Implementability: The first question is whether it is easy enough to implement such a step? It turns out that it is indeed possible to quickly find points x_{k+1} and y_{k+1} that approximately satisfy the above requirements. The reason is that:

- Since $g(\cdot, y)$ is smooth and strongly convex for every $y \in \mathcal{Y}$, we can find ϵ-approximate minimizer for a given y in $O\left(\log \frac{1}{\epsilon}\right)$ iterations.
- Let $x^*(y) := \mathrm{argmin}_{x \in \mathcal{X}} g(x, y)$. The iteration $y^{i+1} = \mathcal{P}_{\mathcal{Y}}\left(w_k + \frac{1}{\beta}\nabla_y g(x^*(y^i), w_k)\right)$ is a $1/2$-contraction with a unique fixed point satisfying the update step requirements (i.e., Step 4 of Algorithm 4 in Appendix B.4). See Lemma 5 in Appendix B.5 for a proof. This means that only $O\left(\log \frac{1}{\epsilon}\right)$ iterations again suffice to find an update that approximately satisfies the requirements.

Convergence rate: Since y_{k+1} and z_{k+1} correspond to an AGD update for $g(x_{k+1}, \cdot)$, we can use the potential function decrease argument for AGD (Lemma 2 in Appendix A) to conclude that $\forall y \in \mathcal{Y}$,

$$\begin{aligned}
&(k+1)(k+2)\left(g(x_{k+1}, y) - g(x_{k+1}, y_{k+1})\right) + 2\beta \cdot \|y - z_{k+1}\|^2 \\
&\leq k(k+1)\left(g(x_{k+1}, y) - g(x_{k+1}, y_k)\right) + 2\beta \cdot \|y - z_k\|^2 \\
&\leq k(k+1)\left(g(x_{k+1}, y) - g(x_k, y)\right) + k(k+1)\left(g(x_k, y) - g(x_k, y_k)\right) + 2\beta \cdot \|y - z_k\|^2,
\end{aligned}$$

where the last step follows from the fact that $x_k = \mathrm{argmin}_x\, g(x, y_k)$ and so $g(x_k, y_k) \leq g(x_{k+1}, y_k)$. Noting that we can further recursively bound $k(k+1)\left(g(x_k, y) - g(x_k, y_k)\right) + 2\beta \cdot \|y - z_k\|^2$ as above, we obtain

$$\begin{aligned}
&(k+1)(k+2)\left(g(x_{k+1}, y) - g(x_{k+1}, y_{k+1})\right) + 2\beta \cdot \|y - z_{k+1}\|^2 \\
&\leq k(k+1)g(x_{k+1}, y) - \sum_{i=1}^{k}(2i) \cdot g(x_i, y) + 2\beta \cdot \|y - z_0\|^2 \\
\Rightarrow &\sum_{i=1}^{k+1}(2i) \cdot g(x_i, y) - (k+1)(k+2)g(x_{k+1}, y_{k+1}) \leq 2\beta \cdot \|y - z_0\|^2.
\end{aligned}$$

Since $g(x_{k+1}, y_{k+1}) \leq g(x, y_{k+1})$ for every $x \in \mathcal{X}$, we have

$$\begin{aligned}
&\sum_{i=1}^{k+1}(2i) \cdot g(x_i, y) - (k+1)(k+2)g(x, y_{k+1}) \leq 2\beta \cdot \|y - z_0\|^2 \\
&\Rightarrow g(\bar{x}_{k+1}, y) - g(x, y_{k+1}) \leq \frac{2\beta \cdot \|y - z_0\|^2}{(k+1)(k+2)},
\end{aligned}$$

where $\bar{x}_{k+1} := \frac{1}{(k+1)(k+2)} \sum_{i=1}^{k+1}(2i) \cdot x_i$. Since x and y are arbitrary above, this gives a $O\left(1/k^2\right)$ convergence rate for the primal dual gap.

3.2 Error analysis

The main issue with Algorithm 4 is that the update step is not exactly implementable. However, as we noted in the previous section, we can quickly find updates that almost satisfy the requirements. Algorithm 1 presents this inexact version. The following theorem states our formal result and a detailed proof is provided in Appendix B.5.

Algorithm 1: Dual Implicit Accelerated Gradient (DIAG) for strongly-convex–concave programming

Input: g, L, σ, $D_{\mathcal{Y}}$, x_0, y_0, K, $\{\varepsilon_{\text{step}}^{(k)}\}_{k=1}^{K}$
Output: $\bar{x}_K, y_K$

Set $\beta \leftarrow 2\frac{L^2}{\sigma}$, $z_0 \leftarrow y_0$
for $k = 0, 1, \ldots, K-1$ **do**
  $\tau_k \leftarrow \frac{2}{(k+2)},\ \eta_k \leftarrow \frac{(k+1)}{2\beta},\ w_k \leftarrow (1-\tau_k)y_k + \tau_k z_k$
  $x_{k+1}, y_{k+1} \leftarrow$ `Imp-STEP(`g, L, σ, x_0, w_k, β, $\varepsilon_{\text{step}}^{(k+1)}$`)`, ensuring:

$$g(x_{k+1}, y_{k+1}) \leq \min_x g(x, y_{k+1}) + \varepsilon_{\text{step}}^{(k+1)},\ \ y_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(w_k + \frac{1}{\beta}\nabla_y g(x_{k+1}, w_k)\right)$$

  $z_{k+1} \leftarrow \mathcal{P}_{\mathcal{Y}}\left(z_k + \eta_k \nabla_y g(x_{k+1}, w_k)\right),\ \ \bar{x}_{k+1} \leftarrow \frac{2}{(k+1)(k+2)}\sum_{i=1}^{k+1} i \cdot x_i$
return $\bar{x}_K, y_K$

`Imp-STEP(`g, L, σ, x_0, w, β, $\varepsilon_{\text{step}}$`):`
  Set $\varepsilon_{\text{mp}} \leftarrow \frac{2\sigma}{5L}\sqrt{\frac{2\varepsilon_{\text{step}}}{L}}$, $R \leftarrow \lceil \log_2 \frac{2D_{\mathcal{Y}}}{\varepsilon_{\text{mp}}} \rceil$, $\varepsilon_{\text{agd}} \leftarrow \frac{\sigma\beta^2\varepsilon_{\text{mp}}^2}{32L^2}$, $y_0 \leftarrow w$
  for $r = 0, 1, \ldots, R$ **do**
    Starting at x_0 use AGD [38] for strongly-convex $g(\cdot, y_r)$, to compute $\hat{x}_r$ such that:

$$g(\hat{x}_r, y_r) \leq \min_x g(x, y_r) + \varepsilon_{\text{agd}}, \tag{7}$$

    $y_{r+1} \leftarrow \mathcal{P}_{\mathcal{Y}}\left(w + \frac{1}{\beta}\nabla_y g(\hat{x}_r, w)\right)$
  return $\hat{x}_R$, y_{R+1}

Theorem 1 (Convergence rate of DIAG). *Let $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a L-smooth, σ-strongly-convex–concave function on $\mathcal{X} = \mathbb{R}^p$ and a convex compact sub-set $\mathcal{Y} \subset \mathbb{R}^q$ (with diameter $D_{\mathcal{Y}}$). Then, after K iterations, DIAG (Algorithm 1) with a tolerance schedule of $\{\varepsilon_{\text{step}}^{(k)}\}_{k=1}^{K}$ for its* `Imp-STEP` *sub-routine, finds $(\bar{x}_K, y_K)$ s.t.:*

$$\max_{\tilde{y}\in\mathcal{Y}} g(\bar{x}_K, \tilde{y}) - \min_{\tilde{x}\in\mathcal{X}} g(\tilde{x}, y_K) \leq \frac{4\frac{L^2}{\sigma}D_{\mathcal{Y}}^2 + \sum_{k=1}^{K} k(k+1)\,\varepsilon_{\text{step}}^{(k)}}{K(K+1)}\,. \tag{8}$$

In particular, setting $\varepsilon_{\text{step}}^{(k)} = \frac{L^2 D_{\mathcal{Y}}^2}{\sigma k^3(k+1)}$ we have: $\max_{\tilde{y}\in\mathcal{Y}} g(\bar{x}_K, \tilde{y}) - \min_{\tilde{x}\in\mathcal{X}} g(\tilde{x}, y_K) \leq \frac{6\frac{L^2}{\sigma}D_{\mathcal{Y}}^2}{K(K+1)}$. Furthermore, for this setting the total first order oracle complexity is given by: $O(\sqrt{\frac{L}{\sigma}}K\log^2(K))$.

Remark 1: Theorem 1 shows that DIAG needs $\tilde{O}((D_{\mathcal{Y}}L/\sqrt{\sigma\varepsilon}) \cdot (\sqrt{L/\sigma}))$ gradient queries for finding a ε-primal-dual-pair, while current best-known rate is $O(1/\varepsilon)$ achieved by Mirror-Prox. This dependence in ε and $D_{\mathcal{Y}}$ is optimal, as it is shown in [41, Theorem 10] that $\Omega(D_{\mathcal{Y}}(L-\sigma)/\sqrt{\sigma\varepsilon})$ gradient queries are necessary to achieve ε error in the primal-dual gap.

Remark 2: Unlike standard AGD for $h(y)$, which only updates y_k in the outer-loop, DIAG's outer-step updates both x_k and y_k thus allowing us to better track the primal-dual gap. However, DIAG's dependence on the condition number L/σ seems sub-optimal and can perhaps be improved if we do not compute `Imp-STEP` nearly optimally allowing for inexact updates; we leave further investigation into improved dependence on the condition number for future work.

4 Nonconvex concave saddle point problem

We study the nonconvex concave minimax problem (1) where $g(x, \cdot)$ is concave, $g(\cdot, y)$ is nonconvex, and $g(\cdot, \cdot)$ is L-smooth, $\mathcal{X} = \mathbb{R}^p$ (such that $\text{Proj}_{\mathcal{X}}(x) = x$) and $\mathcal{Y}$ is a convex compact sub-set of $\mathbb{R}^q$. As mentioned in Section 2, we measure the convergence to an approximate FOSP of this problem (see Definition 6) but it requires weak-convexity of $f(x) := \max_{y\in\mathcal{Y}} g(x, y)$. The following lemma guarantees weak convexity of f given smoothness of g.

Lemma 1. *Let $g(\cdot, y)$ be continuous and $\mathcal{Y}$ be compact. Then $f(x) = \max_{y\in\mathcal{Y}} g(x, y)$ is L-weakly convex, if g is L-weakly convex in x (Definition 1), or if g is L-smooth in x .*

See Appendix B.3 for the proof. The arguments of [18] easily extend to show that applying subgradient method on $f(x)$, [11] gives a convergence rate of $O\left(1/k^{1/5}\right)$. Instead, we exploit the smooth minimax form of $f(\cdot)$ to design a faster converging scheme. The main intuition comes from the proximal viewpoint that gradient descent can be viewed as iteratively forming and optimizing local quadratic upper bounds. As f is weakly convex, adding enough quadratic regularization should ensure that the resulting sequence of problems are all strongly-convex–concave. We then exploit DIAG to efficiently solve such local quadratic problems to obtain improved convergence rates. Concretely, let

$$\widehat{f}(x; x_k) \quad = \quad \max_{y} g(x, y) + L\|x - x_k\|^2 \ . \tag{9}$$

By L-weak-convexity of f, $\widehat{f}(x; x_k)$ is *strongly*-convex–concave (Lemma 3) that can be solved using DIAG up to *certain accuracy* to obtain x_{k+1}. We refer to this algorithm as Prox-DIAG and provide a pseudo-code for the same in Algorithm 2. The following theorem gives convergence guarantees for

---

Algorithm 2: Proximal Dual Implicit Accelerated Gradient (Prox-DIAG) for nonconvex concave programming

---

Input: g, L, ε, x_0, y_0
Output: x_k

Set $\tilde{\varepsilon} \leftarrow \frac{\varepsilon^2}{64\,L}$
for $k = 0, 1, \ldots, K$ **do**
  Using DIAG for strongly convex concave minimax problem,

$$\min_{x} \max_{y\in\mathcal{Y}} \left[\widehat{g}(x, y; x_k) = g(x, y) + L\|x - x_k\|^2\right] \tag{10}$$

  find x_{k+1} such that,

$$\max_{y\in\mathcal{Y}} g(x_{k+1}, y) + L\|x_{k+1} - x_k\|^2 \ \leq \ \min_{x} \max_{y\in\mathcal{Y}} g(x, y) + L\|x - x_k\|^2 + \frac{\tilde{\varepsilon}}{4} \tag{11}$$

  if $\max_{y\in\mathcal{Y}} g(x_k, y) - \frac{3\tilde{\varepsilon}}{4} \leq \max_{y\in\mathcal{Y}} g(x_{k+1}, y) + L\|x_{k+1} - x_k\|^2$ **then**
    return x_k

Prox-DIAG.

Theorem 2 (Convergence rate of Prox-DIAG)**.** *Let $g(x, y)$ be L-smooth, $g(x, \cdot)$ be concave, $\mathcal{X}$ be $\mathbb{R}^p$, $\mathcal{Y}$ be a convex compact subset of $\mathbb{R}^q$, and the minimum value of function $f(x) = \max_{y\in\mathcal{Y}} g(x, y)$ be bounded below, i.e. $f(x) \geq f^* > -\infty$. Then Prox-DIAG (Algorithm 2) after,*

$$K = \left\lceil \frac{4^4 L(f(x_0) - f^*)}{3\varepsilon^2} \right\rceil$$

steps outputs an ε-FOSP. The total first-order oracle complexity to output ε-FOSP is: $O\big(\frac{L^2 D_{\mathcal{Y}}(f(x_0)-f^*)}{\varepsilon^3} \log^2\left(\frac{1}{\varepsilon}\right)\big)$.

A proof is provided in Appendix B.7. Note that Prox-DIAG solves the quadratic approximation problem to higher accuracy of $O(\epsilon^2)$ which then helps bounding the gradient of the Moreau envelope. Also due to the modular structure of the argument, a faster inner loop for special settings, e.g., when $g(x, y)$ is a finite-sum, can ensure more efficient algorithm. While our algorithm is able to significantly improve upon existing state-of-the-art rate of $O(1/\varepsilon^5)$ in general nonconvex-concave setting [18], it is unclear if the rate can be further improved. In fact, precise lower-bounds for this setting are mostly unexplored and we leave further investigation into lower-bounds as a topic of future research.

We also specialize the Prox-DIAG algorithm, as Prox-FDIAG (Algorithm 5 in Appendix C), for the case of minimizing a weakly convex $f(x)$, with the special structure of *finite max-type function*:

$$\min_{x} \ \left[f\left(x\right) = \max_{1\leq i\leq m} f_i(x)\right] , \tag{P3}$$

where f_i's could be nonconvex but are L-smooth, G-Lipschitz and bounded from below. For this case, we improve the current known best rate of $O\left(m/\varepsilon^4\right)$ and obtain a faster rate of $O(m\log^{3/2} m/\varepsilon^3)$ using the Prox-FDIAG algorithm. Please refer to Appendix C for more details.

5 Experiments

We empirically verify the performance of Prox-FDIAG (Algorithm 5 in Appendix C) on a synthetic finite max-type nonconvex minimization problem (P3). We consider the following problem. $\min_{x\in\mathbb{R}^2}\left[f(x)=\max_{1\le i\le m=9} f_i(x)\right]$ where $f_i(x)=q_{(-1,\,(X_i^{(1)},X_i^{(2)}),\,C_i)}(x)$ for all $1\le i\le 8$, where $q_{(a,b,c)}(x)=a\|x-b\|_2^2+c$, $X_i^{(1)}$, $X_i^{(2)}$, and c_i are randomly generated. Thus each f_i is smooth with parameter $L=1$, which implies that f is L-weakly convex. We implement three algorithms: Prox-FDIAG (Algorithm 5, red circles), Adaptive Prox-FDIAG (Algorithm 6, black dots), and subgradient method [11] (blue triangles). Adaptive Prox-FDIAG is a practically faster variant of Prox-FDIAG, with the same first-order oracle complexity guarantee (up to an $O(\log(1/\varepsilon))$ factor). In Figure 1, we plot the norm of gradient of Moreau envelope $\|\nabla f_{\frac{1}{2L}}(x_k)\|_2$ against the number

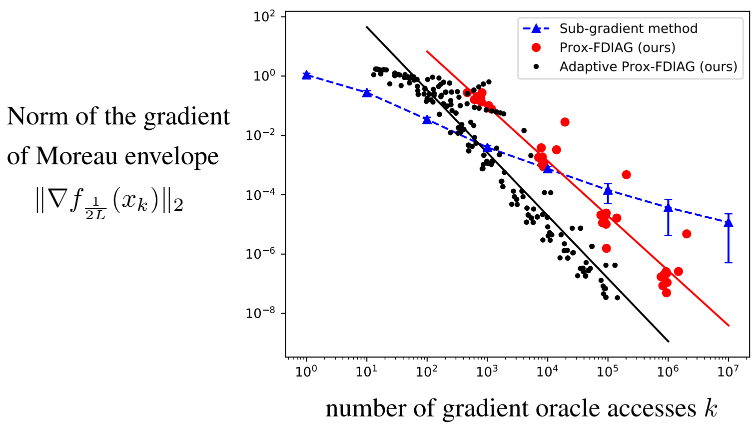

Figure 1: For small target accuracy ε regime, Adaptive Prox-FDIAG (ours) has the fastest convergence rate followed by Prox-FDIAG (ours) and subgradient method.

of iterations k in log-log scale. We see that, Prox-FDIAG and Adaptive Prox-FDIAG have a faster convergence rate than subgradient method, and Adaptive Prox-FDIAG is almost always faster than Prox-FDIAG. We provide more details about the algorithms and the experiments in Appendix E.

6 Conclusion

In this paper, we study smooth minimax problems, where the maximization is concave but the minimization is either strongly convex or nonconvex. In both of these settings, we present new algorithms improving state-of-the-art. The key ideas are i) a novel way to combine Mirror-Prox and Nesterov's AGD for strongly convex case that can tightly bound primal-dual gap and ii) an inexact prox method with good convergence rate to stationary points for the nonconvex case. While we only present our results for the Euclidean setting, generalizing it to non-Euclidean settings with the framework of Bregman divergences should be straight forward. Finally, we showcase the empirical superiority of our nonconvex algorithm over state-of-the-art subgradient method for a case of finite max-type nonconvex minimization problems. Some of the more interesting questions would be to understand the optimality of the rates that we obtain and dependence on the strong convexity parameter. Further extensions of these results to the stochastic setting would also be quite interesting.

Acknowledgement

This work is partially supported by NSF awards CCF-1927712 and RI-1929955.

References

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

[2] Nikhil Bansal and Anupam Gupta. Potential-function proofs for first-order methods. *arXiv preprint arXiv:1712.04581*, 2017.

[3] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[4] Dimitri P Bertsekas. *Convex optimization theory*. Athena Scientific Belmont, 2009.

[5] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[6] Ronald E Bruck Jr. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in hilbert space. *Journal of Mathematical Analysis and Applications*, 61(1):159–164, 1977.

[7] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[8] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

[9] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

[10] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165(1):113–149, 2017.

[11] Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018.

[12] Simon S Du and Wei Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 196–205, 2019.

[13] Jonathan Eckstein and Dimitri P Bertsekas. On the douglas—rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

[14] Tom Goldstein, Brendan O'Donoghue, Simon Setzer, and Richard Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014.

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

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

[17] Yunlong He and Renato DC Monteiro. An accelerated hpe-type algorithm for a class of composite convex-concave saddle-point problems. *SIAM Journal on Optimization*, 26(1):29–56, 2016.

[18] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

[19] Anatoli Juditsky and Arkadi Nemirovski. First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. *Optimization for Machine Learning*, pages 121–148, 2011.

[20] Anatoli Juditsky and Arkadi Nemirovski. First order methods for nonsmooth convex large-scale optimization, II: utilizing problems structure. *Optimization for Machine Learning*, 30(9):149–183, 2011.

[21] Sham Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript*, 2009.

[22] Purushottam Kar, Harikrishna Narasimhan, and Prateek Jain. Surrogate functions for maximizing precision at the top. *arXiv preprint arXiv:1505.06813*, 2015.

[23] David Kinderlehrer and Guido Stampacchia. *An introduction to variational inequalities and their applications*, volume 31. Siam, 1980.

[24] Hidetoshi Komiya. Elementary proof for sion's minimax theorem. *Kodai mathematical journal*, 11(1):5–7, 1988.

[25] Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. Nonconvex optimization for regression with fairness constraints. In *ICML*, pages 2742–2751, 2018.

[26] 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.

[27] A Ya Kruger. On fréchet subdifferentials. *Journal of Mathematical Sciences*, 116(3):3325–3358, 2003.

[28] Songtao Lu, Ioannis Tsaknakis, Mingyi Hong, and Yongxin Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications. *arXiv preprint arXiv:1902.08294*, 2019.

[29] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

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

[31] Roger B Myerson. *Game theory*. Harvard university press, 2013.

[32] Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228, 2009.

[33] Arkadi Nemirovski. Efficient methods for solving variational inequalities. *Ekonomika i Matem.Metody*, 17:344–359, 1981.

[34] Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[35] Yu Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, 16(1):235–249, 2005.

[36] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[37] Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. 1998.

[38] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

[39] Maher Nouiehed, Jason D Lee, and Meisam Razaviyayn. Convergence to second-order stationarity for constrained non-convex optimization. *arXiv preprint arXiv:1810.02024*, 2018.

[40] Maher Nouiehed, Maziar Sanjabi, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *arXiv preprint arXiv:1902.08297*, 2019.

[41] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *arXiv preprint arXiv:1808.02901*, 2018.

[42] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

[43] Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D Lee. On the convergence and robustness of training gans with regularized optimal transport. In *Advances in Neural Information Processing Systems*, pages 7091–7101, 2018.

[44] Maurice Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

[45] Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. *Optimization for machine learning*. Mit Press, 2012.

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

[47] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. 2008.

[48] Zhipeng Xie and Jianwen Shi. Accelerated primal dual method for a class of saddle point problem with strongly convex component. *arXiv preprint arXiv:1906.07691*, 2019.

[49] Yangyang Xu. Iteration complexity of inexact augmented lagrangian methods for constrained convex programming. *arXiv preprint arXiv:1711.05812*, 2017.

[50] Yangyang Xu and Shuzhong Zhang. Accelerated primal–dual proximal block coordinate updating methods for constrained convex optimization. *Computational Optimization and Applications*, 70(1):91–128, 2018.

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

Appendix

A Nesterov's accelerated gradient descent

Algorithm 3: Nesterov's accelerated gradient ascent

Input: Smooth concave function $h(\cdot)$, learning rate $\frac{1}{\beta}$, initial points y_0 and z_0
Output: y_k

for $k = 0, 1, \dots$ **do**
  $w_k \leftarrow (1-\tau_k)y_k + \tau_k z_k,\ y_{k+1} \leftarrow \mathcal{P}_{\mathcal{Y}}\left(w_k + \frac{1}{\beta}\nabla h(w_k)\right),$
  $z_{k+1} \leftarrow \mathcal{P}_{\mathcal{Y}}\left(z_k + \eta_k \nabla h(w_k)\right)$

Nesterov's accelerated gradient descent [38] is an optimal method for minimizing smooth convex functions (or equivalently maximizing smooth concave functions). In order to simplify the exposition in the sequel, we will consider the algorithm for maximizing concave functions. The pseudocode for this is presented in Algorithm 3. Fix any point $y \in \mathcal{Y}$.

A.1 Smooth concave function

Consider the potential function

$$\Phi(k) := k(k+1)\left(h(y) - h(y_k)\right) + 2\beta \cdot \|y - z_k\|^2.$$

The following lemma (from [2]) is the key result that helps us obtain the convergence rate of Algorithm 3. Here $\mathcal{P}_{\mathcal{Y}}(\cdot)$ denotes projection onto $\mathcal{Y}$.

Lemma 2. *[2] Suppose $h(\cdot)$ is an L-smooth concave function and the parameters of Algorithm 3 are chosen so that $\beta > L$, $\eta_k = \frac{k+1}{2\beta}$ and $\tau_k = \frac{2}{k+2}$. Then, we have*

$$\Phi(k+1) \le \Phi(k).$$

Proof of Lemma 2. Writing

$$\begin{aligned}\Phi(k+1) - \Phi(k) =& (k+1)(k+2)\left(h(w_k) - h(y_{k+1})\right) && (12)\\ & - k(k+1)\left(h(w_k) - h(y_k)\right) + 2(k+1)\left(h(y) - h(w_k)\right) \\ & + 2\beta\left(\|z_{k+1} - y\|^2 - \|z_k - y\|^2\right), && (13)\end{aligned}$$

we bound the three terms appearing in separate lines above. Firstly, for the third term, $\|z_{k+1} - y\|^2 \le \|z_k + \eta_k \nabla h(w_k) - y\|^2 - \|z_{k+1} - z_k - \eta_k \nabla h(w_k)\|^2$ due to Pythagoras theorem and so

$$\begin{aligned}\|z_{k+1} - y\|^2 - \|z_k - y\|^2 &\le 2\eta_k\langle \nabla h(w_k), z_k - y\rangle + \eta_k^2 \|\nabla h(w_k)\|^2 - \|z_{k+1} - z_k - \eta_k \nabla h(w_k)\|^2 \\ &\le 2\eta_k \langle \nabla h(w_k), z_{k+1} - y\rangle - \|z_{k+1} - z_k\|^2. && (14)\end{aligned}$$

For the second term, we have

$$\begin{aligned}& - k(k+1)\left(h(w_k) - h(y_k)\right) + 2(k+1)\left(h(y) - h(w_k)\right) \\ & \le -k(k+1)\langle \nabla h(w_k), w_k - y_k\rangle + 2(k+1)\langle \nabla h(w_k), y - w_k\rangle = 2(k+1)\langle \nabla h(w_k), y - z_k\rangle \\ & && (15)\end{aligned}$$

Finally, for the first term, we have $h(y_{k+1}) - h(w_k) \ge \langle \nabla h(w_k), y_{k+1} - w_k\rangle - \frac{\beta}{2}\|y_{k+1} - w_k\|^2$. Since $y_{k+1} = \operatorname{argmax}_{\bar{y} \in \mathcal{Y}} \langle \nabla h(w_k), \bar{y} - w_k\rangle - \frac{\beta}{2}\|\bar{y} - w_k\|^2$, we have for $v := (1-\tau_k)y_k + \tau_k z_{k+1} \in \mathcal{Y}$,

$$\begin{aligned}& h(y_{k+1}) - h(w_k) \ge \langle \nabla h(w_k), y_{k+1} - w_k\rangle - \frac{\beta}{2}\|y_{k+1} - w_k\|^2 \\ & \ge \langle \nabla h(w_k), v - w_k\rangle - \frac{\beta}{2}\|v - w_k\|^2 = \tau_k \langle \nabla h(w_k), z_{k+1} - z_k\rangle - \frac{\beta\tau_k^2}{2}\|z_{k+1} - z_k\|^2, && (16)\end{aligned}$$

where we used $w_k = (1-\tau_k)y_k + \tau_k z_k$ in the last step. Substituting (16), (15) and (14) in (13) proves the lemma. □

B Proofs

B.1 Auxiliary lemma

Lemma 3. *If $f(x)$ is a L-weakly convex function and $\tilde{f}(x)$ is a $\tilde{\sigma}(\geq L)$-strongly convex differentiable function, then $f(x) + \tilde{f}(x)$ is $(\tilde{\sigma} - L)$-strongly convex.*

Proof. Since f is L-weakly convex and $\tilde{f}$ is σ-strongly convex we get that,

$$\begin{aligned} & f(x') \geq f(x) + \langle u_x, x' - x\rangle - \frac{L}{2}\|x' - x\|^2 \,, \\ & \tilde{f}(x') \geq \tilde{f}(x) + \left\langle \nabla \tilde{f}(x), x' - x \right\rangle + \frac{\tilde{\sigma}}{2}\|x' - x\|^2 \,, \\ \implies & f(x') + \tilde{f}(x') \geq f(x) + \tilde{f}(x) + \left\langle u_x + \nabla \tilde{f}(x), x' - x \right\rangle + \frac{\tilde{\sigma} - L}{2}\|x' - x\|^2 \,. \end{aligned} \tag{17}$$

where $u_x \in \partial f(x)$. We finish the proof by noting that $\partial(f + \tilde{f}) = \partial f + \nabla \tilde{f}$ [27, Corollary 1.12.2.]. □

B.2 Properties of Moreau envelope

The following lemma provides some useful properties of the Moreau envelope for weakly convex functions.

Lemma 4. *For an L-weakly convex proper l.s.c. function $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ such that $\mathcal{X} = \mathbb{R}^p$ and $L < 1/\lambda$, the following hold true,*

(a) The minimizer $\hat{x}_\lambda(x) = \arg\min_{x' \in \mathcal{X}} f(x') + \frac{1}{2\lambda}\|x - x'\|^2$ is unique and $f(\hat{x}_\lambda(x)) \leq f_\lambda(x) \leq f(x)$. Furthermore, $\arg\min_x f(x) = \arg\min_x f_\lambda(x)$.

(b) f_λ is $\left(\frac{1}{\lambda} + \frac{1}{\lambda(1-\lambda L)}\right)$-smooth and thus differentiable, and

(c) $\min_{u \in \partial f(\hat{x}_\lambda(x))} \|u\| \leq (1/\lambda)\|\hat{x}_\lambda(x) - x\| = \|\nabla f_\lambda(x)\|$.

Proof. We re-write $f_\lambda(x)$ as minimum value of a $(\frac{1}{\lambda} - L)$-strong convex function $\phi_{\lambda,x}$, as f is L-weakly convex (Definition 3) and $\frac{1}{2\lambda}\|x - x'\|^2$ is differentiable and $\frac{1}{\lambda}$-strongly convex (Lemma 3),

$$f_\lambda(x) = \min_{x' \in \mathcal{X}} \left[\phi_{\lambda,x}(x') = f(x') + \frac{1}{2\lambda}\|x - x'\|^2\right]. \tag{18}$$

Then first part of (a) follows trivially by the strong convexity. For the second part notice the following,

$$\begin{aligned} \min_x f_\lambda(x) &= \min_x \min_{x'} f(x') + \frac{1}{2\lambda}\|x - x'\|^2 \\ &= \min_{x'} \min_x f(x') + \frac{1}{2\lambda}\|x - x'\|^2 \\ &= \min_{x'} f(x') \end{aligned}$$

Thus $\arg\min_x f_\lambda(x) = \arg\min_x f(x)$. For (b) we can re-write the Moreau envelope f_λ as,

$$\begin{aligned} f_\lambda(x) &= \min_{x'} f(x') + \frac{1}{2\lambda}\|x - x'\|^2 \\ &= \frac{\|x\|^2}{2\lambda} - \frac{1}{\lambda}\max_{x'}(x^T x' - \lambda f(x') - \frac{\|x'\|^2}{2}) \\ &= \frac{\|x\|^2}{2\lambda} - \frac{1}{\lambda}\left(\lambda f(\cdot) + \frac{\|\cdot\|^2}{2}\right)^*(x) \end{aligned} \tag{19}$$

where $(\cdot)^*$ is the Fenchel conjugation operator. Since $L < 1/\lambda$, using L-weak convexity of f, it is easy to see that $\lambda f(x') + \frac{\|x'\|^2}{2}$ is $(1 - \lambda L)$-strongly convex, therefore its Fenchel conjugate would be $\frac{1}{(1-\lambda L)}$-smooth [21, Theorem 6]. This, along with $\frac{1}{\lambda}$-smoothness of first quadratic term implies that $f_\lambda(x)$ is $\left(\frac{1}{\lambda} + \frac{1}{\lambda(1-\lambda L)}\right)$-smooth, and thus differentiable.

For (c) we again use the reformulation of $f_\lambda(x)$ as $\min_{x' \in \mathcal{X}} \phi_{\lambda,x}(x')$ (18). Then by first-order necessary condition for optimality of $\hat{x}_\lambda(x)$, we have that $x - \hat{x}_\lambda(x) \in \lambda \partial f(x)$. Further, from proof of part (a) we have that $\phi_{\lambda,x}(x')$ $(1-\lambda L)$-strongly-convex in x' and it is quadratic (and thus convex) in x. Then we can use Danskin's theorem [4, Section 6.11] to prove that, $\nabla f_\lambda(x) = (x - \hat{x}_\lambda(x))/\lambda \in \partial f(x)$. Refer [45, Section B.1] for other proofs of the same result.

□

B.3 Proof of Lemma 1

It is easy to see that $g(\cdot, y)$ is L-weakly convex if it is L-smooth: $g(x', y) \geq g(x, y) + \langle \nabla_x g(x, y), x' - x \rangle - \frac{L}{2}\|x' - x\|^2$. Thus we only need to prove the case of L-weakly convex $g(\cdot, y)$. Since $g(\cdot, y)$ is L-weakly convex we get that,

$$\begin{aligned} g(x', y) &\geq g(x, y) + \langle u_{x,y}, x' - x \rangle - \frac{L}{2}\|x' - x\|^2 \\ \implies g(x', y) + \frac{L}{2}\|x'\|^2 &\geq g(x, y) + \frac{L}{2}\|x\|^2 + \langle u_{x,y} + Lx, x' - x \rangle \end{aligned}$$

where $u_{x,y} \in \partial_x g(x, y)$. This means that $\tilde{g}(x, y) := g(x, y) + \frac{L}{2}\|x\|^2$ is convex, since $\partial_x \tilde{g}(x, y) = \partial_x g(x, y) + Lx$ [27, Corollary 1.12.2.].

Let $\tilde{f}(x) = \max_{y \in \mathcal{Y}} \tilde{g}(x, y)$. Since $\tilde{g}(x, y)$ is convex in x and smooth (Definition 1), and $\mathcal{Y}$ is compact set we use Danskin's theorem [4, Section 6.11] to prove that,

$$\begin{aligned} \partial \tilde{f}(x) &= \mathrm{conv}\{\partial_x \tilde{g}(x, y^*) \,|\, y^* \in \arg\max_{y \in \mathcal{Y}} \tilde{g}(x, y)\}\,, \\ \implies \partial f(x) + Lx &= \mathrm{conv}\{\partial_x g(x, y^*) + Lx \,|\, y^* \in \arg\max_{y \in \mathcal{Y}} g(x, y)\}\,, \\ \implies \partial f(x) &= \mathrm{conv}\{\partial_x g(x, y^*) \,|\, y^* \in \arg\max_{y \in \mathcal{Y}} g(x, y)\}\,. \end{aligned} \tag{20}$$

where the second to last step comes from the facts that $\partial \tilde{f} = \partial f + Lx$, $\partial_x \tilde{g}(x, y) = \partial_x g(x, y) + Lx$ [27, Corollary 1.12.2.], and $\arg\max_{y \in \mathcal{Y}} \tilde{g}(x, y) = \arg\max_{y \in \mathcal{Y}} g(x, y) + \frac{L}{2}\|x\|^2 = \arg\max_{y \in \mathcal{Y}} g(x, y)$. Let $u_{x,y} \in \partial_x g(x, y)$ and $y^* \arg\max_{y \in \mathcal{Y}} g(x, y)$then,

$$\begin{aligned} f(x') \geq g(x', y^*) &\overset{(a)}{\geq} g(x, y^*) + \langle u_{x,y^*}, x' - x \rangle - \frac{L}{2}\|x' - x\|^2 \\ \overset{(b)}{\implies} f(x') &\geq f(x) + \langle v_x, x' - x \rangle - \frac{L}{2}\|x' - x\|^2 \end{aligned}$$

where (a) uses L-weak convexity of $g(\cdot, y)$, and (b) uses (20) and $v_x \in \partial f(x)$.

B.4 Pseudocode for Conceptual DIAG algorithm

The pseudocode for C-DIAG algorithm is presented in Algorithm 4.

B.5 Proof of Theorem 1

A cursory glance of the DIAG (Algorithm 1) reveals that it is a modified version of projected accelerated gradient ascent (Algorithm 3) on some function of y with a modified step given by `Imp-STEP`, which is inspired from the conceptual Mirror-Prox method of [34]. In the following lemma we analyze the `Imp-STEP` sub-routine, which is the most non-trivial step of the algorithm.

Lemma 5. *If $\beta = 2\frac{L^2}{\sigma}$, the sub-routine `Imp-STEP`$(g, L, \sigma, w, \beta, \varepsilon_{\text{step}})$ of Algorithm 1, returns a pair of points $(\hat{x}_R, y_{R+1}) \in \mathcal{X} \times \mathcal{Y}$, such that,*

$$g(\hat{x}_R, y_{R+1}) \leq \min_x g(x, y_R) + \varepsilon_{\text{step}}, \;\; \textit{and,} \;\; y_R = \mathcal{P}_{\mathcal{Y}}\left(w + \frac{1}{\beta}\nabla_y g(\hat{x}_{R-1}, w)\right) \tag{21}$$

in $R = \lceil \log_2\left((5LD_{\mathcal{Y}}/\sigma)\sqrt{L/2\varepsilon_{\text{step}}}\right)\rceil$ iterations with $O\left(\sqrt{L/\sigma}\log\left(1/\varepsilon_{\text{step}}\right)\right)$ gradient computations per iterations.

Algorithm 4: Conceptual Dual Implicit Accelerated Gradient (C-DIAG) for strongly-convex–concave programming

Input: $g, L, \sigma, x_0, y_0, K$
Output: $\bar{x}_K, y_K$
Set $\beta \leftarrow 2\frac{L^2}{\sigma}$, $z_0 \leftarrow y_0$
for $k = 0, 1, \dots, K-1$ **do**
  $\tau_k \leftarrow \frac{2}{(k+2)},\ \eta_k \leftarrow \frac{(k+1)}{2\beta},\ w_k \leftarrow (1-\tau_k)y_k + \tau_k z_k$
  Choose x_{k+1}, y_{k+1} ensuring:
$$g(x_{k+1}, y_{k+1}) = \min_x g(x, y_{k+1}),\ \ y_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(w_k + \frac{1}{\beta}\nabla_y g(x_{k+1}, w_k)\right)$$
  $z_{k+1} \leftarrow \mathcal{P}_{\mathcal{Y}}\left(z_k + \eta_k \nabla_y g(x_{k+1}, w_k)\right),\ \ \bar{x}_{k+1} \leftarrow \frac{2}{(k+1)(k+2)}\sum_{i=1}^{k+1} i \cdot x_i$
return $\bar{x}_K, y_K$

A proof for this lemma is provided in Appendix B.5.1. The above lemma guarantees that the Imp-STEP sub-routine converges fast (linear time), in $O(\log(1/\varepsilon_{\text{step}}))$ steps with $O(\sqrt{L/\sigma}\log^2(1/\varepsilon_{\text{step}}))$ number of gradient computations.

In the rest of the proof we will utilize the recently proposed *potential-function* based proof for accelerated gradient decent (AGD) [2, Section 5.2]. Analyzing AGD using potential-function has an advantage over the standard analysis because, even though AGD does not decrease the function value monotonically the former constructs a potential-function which monotonically decreases over the iterations. Given the guarantees (Lemma 5) for the Imp-STEP sub-routine we can re-write an iteration of the DIAG algorithm by the following steps:

$$\tau_k = \frac{2}{(k+2)},\ \ \eta_k = \frac{(k+1)}{2\beta} \tag{22}$$
$$w_k = (1-\tau_k)y_k + \tau_k z_k \tag{23}$$
$$y_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(w_k + \frac{1}{\beta}\nabla_y h_{x_{k+1}}(w_k)\right) \tag{24}$$
$$z_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(z_k + \eta_k \nabla_y h_{x_{k+1}}(w_k)\right) \tag{25}$$

where $h_{k+1}(y) := g(x_{k+1}, y)$ such that $g(x_{k+1}, y_{k+1}) \le \min_{x\in\mathcal{X}} g(x, y_{k+1}) + \varepsilon_{\text{step}}$. That is at iteration k, DIAG executes the k-th step of the accelerated gradient ascent for the concave function $h_{k+1} = g(x_{k+1}, \cdot)$ (Algorithm 3). As in (12), for the concave function $h_k : \mathcal{Y} \to \mathbb{R}$ and an arbitrary reference point $\tilde{y} \in \mathcal{Y}$, we define the following potential function for iteration j,

$$\Phi^{h_k}(j) = j(j+1)(h_k(\tilde{y}) - h_k(y_j)) + 2\beta\|z_j - \tilde{y}\|^2 \tag{26}$$

Since $g(x, \cdot)$ is L-smooth, it is also $\frac{2L^2}{\sigma}$-smooth ($\sigma \le L$). Then, using Lemma 2 , we see that for a step-size of $\frac{1}{\beta} = \frac{\sigma}{2L^2}$, the potential function $\Phi^{h_k}(k)$ decrease at step of k of the algorithm:

$\Phi^{h_{k+1}}(k+1) \leq \Phi^{h_{k+1}}(k)$. Thus,

$$\begin{aligned}
\Phi^{h_{k+1}}(k+1) &\leq \Phi^{h_{k+1}}(k) \\
&= k(k+1)(h_{k+1}(\tilde{y}) - h_{k+1}(y_k)) + 2\beta\|z_k - \tilde{y}\|^2 \\
&= k(k+1)(h_k(\tilde{y}) - h_k(y_k)) + 2\beta\|z_k - \tilde{y}\|^2 + \\
&\quad\ k(k+1)(h_{k+1}(\tilde{y}) - h_k(\tilde{y})) + k(k+1)(h_k(y_k) - h_{k+1}(y_k)) \\
&= \Phi^{h_k}(k) + k(k+1)(g(x_{k+1}, \tilde{y}) - g(x_k, \tilde{y})) + k(k+1)(g(x_k, y_k) - g(x_{k+1}, y_k)) \\
&\overset{(a)}{\leq} \Phi^{h_k}(k) + k(k+1)(g(x_{k+1}, \tilde{y}) - g(x_k, \tilde{y})) + k(k+1)\varepsilon_{\text{step}}^{(k)} \qquad (27) \\
\overset{(b)}{\implies} \Phi^{h_K}(K) &\leq \Phi^{h_0}(0) + \sum_{k=0}^{K-1} k(k+1)(g(x_{k+1}, \tilde{y}) - g(x_k, \tilde{y})) + \sum_{k=1}^{K-1} k(k+1)\varepsilon_{\text{step}}^{(k)} \\
&\leq \Phi^{h_0}(0) + (K-1)Kg(x_K, \tilde{y}) - \sum_{k=1}^{K-1} 2k\, g(x_k, \tilde{y}) + \sum_{k=1}^{K-1} k(k+1)\varepsilon_{\text{step}}^{(k)} \qquad (28)
\end{aligned}$$

Where (a) follows from Lemma 5 and $g(x_k, y_k) - g(x_{k+1}, y_k) \leq g(x_k, y_k) - \min_x g(x, y_k) \leq \varepsilon_{\text{step}}^{(k)}$, (b) is obtained summing (27) over $k = \{0, \ldots, K-1\}$. Rearranging the terms of (28) we get,

$$\begin{aligned}
\Phi^{h_0}(0) + \sum_{k=1}^{K-1} k(k+1)\varepsilon_{\text{step}}^{(k)} &\geq \sum_{k=1}^{K-1} 2k\, g(x_k, \tilde{y}) + \Phi^{h_K}(K) - (K-1)Kg(x_K, \tilde{y}) \\
&\geq \sum_{k=1}^{K-1} 2k\, g(x_k, \tilde{y}) + K(K+1)(g(x_K, \tilde{y}) - g(x_K, y_K)) + \\
&\quad\ 2\beta\|z_K - \tilde{y}\|^2 - (K-1)Kg(x_K, \tilde{y}) \\
&\geq \sum_{k=1}^{K} 2k\, g(x_K, \tilde{y}) - K(K+1)g(x_K, y_K) \\
&\overset{(a)}{\geq} K(K+1)[g(\bar{x}_K, \tilde{y}) - g(x_K, y_K)] \\
&\overset{(b)}{\geq} K(K+1)[g(\bar{x}_K, \tilde{y}) - g(\tilde{x}, y_K) - \varepsilon_{\text{step}}^{(K)}] \qquad (29)
\end{aligned}$$

where (a) uses the $\bar{x}_K = \frac{1}{K(K+1)} \sum_{k=1}^{K} (2i)\, x_i$ and convexity of $g(\cdot, \tilde{y})$, and (b) uses Lemma 6. Thus we get that,

$$\begin{aligned}
g(\bar{x}_K, \tilde{y}) - g(\tilde{x}, y_K) &\leq \frac{\Phi^{h_0}(0)}{K(K+1)} + \sum_{k=1}^{K} \frac{k(k+1)}{K(K+1)} \varepsilon_{\text{step}}^{(k)} \\
&= \frac{2\beta\|y_0 - \tilde{y}\|^2}{K(K+1)} + \sum_{k=1}^{K} \frac{k(k+1)}{K(K+1)} \varepsilon_{\text{step}}^{(k)} \qquad (30)
\end{aligned}$$

Finally we get the desired general statement by taking minimum and maximum over $\tilde{x}$ and $\tilde{y}$ respectively. By selecting $\varepsilon_{\text{step}}^{(k)} = \frac{L^2 D_{\mathcal{Y}}^2}{\sigma k^3 (k+1)}$ we get,

$$\max_{\tilde{y} \in \mathcal{Y}} g(\bar{x}_K, \tilde{y}) - \min_{\tilde{x} \in \mathcal{X}} g(\tilde{x}, y_K) \leq \frac{6 \frac{L^2}{\sigma} D_{\mathcal{Y}}^2}{K(K+1)} \qquad (31)$$

Further, using Lemma 5 and $\varepsilon_{\text{step}}^{(k)} = \frac{L^2 D_{\mathcal{Y}}^2}{\sigma k^3 (k+1)}$, we get that the total number of gradient computations at iteration k is at most $O\big(\sqrt{\frac{L}{\sigma}} \log^2(k)\big)$:

$$\left\lceil \log_2 5k^2 \sqrt{\frac{L}{\sigma}} \right\rceil O\Big(\sqrt{\frac{L}{\sigma}} \log\left(k^4\right)\Big) \qquad (32)$$

Note that in updating y_{k+1} in Eq. (24) and x_{k+1} in `Imp-STEP` sub-routine, we were applying the principle of conceptual Mirror-Prox, where the update needs to satisfy some fixed point equation. This is critical in proving the above fast convergence rate.

B.5.1 Proof of Lemma 5

For brevity, we define the following operations,

$$x^*(y) = \arg\min_{x \in \mathcal{X}} g(x, y) \tag{33}$$

$$y^+ = \mathcal{P}_{\mathcal{Y}}\left(w + \frac{1}{\beta}\nabla_y g(x^*(y), w)\right) \tag{34}$$

$x^*(y)$ is unique since $g(\cdot, y)$ is strongly convex. We first prove that, $x^*(y)$ is $\frac{L}{\sigma}$-Lipschitz continuous as follows.

$$\begin{aligned}
\sigma\|x^*(y_2) - x^*(y_1)\|^2 &\overset{(a)}{\leq} \langle \nabla_x g(x^*(y_2), y_2) - \nabla_x g(x^*(y_1), y_2), x^*(y_2) - x^*(y_1)\rangle \\
&\overset{(b)}{\leq} \langle -\nabla_x g(x^*(y_1), y_2), x^*(y_2) - x^*(y_1)\rangle \\
&\overset{(c)}{\leq} \langle \nabla_x g(x^*(y_1), y_1) - \nabla_x g(x^*(y_1), y_2), x^*(y_2) - x^*(y_1)\rangle \\
&\overset{(d)}{\leq} L\|y_1 - y_2\|\|x^*(y_2) - x^*(y_1)\|
\end{aligned} \tag{35}$$

where (a) uses σ-strong convexity of $g(\cdot, y)$, (b) and (c) use the necessary first order optimality conditions for $x^*(y_1)$ and $x^*(y_2)$: $\langle \nabla_x g(x^*(y), y), x - x^*(y)\rangle \geq 0$, and (d) uses Cauchy-Schwarz inequality and L-smoothness of g (Definition 1). Next we prove that the operation $(\cdot)^+$ is a contraction as follows,

$$\begin{aligned}
\|y_1^+ - y_2^+\| &= \|\mathcal{P}_{\mathcal{Y}}\left(w + \frac{1}{\beta}\nabla_y g(x^*(y_1), w)\right) - \mathcal{P}_{\mathcal{Y}}\left(w + \frac{1}{\beta}\nabla_y g(x^*(y_2), w)\right)\| \\
&\overset{(a)}{\leq} \frac{1}{\beta}\|\nabla_y g(x^*(y_1), w) - \nabla_y g(x^*(y_2), w)\| \\
&\overset{(b)}{\leq} \frac{L}{\beta}\|x^*(y_1) - x^*(y_2)\| \\
&\overset{(c)}{\leq} \frac{L}{\beta}\frac{L}{\sigma}\|y_1 - y_2\| \overset{(d)}{\leq} 2^{-1}\|y_1 - y_2\|
\end{aligned} \tag{36}$$

where (a) uses Pythagorean theorem and (34), (b) uses L-smoothness of g, (c) uses (35), and (d) uses $\beta \geq 2\frac{LL}{\sigma}$. Therefore, $(\cdot)^+$ is a contraction by Banach's fixed point theorem, and thus it has a unique fixed point $\tilde{y}$: $(\tilde{y})^+ = \tilde{y}$, as $\mathcal{Y}$ is a compact (and hence complete) metric space. Now we will prove that the output of `Imp-STEP`, $(\hat{x}_R, y_{R+1})$ satisfies (21). Notice that if ε_{agd} is small then $\hat{x}_r$ is close to $x^*(y_r)$:

$$\frac{\sigma}{2}\|\hat{x}_r - x^*(y_r)\|^2 \overset{(a)}{\leq} g(\hat{x}_r, y_r) - \min_x g(x, y_r) \overset{(b)}{\Longrightarrow} \|\hat{x}_r - x^*(y_r)\| \leq \sqrt{\frac{2\varepsilon_{\text{agd}}}{\sigma}} = \frac{\beta\varepsilon_{\text{mp}}}{4L} \tag{37}$$

where (a) uses σ-strong convexity and optimality of $x^*(y_r)$, and (b) uses (7), and (c) uses $\varepsilon_{\text{agd}} = \sigma\beta^2\varepsilon_{\text{mp}}/(32L^2)$. Next we see that $\|y_r - \tilde{y}\|$ decreases to ε exponentially fast.

$$\begin{aligned}
\|y_r - \tilde{y}\| &\overset{(a)}{=} \|\mathcal{P}_{\mathcal{Y}}\left(w + \frac{1}{\beta}\nabla_y g(\hat{x}_{r-1}, w)\right) - (\tilde{y})^+\| \\
&\overset{(b)}{\leq} \|y_{r-1}^+ - (\tilde{y})^+\| + \|\mathcal{P}_{\mathcal{Y}}\left(w + \frac{1}{\beta}\nabla_y g(x^*(y_{r-1}), w)\right) - \mathcal{P}_{\mathcal{Y}}\left(w + \frac{1}{\beta}\nabla_y g(\hat{x}_{r-1}, w)\right)\| \\
&\overset{(c)}{\leq} 2^{-1}\|y_{r-1} - \tilde{y}\| + \frac{L}{\beta}\|x^*(y_{r-1}) - \hat{x}_{r-1}\| \\
&\overset{(d)}{\leq} 2^{-1}\|y_{r-1} - \tilde{y}\| + \frac{\varepsilon_{\text{mp}}}{4} && (38) \\
&\overset{(e)}{\leq} 2^{-r}\|y_0 - \tilde{y}\| + \frac{\varepsilon_{\text{mp}}}{2} && (39)
\end{aligned}$$

where (a) uses $y_{r+1} = \mathcal{P}_{\mathcal{Y}}\big(w + \frac{1}{\beta}\nabla_y g(\hat{x}_r, w)\big)$ and the fact that $\tilde{y} = (\tilde{y})^+$ is a fixed point, (b) uses triangular inequality and (34), (c) uses (36), Pythagorean theorem and L-smoothness of g

(Definition 1), (d) uses (37), and (e) just unrolls the recurrence relation in (38) . Next, we prove that the minimizer at y_{R+1}, $x^*(y_{R+1})$ is not far from $\hat{x}_R$.

$$\begin{aligned}
\|x^*(y_{R+1}) - \hat{x}_R\| &\overset{(a)}{\leq} \|x^*(y_{R+1}) - x^*(\tilde{y})\| + \|x^*(\tilde{y}) - x^*(y_R)\| + \|x^*(y_R) - \hat{x}_R\| \\
&\overset{(b)}{\leq} \frac{L}{\sigma}(\|y_{R+1} - \tilde{y}\| + \|y_R - \tilde{y}\|) + \frac{\beta\varepsilon_{\mathrm{mp}}}{4L} \\
&\overset{(c)}{\leq} \frac{L}{\sigma}(\varepsilon_{\mathrm{mp}} + \varepsilon_{\mathrm{mp}}) + \frac{\beta\varepsilon_{\mathrm{mp}}}{4L} = (2\frac{L}{\sigma} + \frac{\beta}{4L})\varepsilon_{\mathrm{mp}} \qquad (40)
\end{aligned}$$

where (a) uses triangle inequality, and (b) uses (35) and 37, and (c) uses (39) and the fact that $R = \lceil \log_2 \frac{2D_{\mathcal{Y}}}{\varepsilon_{\mathrm{mp}}} \rceil$. Finally, we prove that (x_R, y_{R+1}) satisfies (21).

$$\begin{aligned}
g(\hat{x}_R, y_{R+1}) &\overset{(a)}{\leq} g(x^*(y_{R+1}), y_{R+1}) + \langle \nabla_x g(x^*(y_{R+1}), y_{R+1}), \hat{x}_R - x^*(y_{R+1}), \rangle + \frac{L}{2}\|x^*(y_{R+1}) - \hat{x}_R\|^2 \\
&\overset{(b)}{\leq} \min_x g(x, y_{R+1}) + 0 + \frac{25LL^2\varepsilon_{\mathrm{mp}}^2}{8\sigma^2} \overset{(c)}{=} \min_x g(x, y_{R+1}) + \varepsilon_{\mathrm{step}} \qquad (41)
\end{aligned}$$

where (a) uses L-smoothness of $g(\cdot, y)$, (b) uses necessary first order optimality condition: $\langle \nabla_x g(x^*(y), y), x - x^*(y) \rangle = 0$ and (40), and (c) uses $\varepsilon_{\mathrm{mp}} = \frac{2\sigma}{5L}\sqrt{\frac{2\varepsilon_{\mathrm{step}}}{L}}$.

Let the number of gradient computations done per iteration of `Imp-STEP` (a run of accelerated gradient ascent) be T_r and $\kappa = \sqrt{L/\sigma}$. Then, from guarantee on AGD ([2, Eqn. (5.68)]), we get that,

$$\begin{aligned}
g(\hat{x}_r, y_r) - g(x^*(y_r), y_r) &\leq \Big(1 + \frac{1}{\sqrt{\kappa} - 1}\Big)^{-T_r} \Big(g(x_0, y_r) - g(x^*(y_r), y_r) + \frac{\sigma}{2}\|x_0 - x^*(y_r)\|^2\Big) \\
&\leq e^{-T_r/\sqrt{\kappa}}\, 2\left(g(x_0, y_r) - g(x^*(y_r), y_r)\right) \\
&\leq e^{-T_r/\sqrt{\kappa}}\, 2\left(f(x_0) - h(y_r)\right) \\
&\leq e^{-T_r/\sqrt{\kappa}}\, 2\left(f(x_0) - \min_{y' \in D_{\mathcal{Y}}} h(y')\right), \qquad (42)
\end{aligned}$$

where $\min_{y' \in D_{\mathcal{Y}}} h(y')$ is well-defined since $\mathcal{Y}$ is compact and h is smooth (Lemma 6). This means that if we want $g(\hat{x}_r, y_r) - g(x^*(y_r), y_r) \leq \varepsilon_{\mathrm{agd}}$, then required number of steps T_r is at most,

$$\begin{aligned}
\left\lceil \sqrt{\frac{L}{\sigma}} \log \frac{2(f(x_0) - \min_{y' \in D_{\mathcal{Y}}} h(y'))}{\varepsilon_{\mathrm{agd}}} \right\rceil &= \left\lceil \sqrt{\frac{L}{\sigma}} \log \frac{50L(f(x_0) - \min_{y' \in D_{\mathcal{Y}}} h(y'))}{\sigma\varepsilon_{\mathrm{step}}} \right\rceil \\
&= O\Big(\sqrt{\frac{L}{\sigma}} \log\Big(\frac{1}{\varepsilon_{\mathrm{step}}}\Big)\Big) \qquad (43)
\end{aligned}$$

B.6 Smoothness of dual of strongly-convex–concave minimax problem

Lemma 6. *For a σ-strongly-convex–concave L-smooth function $g(\cdot, \cdot)$, $h(u) = \min_{x \in \mathcal{X}} g(x, u)$ is an $\left(L + \frac{L^2}{\sigma}\right)$-smooth concave function.*

Proof. We know that $h(y) = \min_{x \in \mathcal{X}} g(x, y)$, where $g(\cdot, y)$ is σ-strongly convex, $g(x, \cdot)$ is concave, g is L-smooth (Definition 1). Since $g(\cdot, y)$ is strongly convex, the minimizer $x^*(y) = \arg\min_{x \in \mathcal{X}} g(x, y)$ unique. Then by Danskin's theorem [4, Section 6.11], h is differentiable and $\nabla h(y) = \nabla_y g(x^*(y), y)$. Then h can be show to be smooth as follows,

$$\begin{aligned}
\|\nabla h(y_1) - \nabla h(y_1)\| &= \|\nabla_y g(x^*(y_1), y_1) - \nabla_y g(x^*(y_2), y_2)\| \\
&\leq \|\nabla_y g(x^*(y_1), y_1) - \nabla_y g(x^*(y_1), y_2)\| + \|\nabla_y g(x^*(y_1), y_2) - \nabla_y g(x^*(y_2), y_2)\| \\
&\overset{(a)}{\leq} L\|y_1 - y_2\| + L\|x^*(y_1) - x^*(y_2)\| \\
&\overset{(b)}{\leq} L\|y_1 - y_2\| + L\frac{L}{\sigma}\|y_1 - y_2\| = \Big(L + \frac{LL}{\sigma}\Big)\|y_1 - y_2\| \qquad (44)
\end{aligned}$$

where (a) uses L-smoothness of g and (b) uses (35). ☐

B.7 Proof of Theorem 2

We first note that by Lemma 3 and L-weak convexity of $g(\cdot, y)$ and $2L$-strong convexity of $L\|x - x_k\|^2$, $\widehat{g}(x, y; x_k) := g(x, y) + L\|x - x_k\|^2$ is L-strongly-convex. Similarly, $\widehat{f}(\cdot; x_k) := \max_{y \in \mathcal{Y}}[\widehat{g}(x, y; x_k) = g(x, y) + L\|x - x_k\|^2]$ is also L-strongly-convex.

We now divide the analysis of each iteration of our algorithm into two cases:

Case 1: $\widehat{f}(x_{k+1}; x_k) \le f(x_k) - 3\tilde{\varepsilon}/4$. As every instance of Case 1 ensures $f(x_{k+1}) \le \widehat{f}(x_{k+1}; x_k) \le f(x_k) - 3\tilde{\varepsilon}/4$, we can have only $\left\lceil \frac{4(f(x_0) - f^*)}{3\tilde{\varepsilon}} \right\rceil$ Case 1 steps before termination. This claim requires monotonic decrease in $f(x_k)$ which holds until $f(x_{k+1}) \ge f(x_k)$, after which $\widehat{f}(x_{k+1}; x_k) \ge f(x_k)$, which in-turn imply that Prox-DIAG terminates (see termination condition of Prox-DIAG).

Case 2: $\widehat{f}(x_{k+1}; x_k) > f(x_k) - 3\tilde{\varepsilon}/4$: In this case, we show that x_k is already an ε-FOSP and the algorithm returns x_k.

$$f(x_k) - \frac{3\tilde{\varepsilon}}{4} < \widehat{f}(x_{k+1}; x_k) \le \min_x \widehat{f}(x; x_k) + \frac{\tilde{\varepsilon}}{4} \qquad \Longrightarrow \qquad f(x_k) < \min_x \widehat{f}(x; x_k) + \tilde{\varepsilon} \tag{45}$$

Define x_k^* as the point satisfying $x_k^* = \arg\min_x \widehat{f}(x; x_k)$. By L-strong convexity of $\widehat{f}(\cdot; x_k)$ (9), we prove that x_k is close to x_k^*:

$$\widehat{f}(x_k^*; x_k) + \frac{L}{2}\|x_k - x_k^*\|^2 \;\le\; \widehat{f}(x_k; x_k) \;=\; f(x_k) \;\overset{(a)}{<}\; \widehat{f}(x_k^*; x_k) + \tilde{\varepsilon} \implies \|x_k - x_k^*\| < \sqrt{\frac{2\tilde{\varepsilon}}{L}} \tag{46}$$

where (a) uses (45). Now consider any $\tilde{x} \in \mathcal{X}$, such that $4\sqrt{\tilde{\varepsilon}/L} \le \|\tilde{x} - x_k\|$. Then,

$$\begin{aligned} f(\tilde{x}) + L\|\tilde{x} - x_k\|^2 &= \max_{y \in \mathcal{Y}} g(\tilde{x}, y) + L\|\tilde{x} - x_k\|^2 = \widehat{f}(\tilde{x}; x_k) \overset{(a)}{\ge} \widehat{f}(x_k^*; x_k) + \frac{L}{2}\|\tilde{x} - x_k^*\|^2 \\ &\overset{(b)}{\ge} f(x_k) - \tilde{\varepsilon} + \frac{L}{2}(\|\tilde{x} - x_k\| - \|x_k - x_k^*\|)^2 \overset{(c)}{\ge} f(x_k) + \tilde{\varepsilon}, \end{aligned} \tag{47}$$

where (a) uses uses L-strong convexity of $\widehat{f}(\cdot; x_k)$ at its minimizer x_k^*, (b) uses (45), and (b) and (c) use triangle inequality, (46) and $4\sqrt{\tilde{\varepsilon}/L} \le \|\tilde{x} - x_k\|$.

Now consider the Moreau envelope, $f_{\frac{1}{2L}}(x) = \min_{x' \in X} \phi_{\frac{1}{2L}, x}(x')$ where $\phi_{\lambda, x}(x') = f(x') + L\|x - x'\|^2$. Then, we can see that $\phi_{\frac{1}{2L}, x_k}(x')$ achieves its minimum in the ball $\{x' \in \mathcal{X} \mid \|x' - x_k\| \le 4\sqrt{\tilde{\varepsilon}/L}\}$ by (47) and Lemma 4(a). Then, with Lemma 4(b,c) and $\tilde{\varepsilon} = \frac{\varepsilon^2}{64\,L}$, we get that,

$$\|\nabla f_{\frac{1}{2L}}(x_k)\| \le (2L)\|x_k - \hat{x}_{\frac{1}{2L}}(x_k)\| = 8\sqrt{L\tilde{\varepsilon}} = \varepsilon, \tag{48}$$

i.e., x_k is an ε-FOSP.

By combining the above two cases, we establish that $O\left(\left\lceil \frac{4(f(x_0) - f^*)}{3\tilde{\varepsilon}} \right\rceil\right)$ "outer" iterations ensure convergence to a ε-FOSP. We now compute the first-order complexity of each of these "outer" iterations. Recall that we use use the DIAG (Algorithm 1) algorithm for L-strongly-convex concave $2L$-smooth minimax problem to solve the inner optimization problem. So, if for each iteration of inner problem, DIAG algorithm takes K steps then, by $\tilde{\varepsilon} = \frac{\varepsilon^2}{64\,L}$ and Theorem 1,

$$\frac{6(2L)^2 D_{\mathcal{Y}}^2}{LK^2} \le \frac{\tilde{\varepsilon}}{4} = \frac{\varepsilon^2}{2^8 L} \implies O\left(\frac{LD_{\mathcal{Y}}}{\varepsilon}\right) \le K \tag{49}$$

Therefore the number of gradient computations required for each iteration of inner problem is $O\left(\frac{LD_{\mathcal{Y}}}{\epsilon} \log^2\left(\frac{1}{\varepsilon}\right)\right)$ (Theorem 1), which along with the bound on the number of outer iterations establishes the Theorem's upper bound on the number of first-order oracle calls.

C Minimizing finite max-type function with smooth components

As a special case of nonconvex–concave minimax problem, consider minimizing a weakly convex $f(x)$, with a special structure of *finite max-type function*:

$$\min_x \left[f(x) = \max_{1 \le i \le m} f_i(x) \right], \tag{P3}$$

where $x \in \mathbb{R}^p$, the functional components $f_i(x)$'s could be *nonconvex* but are L-smooth and G-Lipschitz. Suppose f itself takes a minimum value $f^* > -\infty$. For this problem, we propose and study a proximal (Prox-FDIAG) algorithm (Algorithm 5 presented in Appendix C.1) that is inspired by Algorithm 2 with the inner problem-solver replaced by Nesterov's finite convex minimax scheme [37, Section 2.3.1] instead of Algorithm 1. Using same proof technique as Theorem 2, we get:

Corollary 1 (Convergence rate of Prox-FDIAG)**.** *If the functional components $f_i(x)$'s are G-Lipschitz and L-smooth, and the optimal solution is bounded below, i.e. $f(x) \geq f^* > -\infty$, then after: $K = \left\lceil \frac{4^4 L(f(x_0)-f^*)}{3\varepsilon^2} \right\rceil$* outer *steps, Prox-FDIAG outputs an ε-FOSP. The total first-order oracle complexity to find ε-FOSP is: $\left\lceil \frac{4^4 L(f(x_0)-f^*)}{3\varepsilon^2} \right\rceil \cdot \left\lceil \frac{2^4 G}{\varepsilon}(m \log^{3/2} m) \right\rceil$.*

See Appendix C.1 for a proof. Current best rate for this problem is achieved by subgradient methods. As the subgradient of a finite minimax function $\nabla_{i^*} f(x)$ is easy to evaluate, where $i^* \in \arg\max_i f_i(x)$, a rate of $O(m/\varepsilon^4)$ first-order oracle and function calls is achieved by the state-of-the-art subgradient method in [11]. We can obtain a similar result using Algorithm 1 but it requires extension to non-Euclidean settings with the framework of Bregman divergences. This is fairly standard and will be updated in the next version of the paper.

Algorithm 5: Proximal Finite Dual Implicit Accelerated Gradient (Prox-FDIAG) for finite nonconvex concave minimax optimization

---

Input: functional components $\{f_i\}_{i=1}^m$, Lipschitzness G, smoothness L, domain $\mathcal{X}$, target accuracy ε, initial point x_0

Output: x_k

1 $\tilde{\varepsilon} \leftarrow \frac{\varepsilon^2}{64\,L}$

2 **for** $k = 0, 1, \ldots$ **do**

3 Using excessive gap technique [35, Problem (7.11)] for strongly convex components, find $x_{k+1} \in \mathcal{X}$ such that,

$$\widehat{f}(x_{k+1}; x_k) \quad \leq \quad \min_x \widehat{f}(x; x_k) + \tilde{\varepsilon}/4 \tag{50}$$

if $f(x_k) - 3\tilde{\varepsilon}/4 < \widehat{f}(x_{k+1}; x_k)$ **then**

4 **return** x_k

C.1 Proof of Corollary 1

Let

$$\widehat{f}(x; x_k) \quad = \quad \max_{1 \leq i \leq m} \; f_i(x_k) + \langle \nabla f_i(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|^2 \tag{51}$$

be a quadratic approximation of the finite max-type function $f(x)$ at x_k. Then, $\widehat{f}(\cdot; x_k)$ is L-strongly convex, since it is a maximum of convex functions and the quadratic term in (51) is independent of i.

Proof is similar to that of Theorem 2. We divide the analysis of each iteration of our algorithm into two cases.

Case 1: $\widehat{f}(x_{k+1}; x_k) \leq f(x_k) - 3\tilde{\varepsilon}/4$**.** This ensure that at iteration k the objective value decreases by at least $3\tilde{\varepsilon}/4$ since, $f(x_{k+1}) \leq \widehat{f}(x_{k+1}; x_k) \leq f(x_k) - 3\tilde{\varepsilon}/4$. One cannot have more than $\left\lceil \frac{4(f(x_0)-f^*)}{3\tilde{\varepsilon}} \right\rceil$ instances of Case 1, before termination.

Case 2: $\widehat{f}(x_{k+1}; x_k) > f(x_k) - 3\tilde{\varepsilon}/4$**:** We show that x_k is an ε-FOSP as follows.

$$f(x_k) - \frac{3\tilde{\varepsilon}}{4} < \widehat{f}(x_{k+1}; x_k) \leq \min_x \widehat{f}(x; x_k) + \frac{\tilde{\varepsilon}}{4} \qquad \Longrightarrow \qquad f(x_k) < \min_x \widehat{f}(x; x_k) + \tilde{\varepsilon} \tag{52}$$

Define x_k^* as the point satisfying $x_k^* = \arg\min_x \widehat{f}(x; x_k)$. By L-strong convexity of $\widehat{f}(\cdot, x_k)$ (51), we prove that x_k is close to x_k^*:

$$\begin{aligned} \widehat{f}(x_k^*; x_k) + \frac{L}{2}\|x_k - x_k^*\|^2 \ &\leq\ \widehat{f}(x_k; x_k) \ = \ f(x_k) \ \overset{(a)}{<} \ \widehat{f}(x_k^*; x_k) + \tilde{\varepsilon} \\ \Longrightarrow \quad \|x_k - x_k^*\| &< \sqrt{\frac{2\tilde{\varepsilon}}{L}} \end{aligned} \tag{53}$$

where (a) uses (52). Now consider any $\tilde{x} \in \mathcal{X}$, such that $4\sqrt{\tilde{\varepsilon}/L} \leq \|\tilde{x} - x_k\|$. Then,

$$\begin{aligned} f(\tilde{x}) + L\|\tilde{x} - x_k\|^2 &= \max_i f_i(\tilde{x}) + L\|\tilde{x} - x_k\|^2 \\ &\overset{(a)}{\geq} \max_i f_i(\tilde{x}) + \langle \nabla f_i(x_k), \tilde{x} - x_k \rangle + \frac{L}{2}\|\tilde{x} - x_k\|^2 \\ &\overset{(b)}{=} \widehat{f}(\tilde{x}; x_k) \\ &\overset{(c)}{\geq} \widehat{f}(x_k^*; x_k) + \frac{L}{2}\|\tilde{x} - x_k^*\|^2 \\ &\overset{(d)}{\geq} f(x_k) - \tilde{\varepsilon} + \frac{L}{2}(\|\tilde{x} - x_k\| - \|x_k - x_k^*\|)^2 \\ &\overset{(e)}{\geq} f(x_k) - \tilde{\varepsilon} + 2\tilde{\varepsilon} \ = \ f(x_k) + \tilde{\varepsilon} \end{aligned} \tag{54}$$

where (a) uses weak convexity of f_i, (b) uses (51), (c) uses L-strong convexity of $\widehat{f}(\cdot; x_k)$ at its minimizer x_k^*, (d) uses (52), and (b) and (e) use triangle inequality, (53) and $4\sqrt{\tilde{\varepsilon}/L} \leq \|\tilde{x} - x_k\|$.

Now consider the Moreau envelope, $f_{\frac{1}{2L}}(x) = \min_{x' \in X} \phi_{\frac{1}{2L}, x}(x')$ where $\phi_{\lambda, x}(x') = f(x') + L\|x - x'\|^2$. Then, we can see that $\phi_{\frac{1}{2L}, x_k}(x')$ achieves its minimum in the ball $\{x' \in \mathcal{X} \mid \|x' - x_k\| \leq 4\sqrt{\tilde{\varepsilon}/L}\}$ by (54) and Lemma 4(a). Thus, with Lemma 4(b,c), we get that,

$$\|\nabla f_{\frac{1}{2L}}(x_k)\| \leq (2L)\|x_k - \hat{x}_{1/2L}(x_k)\| = 8\sqrt{L\tilde{\varepsilon}} = \varepsilon \tag{55}$$

Now we use the excessive gap technique for non-smooth strongly convex functions with max-structure to solve the inner optimization problem in $4G(m \log m)\sqrt{\frac{\log m}{\tilde{\varepsilon} L}}$ computations [35, Problem (7.11)].

Putting these together we see that the total number of inner steps to reach ε-FOSP is,

$$\left\lceil \frac{4(f(x_0) - f^*)}{3\tilde{\varepsilon}} \right\rceil \left\lceil 2G(m \log m)\sqrt{\frac{\log m}{L\tilde{\varepsilon}}} \right\rceil = \left\lceil \frac{4^4 L(f(x_0) - f^*)}{3\varepsilon^2} \right\rceil \left\lceil \frac{2^5 G}{\varepsilon}(m \log^{3/2} m) \right\rceil \tag{56}$$

C.2 Adaptive Prox-FDIAG algorithm

In this section, we provide the Adaptive Prox-FDIAG (Algorithm 6) to find an ε-FOSP of the finite max-type nonconvex minimax problem P3 with L-smooth components. Adaptive Prox-FDIAG is a variation of the Prox-FDIAG (Algorithm 5). Adaptive Prox-FDIAG uses Prox-FDIAG as a sub-routine and successively finds ε'-FOSPs, for geometrically decreasing values of ε' starting from ε_0 $(\geq \varepsilon)$ until ε' becomes equal to ε. It uses the ε'-FOSP as the starting point to find an $\varepsilon'/2$-FOSP. In the following corollary, we show that Adaptive Prox-FDIAG has the same the first-order oracle complexity (up to a $O(\log(\frac{1}{\varepsilon}))$ factor) as the Prox-FDIAG.

Corollary 2 (Convergence rate of Adaptive Prox-FDIAG)**.** *If the functional components $f_i(x)$'s are G-Lipschitz and L-smooth, and the optimal solution is bounded below, i.e. $f(x) \geq f^* > -\infty$, then after: $K = \left\lceil \log_2 \frac{\varepsilon_0}{\varepsilon} \right\rceil$* outer *steps, Adaptive Prox-FDIAG outputs an ε-FOSP. The total first-order oracle complexity to find ε-FOSP is: $\left\lceil \log_2 \frac{\varepsilon_0}{\varepsilon} \right\rceil \left\lceil \frac{4^4 L(f(x_0) - f^*)}{3\varepsilon^2} \right\rceil \cdot \left\lceil \frac{2^4 G}{\varepsilon}(m \log^{3/2} m) \right\rceil$.*

Proof. Notice that, each iteration of Adaptive Prox-FDIAG for finding an ε'-FOSP, is a run of Prox-FDIAG (Algorithm 5), which has a maximum first-order oracle complexity of $\left\lceil \frac{4^4 L(f(x_0) - f^*)}{3\varepsilon^2} \right\rceil \cdot$

$\left\lceil \frac{2^4 G}{\varepsilon}(m \log^{3/2} m) \right\rceil$ for finding an ε'-FOSP (Corollary 1), as $\varepsilon \leq \varepsilon'$. Further, since ε' starts at ε_0 and halves after each iteration until ε' becomes less than or equal to ε, the total number of outer iterations is $K = \left\lceil \log_2 \frac{\varepsilon_0}{\varepsilon} \right\rceil$. □

Therefore, Adaptive Prox-FDIAG has the same first-order oracle complexity as Prox-FDIAG, up to a $O(\log(\frac{1}{\varepsilon}))$ factor. However, we observe that Adaptive Prox-FDIAG converges faster than Prox-FDIAG in our experiments.

Algorithm 6: Adaptive Proximal Finite Dual Implicit Accelerated Gradient (Adaptive Prox-FDIAG) for finite nonconvex concave minimax optimization

Input: functional components $\{f_i\}_{i=1}^m$, Lipschitzness G, smoothness L, domain $\mathcal{X}$, target accuracy ε, initial point x_0, initial accuracy ε_0

Output: x_k

```
ε' ← max(ε_0, ε)
for k = 0, 1, ... do
    Using Prox-FDIAG (Algorithm 5) initialized at x_k, find x_{k+1} ∈ 𝒳 such that x_{k+1} is an
      ε'-FOSP (Definition 6) of the function f(x) = max_{1≤i≤m} f_i(x)
    if ε = ε' then
        k ← k + 1
        return x_k
    else
        ε' ← max(ε'/2, ε)
```

D Smoothing technique for strongly-convex–concave minimax problem

In this section we propose and analyze a smoothing technique [36] based indirect algorithm for solving the L-smooth σ-strongly-convex–concave minimax problem. The basic idea is to solve a smoothed (perturbed) version of the original function, $\tilde{g}(x, y) = g(x, y) - \varepsilon \|y\|^2 / 2\, D_{\mathcal{Y}}^2$, which would be a strongly-convex–strongly-concave minimax problem. [1] proposes solving a strongly-convex–strongly-concave problem in linear rate using inexact accelerated gradient descent on its dual, whose main guarantee is given in the theorem below.

Theorem 3. *[1] Inexact accelerated gradient ascent on the dual problem can find an ε-primal dual pair of an L-smooth σ_x-strongly-convex–σ_y-strongly-concave problem:* $\min_x \max_y g(x, y)$*, with* $\widetilde{O}\left(\sqrt{\frac{L + \frac{L^2}{\sigma_x}}{\sigma_y}} \sqrt{\frac{L}{\sigma_x}}\right)$ *calls to the first order gradient oracle of g.*

Now using this algorithm on the function $\tilde{g}$ can recover the same rate as DIAG method as follows. Plugging in $L = O(L)$, $\sigma_x = \sigma$, and $\sigma_y = \varepsilon / D_{\mathcal{Y}}^2$ into the algorithm complexity of Theorem 3 gives you a complexity of,

$$\widetilde{O}\left(\frac{L D_{\mathcal{Y}}}{\sqrt{\sigma \varepsilon}} \sqrt{\frac{L}{\sigma}}\right),$$

finding an ε-primal dual pair, $(\bar{x}, \bar{y})$, of $\tilde{g}$. Since $\max_{y \in \mathcal{Y}} g(\bar{x}, y) \leq \max_{y \in \mathcal{Y}} \tilde{g}(\bar{x}, y) + \varepsilon/2$ and $\tilde{g}(x, \bar{y}) \leq g(x, \bar{y})$, we get that,

$$\max_{y \in \mathcal{Y}} g(\bar{x}, y) - \min_{x \in \mathcal{X}} g(x, \bar{y}) \leq \max_{y \in \mathcal{Y}} \tilde{g}(\bar{x}, y) - \min_{x \in \mathcal{X}} \tilde{g}(x, \bar{y}) + O(\varepsilon) .$$

Using these two facts, we see that smoothing technique has the same algorithmic complexity, $\widetilde{O}\left(\frac{L D_{\mathcal{Y}}}{\sqrt{\sigma \varepsilon}} \sqrt{\frac{L}{\sigma}}\right)$, as that of DIAG. However the drawback for this method over the direct DIAG is that smoothing technique requires a prefixed tolerance parameter ε.

E Experimental details

We consider the following problem.

$$\min_{x\in\mathbb{R}^2}\left[f(x)=\max_{1\le i\le m=9} f_i(x)\right] \tag{57}$$

where $f_i(x)=q_{(-1,\,(X_i^{(1)},X_i^{(2)}),\,C_i)}(x)$ for all $1\le i\le 8$, where $q_{(a,b,c)}(x)=a\|x-b\|_2^2+c$, $X_i^{(1)}$ and $X_i^{(2)}$ are generated from the interval $[-3.0,3.0]$ uniformly at random, and C_i is generated from the interval $[1.0,5.0]$ uniformly at random. We fix the last component $f_9(x)=q_{(0.5,\,(0,0),\,0)}(x)$. Each f_i is smooth with parameter $L=1$, which implies that f is L-weakly convex.

We implement three algorithms: Prox-FDIAG (Algorithm 5), Adaptive Prox-FDIAG (Algorithm 6), and subgradient method [11]. In Prox-FDIAG, we use excessive gap technique [35, Problem (7.11)] (a primal-dual algorithm) to solve the inner sub-problem. As the stopping criteria $\widehat{f}(x_{k+1};x_k)\le \min_x \widehat{f}(x;x_k)+\tilde{\varepsilon}/4$ cannot be directly checked, we instead check a sufficient condition; we stop the excessive gap technique when the primal-dual gap is less than $\tilde{\varepsilon}/4$, which can be checked efficiently. Adaptive Prox-FDIAG is a variant of Prox-FDIAG, where we adaptively and successively decrease the tolerance parameter ε' starting from a large tolerance ε_0. It has the same first-order oracle complexity guarantee as Prox-FDIAG (up to an $O(\log(1/\varepsilon))$ factor). However, in Figure 1, we observe that Adaptive Prox-FDIAG can converge faster in practice. We set the initial tolerance ε_0 as 10.0. For a description of the algorithm we refer to Appendix C.2.

All the algorithms are initialized with the point $x_0=(4,4)$ and are given a Lipschitzness parameter of $G=2\,L\,\|x_0\|_2$. We run the algorithms ten times with randomly generated instances of the objective function $f(x)$. In Figure 1, we plot the norm of gradient of Moreau envelope $\|\nabla f_{\frac{1}{2L}}(x_k)\|_2$ against the number of iterations k in log-log scale. We compute the gradient of the Moreau envelope at any point x, by solving the corresponding convex-concave saddle point problem (18) using Mirror-Prox [34] method with appropriate primal-dual gap based stopping criteria and then using Lemma 4(c). For Prox-FDIAG (red circles), we show in a scatter plot the gradient norm $\|\nabla f_{\frac{1}{2L}}(x_{K(\varepsilon)})\|_2$ at the final output of Prox-FDIAG $x_{K(\varepsilon)}$ versus the total number of inner iterations (of excessive gap technique) taken, for $\varepsilon=10^0,10^{-1},10^{-2},10^{-3}$ over the 10 functions. For Adaptive Prox-FDIAG (black dots) in a scatter plot, we plot the gradient norm $\|\nabla f_{\frac{1}{2L}}(x')\|_2$ at the output x' of each inner sub-problem (excessive gap technique) of each inner Prox-FDIAG step versus the total number of inner iterations (of excessive gap technique) taken to reach that point from the beginning, for $\varepsilon=10^{-7}$ over the 10 functions. For Prox-FDIAG and Adaptive Prox-FDIAG, using solid red and black (respectively) lines we also plot the best linear function (in log-scale) which fits the scatter points (using default parameters of `scipy.stats.linregress`[2]). For the subgradient method (blue triangles), we plot the mean and standard error of gradient norm $\max_{0\le k'\le k}\|\nabla f_{\frac{1}{2L}}(x_{\hat{k}(k')})\|_2$ over the 10 instances at iterations $k=10^0,10^1,\ldots,10^7$. The estimate at each iteration is the best one so far in the function value, i.e. $\hat{k}(k)\in\arg\min_{0\le k'\le k} f(x_{k'})$. We see that, Prox-FDIAG and Adaptive Prox-FDIAG have a faster convergence rate than subgradient method. Further, in the same vein as analogous variants in convex non-smooth optimization, Adaptive Prox-FDIAG is faster than Prox-FDIAG almost always.

Subgradient method has a theoretical convergence rate of $O(\frac{1}{\sqrt{K}})$ for a fixed number of iterations K and a constant step-size $\gamma/\sqrt{K+1}$ [11, Corollary 2.2]. However, similar to the case of convex non-smooth problems, we observe that fixed step-size results in a slow convergence. In our experiments, we achieve a faster convergence for the subgradient method by using a diminishing, non-summable but square-summable step-size, $\gamma/\sqrt{k+1}$, which varies with the iteration number k. This step-size has convergence rate of $O(\frac{\log(k)}{\sqrt{k}})$ [11, Theorem 2.1], but in practice we observe a faster convergence rate than the constant step-size. After a very simple parameter search, we set γ as $0.1\times G\times L^{3/2}$. We ran subgradient method for a total of $K=10^7$ number of iterations. Since, subgradient method is not a descent method, at any iteration k, we keep track of the best point among all the points we have observed so far, $\{x_0,\cdots,x_{k-1}\}$. Ideally, we should keep track of the point with the minimum norm for the gradient of the Moreau envelope, $\|\nabla f_{\frac{1}{2L}}(x_k)\|_2$, but since the computation of the gradient of Moreau envelope is costly, we only keep track of the point with the minimum function value we have observed so far.

[2] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.linregress.html`