Large-scale Minimax Optimization Problems

by

Saeid Hajizadeh
B.Sc., Ferdowsi University of Mashahd, 2011

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Mathematics
in the Graduate College of the
University of Illinois at Chicago, 2023

Chicago, Illinois

Defense Committee:
Lev Reyzin, Chair and Advisor
Gyorgy Turan
Negar Soheili Azad, Information and Decision Sciences
Haihao Lu, University of Chicago Booth School of Business
Saeed Seddighin, TTIC

Dedicated to my parents, my sister, and my lovely nephew, Radman

ACKNOWLEDGMENT

I am grateful for the opportunity to pursue a PhD degree in Mathematics, and I acknowledge the invaluable support and guidance that I received from many individuals during my academic journey.

First and foremost, I would like to express my sincere gratitude to my advisor, Lev Reyzin, for his unwavering support, encouragement, and guidance throughout my PhD program. Lev's expertise, passion for research, and commitment to excellence were instrumental in shaping my research skills and academic growth. I am grateful for the countless hours that he spent mentoring me, and for his willingness to share his vast knowledge and expertise.

I would also like to thank my research mentors, Haihao Lu and Benjamin Grimmer, for their invaluable insights on large-scale optimization professional research. Their guidance and support have been instrumental in shaping my research and enhancing my technical skills. I am grateful for the opportunities they provided to collaborate on research projects and for the valuable feedback that they provided on my work.

I would like to acknowledge the then DGS of the math department, Julius Ross, for his support of inter-institution PhD research, which provided me with the opportunity to pursue my doctoral studies working on the subject of my interest, that is, theoretical optimization.

Finally, I would like to express my heartfelt gratitude to my wonderful family for their unwavering support, love, and encouragement throughout my PhD journey. Their sacrifices,

ACKNOWLEDGMENT (Continued)

patience, and understanding have been instrumental in enabling me to achieve my academic goals.

To all of these individuals, I offer my deepest gratitude for their invaluable support, guidance, and encouragement. Without their help, I would not have been able to achieve this important milestone in my career.

PES

CONTRIBUTIONS OF AUTHORS

Chapter 1 is the review of minimax optimization problem, various first-order algorithms to solve it, and the literature of primal-dual methods and their convergence rates.

Chapter 2 is based on the work "On the Linear Convergence of Extra-Gradient Methods for Nonconvex-Nonconcave Minimax Problems", Saeed Hajizadeh, Haihao Lu, and Benjamin Grimmer, 2022, and covers the analysis of a first order method known as Extra-Gradient Method to solve structured nonconvex-nonconcave minimax optimization problems.

Chapter 3 is based on the (ongoing) work "On Nonconvex-Nonconcave Nonsmooth Minimax Problems", Saeed Hajizadeh, Haihao Lu, and Benjamin Grimmer, which covers the presence of constraints in minimax optimization problems, current results, and challenges.

Chapters 2 and 3 are exclusively the results of the work done by the author in collaboration with Haihao Lu and Benjamin Grimmer. Chapter 1 is the overview of primal-dual methods and their literature.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (Continued)

FIGURE **PAGE**

SUMMARY

Every problem in real life is an optimization problem. From choosing the market from which we buy certain items to the task of the day we choose to focus on first, to which job offer we accept, all can be formulated as an optimization problem. Most of these optimization problems, however, are either very hard, or even impossible, to solve. Modern large-scale optimization aims at finding the class of these problems that are efficiently solvable.

This thesis focuses on a class of optimization problems called minimax optimization. In chapter 1, we will review the literature and introduce somee of the more well-known methods, called "primal-dual algorithms", that are commonly used to solve these methods. In Chapter 2, we propose a variant of one of these methods, that solves the nonconvex-nonconcave minimax optimization problems efficiently upon assuming some structural assumption. We also showcase a class of problems that shows that our structural assumption is tight for the convergence of the said first order method. Chapter 3, addresses the problem of minimax optimization in the presence of constraints. We will present the convergence of damped proximal point methods under an equivalent structure when neither convexity nor concavity is present in our objective function. We also furnish some properties for the constrained saddle envelope, a generalization of Moreau envelope to primal-dual problems, that we believe will be of independent interest in future research endeavors.

CHAPTER 1

INTRODUCTION TO MINIMAX OPTIMIZATION

This dissertation studies minimax optimization problems which have always been an essential part of the optimization due to its wide range of applications in game theory (Başar and Olsder, 1998) and control theory (Hast et al., 2013). Recently, with the advent of General Adversarial Networks (GANs) (Goodfellow et al., 2014), reinforcement learning (Dai et al., 2018; Sutton and Barto, 2018), in particular, the interests in minimax optimization have seen a considerable increase.

Minimax optimization problems are formulated as the following,

$$\min_{x \in C} \max_{y \in D} f(x, y), \tag{1.1}$$

where $C \subset \mathbb{R}^n$ and $D \subset \mathbb{R}^m$ are the constraints of the problem. The case of $C = \mathbb{R}^n$ and $D = \mathbb{R}^m$ is called the unconstrained minimax problem.

The problem (Equation 1.1) is called a convex-concave problem if $x \mapsto f(x, y)$ is convex for every $y \in D$ and $y \mapsto f(x, y)$ is concave for every $x \in C$. Various combinations of these classes are thus defined.

A point $(x^*, y^*) \in C \times D$ is called a Nash equilibrium of problem (Equation 1.1) if

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*), \qquad \text{for all } x \in C, y \in D. \tag{1.2}$$

The following sections of this Chapter are as follows: section 1.1 goes over two of the recent robust machine learning models that motivate the study of minimax problems. Section 1.2 introduces a few of most common first-order methods used in solving minimax optimization problems and argues why the most simple method, gradient descent-ascent, fails to converge to the solution of bilinear (convex-concave) problems. Section 1.3 goes over the literature of primal-dual methods and further covers more particular primal-dual methods.

1.1 Motivations and Applications

In GANs, where the goal is to predict the class a data point belongs to by learning the distribution of the data, two models are simultaneously trained: the generative network G and the discriminator network D. The generative network G tries to generate data from noise prior $p_W(w)$ via a differentiable function $\mathsf{G}(w, X_g)$, where X_g represents the generator's parameters, and the discriminator maps the data point ν to a scalar in $\mathsf{D}(\nu, Y_d) \in (0, 1)$, the probability the data came from data rather than the generator's fake output, where Y_d represents the generator's parameters. This is a zero-sum game and the whole model is trained with D trying to maximize $\log \mathsf{D}(\nu, Y_d)$ and the generator trying to minimize $\log(1 - \mathsf{D}(\mathsf{G}(w, X_g)))$. The game that G and D play simultaneously leads, with some abuse of notation, to the following formulation

$$\min_{X_g \in \mathsf{G}} \max_{Y_d \in \mathsf{D}} \mathbb{E}_{\nu \sim \mathbb{P}_{\text{data}}(\nu)} \mathsf{D}(\nu, Y_d) + \mathbb{E}_{w \in \mathbb{P}_W(w)} \log\left(1 - \mathsf{D}(\mathsf{G}(w, X_g), Y_d)\right) \tag{1.3}$$

where $\mathbb{P}_{\text{data}}$ is the true data distribution.

GANs have been very successful in machine learning community. However, GANs can be succeptible to adversarial attacks. Alternatives (Madry et al., 2018) involve solving a minimax problem with adversarial corruptions t. For example, let (u, v) represent the feature vector and its label, respectively, x represent the model's parameters, and t represent the adversarial corruption parameter. In that case, the robust minimax formulation of the machine learning model $\min_x \mathbb{E}_{(u,v)} l(u, v, x)$ becomes

$$\min_x \mathbb{E}_{(u,v)} \left[\max_{t \in \mathcal{T}} l(u + t, v, x) \right], \tag{1.4}$$

where $\mathcal{T}$ is the set of all corruptions, and l is the loss function.

While, for instance, GANs are formulated as a simultaneous zero-sum game, the robust training example is a general-sum game that has a different solution concept known as Stackelberg equilibrium (von Stackelberg, 2010). In the robust training example, the maximizing player t, also sometimes known as the leader, takes an action t_0, knowing that the minimizing player, also known as the follower, plays the best response $x_0(t_0) := \operatorname{argmin}_x \mathbb{E}_{(u,v)} l(u + t_0, v, x)$.

1.2 First-Order Methods for Solving Minimax Problems

Perhaps the simplest method to solve minimax optimization problems is the (vanilla) gradient descent-ascent (GDA) that goes from each iteration to the next by going in the negative direction of the gradient oracle. These methods are applied when the bivariate objective function $f(x, y)$ is in class $\mathcal{C}^1$, i.e. the class of continuously differentiable functions. At each iteration (x_k, y_k), the next iterate is found via

$$\begin{aligned} x_{k+1} &= x_k - s_k \nabla_x f(x_k, y_k) \\ y_{k+1} &= y_k + s_k \nabla_y f(x_k, y_k), \end{aligned} \tag{1.5}$$

where $s_k > 0$ is the step-size at iteration k.

As simple and as computationally inexpensive as this method is, and contrary to its minimization counterpart, GDA fails to converge even for convex-concave minimax optimization problems. In fact, consider in problem (Equation 1.1) that $f(x, y) = x^T A y$ where $A \in \mathbb{R}^{n \times m}$ is full-rank. This problem has a Nash equilibrium at the origin. Applying GDA to this problem yields

$$\begin{aligned} x_{k+1} &= x_k - s_k A y_k \\ y_{k+1} &= y_k + s_k A^T x_k \end{aligned}$$

so that

$$\frac{\left\| \begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} \right\|^2}{\left\| \begin{bmatrix} x_k \\ y_k \end{bmatrix} \right\|^2} = 1 + s_k^2 \frac{\left\| \begin{bmatrix} A y_k \\ A^T x_k \end{bmatrix} \right\|^2}{\left\| \begin{bmatrix} x_k \\ y_k \end{bmatrix} \right\|^2} > 1,$$

where the last ineqluality follows since A is full-rank. Therefore, starting from any initial point $(x_k, y_k) \in \mathbb{R}^{n+m}$, GDA diverges away from the unique solution at the origin, even if the algorithm starts from any of the vertices of the $n+m-1$-dimensional standard simplex. This is, as stated before, in contrast to smooth convex minimization where gradient descent efficiently converges to the minimum.

A more sophisticated method, called Extra-Gradient Method (EGM), initially introduced by (Korpelevich, 1976), implies a GDA step to obtain a middle point (x'_{k+1}, y'_{k+1}), and then subsequently moves in the negative direction of the gradients evaluated at the middle point to reap the next iterate. In other words,

$$\begin{aligned} x'_{k+1} &= x_k - s_k \nabla_x f(x_k, y_k) \\ y'_{k+1} &= y_k + s_k \nabla_y f(x_k, y_k), \end{aligned} \quad \Rightarrow \quad \begin{aligned} x_{k+1} &= x_k - s_k \nabla_x f(x'_{k+1}, y'_{k+1}) \\ y_{k+1} &= y_k + s_k \nabla_y f(x'_{k+1}, y'_{k+1}), \end{aligned} \tag{1.6}$$

We will study, in Chapter 2, that a variant of this method, called damped EGM, converges for a structural class of nonconvex-nonconcave minimax optimization problems.

The other method, Optimistic Gradient Descent-Ascent (OGDA), shares some similarity with EGM,

$$\begin{aligned} x'_{k+1} &= x_k - s_k \nabla_x f(x_k, y_k) \\ y'_{k+1} &= y_k + s_k \nabla_y f(x_k, y_k), \end{aligned} \quad \Rightarrow \quad \begin{aligned} x_{k+1} &= x'_{k+1} - s_k \left(\nabla_x f(x_k, y_k) - \nabla_x f(x'_{k-1}, y'_{k-1}) \right) \\ y_{k+1} &= y'_{k+1} + s_k \left(\nabla_y f(x_k, y_k) - \nabla_y f(x'_{k-1}, y'_{k-1}) \right), \end{aligned} \tag{1.7}$$

The fourth method that can solve these minimax problems is the proximal point method (PPM) that was first introduced by (Martinet, 1970). PPM calculates the next iterate in the following manner

$$(x_{k+1}, y_{k+1}) = \operatorname*{argminimax}_{\substack{u \in C \\ v \in D}} \left\{ f(u, v) + \frac{1}{2s_k}\|u - x_k\|^2 - \frac{1}{2s_k}\|v - y_k\|^2 \right\} =: \mathrm{Prox}_{s_k f}(x_k, y_k). \tag{1.8}$$

This method is the most theoretically appealing of all the methods so far stated. If the step-size $s_k > 0$ is small enough, the subproblem is strongly-convex-strongly-concave[1]. Many problems are known to approximate PPM, such as EGM (Tseng, 1995; Nemirovski, 2004), and OGDA (Daskalakis and Panageas, 2018). Despite enjoying favorable theoretical properties, PPM is an implicit method (also known in the literature as a "conceptual" method). Each iteration requires solving a computationally nontrivial minimax problem, whence not an "implementable" method, at least not the exact version, in practice.

1.3 Literature

The convex-concave minimax optimization problem (Equation 1.1) has been well-studied in the literature. PPM (Equation 1.8) was first introduced by (Martinet, 1970) in the context of

[1] This means the mapping $x \mapsto f(x, y)$ is strongly convex for every $y \in D$, and $y \mapsto f(x, y)$ is strongly concave for every $x \in C$.

maximal monotone variational inequality (VI) with $s_k \equiv s > 0$, i.e. constant step-size[1] This exact PPM was extended by (Rockafellar, 1976) to inexact PPM with varying step-sizes s_k, i.e. when each iteration of (Equation 1.8) is replaced by $(x_{k+1}, y_{k+1}) \approx \mathrm{Prox}_{s_k f}(x_k, y_k)$, that is

$$\|(x_{k+1}, y_{k+1}) - \mathrm{Prox}_{s_k f}(x_k, y_k)\| \leq e_k, \qquad \sum_{k=1}^{\infty} e_k < \infty,$$

$$\|\,(x_{k+1}, y_{k+1}) - \mathrm{Prox}_{s_k f}(x_k, y_k)\| \leq d_k \|\,(x_{k+1}, y_{k+1}) - (x_k, y_k)\,\|, \qquad \sum_{k=1}^{\infty} d_k < \infty$$

Results in (Rockafellar, 1976) imply that if further (i) the solution to the problem (Equation 1.1) is unique, (ii) and that the inverse of the gradient operator (also known as oracle) $F(x, y) := \begin{bmatrix} \nabla_x f(x, y) \\ -\nabla_y f(x, y) \end{bmatrix}$ is single-valued around the unique solution, and (iii) that F is metrically sub-regular around the solution, then local linear convergence rate can be attained by PPM. Later on, (Tseng, 1995) showed that both PPM and EGM converge linearly to the solution of a variational inequality granted a projection-type error bound holds, i.e. one where the distance to the solution is bounded for all iterations. Nemirovski (Nemirovski, 2004), later, showed that the mirror prox algorithm, and therefore EGM, converges sublinearly $O\left(\frac{1}{\epsilon}\right)$ to the solution of the minimax problem when the objective is convex-concave.

There are quite a few other algorithms that have been proposed to solve convex-concave minimax optimization problems.

[1] An operator $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is monotone if for every $x, x' \in \mathbb{R}^n$ and every $y \in T(x), y' \in T(x)$, we have $\langle x - x', y - y' \rangle \geq 0$. This mapping is, further, maximally monotone if its graph gph $T := \{(x, y) \mid y \in T(x)\}$ is not properly contained in the graph of another monotone operator.

(Solodov and Svaiter, 1999a; Solodov and Svaiter, 1999b; Solodov and Svaiter, 2000; Solodov and Svaiter, 2001) furnished new versions of approximate PPM using an error criteria relative to that of (Rockafellar, 1976). In particular, for solving the maximally monotone inclusion problem $0 \in T(x)$, (Solodov and Svaiter, 1999a) introduced a hybrid approximate proximal extragradient method, using ϵ-enlargement T^ϵ of T. More precisely, let the ϵ-enlargement for $\epsilon > 0$ of T at $x \in \mathbb{R}^n$ be defined as

$$T^\epsilon(x) := \{v \in \mathbb{R}^n \mid \langle w - v, y - x\rangle \geq -\epsilon \text{ for all } y \in \mathbb{R}^n, w \in T(y)\},$$

with $T^0 := T$. Using this approximation of the maximally monotone T, define now for $x \in \mathbb{R}^n$, $s > 0$ and $\delta \in [0, 1)$, the δ-approximate solution to the proximal subproblem $0 \in T(.) + \frac{1}{s}(. - x)$ as the pair (x', v) if there exists an $\epsilon > 0$ such that $v \in T^\epsilon(x')$, $v + \frac{1}{s}(x' - x) = e(\epsilon) \neq 0$ with $\|e(\epsilon)\|^2 + 2s\epsilon \leq \delta^2\|x' - x\|^2$. Now define the hybrid approximate proximal extragradient iterates with step $s_k > 0$ as follows

1. Find the δ-approximate solution (x'_{k+1}, v'_{k+1}) to the subproblem

$$0 \in T(.) + \frac{1}{s_k}(. - x_k)$$

2. Define

$$x_{k+1} = x_k - s_k v'_{k+1}$$

(Solodov and Svaiter, 1999a) showed the above algorithm, called Hybrid Approximate Proximal Extragradient Method (HPEM) converges sublinearly to the solution of a maximally monotone inclusion problem.

(Monteiro and Svaiter, 2010) modified the termination criteria for HPEM to be of the following form: given $\epsilon > 0$, the algorithm terminates whenever it finds a tuple (x^*, ν^*, ϵ^*) such that

$$\nu^* \in T^{\epsilon}(y^*), \qquad \text{and} \quad \max\{\|\nu^*\|, \epsilon^*\} \leq \epsilon$$

More precisely, (Solodov and Svaiter, 1999a) showed the sequence $(x_k, y_k, x'_k, y'_k, \delta_k)$ generated via HPEM converges to a solution of monotone inclusion problem as $(x_k, y_k) \xrightarrow[O\left(\frac{1}{\sqrt{k}}\right)]{} (x^*, y^*)$ and $\delta_k \xrightarrow[O\left(\frac{1}{k}\right)]{} 0$.

Another algorithm in the literature to solve monotone variational inequalities is the Douglas-Rachford splitting method (DRSM). Motivated by the difficulty of calculating the proximal step $(I+s_kT)^{-1}$, an alternative is to find maximal montone operators A and B with $A+B = T$ whereas calculating $(I + s_kA)^{-1}$ and $(I + s_kB)^{-1}$ is easier. Initially motivated by problems in numerical linear algebra (Douglas and Rachford, 1956), DRSM was introduced in the optimization community by (Eckstein and Bertsekas, 1992).

Closely related to DRSM is the Primal-Dual Hybrid Gradient (PDHG) method introduced in (Chambolle and Pock, 2011), where they analyzed PDHG to solve nonsmooth convex-concave minimax optimization problems with bilinear interaction, i.e. problems of the form,

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m} f_1(x) + \langle A^T x, y\rangle - f_2(y),$$

The algorithm runs a GDA using only the interaction term to get to the "middle point", and then runs two proximal step from that middle point using f_1 and f_2 to get to the next iterate. Assuming simple structure for f_1 and f_2, these proximal steps are easy to compute. It was shown in (Chambolle and Pock, 2011) that PDHG converges sublinearly to the solution of the bilinear problem mentioned above. The close relationship between PDHG and DRSM was recently shown in (O'Connor and Vandenberghe, 2018) to be an equivalence relation under preconditioning. While PDHG and DRSM are known to converge linearly under further regularity conditions, and that these methods are not designed to be applied to generic minimax problems, the main difference they have with the algorithms that we cover in this Thesis like EGM and PPM is that PDHG and DRSM update the primal and dual sequentially whereas EGM and PPM simultaneously update the primal and the dual.

(Nesterov, 2005) introduced a smoothing scheme to solve minimization problems, which is also one of the most influential algorithms to solve convex-concave minimax problems with bilinear interaction term, i.e. functions in the class $f(x,y) = f_1(x) + x^T A y - f_2(y)$, where f_1, f_2 are continuous and convex. More precisely, motivated by the constrained minimization

$$\min_{x \in C} f(x) \tag{1.9}$$

of a continuous, but not necessarily differentiable, function, (Nesterov, 2005) proposed a smooth reformulation of (Equation 1.9) as

$$\min_{x \in C} \max_{y \in D} \hat{f}(x) + \langle Ax, y \rangle - \hat{g}(y) - \frac{\sigma}{2} \|y - y_0\|^2,$$

for some $y_0 \in \mathbb{R}^m$ with the structure of $D \subset \mathbb{R}^m$ was assumed to be "simple", and $\hat{g}(.)$ to be an affine function. He thus furnished a rate of $O\left(\frac{1}{\sqrt{\epsilon}}\right)$ to solve the outer minimization problem, thereby achieving, in this way, an overall rate of $O\left(\frac{1}{\epsilon}\right)$ if both $C \subset \mathbb{R}^n, D \subset \mathbb{R}^m$ are simple and the objective is affine in both x and y, with bilinear interaction term.

More recently, the advent of GANs (Goodfellow et al., 2014), and, subsequently, Wasserstein GANs (Arjovsky et al., 2017), minimax optimization problems have gained new attraction in the machine learning community. In one of the very first papers in this line of work, (Syrgkanis et al., 2015) proved the OGDA[1] converges faster to the solution of the convex-concave general normal form games. Later on, (Daskalakis et al., 2018) showed that OGDA exhibits last-

[1] Also known in the literature as Optimistic Mirror Descent-Ascent (OMDA)

iterate convergence to the solution of the specific class of bilinear minimax games, whence linear convergence, if the bilinear interaction matrix is full-rank.

(Mokhtari et al., 2020) furnished a unified analysis of different primal-dual algorithms to solve convex-concave minimax optimization problems. In particular, they showed that OGDA and EGM are approximations of PPM (the case of EGM was, indeed, shown earlier by (Nemirovski, 2004)), and obtained linear rates for smooth and strongly-convex-strongly-concave problems.

A different perspective taken in the literature with respect to minimax optimization problems, in particular, and the problem of (non-)monotone inclusion more generally, is analyzing various algorithms using dynamical systems perspective. (Su et al., 2016) derived a second order ordinary differential equation that behaves asymptotically the same as Nesterov's accelerated gradient descent (Nesterov, 1983) as the (constant) step-size s converges to 0, and thus, in this way, furnished an explanation why Nesterov's acceleration scheme speeds up convergence rates. One drawback of this approach is that various algorithms behave distinctly under different step-sizes, thereby making the explanation, under vanishing step-size s, restrictive. Recently, (Shi et al., 2022) has furnished a higher $O(s)$-resolution ODE explanation for primal-dual algorithms, which was further generalized to $O(s^r)$-resolution in (Lu, 2022).

However, most of the applications of minimax optimization in machine learning, such as GANs and reinforcement learning, in particular, involve a nonconvex-concave or even nonconvex-nonconcave objective, which is much more challenging than convex-concave minimax problems.

When the objective is nonconvex-concave, a basic technique to solve the minimax optimization problem (Equation 1.1) is to do a minimization on the outer function $\Psi(x) := \max_{y \in D} f(x, y)$. The function Ψ in this case is well-defined since the function $f(x, .)$ is always concave. Then one can use the recent developments in nonconvex minimization problems. (Rafique et al., 2022) considered a weakly-convex concave bivariate function and showed a convergence rate of $O\left(\frac{1}{\epsilon^6} \log\left(\frac{1}{\epsilon}\right)\right)$ to the approximate stationary point of the saddle function using a proximally-giuded stochastic mirror descent, and further showed this rate is improved to $O\left(\frac{\kappa}{\epsilon^2} \log\left(\frac{1}{\epsilon}\right)\right)$ when the objective is weakly-convex-strongly-concave with κ being the condition number of $f(x, .)$. (Lin et al., 2020a) used GDA with two different step-sizes, called two-timescale GDA, for solving a nonconvex-concave minimax optimization with one constraint on the maximization domain. They showed the efficient convergence of two-timescale GDA to a stationary point of $\Psi(.) = \max_{y \in D} f(., y)$ in $O\left(\frac{1}{\epsilon^6}\right)$ iterations, thereby, also improving the convergence rates of (Rafique et al., 2022) for nonconvex-strongly-concave to $O\left(\frac{\kappa}{\epsilon^2}\right)$. (Thekumparampil et al., 2019) furnished a better rate of $O\left(\frac{1}{\epsilon^3} \log^2\left(\frac{1}{\epsilon}\right)\right)$ earlier for this setting by applying inexact proximal point algorithms, which is, in fact, an implicit method. The Minimax-PPA algorithm proposed by (Lin et al., 2020b) achieves $O\left(\frac{1}{\epsilon^3} \log^2\left(\frac{1}{\epsilon}\right)\right)$ that matches that of (Thekumparampil et al., 2019). For completeness, we would like to mention that Minimax-PPA runs accelerated proximal point algorithm on the smooth nonconvex $f(., y)$ to find $\hat{x}$ followed by accelerated gradient ascent on $f(\hat{x}, .)$.

These techniques, however, are not applicable to nonconvex-nonconcave minimax optimization problems as calculating $\Psi(.)$ is no longer tractable. There are quite a few very recent research results that address this problem under a wide array of structural assumptions. Before reviewing the state-of-the-art, let us specify the issues of nonconvex-nonconcave minimax optimizaiton.

Cycling. Computational tractability and the scale of motivating applications, such as GANs, require one to largely consider first-order methods. One major challenge in nonconvex-nonconcave minimax optimizaiton using first order-methods is that all algorithms known do cycle, that is, the algorithm's trajectories converge to limit cycles that do not contain the equilibrium of the problem. This phenomenon, more precisely described in chapter 2, is currently the biggest challenge in nonconvex-nonconcave minimax literature. While many efforts have been made to describe this phenomenon (Lu, 2022; Grimmer et al., 2022b; Pethick et al., 2022), it is largely an unknown issue in this line of research.

Nonconvex Minimization does not translate well into nonconvex-nonconcave minimaxation. For one, the nonconvexity-nonconcavity, to the best of our knowledge, revolves around weak-convexity-weak-concavity in all literature works. This condition, while sufficient in minimization problems to obtain convergence (Davis and Drusvyatskiy, 2019), is generally insufficient in minimax optimization. An additional structure, such as Polyak-Łojasiewicz, nonnegative interaction dominance, negative comonotonicity, or weak minty variational inequality, are necessary to show convergence for various first-order methods. On the other hand, algorithms that solve nonconvex minimization are generally hard to get to converge in nonconvex minimax problems.

Lack of strong progress measures. While minimization problems have a clear progress measure of the value of the function, with which one can compare different optimum solutions, nonconvex-nonconcave minimax optimization problems lack such measures which makes it very hard to show the convergence of an algorithm to the solution. The most well-known progress measure, the saddle gap, does not decay monotonically, and, at times, can be infinite, even in bilinear minimax (Applegate et al., 2022). However, in smooth minimization, a move in the negative direction of the gradient always leads to progress.

In one of the earliest research efforts, (Liu et al., 2021) considered the class of weakly-convex-weakly-concave objectives that satisfy the Minty Variational Inequality (MVI) regularity condition at the solution. Leveraging the inexact PPM results in (Davis and Grimmer, 2019), they showed a $O\left(\epsilon^{-2}\right)$ convergence rate to an approximate solution of the MVI. Other research papers studying MVI and similar conditions include (Malitsky, 2020; Mertikopoulos et al., 2019; Song et al., 2020).

The issue, however, with this structural assumption is that it is relatively strong, and almost convex-concave like. In an attempt to address this issue, (Grimmer et al., 2022a) considered the class of nonconvex-nonconcave $f(x, y)$ with strong interaction[1] between the two agents in the zero-sum game. They showed that under such sufficient interaction, the damped PPM run on $f(x, y)$ is equivalent to running GDA on the envelope and achieving, in this way, linear convergence rate for such problems. Damped PPM, that is closely related to our algorithm

[1] This will be made precise in the second chapter as it is our assumption in pursuing the convergence behavior of EGM.

discussed in Chapter 2, (carefully) chooses a damping parameter $\lambda \in (0, 1]$ and thereby computes the next iterate as

$$(x_{k+1}, y_{k+1}) = \lambda \mathrm{Prox}_{sf}(x_k, y_k) + (1 - \lambda)(x_k, y_k). \tag{1.10}$$

Later on, or perhaps at the same time as (Grimmer et al., 2022a), (Diakonikolas et al., 2021) considered the more general class of nonconvex-nonconcave $f(x, y)$ that satisfy a weak MVI, i.e. there exists a point $z^* := (x^*, y^*)$ such that the gradient oracle of the (smooth) objective function $F(z) = \begin{bmatrix} \nabla_x f(z), \\ \\ -\nabla f(z) \end{bmatrix}$, where $z := (x, y)$, satisfies, for some $\sigma > 0$ "small enough"[1],

$$\langle F(z), z - z^* \rangle \geq -\frac{\sigma}{2} \|F(z)\|^2, \qquad \forall z \in \mathbb{R}^{n+m} \tag{1.11}$$

(Diakonikolas et al., 2021) showed that EG+, a damped version of EGM, converges sublinearly to the solution of such structured problem. To this day, this is the most general setting a result is known in the literature for nonconvex-nonconcave minimax problems using first-order methods. Other notable work in this line of research include (Lee and Kim, 2021; Pethick et al., 2022)

[1] Assuming that the gradient oracle is further globally Lipschitz with modulus $L > 0$, i.e. $\langle F(z') - F(z), z' - z \rangle \leq L\|z' - z\|$, for all $z, z' \in \mathbb{R}^{n+m}$, "small enough" here means $\sigma \in \big[0, \frac{1}{4L}\big)$.

CHAPTER 2

ON THE LINEAR CONVERGENCE OF EXTRA-GRADIENT METHODS FOR MINIMAX OPTIMIZATION

Materials in this chapter are published in (Hajizadeh et al., 2023)

2.1 Introduction

Recently, minimax optimization received renewed focus due to modern applications in machine learning, robust optimization, and reinforcement learning. The scale of these applications naturally leads to the use of first-order methods. However, the nonconvexities and nonconcavities present in these problems, prevents the application of typical Gradient Descent-Ascent, which is known to diverge even in bilinear problems. Recently, it was shown that PPM converges linearly for a family of nonconvex-nonconcave problems. In this chapter, we study the convergence of a damped version of EGM which avoids potentially costly proximal computations, only relying on gradient evaluation. We show that EGM converges linearly for smooth minimax optimization problem satisfying the same nonconvex-nonconcave condition needed by PPM.

Formally, we consider unconstrained minimiax optimization problem of interest in this chapter in the following form

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m} L(x,y) \ , \tag{2.1}$$

where L is β-smooth, twice-differentiable, ρ-weakly-convex-weakly-concave, and is positive interaction dominance.[1]

2.2 State-of-the-Art

The introduction of General Adversarial Networks (GANs) (Goodfellow et al., 2014) has shifted a ton of attention towards nonconvex-nonconcave minimax problems. Recently, Grimmer et. al (Grimmer et al., 2022a) studied PPM and its global convergence on a class of nonconvex-nonconcave minimax problems that enjoy sufficiently strong interaction between the two underlying agents. (Yang et al., 2020) furnished another example of global convergence of Alternating Gradient Descent-Ascent (AGDA) for the class of objective functions that satisfy the *two-sided Polyak-Lojasiewicz* inequality, known to be a weaker condition than strongly-convex-strongly-concave. (Jin et al., 2020) studied the various notions of optimality in nonconvex-nonconcave minimax problems. (Diakonikolas et al., 2021) furnished the proof for the ergodic convergence of a special class of our damped EGM, that is, one with a damping parameter of $\lambda = \frac{1}{2}$, when applied to a nonconvex-nonconcave minimax problems. Lee and Kim (Lee and Kim, 2021) furnished the convergence of the so-called *two-time-scale anchored extragradient method* (FEG) to a stationary point (Definition 2.4.1) of an objective function with a negatively ρ-comonotone oracle. This is also known in literature as $|\rho|$-cohypomonotonicity (see (Bauschke1 et al., 2021, Remark 2.5 (ii))). Each iteration of FEG iterates from a convex combination of the "current iterate" z_k and the "initial point" z_0, and moves along the direction

[1] These assumptions will be made precise in Section 2.3

Algorithm	Explicit Method	Setting	Constraints	Convergence Rate
AGDA (Yang et al., 2020)	✓	two-sided PŁ	✘	$O\left(\log\left(\frac{1}{\epsilon}\right)\right)$
PPM (Grimmer et al., 2022a)	✘	Positive Interaction Dominance	✘	$O\left(\log\left(\frac{1}{\epsilon}\right)\right)$
Damped EGM (Diakonikolas et al., 2021)	✓	weak MVI	✘	$O\left(\frac{1}{\epsilon}\right)$
CEG+ (Pethick et al., 2022)	✘	weak MVI	✓	$O\left(\frac{1}{\epsilon}\right)$
EGM Variant (Lee and Kim, 2021)	✓	Negatively Comontone	✘	$O\left(\frac{1}{\epsilon}\right)$
Damped EGM[**This chapter**]	✓	Positive Interaction Dominance	✘	$O\left(\log\left(\frac{1}{\epsilon}\right)\right)$

TABLE I: Comparison of the algorithms studied in recent papers on nonconvex-nonconcave minimax optimization. Any method that used a resolvent or proximal step is not considered an explicit method and may thus have an expensive cost per iteration.

of a linear combination of gradient information in its transition from the mid-point to the next iteration. This was shown to converge sublinearly to a stationary point of an objective function that admits negative comonotonicity.

The convergence rate of various implicit and explicit methods for solving nonconvex-nonconcave minimax problems, their convergence rate, are illustrated in Table Table I. We note that positive interaction dominance, a slight strengthening of negative comonotonicity, is, to the best of the authors' knowledge, the most general setting for which linear convergence is known in the literature.

2.3 Notations and Assumption

Throughout this section, we use x and y to denote the minimizing and maximizing variables, respectively. We use I to denote identity matrix of appropriate dimension. The symbols ∇,

∇^2, ∇_x, and ∇^2_{xx} are used to denote the gradient, Hessian, partial gradient, and partial Hessian of a function following the symbol. Let $U \subset \mathbb{R}^n \times \mathbb{R}^m$. We say a mapping $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is ξ-Lipschitz on U if for any pair of $z, z' \in U$, $\|L(z) - L(z')\| \leq \xi\|z - z'\|$. We are primarily interested in twice differentiable functions L. We say L is β-smooth on U if its gradient is β-Lipschitz in U, or equivalently, when L is twice continuously differentiable and its Hessian satisfies $\|\nabla^2 L(z)\| \leq \beta$ for all $z \in U$. We say that L is μ-strongly convex-strongly concave on U for $\mu > 0$ if for any $z = (x, y) \in U$, $\nabla^2_{xx} L(z) \succeq \mu I$ and $-\nabla^2_{yy} L(z) \succeq \mu I$. When $\mu = 0$, this is equivalent to convexity and concavity in x and y, respectively. However, our primary interest is in nonconvex-nonconcave objectives. We quantify the level of negative curvature in x and positive curvature in y as follows: We say L is ρ-weakly convex-weakly concave on U if for all $z = (x, y) \in U$,

$$\nabla^2_{xx} L(z) \succeq -\rho I, \qquad -\nabla^2_{yy} L(z) \succeq -\rho I \ .$$

Moreover, we denote the first-order oracle for the problem (Equation 2.1) by $F(z) = \begin{bmatrix} \nabla_x L(z) \\ -\nabla_y L(z) \end{bmatrix}$ for any $z \in \mathbb{R}^n \times \mathbb{R}^m$. For any ρ-weakly-convex-weakly-concave function L, we denote the prox operator with stepsize $0 < s \leq 1/\rho$ by

$$(x, y) = \mathrm{Prox}_{s.f}(u, v) := \underset{u \in \mathbb{R}^n, v \in \mathbb{R}^m}{\mathrm{argminimax}}\, f(u, v) + \frac{1}{2s}\|u - x\|^2 - \frac{1}{2s}\|v - y\|^2 \ .$$

For any $0 < s \leq 1/\rho$, we say a function L is $\alpha(s)$-interaction dominant in x if for all $z \in \mathbb{R}^{n+m}$,

$$\nabla^2_{xx}L(z) + \nabla^2_{xy}L(z)\left(s^{-1}I - \nabla^2_{yy}L(z)\right)^{-1}\nabla^2_{yx}L(z) \succeq \alpha(s)I \tag{2.2}$$

and $\alpha(s)$-interaction dominant in y if for any $z \in \mathbb{R}^{n+m}$

$$-\nabla^2_{yy}L(z) + \nabla^2_{yx}L(z)\left(s^{-1}I + \nabla^2_{xx}L(z)\right)^{-1}\nabla^2_{xy}L(z) \succeq \alpha(s)I\ . \tag{2.3}$$

Throughout our analysis of the Extragradient Method, we assume the following regularity conditions on the objective. The first two conditions (Lipschitz continuity, smoothness, and weak convexity-concavity) are relatively standard. The third regularity condition is positive interaction dominance (see Assumption 2.3.1) and is equivalent to the settings considered for the proximal point method in (Grimmer et al., 2022a) and the negative comonotone setting where accelerated, sublinear rates were recently derived by (Lee and Kim, 2021).

Assumption 2.3.1. *The objective function* $L\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *satisfies the following four conditions*

1. L *is continuously twice differentiable and* β*-smooth on* $\mathbb{R}^n \times \mathbb{R}^m$*.*
2. L *is* ρ*-weakly convex in* x *and* ρ*-weakly concave in* y *on* $\mathbb{R}^n \times \mathbb{R}^m$*.*
3. *For some* $s \in \left(0, \frac{1}{\rho}\right)$, L *satisfies the interaction dominance conditions* (Equation 2.2) *and* (Equation 2.3) *where* $\alpha > 0$ *is a positive function.*

The smoothing constant β, *weak-convexity-weak-concavity constant* ρ, *and a pair of values* s *and* α *satisfying interaction dominance are needed to select stepsize parameters where our theory applies.*

Positive Interaction-Dominance: Provided $0 < s \leq 1/\rho$, the second terms

$$\nabla^2_{xy}L(z)\left(s^{-1}I - \nabla^2_{yy}L(z)\right)^{-1}\nabla^2_{yx}L(z)$$
$$\nabla^2_{yx}L(z)\left(s^{-1}I + \nabla^2_{xx}L(z)\right)^{-1}\nabla^2_{xy}L(z)$$

in (Equation 2.2) and (Equation 2.3) are always positive semi-definite. We thus see that any convex-concave objective function is always nonnegative interaction dominant in both x and y. A ρ-weakly-convex-weakly-concave function is $\alpha(s)$-interaction dominant with $\alpha(s) \geq -\rho$. For this weakly-convex-weakly-concave function $L(x, y)$ to have nonnegative interaction dominance, L must have "large enough" interaction between x and y in the Hessian in order to "overcome" the effect of the smallest (negative) eigenvalue of partial Hessian $\nabla^2_{xx}L$ and the largest (positive) eigenvalue of partial Hessian $\nabla^2_{yy}L$.

2.4 Preliminaries

(Grimmer et al., 2022a) used a generalization of the Moreau envelope, called *saddle envelope*, introduced by (Attouch et al., 1986), to study the behavior of a certain class of nonconvex-nonconcave objective functions. More precisely, given any ρ-weakly-convex-weakly-concave objective function $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ with $L \in \mathcal{C}^2$, and any $s > 0$, the saddle envelope $\mathcal{L}_s$ is defined as

$$\mathcal{L}_s(x, y) := \min_{u \in \mathbb{R}^n} \max_{v \in \mathbb{R}^m} \left\{ L(u, v) + \frac{1}{2s}\|u - x\|^2 - \frac{1}{2s}\|v - y\|^2 \right\} . \tag{2.4}$$

Suppose $s < \frac{1}{\rho}$, then it is easy to see that $M(u, v) := L(u, v) + \frac{1}{2s}\|u - x\|^2 - \frac{1}{2s}\|v - y\|^2$ is $\left(\frac{1}{s} - \rho\right)$-strongly-convex-strongly-concave so that the saddle envelope (Equation 2.4) is well-defined. Corollary 2.2 in (Grimmer et al., 2022a) implies that to find *any* stationary point of L, one only needs to find those of $\mathcal{L}_s$. Imposing nonnegative interaction dominance paves the way for proving more powerful properties of the saddle envelope.

If an objective function is nonnegative interaction dominant in both x and y for some $s \in \left(0, \frac{1}{\rho}\right)$, its saddle envelope $\mathcal{L}_s(x, y)$ would then be convex-concave ((Grimmer et al., 2022a), Proposition 2.6). If for some $s \in \left(0, \frac{1}{\rho}\right)$ the objective function is further positive interaction dominance, then the saddle envelope $\mathcal{L}_s(x, y)$ is strongly-convex-strongly-concave.

These interaction dominance conditions can be equivalently characterized in terms of the convexity and concavity of the saddle envelop (Equation 2.4) and in terms of the monotonicity of the saddle gradient $F(z) = \begin{bmatrix} \nabla_x L(z) \\ -\nabla_y L(z) \end{bmatrix}$. This is formalized in Proposition 2.4.1.

Recently, (Lee and Kim, 2021) presented algorithms with sublinear rate for nonconvex-nonconcave minimax optimization problems with ρ-comonotone gradient oracle for some negative ρ. A set-valued mapping $T : \mathbb{R}^{n+m} \rightrightarrows \mathbb{R}^{n+m}$ is said to be **ρ-monotone** on $\mathbb{R}^{n+m}$ if for every $\bar{z} \in \mathbb{R}^{n+m}$ there exists a neighborhood V of $\bar{z}$ such that

$$\langle w - w', z - z' \rangle \geq \rho \|z - z'\|^2 \qquad \text{for all } z, z' \in V \text{and all } w \in T(z),\ w' \in T\left(z'\right) \ .$$

A set-valued mapping[1] $T : \mathbb{R}^{n+m} \rightrightarrows \mathbb{R}^{n+m}$ is said to be **ρ-comonotone** if for every $\bar{z} \in \mathbb{R}^{n+m}$ there exists a neighborhood V of $\bar{z}$ such that

$$\langle w - w', z - z' \rangle \geq \rho \|w - w'\|^2 \qquad \text{for all } z, z' \in V \text{and all } w \in T(z),\ w' \in T\left(z'\right) \ .$$

The following proposition, in particular, establishes the equivalence of our assumptions and the negative comonotone setting recently considered by (Lee and Kim, 2021).

Proposition 2.4.1. *Let $L(x, y)$ be a twice-differentiable, ρ-weakly-convex-weakly-concave objective function, and $s \in \left(0, \frac{1}{\rho}\right)$. Then the following statements are equivalent:*

(i) $L(x, y)$ *is* $\alpha(s) \geq 0$*-interaction dominance in both* x *and* y*,*

(ii) The saddle envelope $\mathcal{L}_s(x, y)$ *of* L *is convex-concave,*

(iii) The oracle $F(x, y) = \begin{bmatrix} \nabla_x L(x, y) \\ \nabla_y L(x, y) \end{bmatrix}$ *of* $L(x, y)$ *is* $-s$*-comonotone.*

[1] The use of set-valued operators is typical here, allowing these definitions to be applied to cases where the objective function is not smooth whence only admitting subdifferentials. However, such nonsmooth optimization is beyond the scope of this chapter. We will discuss Nonsmooth Minimax optimization in Chapter 3.

Proof. $(\mathrm{i}) \iff (\mathrm{ii})$: Observe that the conditions (Equation 2.2) and (Equation 2.3) hold with nonnegative α exactly when

$$\begin{aligned}\nabla^2_{xx}L(z) + \nabla^2_{xy}L(z)\left(s^{-1}I - \nabla^2_{yy}L(z)\right)^{-1}\nabla^2_{yx}L(z) \succeq 0\\ -\nabla^2_{yy}L(z) + \nabla^2_{yx}L(z)\left(s^{-1}I + \nabla^2_{xx}L(z)\right)^{-1}\nabla^2_{xy}L(z) \succeq 0\ .\end{aligned}$$

Adding an identity matrix $\frac{1}{s}I$ above and inverting the resulting positive definite matrix yields the following equivalent characterization

$$\begin{aligned}s^{-1}I - s^{-2}\left(s^{-1}I + \nabla^2_{xx}L(z) + \nabla^2_{xy}L(z)\left(s^{-1}I - \nabla^2_{yy}L(z)\right)^{-1}\nabla^2_{yx}L(z)\right)^{-1} \succeq 0\\ s^{-1}I - s^{-2}\left(s^{-1}I - \nabla^2_{yy}L(z) + \nabla^2_{yx}L(z)\left(s^{-1}I + \nabla^2_{xx}L(z)\right)^{-1}\nabla^2_{xy}L(z)\right)^{-1} \preceq 0.\end{aligned}$$

By Lemma 3 in (Grimmer et al., 2022a), these are exactly the xx and yy components of the saddle envelope's Hessian, i.e. $\nabla^2_{xx}\mathcal{L}_s(z)$ and $\nabla^2_{yy}\mathcal{L}_s(z)$, whence the assertion is proved.

$(\mathrm{ii}) \iff (\mathrm{iii})$: In (Lee and Kim, 2021, Appendix A.1), Lee and Kim showed that the saddle envelope's gradient mapping $F_s(.) = \begin{bmatrix}\nabla_x\mathcal{L}_s(.)\\ -\nabla_y\mathcal{L}_s(.)\end{bmatrix}$ is monotone if and only if F is $-s$-comonotone. Recalling that a gradient mapping $(\nabla_x L, -\nabla_y L)$ is monotone if and only if the associated function is convex-concave, the proof is complete. □

Let us now state the definition of a stationary point of a bifunction.

Definition 2.4.1. *A point* $(x^*, y^*) \in \mathbb{R}^{n+m}$ *is a stationary point of an objective function* $L(x, y)$ *if*

$$\nabla_x L(x^*, y^*) = 0, \quad \textit{and} \quad \nabla_y L(x^*, y^*) = 0\ .$$

We now provide a result that is all but stated in (Grimmer et al., 2022a):

Lemma 2.4.1. *A ρ-weakly-convex-weakly-concave objective function L that is positive interaction dominance in both x and y has a unique stationary point.*

Proof. By hypothesis and Proposition 2.6 in (Grimmer et al., 2022a), the saddle envelope is strongly-convex-strongly-concave whence adopting a unique saddle point (x^*, y^*) which is also its unique stationary point. By Corollary 2.2 in (Grimmer et al., 2022a), then (x^*, y^*) is the unique stationary point of L; because otherwise, if L has any other stationary point $(\tilde{x}^*, \tilde{y}^*) \neq (x^*, y^*)$, it would clearly contradict Corollary 2.2 of (Grimmer et al., 2022a). □

2.5 Damped EGM

Damped PPM, as the name suggests, introduces a damping parameter $\lambda \in (0, 1]$ in each iteration of PPM. The iteration update is

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = (1-\lambda) \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \lambda \operatorname*{Prox}_{sf} \left(\begin{bmatrix} x_k \\ y_k \end{bmatrix} \right) . \tag{2.5}$$

The damping, illustrated in (Equation 2.5), decreases the size of the proximal step in each iteration. The inclusion of $\lambda = 1$ allows taking a full proximal step in each iteration.

In this section, we introduce the damped EGM and show that it is approximating the damped PPM introduced in (Grimmer et al., 2022a). We first recall that EGM is an approximation of PPM (Nemirovski, 2004). Next, we will extend this result and show that the damped EGM is an approximation to damped PPM (Equation 2.5). Notice that damped EGM does not need to solve an implicit step, thus the update is computationally cheaper than that of damped PPM (Equation 2.5). The damped EGM is presented in Algorithm 1.

Algorithm 1 Damped EGM

```
    **Input**: $z_0 := (x_0, y_0)$, step-size $s > 0$, damping parameter $\lambda \in (0, 1]$, and tolerance $\epsilon > 0$

k = 0
**while** $\left\| \begin{bmatrix} \nabla_x L(x_k, y_k) \\ -\nabla_y L(x_k, y_k) \end{bmatrix} \right\| \geq \epsilon$ **do**
    Find the mid-point: $\begin{bmatrix} x'_{k+1} \\ y'_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - s \begin{bmatrix} \nabla_x L(x_k, y_k) \\ -\nabla_y L(x_k, y_k) \end{bmatrix}$,
    Find the next-iterate: $\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \lambda s \begin{bmatrix} \nabla_x L\left(x'_{k+1}, y'_{k+1}\right) \\ -\nabla_y L\left(x'_{k+1}, y'_{k+1}\right) \end{bmatrix}$
    $k \leftarrow k + 1$
**end while**
```

Damped EGM is a generalization of traditional EGM, where the step-size of the two steps can be chosen differently. This difference in steps comes from the damping parameter λ that controls the length of the second step. Intuitively, it is natural to think that the lack of convexity and concavity would require an algorithm to take a type of “cautiously aggressive” steps to avoid divergence and cycling that are common phenomena in nonconvex-nonconcave minimax optimization (Grimmer et al., 2022b). The following Proposition establishes that the damped EGM in Algorithm 1 approximates damped PPM.

Proposition 2.5.1. *Damped EGM update in Algorithm 1, when applied to the* ρ*-weakly-convex-weakly-concave objective function* $L(x, y)$*, is an approximation to the update for damped PPM (Equation 2.5).*

Proof. We write the Taylor expansion of the update for x in Algorithm 1, which gives us,

$$\begin{aligned} x_{k+1} &= x_k - \lambda s \nabla_x L\left(x_k - s\nabla_x L(z_k), y_k + s\nabla_y L(z_k)\right) \\ &= x_k - \lambda s \left[\nabla_x L(z_k) - s\nabla^2_{xx} L(z_k)\nabla_x L(z_k) + s\nabla^2_{xy} L(z_k)\nabla_y L(z_k) + o(s)\right] \\ &= x_k - \lambda s \nabla_x L(z_k) + \lambda s^2 \nabla^2_{xx} L(z_k)\nabla_x L(z_k) - \lambda s^2 \nabla^2_{xx} L(z_k)\nabla_y L(z_k) + o\left(s^2\right) . \end{aligned} \tag{2.6}$$

Similarly, one can find the update of y as

$$y_{k+1} = y_k + \lambda s \nabla_y L(z_k) - \lambda s^2 \nabla^2_{yx} L(z_k)\nabla_x L(z_k) + \lambda s^2 \nabla^2_{yy} L(z_k)\nabla_y L(z_k) + o\left(s^2\right) . \tag{2.7}$$

Let now $z_k^+ := \text{Prox}_{s,L}(z_k)$ be one proximal step of size from the current iterate z_k. We then have,

$$\begin{aligned} x_k^+ &= x_k - s\nabla_x L\left(x_k - s\nabla_x L(z_k^+), y_k + s\nabla_y L(z_k^+)\right) \\ &= x_k - s\left[\nabla_x L(z_k) - s\nabla^2_{xx} L(z_k)\nabla_x L(z_k^+) + s\nabla^2_{xy} L(z_k)\nabla_y L(z_k^+) + o(s)\right] \\ &= x_k - s\nabla_x L(z_k) + s^2\nabla^2_{xx} L(z_k)\nabla_x L(z_k) - s^2\nabla^2_{xy} L(z_k)\nabla_y L(z_k) + o(s^2) . \end{aligned} \tag{2.8}$$

where the second equality follows from the local Lipschitzness of partial gradients.

Similarly,

$$y_k^+ = y_k + s\nabla_y L(z_k) - s^2\nabla^2_{yx}L(z_k)\nabla_x L(z_k) + s^2\nabla^2_{yy}L(z_k)\nabla_y L(z_k) + o(s^2)\ . \tag{2.9}$$

We know from (Equation 2.5) that the update iterate of a damped PPM with constant λ is given by

$$\tilde{z}_{k+1} = (1-\lambda)z_k + \lambda z_k^+\ . \tag{2.10}$$

Combining (Equation 2.8) and (Equation 2.9) with (Equation 2.10) and comparing with (Equation 2.6) and (Equation 2.7), one can observe that

$$\|\tilde{z}_{k+1} - z_{k+1}\| = o\left(s^2\right)\ .$$

This concludes the proof of the claim that the damped EGM update as defined in Algorithm 1 is an approximation to that of damped PPM. □

In the next section, we show that if $L(x, y)$ is interaction dominance, then the damped EGM with proper step-size converges linearly to a stationary point. Furthermore, we show that damped EGM may diverge without interaction dominance, which showcases the tightness of using interaction dominance to characterize the convergence of damped EGM.

2.6 Convergence Result

First, let us state our main convergence result:

Theorem 1. *Suppose the objective function* $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *in Problem (Equation 2.1) satisfies Assumption 2.3.1, and let* $z^* := (x^*, y^*)$ *be its saddle point. Choose the parameters* s *and* λ *such that they satisfy*

$$\frac{2}{s^3\left(\frac{1}{s\alpha(s)}+1\right)} > \beta^3, \qquad \lambda < \min\left\{1, \left(\frac{1}{s\rho}-1\right)^2\right\}\left[\frac{2}{\frac{1}{s\alpha(s)}+1} - s^3\beta^3\right] \tag{2.11}$$

The damped EGM with step-size $s \in \left(0, \frac{1}{\rho}\right)$ *applied to the problem* $\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m} L(x,y)$ *with the constant* $\lambda \in [0,1]$ *linearly converges to the unique stationary point of* L. *More precisely, for any iterate* (x_k, y_k) *and starting point* (x_0, y_0) *one has*

$$\left\|\begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix}\right\| \le \left(1 - \frac{2\lambda}{\frac{1}{s\alpha(s)}+1} + \frac{\lambda^2}{\min\left\{1, \left(\frac{1}{s\rho}-1\right)^2\right\}} + \lambda s^3\beta^3\right)^k \left\|\begin{bmatrix} x_0 - x^* \\ y_0 - y^* \end{bmatrix}\right\|.$$

We now state some of remarks to better understand the statement of the theorem, simplify the conditions under which the statements hold, and observe what the Theorem translates into when considering special cases,

Remark 1. *For any* μ*-strongly-concave-strongly-concave, Theorem 1 has a linear convergence rate of* $O\left(\frac{1}{s\lambda\mu}\log\left(\frac{1}{\epsilon}\right)\right)$. *This is evident by plugging* $\alpha = \mu$ *and* $\rho = -\mu$ *in the convergence rate in the statement of the Theorem. Assuming that* $s = O\left(\beta^{-1}\sqrt{\frac{\mu}{\beta}}\right)$, *which results in a step-size smaller than* $\frac{1}{\beta}$ *by a factor of* $\sqrt{\frac{\mu}{\beta}}$, *we obtain a convergence rate of* $O\left(\sqrt{\frac{\beta^3}{\lambda^2\mu^3}}\log\left(\frac{1}{\epsilon}\right)\right)$ *which is a reasonable linear rate. The best rate known for EGM in the convex-concave case* *is* $O\left(\frac{\beta}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$ *(see for example (Gidel et al., 2020, Theorem 1), (Mokhtari et al., 2020, Theorem 7), (Tseng, 1995, Lemma 3.1), or (Alves et al., 2016, Proposition 2.2))*

Remark 2. *The selection of the pair of parameters* λ *and* s *satisfying (Equation 2.11) given* ρ *and* β*, and* α *as a function of* s *is not difficult. One observes that, plugging* $s = \frac{t}{\beta}$ *in the condition on* s *on the left-hand side of (Equation 2.11), one would get an inequality in* t *and solve for* t*. In many cases, e.g. quadratic problems, this inequality entails solely a polynomial and is trivial to solve. This would enlighten one on how smaller than* β *should the step-size* $s > 0$ *be taken. We would like to point out that the possible values for* s *usually involve an interval as opposed to arbitrarily small values. We also note that the damping introduced in our method is sometimes necessary for the convergence of EGM.*

For instance, consider the problem

$$\min_x \max_y L(x,y) = f(x) + \bar{A}xy - f(y), \qquad \textit{with } f(x) = (x-1)(x+1)(x-3)(x+3) \tag{2.12}$$

that is a nonconvex-nonconcave problem. Let $\bar{A} = 100$*. A simple examination of our conditions in the Theorem as described above implies that any* $s \in (0.00245, 0.00651)$ *with a damping factor* $\lambda \in (0, 0.06)$ *would guarantee convergence. Choosing* $s^* = 200$ *and* $\lambda^* = 0.01$ *we observe convergence for any starting point the box* $[-4, 4] \times [-4, 4]$ *as in Fig. 1a. Choosing, instead,* $\lambda^* = 1$*, whence recovering undamped EGM, results in cycling as shown in Fig. 1b.*

Remark 3. *Let us see through the inequality on the left-hand side of (Equation 2.11) more explicitly under some assumptions. Suppose we are given a problem that is nonnegative interaction dominant and we further restrict* s *to satisfy*

$$\frac{1}{s} \geq \frac{\max\left\{\|\nabla^2_{xx}L\|^2, \|\nabla^2_{xx}L\|^2\right\}}{\rho} > \frac{\rho^2}{\rho} = \rho \ . \tag{2.13}$$

On the other hand, we have

$$\begin{aligned} \left(s^{-1}I - \nabla^2_{yy}L(z)\right)^{-1} &\succeq \left(s^{-1} + \max\left\{\|\nabla^2_{xx}L\|, \|\nabla^2_{xx}L\|\right\}\right)^{-1} I \\ \left(s^{-1}I + \nabla^2_{xx}L(z)\right)^{-1} &\succeq \left(s^{-1} + \max\left\{\|\nabla^2_{xx}L\|, \|\nabla^2_{xx}L\|\right\}\right)^{-1} I \ . \end{aligned}$$

so that α *can be lower-bounded as*

$$\alpha \geq -\rho + \frac{s.\lambda_{\min}\left(\nabla^2_{xy}L(z)\nabla^2_{yx}L(z)\right)}{1 + s.\max\{\|\nabla^2_{xx}L\|, \|\nabla^2_{xx}L\|\}} \ . \tag{2.14}$$

Clearly, s *can not be chosen arbitrarily small as that would mar interaction dominance. From (Equation 2.14), a sufficient upper bound on* s *to preserve nonnegative interaction dominance is*

$$s \geq \frac{\rho}{\lambda_{\min}\left(\nabla^2_{xy}L(z)\nabla^2_{yx}L(z)\right) - \rho.\max\{\|\nabla^2_{xx}L\|, \|\nabla^2_{xx}L\|\}} . \tag{2.15}$$

The lower bound (Equation 2.15), as mentioned, preserves nonnegative interaction dominance. It further illustrates what the "sufficiently large" interaction requirement means. To further illustrate the explicit restrictions on s*, in light of the inequality on the left-hand side of (Equation 2.11), let us further suppose* $s < \frac{1}{\beta}$[1]*. This further assumption and (Equation 2.14) imply*

$$\alpha(s) \geq -\rho + \frac{1 + s.\max\left\{\|\nabla^2_{xx}L\|, \|\nabla^2_{xx}L\|\right\}}{s.(\beta + \max\{\|\nabla^2_{xx}L\|, \|\nabla^2_{xx}L\|\})} \cdot \rho = \frac{1 - s \cdot \beta}{s \cdot (\beta + \max\{\|\nabla^2_{xx}L\|, \|\nabla^2_{xx}L\|\})} \cdot \rho \ . \qquad (2.16)$$

Combining (Equation 2.13) and (Equation 2.16) with the inequality on the left-hand side of (Equation 2.11) one would get

$$\frac{2}{s^3\left(\frac{1}{s\alpha(s)} + 1\right)} \geq \frac{2^{\frac{\max\{\|\nabla^2_{xx}L\|^6, \|\nabla^2_{xx}L\|^6\}}{\rho^3}}}{1 + \frac{(\beta + \max\{\|\nabla^2_{xx}L\|, \|\nabla^2_{xx}L\|\})}{\rho(1 + s\cdot\beta)}} \geq \frac{2^{\frac{\max\{\|\nabla^2_{xx}L\|^6, \|\nabla^2_{xx}L\|^6\}}{\rho^3}}}{1 + \frac{1}{s\cdot\rho}} > \beta^3.$$

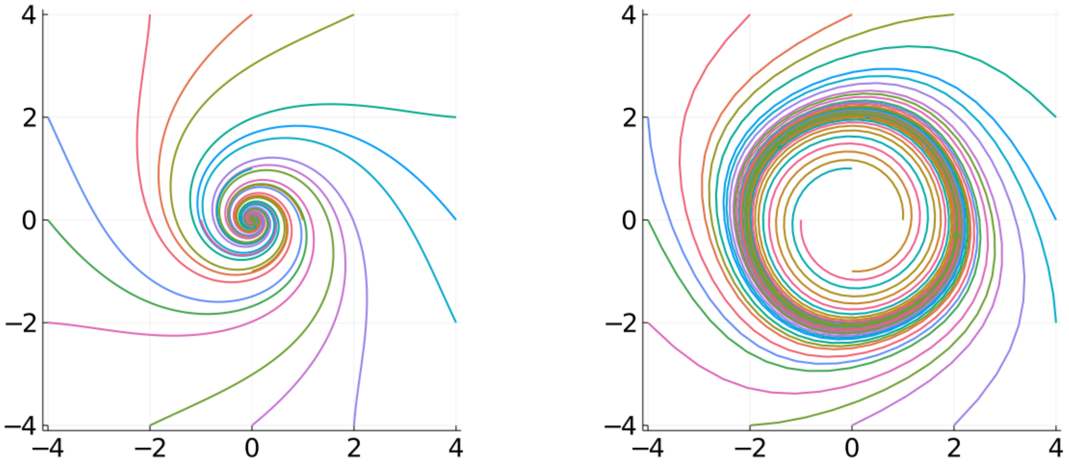
Therefore, the lower bound

$$s > \frac{1}{\rho\left[2\left(\frac{\max\{\|\nabla^2_{xx}L\|^2, \|\nabla^2_{yy}L\|^2\}}{\rho\beta}\right)^3 - 1\right]}. \qquad (2.17)$$

is a sufficient to guarantee the inequality condition on the left-hand side of (Equation 2.11) is satisfied.

The reader notes all at once that (Equation 2.15) and (Equation 2.17) are two ***explicit*** *lower bounds on* s *illustrating how small could one select the step-size while ensuring the conditions*

[1] *This condition is consistent with the classical smooth optimization literature choosing a step-size smaller than the reciprocal of the Lipschitz constant of the oracle.*



(a) Convergence of damped EGM

(b) Cycling of vanilla EGM

Figure 1: The convergence of damped EGM to the unique stationary point of the nonconvex-nonconcave problem (Equation 2.12) with $\bar{A} = 100$ from any starting point within the box $[-4, 4] \times [-4, 4]$ and the cycling of EGM.

of the Theorem, nonnegative interaction dominance and the inequality condition on the left-hand side of (Equation 2.11) are satisfied. We have been conservative to preserve the implicit condition in the main theorem to preserve generality as much as is achievable.

Remark 4. *One observes that damped EGM has a slower convergence rate* $1 - \frac{2\lambda}{\frac{1}{s\cdot\alpha(s)}+1} + \frac{\lambda^2}{\min\left\{1,\left(\frac{1}{s\rho}-1\right)^2\right\}}+\lambda s^3\beta^3$ *than damped PPM introduced in (Grimmer et al., 2022a) which converges at a rate of* $1 - \frac{2\lambda}{\frac{1}{s\alpha(s)}+1} + \frac{\lambda^2}{\min\left\{1,\left(\frac{1}{s\rho}-1\right)^2\right\}}$. *This difference aligns with the perspective of EGM approximating the proximal step via two cheaper gradient descent-ascent steps with accuracy depending on the smoothness of the objective function* β.

Remark 5. *Our main result also recovers the setting and results of (Zhang et al., 2022). In particular, Theorem* 5.14 *in (Zhang et al., 2022) asserts that one does not lose convergence by shrinking the damping parameter. This fact follows as well from our theorem, which additionally quantifies the rate of convergence rate one would see as the damping parameter shrinks.*

2.7 Proof of Theorem 1

Having noticed the remarks and ramifications of our main theorem, we now furnish a proof for the theorem.

Proof of Theorem 1. Given any iterate z_k, let us first find the exact upper bound to $\|\tilde{z}_{k+1} - z_{k+1}\|$ where, as noted before, $\tilde{z}_{k+1}$ is the update of damped PPM and z_{k+1} is the update of the damped EGM. One notes that

$$z_{k+1} = z_k - \lambda s F\left(z_k - sF(z_k)\right) \ ,$$

and

$$\tilde{z}_{k+1} = (1-\lambda)z_k + \lambda z_k^+ = (1-\lambda)z_k + \lambda\left[z_k - sF\left(z_k^+\right)\right] = z - \lambda s F\left(z_k^+\right) \ .$$

One also would observe that $z_k^+ = z_k - sF\left(z_k^+\right)$, whence $(I + sF)\left(z_k^+\right) = z_k$. We need the following lemma in order for further proceeding with the proof.

Lemma 2.7.1. *The operator* $I + sF : \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$ *for any* $0 < s < \frac{1}{\rho}$ *is invertible.*

Proof. Since L is β-smooth, the operator $(\mathrm{I}+s\mathsf{F})(.)$ is continuous. On the other hand, at any $z \in \mathbb{R}^{n+m}$ we have

$$
\begin{aligned}
|\mathrm{I}+s\nabla \mathsf{F}(z)| &= \left|\begin{bmatrix} \mathrm{I}+s\nabla^2_{xx}\mathsf{L}(z) & \nabla^2_{xy}\mathsf{L}(z) \\ -\nabla^2_{xy}\mathsf{L}(z)^\mathsf{T} & \mathrm{I}-s\nabla^2_{yy}\mathsf{L}(z) \end{bmatrix}\right| \\
&= \left|\mathrm{I}+s\nabla^2_{xx}\mathsf{L}(z)\right| . \left|\mathrm{I}-s\nabla^2_{yy}\mathsf{L}(z)+\nabla^2_{xy}\mathsf{L}(z)^\mathsf{T}(\mathrm{I}+s\nabla^2_{xx}\mathsf{L}(z))^{-1}\nabla^2_{xy}\mathsf{L}(z)\right| \ ,
\end{aligned}
$$

where the second equality follows from Schur complement. Moreover, by hypothesis, $\mathrm{I}+s\nabla^2_{xx}\mathsf{L}(z) \succ 0$, $\mathrm{I}-s\nabla^2_{yy}\mathsf{L}(z) \succ 0$, and

$$
\nabla^2_{xy}\mathsf{L}(z)^\mathsf{T}(\mathrm{I}+s\nabla^2_{xx}\mathsf{L}(z))^{-1}\nabla^2_{xy}\mathsf{L}(z) \succeq 0 \ ,
$$

whence the Jacobian has always nonzero determinant, i.e. $|\mathrm{I}+s\nabla \mathsf{F}(z)| \neq 0$. Therefore, by inverse function Theorem, the operator $(\mathrm{I}+s\mathsf{F})(.)$ is invertible with a continuously differentiable inverse. □

One observes that the proximal step in (Equation 2.10) is equivalent to that of (Grimmer et al., 2022a) with $\eta = \frac{1}{s}$. For any z, the inverse $(\mathrm{I}+s\mathsf{F})^{-1}(z)$ of the operator $\mathrm{I}+s\mathsf{F}$ is given by

$$
(\mathrm{I}+s\mathsf{F})^{-1}(z) = z - s\mathsf{F}(z) + s^2\nabla\mathsf{F}(z)\mathsf{F}(z) + o\left(s^2\right) \ . \tag{2.18}
$$

For details, one may refer to Appendix B of (Lu, 2022). Therefore, we can write

$$\begin{aligned}
\|\tilde{z}_{k+1} - z_{k+1}\| &= \lambda s \|F\left((I + sF)^{-1}(z_k)\right) - F\left(z_k - sF(z_k)\right)\| \\
&\leq \lambda s \beta \left\| \frac{1}{1!} \int_0^s \frac{\partial^2}{\partial \tau^2} (I + \tau F)^{-1}(z_k) \bigg|_{\tau = t} (s - t)\ dt \right\| \\
&\leq \lambda s \beta \int_0^s \left\| \frac{\partial^2}{\partial \tau^2} (I + \tau F)^{-1}(z_k) \bigg|_{\tau = t} \right\| (s - t)\ dt\ , \qquad (2.19)
\end{aligned}$$

where β-smoothness of L, Taylor expansion with the integral form of the remainder of the inverse operator $(I + sF)^{-1}(z)$, and Cauchy-Schwartz inequality are invoked.

We now turn to evaluate $\frac{\partial^2}{\partial \tau^2}(I + \tau F)^{-1}(z_k)\Big|_{\tau = t}$. For that matter, first note that by Appendix B in (Lu, 2022) and by Lemma 2.7.1, we have for any $t < \frac{1}{\rho}$

$$(I + tF)^{-1}(z) = z - tF(z) + t^2 \nabla F(z) F(z) + t^3 \left(-(\nabla F(z))^2 F(z) - \frac{1}{2} \nabla^2 F(z) \otimes_2 F(z) \right) + o\left(t^3\right).$$

where $\otimes_2$ refers to the 2-times tensor product of a 2-dimensional tensor with a vector. Now for any z, let $h_z : \mathbb{R} \to \mathbb{R}^{n+m}$ be the mapping $h_z : t \mapsto (I + tF)^{-1}(z)$. By Lemma 2.7.1, for any $t < \frac{1}{\rho}$ the derivative,

$$\begin{aligned}
h_z'(t) &= \frac{\partial}{\partial t} (I + tF)^{-1}(z) \\
&= -F(z) + t(2 \nabla F(z) F(z)) + t^2 \left(-3(\nabla F(z))^2 F(z) - \frac{3}{2} \nabla^2 F(z) \otimes_2 F(z) \right) + o\left(t^2\right)
\end{aligned}$$

is continuous. Invoking, in addition, the mean value Theorem, for any $t < \frac{1}{\rho}$, one would have a well-defined mapping $f : (0, t] \to (0, t]$, $f : t \mapsto f(t) \in (0, t)$ such that

$$h_z^{'}(t) = -F(z) + 2f(t)\nabla F(z)F(z) \ .$$

This mapping f is continuous in $(0, \frac{1}{\rho})$, because otherwise there would exist an $\epsilon > 0$, $t_0 \in (0, \frac{1}{\rho})$ such that for any $\delta > 0$,

$$\inf_{t\in\mathbb{B}(t_0,\delta)\backslash\{t_0\}} |h_z^{'}(t) - h_z^{'}(t_0)| = \inf_{t\in\mathbb{B}(t_0,\delta)\backslash\{t_0\}} |2\nabla F(z)F(z)| \cdot |f(t) - f(t_0)| = |2\nabla F(z)F(z)| \cdot \epsilon \ ,$$

which contradicts the continuity of $h_z^{'}(t)$.

Hence, for any $t < 1/\rho$, there exists a sequence $\delta_n \downarrow 0$ such that

$$h_z^{''}(t) = \limsup_{n\to\infty} \frac{h_z^{'}(t+\delta_n) - h_z^{'}(t)}{\delta_n} = 2\nabla F(z)F(z) \limsup_{n\to\infty} \frac{f(t+\delta_n) - f(t)}{\delta_n} \ ,$$

with $\limsup_{n\to\infty} \frac{f(t+\delta_n) - h_z^{'}(t)}{\delta_n} \leq f(t) \leq 1$ by construction. Therefore, we can bound (Equation 2.19) as follows

$$\|\tilde{z}_{k+1} - z_{k+1}\| \leq \lambda s^3 \beta \|\nabla F(z_k)\|.\|F(z_k)\| \leq \lambda s^3 \beta^3 \|z_k - z^*\|$$

by using $F(z^*) = 0$.

Let $c := 1 - \frac{2\lambda}{\frac{1}{s\cdot\alpha(s)}+1} + \frac{\lambda^2}{\min\left\{1,\left(\frac{1}{s\rho}-1\right)^2\right\}} \in (0,1)$ be the shrinking constant of the distance to the unique stationary point of L on one iteration of damped PPM as in (Grimmer et al., 2022a). Given the upper bound for $\|\tilde{z}_{k+1} - z_{k+1}\|$ just furnished, one has

$$\begin{aligned}\|z_{k+1} - z^*\|^2 &= \|z_{k+1} - \tilde{z}_{k+1}\|^2 + \|\tilde{z}_{k+1} - z^*\|^2 + 2(z_{k+1} - \tilde{z}_{k+1})^T(\tilde{z}_{k+1} - z^*)\\ &\leq \left(\lambda^2\beta^6 s^6 + c^2 + 2\lambda c s^3\beta^3\right)\|z_k - z^*\|^2\\ &= \left(\lambda\beta^3 s^3 + c\right)^2\|z_k - z^*\|^2 .\end{aligned}$$

Should one select a value of s and λ such that

$$\lambda\beta^3 s^3 + c < 1$$

linear convergence can be claimed. One would attain convergence if λ is chosen small enough, that is,

$$\lambda < \min\left\{1,\left(\frac{1}{s\rho}-1\right)^2\right\}\left[\frac{2}{\frac{1}{s\alpha(s)}+1} - s^3\beta^3\right] . \tag{2.20}$$

One notices that (Equation 2.20) indicates an implicit lower bound on s, that is already needed to be smaller than $\frac{1}{\beta}$. Given $\alpha(s) > 0$, one would need s to be small enough to make the upper bound on λ positive (i.e. so that for some $\lambda \in (0,1)$ convergence can be attained). □

2.8 Tightness of Nonnegative Interaction Dominance

We observed that damped EGM converges to the saddle point of an objective function in problem (Equation 2.1) if L is nonnegative interaction dominance in both variables x and y. This interaction dominance does not hold if s is too small. This is an interesting observation which is in contradiction with classic minimization problems where every step-size smaller than the reciprocal of the smoothing modulus is acceptable for convergence. An interesting question, therefore, is whether there is a class of nonconvex-nonconcave problems for which it is "necessary" to have interaction dominance in both variables for the convergence to the saddle point of (Equation 2.1). We answer this question in the affirmative by exploring the class of nonconvex-nonconcave quadratic saddle problems and showing that the nonnegative interaction dominance is necessary for convergence so that our main result in Theorem 1 is tight. This would affirm that nonconvex-nonconcave minimax problems are in contrast to classical optimization problems where choosing any step-size smaller than the inverse of the Hessian norm would suffice to guarantee convergence.

We show that a slight nonconvexity-nonconcavity in a given quadratic saddle problem would necessitate the nonnegative, in fact, even positive, interaction dominance to hold for convergence to occur. More precisely, consider the following nonconvex-nonconcave quadratic problem with interaction $\bar{A}$[1],

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^n} L(x, y) := -\frac{\rho}{2} x^T x + \bar{A} x^T y + \frac{\rho}{2} y^T y \ . \tag{2.21}$$

It is observed in (Grimmer et al., 2022a) that for the very specific problem of quadratic minimax optimization problem (Equation 2.21), interaction dominance is a necessary condition for the convergence of damped PPM on that problem. Since damped EGM and damped PPM only differ in terms concerning derivatives of higher order, it is natural to think that damped EGM too accedes the positive interaction dominance as a necessary condition in converging to the solution of (Equation 2.21). We show that this, indeed, is the case. One can observe this by plugging the objective function of problem (Equation 2.21) in the update iterations (Equation 2.6)-(Equation 2.7) of damped EGM. More precisely, first notice that the largest α that satisfies the interaction dominance conditions (Equation 2.2)-(Equation 2.3) for the problem (Equation 2.21) is $\alpha = -\rho + \frac{s \cdot \bar{A}^2}{1 - s \cdot \rho}$. The update on x is given by

$$\begin{aligned} x_{k+1} &= x_k - \lambda s \nabla_x L(z_k) + \lambda s^2 \nabla^2_{xx} L(z_k) \nabla_x L(z_k) - \lambda s^2 \nabla^2_{xy} L(z_k) \nabla_y L(z_k) \\ &= \left(1 + s\lambda\rho + s^2\lambda\rho^2 - \lambda s^2 \bar{A}^2\right) x_k - \left(s\lambda\bar{A} + 2s^2\lambda\rho\bar{A}\right) y_k \ . \end{aligned}$$

[1] For simplicity, we are considering equal dimensions for both x and y.

Applying similar calculations for y_{k+1} and stacking the equations yields,

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} \Theta I & -\Sigma I \\ \Sigma I & \Theta I \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix}, \tag{2.22}$$

where $\Theta := \left(1 + s\lambda\rho + s^2\lambda\rho^2 - \lambda s^2\bar{A}^2\right)$ and $\Sigma := s\lambda\bar{A}\,(1 + 2s\rho)$. Taking the norm of both sides in (Equation 2.22) and simplifying, one gets

$$\|z_{k+1}\|^2 = (\Theta^2 + \Sigma^2)\|z_k\|^2 .$$

Now observing the unique stationary point of the objective function of the quadratic problem (Equation 2.21) is $z = 0$, it follows that damped EGM converges if and only if $\Theta^2 + \Sigma^2 < 1$ holds.

Suppose now that in problem (Equation 2.21) we have $\bar{A} = 10$ and $\rho > 0$ a very small positive value. In other words, the problem is nonconvex-nonconcave with a small negative curvature in partial Hessian $\nabla^2_{xx}L(z)$ and a small positive curvature in partial Hessian $\nabla^2_{yy}L(z)$. Therefore, one can write the convergence condition of EGM on problem (Equation 2.21) as follows

$$\Theta^2 + \Sigma^2 = \Big(1 + s\lambda\rho + \lambda s^2(\rho^2 - 100)\Big)^2 + (10\lambda s\,(1 + 2s\rho))^2 < 1.$$

It is easy to notice that one must have $s\lambda\rho + \lambda s^2(\rho^2 - 100) < 0$ for convergence because otherwise the term $\Theta > 1$ so that $\Theta^2 + \Sigma^2 > 1$, and by (Equation 2.22) the iterations diverge away from the origin. This implies

$$s > \frac{\rho}{100 - \rho^2} \ .$$

Since $\alpha(s) = -\rho + \frac{100s}{1 - s\rho}$ the convergence condition on $\alpha(s)$ simplifies to

$$\alpha(s) > \frac{\rho^2(\rho + 1)}{100 - 2\rho^2} \ . \tag{2.23}$$

The condition (Equation 2.23) implies that even a small nonconvexity-nonconcavity in the problem would necessitate the value of $\alpha(s)$ to be positive in order to attain convergence. Hence even for simple quandratic minimax optimization, our guarantees based on the positive interaction dominance condition are essentially tight, as simple counter examples exist just beyond this regime.

CHAPTER 3

ON NONCONVEX-NONCONCAVE NONSMOOTH MINIMAX PROBLEMS

3.1 Introduction

In this chapter, we aim to study the problem of nonsmooth nonconvex-nonconcave minimax optimization problems, i.e. problems of the form (Equation 1.1):

$$\min_{x \in C} \max_{y \in D} L(x, y)$$

where $C \subset \mathbb{R}^n$ and $D \subset \mathbb{R}^m$, and $f(., y)$ and $f(x, .)$ are generally nonconvex and nonconcave, respectively, and nonsmooth in general.

We will first prove some properties of constrained saddle envelope which we believe could be of independent interest. Then we show that while exact PPM only attains nonexpansivity, the damped PPM operator is quasi-nonexpansive, which will thus be used to show the sublinear convergence of the damped PPM for constrained nonconvex-nonconvex nonsmooth minimax problems under negative comonotonicity.

3.2 Preliminaries

In this section, we cover the preliminaries from variational analysis of nonsmooth analysis, and proximal calculus, in particular, to set the stage for the arguments that follow.

Given a closed set $\Omega \subset \mathbb{R}^n$ and a point $\bar{\omega} \in \Omega$, the regular normal cone of Ω at $\bar{\omega}$, denoted by $\widehat{N}_\Omega(\bar{\omega})$, is the set

$$\widehat{N}_\Omega(\bar{\omega}) = \{\nu \mid \langle \nu, \omega - \bar{\omega} \rangle \leq o(|\omega - \bar{\omega}|), \quad \text{for all } \omega \in \Omega\} \tag{3.1}$$

(Equation 3.1) is equivalent to

$$\widehat{N}_\Omega(\bar{\omega}) = \left\{ \nu \;\middle|\; \limsup_{\substack{\omega \xrightarrow[\Omega]{} \bar{\omega} \\ \omega \neq \bar{\omega}}} \frac{\langle \nu, \omega - \bar{\omega} \rangle}{|\omega - \bar{\omega}|} \leq 0 \right\} \tag{3.2}$$

where $\omega \xrightarrow[\Omega]{} \bar{\omega}$ denotes a sequence converging to $\bar{\omega}$ from within the set Ω. The (limiting) normal cone to Ω at $\bar{\omega}$ defined via $N_\Omega(\bar{\omega})$ is the set of all vectors $\nu \in \mathbb{R}^n$ such that there are sequences $\omega^\nu \xrightarrow[\Omega]{} \bar{\omega}$ and $\nu^\nu \to \nu$ with $\nu^\nu \in \widehat{N}_\Omega(\omega^\nu)$. For a regular set, in the sense of Clarke, (e.g. one that is convex), the regular and limiting normal cones coincide. The distance function $d(.;\Omega)$ to the set Ω is defined as $d(x;\Omega) = \min_{\omega \in \Omega} \|x - \omega\|$. The Proximal Normal to Ω at $\bar{\omega} \in \Omega$ is defined as

$$N^P_\Omega(\bar{\omega}) = \{\zeta \in \mathbb{R}^n \mid \exists\, \tau > 0 \text{ so that } d(\bar{\omega} + \tau\zeta) = \tau\|\zeta\|\} \tag{3.3}$$

The *epigraph* epi f of a function $f : \mathbb{R}^n \to (-\infty, \infty]$ is defined as epi f := $\{(x, \alpha) \mid f(x) \leq \alpha\}$ and the *hypograpgh* hypo f is defined as hypo f := $\{(x, \alpha) \mid f(x) \geq \alpha\}$. The function f is called *lower semicontinuous* at $\bar{x}$ if $\liminf_{x \to \bar{x}} f(x) \geq f(\bar{x})$, and *upper semicontinuous* at $\bar{x}$ if $\limsup_{x \to \bar{x}} f(x) \leq f(\bar{x})$. Given a lower semicontinuous function f and $\bar{x} \in \operatorname{dom} f$, we say a vector ν is a *proximal* *subgradient* to f at $\bar{x}$ if $(v, -1) \in N^P_{\text{epi}\, f}(\bar{x}, f(\bar{x}))$. The collection of all proximal subgradients of f at $\bar{x}$ is the proximal subdifferential $\partial_P f(\bar{x})$ of f at $\bar{x}$. For an upper semicontinuous function f and $\bar{x} \in \operatorname{dom} f$, we define the proximal supergradient $\partial^P f(\bar{x})$ of f at $\bar{x}$ as $-\partial_P(-f)(\bar{x})$. Therefore, $w \in \partial^P f(\bar{x})$ if and only if $(-w, 1) \in N^P_{\text{hypo}\, f}(\bar{x}, f(\bar{x}))$.

Below is a preliminary statement that will be later used in our analysis. The proof is included for completeness.

Proposition 3.2.1. *The* ρ*-weakly-convex-weakly-concave twice continuously differentiable* $\Lambda(x, y)$ *has a* ρ*-weakly monotone gradient oracle.*

Proof. Notice that since $x \mapsto \Lambda(x, y) + \frac{\rho}{2}\|x\|^2$ is convex, we have $\nabla^2_{xx}\Lambda(x, y) \succeq -\rho I_n$. On the other hand, fix $\bar{y}$ and let $(x_1, \bar{y}), (x_2, \bar{y}) \in \operatorname{dom} \Lambda$, we obtain, for every $x, x' \in \operatorname{dom} \Lambda \cap (\mathbb{R}^n \times \{\bar{y}\})$,

$$\begin{aligned}\Lambda(x, \bar{y}) &= \Lambda(x_1, \bar{y}) + \langle \nabla\Lambda_x(x_1, \bar{y}), x - x_1\rangle + \frac{1}{2!}\langle x - x_1, \nabla^2_{xx}\Lambda(x_1 + t(x - x_1))(x - x_1)\rangle \quad t \in (0, 1)\\ &\geq \Lambda(x_1, \bar{y}) + \langle \nabla\Lambda_x(x_1, \bar{y}), x - x_1\rangle - \frac{\rho}{2!}\|x - x_1\|^2 \qquad (3.4)\end{aligned}$$

$$\begin{aligned}\Lambda(x', \bar{y}) &= \Lambda(x_2, \bar{y}) + \langle \nabla\Lambda_x(x_2, \bar{y}), x' - x_2\rangle + \frac{1}{2!}\langle x' - x_2, \nabla^2_{xx}\Lambda(x_2 + t(x' - x_2))(x' - x_2)\rangle \quad t' \in (0, 1)\\ &\geq \Lambda(x_2, \bar{y}) + \langle \nabla\Lambda_x(x_2, \bar{y}), x' - x_2\rangle - \frac{\rho}{2!}\|x' - x_2\|^2 \qquad (3.5)\end{aligned}$$

Evaluating (Equation 3.4) at $x = x_2$ and (Equation 3.5) at $x' = x_1$ and adding the inequalities, we obtain

$$\langle \nabla_x \Lambda(x_2, \bar{y}) - \nabla_x \Lambda(x_1, \bar{y}), x_2 - x_1 \rangle \geq -\rho \|x_2 - x_1\|^2$$

Similarly, one can show, for fixed $\bar{x}$, and $y_1, y_2 \in \operatorname{dom} \Lambda \cap (\{\bar{x}\} \times \mathbb{R}^m)$, that

$$\langle \nabla_y \Lambda(\bar{x}, y_1) - \nabla_y \Lambda(\bar{x}, y_2), y_2 - y_1 \rangle \geq -\rho \|y_2 - y_1\|^2$$

Therefore, the oracle of Λ is $-\rho$-monotone. □

3.3 Properties of Constrained Saddle Envelope

In optimization, Moreau Envelope is widely used to make the objective smooth, i.e. "regularize" it. First introduced and studied by (Moreau, 1965), the Moreau envelope $(e_s f)(.)$ of a function $f : \mathbb{R}^n \to (-\infty, \infty]$ is defined by

$$(e_s f)(x) = \inf_{u \in \mathbb{R}^n} \left\{ f(u) + \frac{1}{2s} \|u - x\|^2 \right\} \tag{3.6}$$

which is a $\mathcal{C}^{1,1}$ function, i.e. continuously differentiable with Lipschitz gradients. This function has many applications in optimization and is fundamental to the study of weakly convex minimization problems. The generalization of the Moreau envelope to one that is applicable in primal-dual methods was first undertaken in (Attouch et al., 1986). More precisely, they considered the saddle envelope[1]

$$f_s(x, y) = \inf_{u \in \mathbb{R}^n} \sup_{v \in \mathbb{R}^m} f(u, v) + \frac{1}{2s}\|u - x\|^2 - \frac{1}{2s}\|v - y\|^2 \tag{3.7}$$

The saddle envelope was further studied in (Grimmer et al., 2022a) and it was shown that for the envelope to be convex-concave, it is only necessary for the original function to have certain structure that is weaker than convexity-concavity.

It is a well-known practice in minimization problems to capture constraints through a lower semicontinuous, proper, convex (LCP) function, say, $f_0(.)$, inside the Moreau envelope. In moving from unconstrained minimax optimization to constrained minimax, one is tempted to capture constraints in this way. However, using the difference of two LCP functions, say, $f_0(x) - g_0(y)$, which can create ambiguities such as $\infty - \infty$, under which one has no choice but to make a convention which benefits one player more than the other. Therefore, it is vital to generalize the properties of the saddle envelope to the one in which the constraints are dealt separately, i.e.

$$f_s(x, y) = \inf_{u \in C} \sup_{v \in D} f(u, v) + \frac{1}{2s}\|u - x\|^2 - \frac{1}{2s}\|v - y\|^2 \tag{3.8}$$

[1] The nomenclature used for the saddle envelope in (Attouch et al., 1986) was “Moreau-Yosia Approximation”.

In the first part of this chapter, we develop the calculus for the constrained saddle envelope setting in (Equation 3.8).

Theorem 2 below shows the convex-concavity of the saddle envelope in the presence of constraints granted the original function is convex-concave itself. For so doing, we will furnish a first-principle proof of convexity-concavity.

Theorem 2. *Suppose* $\Psi : \mathbb{R}^n \times \mathbb{R}^m \to \bar{\mathbb{R}}$ *is a convex-concave lsc-usc bivariate function and* $C \subset \mathbb{R}^n$ *and* $D \subset \mathbb{R}^m$ *be closed convex functions. The constrained saddle envelope*

$$\Psi_s(x, y) = \min_{u \in C} \max_{v \in D} \left\{ \Psi(u, v) + \frac{1}{2s}\|u - x\|^2 - \frac{1}{2s}\|v - y\|^2 \right\}$$

of Ψ *is convex-concave on the domain* $D := \operatorname{dom} \Psi = \{(x, y) \mid -\infty < \Psi(x, y) < \infty\}$ *of* Ψ*.*

First-Principle Proof. Let $y \in \operatorname{dom} \Psi \cap \mathbb{R}^m$ be fixed. Let $(x_1, y), (x_2, y) \in \operatorname{dom} \Psi \cap (\mathbb{R}^n \times \{y\})$. Since $\Psi(., y)$ is convex function, the set $\operatorname{dom} \Psi \cap (\mathbb{R}^n \times \{y\})$ is convex in $\mathbb{R}^n \times \mathbb{R}^m$. Hence, for any $\lambda \in (0, 1)$, $(x_\lambda, y) := (\lambda x_1 + (1 - \lambda) x_2, y) \in \operatorname{dom} \Psi \cap (\mathbb{R}^n \times \{y\})$. Let

$$(x_1^+, y_1^+) := \operatorname{Prox}_{s\Psi}(x_1, y),$$

$$(x_2^+, y_2^+) := \operatorname{Prox}_{s\Psi}(x_2, y),$$

$$(x_\lambda^+, y^+) := \operatorname{Prox}_{s\Psi}(x_\lambda, y).$$

Both x_1^+ and x_2^+ are inside the set $\operatorname{dom}\Psi \cap \mathbb{R}^n$. We can now write

$$\Psi_s(x_\lambda, y) = \min_{u \in C} \max_{v \in D} \Psi(u, v) + \frac{1}{2s}\|u - x_\lambda\|^2 - \frac{1}{2s}\|v - y\|^2 \tag{3.9}$$

Let $u' := u - \lambda x_1^+ - (1-\lambda)x_2^+$. Then one can write (Equation 3.9) as

$$\begin{aligned}
\text{(Equation 3.9)} &= \min_{u' \in C + \lambda x_1^+ + (1-\lambda)x_2^+} \max_{v \in D} \Psi(\lambda(x_1^+ + u') + (1-\lambda)(x_2^+ + u'), v) \\
&\qquad + \frac{1}{2s}\|u' + \lambda x_1^+ + (1-\lambda)x_2^+ - x_\lambda\|^2 - \frac{1}{2s}\|v - y\|^2 \\
&= \max_{v \in D} \min_{u' \in C + \lambda x_1^+ + (1-\lambda)x_2^+} \Psi(\lambda(x_1^+ + u') + (1-\lambda)(x_2^+ + u'), v) \\
&\qquad + \frac{1}{2s}\|u' + \lambda x_1^+ + (1-\lambda)x_2^+ - x_\lambda\|^2 - \frac{1}{2s}\|v - y\|^2 \\
&\le \max_{v \in D} \min_{u' \in C + \lambda x_1^+ + (1-\lambda)x_2^+} \lambda\Psi(x_1^+ + u', v) + (1-\lambda)\Psi(x_2^+ + u', v) \\
&\qquad + \frac{1}{2s}\|u' + \lambda x_1^+ + (1-\lambda)x_2^+ - x_\lambda\|^2 - \frac{1}{2s}\|v - y\|^2 \\
&= \max_{v \in D} \min_{u' \in \Theta} \lambda\Psi(x_1^+ + u', v) + (1-\lambda)\Psi(x_2^+ + u', v) \\
&\qquad + \frac{1}{2s}\|u' + \lambda x_1^+ + (1-\lambda)x_2^+ - x_\lambda\|^2 - \frac{1}{2s}\|v - y\|^2 \\
&\le \max_{v \in D} \inf_{u' \in \Theta^\circ} \lambda\Psi(x_1^+ + u', v) + (1-\lambda)\Psi(x_2^+ + u', v) \\
&\qquad + \frac{1}{2s}\|u' + \lambda x_1^+ + (1-\lambda)x_2^+ - x_\lambda\|^2 - \frac{1}{2s}\|v - y\|^2
\end{aligned} \tag{3.10}$$

where

$$\Theta := \left\{ u' \;\middle|\; \begin{matrix} (u'+x_1^+, v) \in (\operatorname{dom}\Psi)\cap(C\times D) \\ (u'+x_2^+, v) \in (\operatorname{dom}\Psi)\cap(C\times D) \end{matrix} \right\}, \qquad \Theta^\circ := \left\{ u' \;\middle|\; \begin{matrix} (u'+x_1^+, v) \in \operatorname{int}\{(\operatorname{dom}\Psi)\cap(C\times D)\} \\ (u'+x_2^+, v) \in \operatorname{int}\{(\operatorname{dom}\Psi)\cap(C\times D)\} \end{matrix} \right\},$$

and int implies the interior of the set following it.

Notice that, in general, Ψ is lsc-usc but is locally Lipschitz continuous on the interior of its domain. Now since $(u'+x_1^+, v) \in \operatorname{int}\{(\operatorname{dom}\Psi) \cap (C \times D)\}$ and $(u'+x_2^+, v) \in \operatorname{int}\{(\operatorname{dom}\Psi) \cap (C \times D)\}$, we can say that the restriction to Θ° of the set-valued mapping $u \mapsto \partial_C \Psi(u, v)$ is bounded where ∂_C is the partial convex subdifferential (includes proximal subdifferential since we're in convex world). Therefore, $\partial_C \Psi(x_1^+ + u', v)$ and $\partial_C \Psi(x_2^+ + u', v)$ are bounded sets in $\mathbb{R}^n$. Choose $w_1 \in \partial_C \Psi(x_1^+ + u', v)$ and $w_2 \in \partial_C \Psi(x_2^+ + u', v)$. By convexity of $\Psi(., v)$ we have

$$\begin{aligned} \Psi(x_1^+, v) &\geq \Psi(x_1^+ + u', v) - \langle w_1, u' \rangle \\ \Psi(x_2^+, v) &\geq \Psi(x_2^+ + u', v) - \langle w_2, u' \rangle \end{aligned} \tag{3.11}$$

Plugging inequalities (Equation 3.11) in (Equation 3.10) implies that $\Psi_s(x_\lambda, y)$ is upper bounded by

$$\max_{v \in D} \min_{u' \in \Theta^\circ} \lambda \Psi(x_1^+, v) + (1-\lambda)\Psi(x_2^+, v) + \langle w_\lambda, u' \rangle + \frac{1}{2s}\|u' + \lambda x_1^+ + (1-\lambda)x_2^+ - x_\lambda\|^2 - \frac{1}{2s}\|v - y\|^2 \tag{3.12}$$

Notice now that there exists a sequence $\tilde{u}^\nu \downarrow 0$ such that eventually $(\tilde{u}^\nu + x_1^+, v) \in \operatorname{int}\operatorname{dom}\Psi$ and $(\tilde{u}^\nu + x_2^+, v) \in \operatorname{int}\operatorname{dom}\Psi$, whence eventually $\tilde{u}^\nu \in \Theta^\circ$. Let $\epsilon > 0$ be arbitrary. Choose ν

large enough such that $\tilde{u}^\nu \in \Theta^\circ$ and $\|\tilde{u}^\nu\| < \min\left\{\frac{\epsilon}{2\|w_\lambda\|}, \sqrt{\frac{s\epsilon}{2}}\right\}$. We can then further loosen the upper bound (Equation 3.12) as

$$\begin{aligned}
&\max_v \Big\{\lambda\Psi(x_1^+, v) + (1-\lambda)\Psi(x_2^+, v) + \|w_\lambda\|\|\tilde{u}^\nu\| + \frac{1}{2s}\|\tilde{u}^\nu\|^2 \\
&\qquad + \frac{\lambda^2}{2s}\|x_1^+ - x_1\|^2 + \frac{(1-\lambda)^2}{2s}\|x_2^+ - x_2\|^2 - \frac{1}{2s}\|v - y\|^2\Big\} \\
&\le \max_v \Big\{\lambda\Psi(x_1^+, v) + (1-\lambda)\Psi(x_2^+, v) + \|w_\lambda\|\|\tilde{u}^\nu\| + \frac{1}{2s}\|\tilde{u}^\nu\|^2 \\
&\qquad + \frac{\lambda}{2s}\|x_1^+ - x_1\|^2 + \frac{1-\lambda}{2s}\|x_2^+ - x_2\|^2 - \frac{1}{2s}\|v - y\|^2\Big\} \\
&\le \lambda \ \max_v \Big\{\Psi(x_1^+, v) + \frac{1}{2s}\|x_1^+ - x_1\|^2 - \frac{1}{2s}\|v - y\|^2\Big\} \\
&\qquad + (1-\lambda) \ \max_v \Big\{\Psi(x_2^+, v) + \frac{1}{2s}\|x_2^+ - x_2\|^2 - \frac{1}{2s}\|v - y\|^2\Big\} + \|w_\lambda\|\|\tilde{u}^\nu\| + \frac{1}{2s}\|\tilde{u}^\nu\|^2 \\
&= \lambda \ \Big[\Psi(x_1^+, y_1^+) + \frac{1}{2s}\|x_1^+ - x_1\|^2 - \frac{1}{2s}\|y_1^+ - y\|^2\Big] \\
&\qquad + (1-\lambda) \ \Big[\Psi(x_2^+, y_2^+) + \frac{1}{2s}\|x_2^+ - x_2\|^2 - \frac{1}{2s}\|y_2^+ - y\|^2\Big] + \|w_\lambda\|\|\tilde{u}^\nu\| + \frac{1}{2s}\|\tilde{u}^\nu\|^2 \\
&\qquad = \lambda\Psi_s(x_1, y) + (1-\lambda)\Psi_s(x_2, y) + \|w_\lambda\|\|\tilde{u}^\nu\| + \frac{1}{2s}\|\tilde{u}^\nu\|^2 \\
&\qquad < \lambda\Psi_s(x_1, y) + (1-\lambda)\Psi_s(x_2, y) + \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&\qquad = \lambda\Psi_s(x_1, y) + (1-\lambda)\Psi_s(x_2, y)\epsilon
\end{aligned} \tag{3.13}$$

Since $\epsilon > 0$ was arbitrary, (Equation 3.13) implies that

$$\text{(Equation 3.12)} \le \lambda\Psi_s(x_1, y) + (1-\lambda)\Psi_s(x_2, y) \tag{3.14}$$

Notice that (Equation 3.12) is,in itself, an upper bound for $\Psi_s(x_\lambda, y)$. Combining this with (Equation 3.14) implies

$$\Psi_s(x_\lambda, y) \leq \lambda \Psi_s(x_1, y) + (1-\lambda)\Psi_s(x_2, y),$$

Since $y \in \operatorname{dom} \Psi \cap \mathbb{R}^m$ was arbitrary, this establishes the convexity, for any such y, of $x \mapsto \Psi_s(x, y)$. The symmetry of the problem concludes the proof of the Theorem. □

We now show that if the objective is composite with one element ρ-weakly-convex-weakly-concave, then so is the saddle envelope.

Theorem 3. *Suppose* $L(x, y) = \Lambda(x, y) + \Psi(x, y)$ *is a real-valued function on* $\mathbb{R}^n \times \mathbb{R}^m$ *where* Ψ *is convex-concave lsc-usc, and* Λ *is a* ρ*-weakly-convex-weakly-concave* $\mathcal{C}^{1+}$ *function, and* $C \subset \mathbb{R}^n$ *and* $D \subset \mathbb{R}^m$ *be closed convex sets. The constrained saddle envelope*

$$L_s(x, y) = \min_{u \in C} \max_{v \in D} \left\{ \Lambda(u, v) + \Psi(u, v) + \frac{1}{2s}\|u - x\|^2 - \frac{1}{2s}\|v - y\|^2 \right\}$$

of L *is* $\frac{\rho}{(1-s\rho)}$*-weakly-convex-weakly-concave on* $\operatorname{dom} L$.

Proof. Suppose $(x_1, y), (x_2, y) \in \operatorname{dom} L = \operatorname{dom} \Lambda \cap \operatorname{dom} \Psi$ and $\lambda \in (0, 1)$. Let

$$(x_1^+, y_1^+) := \operatorname{Prox}_{sL}(x_1, y),$$

$$(x_2^+, y_2^+) := \operatorname{Prox}_{sL}(x_2, y),$$

$$(x_\lambda^+, y^+) := \operatorname{Prox}_{sL}(x_\lambda, y).$$

Let $u' := u - (\lambda x_1^+ + (1-\lambda)x_2^+)$. We have

$$\begin{aligned}
L_s(x_\lambda, y) &= \min_{u \in C} \max_{v \in D} L(u,v) + \frac{1}{2s}\|u - x_\lambda\|^2 - \frac{1}{2s}\|v - y\|^2 \\
&= \min_{u' \in C} \max_{v \in D} \; L(\lambda(u' + x_1^+) + (1-\lambda)(u' + x_2^+), v) + \frac{1}{2s}\|u' + \lambda x_1^+ + (1-\lambda)x_2^+ - x_\lambda\|^2 \\
&\qquad - \frac{1}{2s}\|v - y\|^2 \\
&\le \min_{u' \in C} \max_{v \in D} \; \lambda L(u' + x_1^+, v) + (1-\lambda)L(u' + x_2^+, v) + \frac{\lambda(1-\lambda)\rho}{2}\|x_1^+ - x_2^+\|^2 \\
&\qquad + \frac{1}{2s}\|u' + \lambda x_1^+ + (1-\lambda)x_2^+ - x_\lambda\|^2 - \frac{1}{2s}\|v - y\|^2 \\
&= \max_{v \in D} \min_{u' \in C} \; \lambda L(u' + x_1^+, v) + (1-\lambda)L(u' + x_2^+, v) + \frac{\lambda(1-\lambda)\rho}{2}\|x_1^+ - x_2^+\|^2 \\
&\qquad + \frac{1}{2s}\|u' + \lambda x_1^+ + (1-\lambda)x_2^+ - x_\lambda\|^2 - \frac{1}{2s}\|v - y\|^2 \\
&\le \max_{v \in D} \; \lambda L(x_1^+, v) + (1-\lambda)L(x_2^+, v) + \frac{\lambda(1-\lambda)\rho}{2}\|x_1^+ - x_2^+\|^2 \\
&\qquad + \frac{1}{2s}\|\lambda x_1^+ + (1-\lambda)x_2^+ - x_\lambda\|^2 - \frac{1}{2s}\|v - y\|^2 \\
&= \max_{v \in D} \; \lambda L(x_1^+, v) + (1-\lambda)L(x_2^+, v) + \frac{\lambda(1-\lambda)\rho}{2}\left[\|x_1^+ - x_1\|^2 + \|x_1 - x_2\|^2 + \|x_2 - x_2^+\|^2\right] \\
&\qquad - \lambda(1-\lambda)\rho\langle x_1^+ - x_1, x_2^+ - x_2\rangle + \lambda(1-\lambda)\rho\langle x_1^+ - x_1, x_1 - x_2\rangle \\
&\qquad + \lambda(1-\lambda)\rho\langle x_1 - x_2, x_2 - x_2^+\rangle + \frac{\lambda^2}{2s}\|x_1^+ - x_1\|^2 + \frac{(1-\lambda)^2}{2s}\|x_2^+ - x_2\|^2 \\
&\qquad + \frac{\lambda(1-\lambda)}{s}\langle x_1^+ - x_1, x_2^+ - x_2\rangle - \frac{1}{2s}\|v - y\|^2 \\
&\le \lambda \max_{v \in D}\left\{L(x_1^+, v) + \frac{1}{2s}\|x_1^+ - x_1\|^2 - \frac{1}{2s}\|v - y\|^2\right\} \\
&\qquad + (1-\lambda)\max_{v}\left\{L(x_2^+, v) + \frac{1}{2s}\|x_2^+ - x_2\|^2 - \frac{1}{2s}\|v - y\|^2\right\} \\
&\qquad + \frac{\lambda(1-\lambda)\rho}{2}\left[\|x_1^+ - x_1\|^2 + \|x_1 - x_2\|^2 + \|x_2 - x_2^+\|^2\right] \\
&\qquad - \lambda(1-\lambda)\rho\langle x_1^+ - x_1, x_2^+ - x_2\rangle + \lambda(1-\lambda)\rho\langle x_1^+ - x_1, x_1 - x_2\rangle \\
&\qquad + \lambda(1-\lambda)\rho\langle x_1 - x_2, x_2 - x_2^+\rangle - \frac{\lambda(1-\lambda)}{2s}\|x_1^+ - x_1\|^2 - \frac{\lambda(1-\lambda)}{2s}\|x_2^+ - x_2\|^2 \\
&\qquad + \frac{\lambda(1-\lambda)}{s}\langle x_1^+ - x_1, x_2^+ - x_2\rangle
\end{aligned}$$

$$
\begin{aligned}
&= \lambda L_s(x_1, y) + (1-\lambda) L_s(x_2, y) - \frac{\lambda(1-\lambda)}{2}\left(\frac{1}{s} - \rho\right)\left[\|x_1^+ - x_1\|^2 + \|x_2^+ - x_2\|^2\right] \\
&\qquad + \frac{\lambda(1-\lambda)\rho}{2}\|x_1 - x_2\|^2 + \lambda(1-\lambda)\rho\langle x_1^+ - x_1 + x_2 - x_2^+, x_1 - x_2\rangle \\
&\qquad + \lambda(1-\lambda)\left(\frac{1}{s} - \rho\right)\langle x_1^+ - x_1, x_2^+ - x_2\rangle \\
&= \lambda L_s(x_1, y) + (1-\lambda) L_s(x_2, y) - \frac{\lambda(1-\lambda)}{2}\left(\frac{1}{s} - \rho\right)\|x_1^+ - x_1 - (x_2^+ - x_2)\|^2 \\
&\qquad - \frac{\lambda(1-\lambda)\rho}{2}\|x_1 - x_2\|^2 + \lambda(1-\lambda)\rho\langle x_1^+ - x_2^+, x_1 - x_2\rangle \\
&= \lambda L_s(x_1, y) + (1-\lambda) L_s(x_2, y) - \frac{\lambda(1-\lambda)\rho}{2}\|x_1 - x_2\|^2 + \lambda(1-\lambda)\rho\langle x_1^+ - x_2^+, x_1 - x_2\rangle \\
&\qquad - \frac{\lambda(1-\lambda)}{2}\left(\frac{1}{s} - \rho\right)\left[\|x_1^+ - x_2^+\|^2 + \|x_1 - x_2\|^2 - 2\langle x_1^+ - x_2^+, x_1 - x_2\rangle\right] \\
&= \lambda L_s(x_1, y) + (1-\lambda) L_s(x_2, y) - \frac{\lambda(1-\lambda)}{2s}\|x_1 - x_2\|^2 + \frac{\lambda(1-\lambda)}{s}\langle x_1^+ - x_2^+, x_1 - x_2\rangle \\
&\qquad - \frac{\lambda(1-\lambda)}{2}\left(\frac{1}{s} - \rho\right)\|x_1^+ - x_2^+\|^2 \\
&= \lambda L_s(x_1, y) + (1-\lambda) L_s(x_2, y) - \frac{\lambda(1-\lambda)}{2s}\|x_1 - x_2 - (x_1^+ - x_2^+)\|^2 + \frac{\lambda(1-\lambda)\rho}{2}\|x_1^+ - x_2^+\|^2 \\
&= \lambda L_s(x_1, y) + (1-\lambda) L_s(x_2, y) + \frac{\lambda(1-\lambda)\rho}{2(1-s\rho)}\|x_1 - x_2\|^2
\end{aligned}
$$

so that the mapping $x \mapsto L_s(x, y)$ is $\frac{\rho}{(1-s\rho)}$-weakly-convex. Similarly, by the symmetry of the problem, we claim that $y \mapsto L_s(x, y)$ is $\frac{\rho}{(1-s\rho)}$-weakly-concave for every x. □

The Theorem above generalizes the result (Poliquin and Rockafellar, 1996, Theorem 5.2), stated about the unconstrained Moreau envelopes, to constrained saddle envelopes.

We now move on to show, in multiple steps, that the constrained saddle envelope is further constinuously differentiable. We start with a Proposition showing that a constrained proximal mapping is Lipschitz continuous.

Proposition 3.3.1. *Consider our objective* $L = \Lambda + \Psi$ *with* $\Lambda \in \mathcal{C}^{1+}$ ρ*-weakly-convex-weakly-concave and* Ψ *convex-concave lsc-usc and* $C \subset \mathbb{R}^n$ *and* $D \subset \mathbb{R}^m$ *closed and convex. The constrained proximal operator*

$$(x^+, y^+) = \mathrm{Prox}_{sL}(x, y) = \underset{\substack{u \in C \\ v \in D}}{\mathrm{argminimax}}\, L(u, v) + \frac{1}{2s}\|u - x\|^2 - \frac{1}{2s}\|v - y\|^2$$

is $\frac{1}{1-s\rho}$ *Lipschitz continuous.*

Proof. Let F_L be the oracle of L and define $F := F_L + N_{C \times D}$. By hypothesis and Proposition 3.2.1, the operator $F_L + \rho I$ is monotone, so that $F + \rho I$ is maximally monotone by the pioneering result of (Rockafellar, 1970). Let $z_i^+ = (I + sF)^{-1}(z_i)$, $i = 1, 2$, and note that $\frac{1}{s}\left(z_i - z_i^+\right) \in F\left(z_i^+\right)$, $i = 1, 2$, so that

$$\left\langle \frac{1}{s}\left(z_1 - z_1^+\right) - \frac{1}{s}\left(z_2 - z_2^+\right), z_1^+ - z_2^+ \right\rangle \geq -\rho\|z_1^+ - z_2^+\|^2$$

Rearranging both sides and invoking Cauchy-Schwartz inequality yields

$$\|z_1^+ - z_2^+\| \leq \frac{1}{1 - s\rho}\|z_1 - z_2\|,$$

the desired result. □

Lemma 3.3.1. *Assume the following for the objective function*

(i) $L(., y)$ *is lsc for every* y *and* $L(x, .)$ *is usc for every* x*,*

(ii) The epigraph of $\mathsf{L}(.,y)$ *and the hypograph of* $\mathsf{L}(x,.)$ *are Clarke-regular on* C *and* D*, respectively.*

For any point $(x,y) \in \mathsf{C} \times \mathsf{D}$*, the constrained saddle envelope satisfies*

$$\partial_P \mathsf{L}_s(.,y)(x) \times \partial^P \mathsf{L}_s(x,.)(y) \subset \left\{ \left(\frac{1}{s}(x - x^+), \frac{1}{s}(y^+ - y) \right) \;\middle|\; (x^+, y^+) \in \mathrm{Prox}_{s\mathsf{L}}(x,y) \right\}$$

for every $s > 0$*.*

Proof. Let $(x^+, y^+) \in \mathrm{Prox}_{s\mathsf{L}}(x,y)$. For every point $x' \in \mathsf{C}$, we have

$$\mathsf{L}(x', y^+) + \frac{1}{2s}\|x' - x\|^2 - \frac{1}{2s}\|y^+ - y\|^2 \geq \mathsf{L}(x^+, y^+) + \frac{1}{2s}\|x^+ - x\|^2 - \frac{1}{2s}\|y^+ - y\|^2.$$

Therefore,

$$\begin{aligned} \mathsf{L}(x', y^+) - \mathsf{L}(x^+, y^+) &\geq \frac{1}{2s}\left\{ \|x^+ - x\|^2 - \|x' - x\|^2 \right\} \\ &= \left\langle \frac{1}{s}(x - x^+), x' - x^+ \right\rangle - \frac{\|x' - x^+\|^2}{2s}, \end{aligned}$$

this being true for every $x' \in \mathsf{C}$, and in particular, for x' close enough to x so that $\frac{1}{s}(x - x^+) \in \partial_P \mathsf{L}(.,y)(x)$. Similarly, one can show $\frac{1}{s}(y^+ - y) \in \partial^P \mathsf{L}(x,.)(y)$. Therefore, for every $(x^+, y^+) \in \mathrm{Prox}_{s\mathsf{L}}(x,y)$ we have

$$\left(\frac{1}{s}(x - x^+), \frac{1}{s}(y^+ - y) \right) \in \partial_P \mathsf{L}(.,y)(x) \times \partial^P \mathsf{L}(x,.)(y).$$

Suppose $(v, w) \in \partial_P L_s(., y)(x) \times \partial^P L_s(x, .)(y)$ so that $v \in \partial_P L_s(., y)(x)$. Then there exists $\sigma > 0$ and $r > 0$ such that whenever $x' \in \mathbb{B}(x, r)$ we have

$$L_s(x', y) \geq L_s(x, y) + \langle v, x' - x \rangle - \sigma \|x' - x\|^2. \tag{3.15}$$

Let $(x^+, y^+) \in \text{Prox}_{sL}(x, y)$ and $(x'^+, y'^+) \in \text{Prox}_{sL}(x', y)$. Therefore,

$$\begin{aligned} L(x^+, y'^+) + \frac{1}{2s}\|x^+ - x'\|^2 - \frac{1}{2s}\|y'^+ - y\|^2 &\geq L(x'^+, y'^+) + \frac{1}{2s}\|x'^+ - x'\|^2 - \frac{1}{2s}\|y'^+ - y\|^2 \\ &= L_s(x', y) \\ &\geq L_s(x, y) + \langle v, x' - x \rangle - \sigma\|x' - x\|^2 \\ &= L(x^+, y^+) + \frac{1}{2s}\|x^+ - x\|^2 - \frac{1}{2s}\|y^+ - y\|^2 \\ &\qquad + \langle v, x' - x \rangle - \sigma\|x' - x\|^2 \\ &\geq L(x^+, y'^+) + \frac{1}{2s}\|x^+ - x\|^2 - \frac{1}{2s}\|y'^+ - y\|^2 \\ &\qquad + \langle v, x' - x \rangle - \sigma\|x' - x\|^2 \end{aligned}$$

where the first inequality holds since $(x'^+, y'^+) \in \text{Prox}_{sL}(x', y)$, the second inequality is due to (Equation 3.15), and the last inequality holds since $(x^+, y^+) \in \text{Prox}_{sL}(x, y)$. We thus have,

$$\frac{1}{2s}\|x^+ - x'\|^2 \geq \frac{1}{2s}\|x^+ - x\|^2 + \langle v, x' - x \rangle - \sigma\|x' - x\|^2,$$

so that for $x' \neq x$,

$$\left\langle v - \frac{1}{s}(x - x^+), \frac{x' - x}{\|x' - x\|} \right\rangle \leq \left(\frac{1}{2s} + \sigma\right) \|x' - x\|,$$

this being true for all $x' \in \mathbb{B}(x, r) \setminus \{x\}$. Therefore, we must have $v - \frac{1}{s}(x - x^+) = 0$. A similar argument yields $w - \frac{1}{s}(y^+ - y) = 0$ so that

$$(v, w) \in \left\{ \left(\frac{1}{s}(x - x^+), \frac{1}{s}(y^+ - y)\right) \;\middle|\; (x^+, y^+) \in \mathrm{Prox}_{sL}(x, y) \right\}.$$

This establishes the Lemma. □

One notes that weak-convexity-weak-concavity, and, therefore, strong-convexity-strong-concavity of the subproblem, is not needed for Lemma 3.3.1 to hold. The following Lemma strengthens the guarantees of Lemma 3.3.1.

Lemma 3.3.2. *Let $L(x, y)$ be proper, lsc-usc, level$_\leq$ bounded in x locally uniformly in y, level$_\geq$ bounded in y locally uniformly in x, $L(., y)$ is bounded from below for every $y \in D$, $L(x, .)$ is bounded from above for every $x \in C$, and ρ-weakly-convex-weakly-concave. Then the constrained saddle envelope for $0 < s < \frac{1}{\rho}$ is Fréchet differentiable at every (x, y) with*

$$\nabla_x L_s(x, y) = \frac{1}{s}(x - x^+), \qquad \nabla_y L_s(x, y) = \frac{1}{s}(y^+ - y).$$

Moreover, the gradients are $\frac{2 - s\rho}{s(1 - s\rho)}$ Lipschitz continuous.

Proof. We first begin by arguing that if the constrained saddle envelope is differentiable, its gradients are Lipschitz continuous. Suppose $(x_1, y_1), (x_2, y_2)$ are two arbitrary points in $C \times D$. We then have

$$\begin{aligned}
\left\| \begin{bmatrix} \nabla_x L_s(x_2, y_2) - \nabla_x L_s(x_1, y_1) \\ \nabla_y L_s(x_1, y_1) - \nabla_y L_s(x_2, y_2) \end{bmatrix} \right\| &= \frac{1}{s} \left\| \begin{bmatrix} x_2 - x_2^+ - (x_1 - x_1^+) \\ y_2 - y_2^+ - (y_1 - y_1^+) \end{bmatrix} \right\| \\
&\leq \left\| \begin{bmatrix} x_2 - x_1 \\ y_2 - y_1 \end{bmatrix} \right\| + \left\| \begin{bmatrix} x_2^+ - x_1^+ \\ y_2^+ - y_1^+ \end{bmatrix} \right\| \\
&\leq \frac{1}{s} \left\{ 1 + \frac{1}{1 - s\rho} \right\} \left\| \begin{bmatrix} x_2 - x_1 \\ y_2 - y_1 \end{bmatrix} \right\|
\end{aligned}$$

where the last inequality follows from Proposition 3.3.1. We now prove the differentiability of the constrained saddle envelope.

By hypothesis, for any (x, y), $\mathrm{Prox}_{sL}(x, y)$ is a singleton, and thus so too is the set

$$\left\{ \left(\frac{1}{s}(x - x^+), \frac{1}{s}(y^+ - y) \right) \,\middle|\, (x^+, y^+) \in \mathrm{Prox}_{sL}(x, y) \right\} \tag{3.16}$$

We now move on to show that the epigraph $E := \mathrm{epi}\{g_y \ : \ x \mapsto L_s(x, y)\}$ is everywhere Clarke regular.

The first step is to show that g_y is lsc everywhere. To that end, let $(\bar{x}, \bar{y})$ be arbitrary. By the level$_\leq$ boundedness in x of $L(x, y)$ locally uniformly in y, we conclude from (Rockafellar and Wets, 1998, Theorem 1.17) the existence of the solution $(\bar{x}^+, \bar{y}^+) := \mathrm{Prox}_{sL}(\bar{x}, \bar{y})$ in $C \times D$. Therefore,

$$g_{\bar{y}}(\bar{x}) = L(\bar{x}^+, \bar{y}^+) + \frac{1}{2s}\|\bar{x}^+ - \bar{x}\|^2 - \frac{1}{2s}\|\bar{y}^+ - \bar{y}\|^2.$$

Let now that $\{x^\nu\}_{\nu \in \mathbb{N}}$ be an arbitrary sequence converging to $\bar{x}$. Thus, $x^\nu \in \bar{x} + \tau^\nu \mathbb{B}$ for some $\tau^\nu \downarrow 0$ and we can write $x^\nu = \bar{x} + \tau^\nu d^\nu$ for some $d^\nu \in \mathbb{B}$. We have

$$\begin{aligned} g_{\bar{y}}(x^\nu) &= \min_{u \in C} \max_{v \in D} L(u, v) + \frac{1}{2s}\|u - \bar{x} - \tau^\nu d^\nu\|^2 - \frac{1}{2s}\|v - \bar{y}\|^2 \\ &\geq \min_{u \in C} \left\{ L(u, \bar{y}^+) + \frac{1}{2s}\|u - \bar{x} - \tau^\nu d^\nu\|^2 \right\} - \frac{1}{2s}\|\bar{y}^+ - \bar{y}\|^2 \\ &=: \left(e_s L(., \bar{y}^+)\right)(x^\nu) - \frac{1}{2s}\|\bar{y}^+ - \bar{y}\|^2 \end{aligned}$$

where $(e_s L(., \bar{y}^+))(x^\nu)$ is the Moreau envelope of $L(., \bar{y}^+)$ at x^ν. By the level$_\leq$ boundedness in x locally uniformly in y, the said Moreau envelope has a unique bounded solution $x^{+^\nu} \in C$. Therefore,

$$g_{\bar{y}}(x^\nu) \geq L(x^{+^\nu}, \bar{y}^+) + \frac{1}{2s}\|x^{+^\nu} - x^\nu\|^2 - \frac{1}{2s}\|\bar{y}^+ - \bar{y}\|^2 \tag{3.17}$$

The boundedness from below of $L(., \bar{y}^+)$, implies it is prox-bounded with threshold $\bar{s} = \infty$ (Rockafellar and Wets, 1998, Exercise 1.24). Since now $L(., \bar{y}^+)$ is lsc, proper, and prox-bounded, and $x^\nu \to \bar{x}$, (Rockafellar and Wets, 1998, Theorem 1.25) implies every cluster point of the sequence $x^{+\nu}$ lies in $\operatorname{argmin}_{u \in C} L(u, \bar{y}^+) + \frac{1}{2s}\|u - \bar{x}\|^2 = \{\bar{x}^+\}$. Thus, $x^{+\nu} \to \bar{x}$. Therefore,

$$\begin{aligned}
\liminf_{\nu \to \infty} g_{\bar{y}}(x^\nu) &\geq \liminf_{\nu \to \infty} \left\{ L(x^{+\nu}, \bar{y}^+) + \frac{1}{2s}\|x^{+\nu} - x^\nu\|^2 - \frac{1}{2s}\|\bar{y}^+ - \bar{y}\|^2 \right\} \\
&\geq \liminf_{\nu \to \infty} \left\{ L(x^{+\nu}, \bar{y}^+) \right\} + \frac{1}{2s}\|\bar{x}^+ - \bar{x}\|^2 - \frac{1}{2s}\|\bar{y}^+ - \bar{y}\|^2 \\
&\geq \liminf_{x \to \bar{x}^+} \left\{ L(x, \bar{y}^+) \right\} + \frac{1}{2s}\|\bar{x}^+ - \bar{x}\|^2 - \frac{1}{2s}\|\bar{y}^+ - \bar{y}\|^2 \\
&\geq L(\bar{x}^+, \bar{y}^+) + \frac{1}{2s}\|\bar{x}^+ - \bar{x}\|^2 - \frac{1}{2s}\|\bar{y}^+ - \bar{y}\|^2 = g_{\bar{y}}(\bar{x})
\end{aligned}$$

where the first inequality follows from (Equation 3.17), the second inequality follows by the minimizing sequence argument just furnished, the third inequality follows from (Rockafellar and Wets, 1998, Lemma 1.7), and the last inequality follows from the fact that $L(., y)$ itself is lsc for every $y \in D$. Since $x^\nu \to \bar{x}$ was arbitrary, we conclude

$$\liminf_{x \to \bar{x}} g_{\bar{y}}(x) \geq g_{\bar{y}}(\bar{x}),$$

the lower semi-continuity of $g_{\bar{y}}(.)$, which thus, since $\bar{x}$ was arbitrary, implies that E is closed everywhere. In particular, then E is locally closed everywehere.

We start the second step of proving Clarke regularity of E by arguing that the mapping $g_{\bar{y}}(.)$ must be usc as well. Suppose not, then

$$L(\bar{x}^+, \bar{y}^+) + \frac{1}{2s}\|\bar{x}^+ - \bar{x}\|^2 - \frac{1}{2s}\|\bar{y}^+ - \bar{y}\|^2 = g_{\bar{y}}(\bar{x}) < \limsup_{x \to \bar{x}} = \inf_{\tau > 0} \left[\sup_{x \in \mathbb{B}(\bar{x}, \tau)} g_{\bar{y}}(x) \right]$$

This implies that for every sequence $\{x^\nu\}_{\nu\in\mathbb{N}}$ converging to $\bar{x}$, we eventually have (for large ν)

$$\begin{aligned} g_{\bar{y}}(\bar{x}) < g_{\bar{y}}(x^\nu) &= \min_{u\in C}\max_{v\in D} L(u,v) + \frac{1}{2s}\|u - x^\nu\|^2 - \frac{1}{2s}\|v - y\|^2 \\ &\leq \max_{v\in D} L(\bar{x}^+, v) + \frac{1}{2s}\|\bar{x}^+ - x^\nu\|^2 - \frac{1}{2s}\|v - y\|^2 \\ &= L(\bar{x}^+, \bar{y}^+) + \frac{1}{2s}\|\bar{x}^+ - x^\nu\|^2 - \frac{1}{2s}\|\bar{y}^+ - \bar{y}\|^2. \end{aligned}$$

Since this strict inequality holds for every large enough $\nu \in \mathbb{N}$, passing to a limit implies $g_{\bar{y}}(\bar{x}) < g_{\bar{y}}(\bar{x})$, a contradiction. Therefore, $g_{\bar{y}} : x \mapsto L_s(x,\bar{y})$ is continuous on $\mathbb{R}^n$ for every $\bar{y} \in \mathbb{R}^m$.

To complete the proof of Clarke-regularity of E, we now need to show that every (limiting) normal vector to E is a regular normal vector. Suppose now that $v \in N_E((\bar{x},\bar{\alpha}))$. We need to show $v \in \widehat{N}_E((\bar{x},\bar{\alpha}))$. The case for when $(\bar{x},\bar{\alpha}) \in \operatorname{int} E$ is trivial because then $v = 0$ and both N_E and $\widehat{N}_E$ contain 0 as they are "cones". Therefore, suppose $(\bar{x},\bar{\alpha}) = (\bar{x}, L_s(\bar{x},\bar{y})) \in \operatorname{bdry} E$ and that $0 \neq v \in N_E((\bar{x},\bar{y}))$. This implies the existence of a sequence $(x^\nu,\alpha^\nu) \in E$, $v^\nu \in \widehat{N}_E((x^\nu,\alpha^\nu))$ with $x^\nu \to \bar{x}$, $\alpha^\nu \to L_s(\bar{x},\bar{y})$, and $v^\nu \to v$. Since $v \neq 0$, the sequence (x^ν,α^ν) must eventually lie on the boundary of E. Otherwise, $v^\nu \equiv 0$, a contradiction. Thus, we eventually can write $(x^\nu,\alpha^\nu) = (x^\nu, L_s(x^\nu,\bar{y}))$ with $v^\nu \in \widehat{N}_E((x^\nu, L_s(x^\nu,\bar{y}))$ with $v^\nu \to v$. We need to show that $v \in \widehat{N}_E((\bar{x}, L_s(\bar{x}, \bar{y}))$. Let $(x, \alpha) \in E$ be arbitrary. The inclusion $v^\nu \in \widehat{N}_E((x^\nu, L_s(x^\nu, \bar{y}))$ hypothesized implies for large ν

$$\left\langle v^\nu, \begin{bmatrix} x - x^\nu \\ \alpha - L_s(x^\nu, \bar{y}) \end{bmatrix} \right\rangle \leq o\left(\left| \begin{bmatrix} x - x^\nu \\ \alpha - L_s(x^\nu, \bar{y}) \end{bmatrix} \right| \right)$$

so that

$$\limsup_{\substack{(x,\alpha)\to(x^\nu, L_s(x^\nu,\bar{y})) \\ (x,\alpha)\in E \\ (x,\alpha)\neq(x^\nu, L_s(x^\nu,\bar{y}))}} \frac{\left\langle v^\nu, \begin{bmatrix} x - x^\nu \\ \alpha - L_s(x^\nu, \bar{y}) \end{bmatrix} \right\rangle}{\left| \begin{bmatrix} x - x^\nu \\ \alpha - L_s(x^\nu, \bar{y}) \end{bmatrix} \right|} \leq 0, \tag{3.18}$$

this last inequality holding for all large ν. Passing (Equation 3.18) to a limit as $\nu \to \infty$ and invoking the continuity of $L_s(., \bar{y})$ implies

$$\limsup_{\substack{(x,\alpha)\to(\bar{x}, L_s(\bar{x},\bar{y})) \\ (x,\alpha)\in E \\ (x,\alpha)\neq(\bar{x}, L_s(\bar{x},\bar{y}))}} \frac{\left\langle v, \begin{bmatrix} x - \bar{x} \\ \alpha - L_s(\bar{x}, \bar{y}) \end{bmatrix} \right\rangle}{\left| \begin{bmatrix} x - \bar{x} \\ \alpha - L_s(\bar{x}, \bar{y}) \end{bmatrix} \right|} \leq 0$$

Hence, $\left\langle v, \begin{bmatrix} x - \bar{x} \\ \alpha - L_s(\bar{x}, \bar{y}) \end{bmatrix} \right\rangle \leq o\left(\left| \begin{bmatrix} x - \bar{x} \\ \alpha - L_s(\bar{x}, \bar{y}) \end{bmatrix} \right| \right)$ which implies $v \in \widehat{N}_E((\bar{x}, \bar{\alpha}))$. This, along with E being locally closed everywhere, implies that E is Clarke regular everywhere.

By (Clarke, 1990, Theorem 2.4.9) we conclude that $g_{\bar{y}}(.) = L_s(.\bar{y})$ is a regular function so that it it admits directional derivatives at $\bar{x}$ for all $\bar{x} \in \partial_P L_s(., \bar{y})(\bar{x})$. This means $\partial_P L_s(., y)(x) \neq \emptyset$ everywhere.

Similarly, one can show that $\partial^P L_s(x, .)(y) \neq \emptyset$ everywhere. Therefore, we have shown that $\partial_P L_s(., y)(x) \times \partial^P L_s(x, .)(y)$ is nonempty everywhere and is a subset of a singleton. Combining all these facts with (Clarke, 1990, Proposition 2.2.4) imply the Fréchet differentiability of $L_s(x, y)$.

□

3.4 Sublinear Convergence of Damped Proximal Point Method

In this section, we work on the negative comontonicity direction. We recall that Proposition 3.3.1 showed that the constrained proximal operator is $\frac{1}{1-s\rho}$ Lipschitz continuous, which lacks nonexpansivity. We now expand that to the case where the operator is negatively comonotone.

Proposition 3.4.1. *Consider the objective* $L(x, y)$ *be lsc-usc,* ρ*-weakly-convex-weakly-concave, and* $C \subset \mathbb{R}^n$ *and* $D \subset \mathbb{R}^m$ *be closed and convex. Suppose further that the operator*

$$F = \partial_P L(., y) \times -\partial^P L(x, .) + N_{C\times D}$$

is $-\xi$*, comonotone with* $\xi \in \left(0, \frac{1}{2\rho}\right)$*.Then the constrained proximal operator*

$$(x^+, y^+) = \mathrm{Prox}_{sL}(x, y) = \operatorname*{argminimax}_{\substack{u\in C\\ v\in D}} L(u, v) + \frac{1}{2s}\|u - x\|^2 - \frac{1}{2s}\|v - y\|^2$$

is nonexpansive whenever $s \in \left(2\xi, \frac{1}{\rho}\right)$.

Proof. Let $z_1, z_2 \in C \times D$ and $z_i^+ = \text{Prox}_{sL}(x_i, y_i), i = 1, 2$. Optimality conditions of the proximal operator imply $\frac{1}{s}\left(z_i - z_i^+\right) \in F(z_i^+)$. Negative comonotonicity of F thus imply

$$\begin{aligned}\frac{1}{s}\langle z_1 - z_2 - (z_1^+ - z_2^+), z_1^+ - z_2^+\rangle &\geq -\frac{\xi}{s^2}\|z_1 - z_2 - (z_1^+ - z_2^+)\|^2 \\ &= -\frac{\xi}{s^2}\|z_1^+ - z_2^+\|^2 - \frac{\xi}{s^2}\|z_1 - z_2\|^2 + \frac{2\xi}{s^2}\langle z_1^+ - z_2^+, z_1 - z_2\rangle,\end{aligned}$$

the rearranging of which implies

$$\begin{aligned}(s - \xi)\|z_1^+ - z_2^+\|^2 &\leq \xi\|z_1 - z_2\|^2 + (s - 2\xi)\langle z_1^+ - z_2^+, z_1 - z_2\rangle \\ &\leq \xi\|z_1 - z_2\|^2 + (s - 2\xi)\|z_1^+ - z_2^+\|\|z_1 - z_2\| \qquad (3.19) \\ &\leq \xi\|z_1 - z_2\|^2 + \frac{s - 2\xi}{1 - s\rho}\|z_1 - z_2\|^2 \\ &= \frac{s - \xi(1 + s\rho)}{1 - s\rho}\|z_1 - z_2\|^2,\end{aligned}$$

where the last inequality follows from Proposition 3.3.1. We thus so far have the inequality

$$\|z_1^+ - z_2^+\| \leq \sqrt{\frac{s - \xi(1 + s\rho)}{(s - \xi)(1 - s\rho)}}\|z_1 - z_2\|.$$

Plugging this inequality into (Equation 3.19) further implies,

$$\|z_1^+ - z_2^+\| \leq \sqrt{\frac{1}{s - \xi}\left[\xi + (s - 2\xi)\sqrt{\frac{s - \xi(1 + s\rho)}{(s - \xi)(1 - s\rho)}}\right]}\|z_1 - z_2\|.$$

Iteratively continuing this process yields,

$$\|z_1^+ - z_2^+\| \leq \sqrt{\frac{1}{s-\xi}\left[\xi + (s-2\xi)\sqrt{\frac{1}{s-\xi}\left[\xi + (s-2\xi)\sqrt{\ldots}\right]}\right]}\|z_1 - z_2\| =: C\|z_1 - z_2\|,$$

Simple algebraic observation over this identity implies,

$$C^2 = \frac{1}{s-\xi}\left[\xi + (s-2\xi)C\right],$$

so that $C = 1$.

Therefore,

$$\|z_1^+ - z_2^+\| \leq \|z_1 - z_2\|$$

□

Therefore, the prox operator is nonexpansive. Let now $\lambda \in (0,1)$ and write $T(z) := \lambda z^+ + (1-\lambda)z$, i.e. $T = \lambda \text{Prox}_{sL} + (1-\lambda)I$. Since Prox_{sL} is nonexpansive, (Bauschke and Combettes, 2017, Proposition 4.35) implies that T is λ-averaged and satisfies

$$\|T(z) - T(z')\|^2 \leq \|z - z'\|^2 - \frac{1-\lambda}{\lambda}\|(I-T)(z) - (I-T)(z')\|^2, \quad \text{for all } z, z' \in C \times D \tag{3.20}$$

We now furnish the sublinear convergence of damped PPM to the solution of a structured nonconvex-nonconcave minimax optimization problem.

Theorem 4. *Let* $C \subset \mathbb{R}^n$ *and* $D \subset \mathbb{R}^m$ *be closed and convex sets. Consider the problem* $\min_{x\in C}\max_{y\in D} L(x,y)$ *where* L *is lsc-usc,* ρ*-weakly-convex-weakly-concave, and suppose further that the operator* $F = \partial_P L(.,y) \times -\partial^P L(x,.) + N_{C\times D}$ *is* $-\xi$ *comonotone with* $\xi \in \left(0, \frac{1}{2\rho}\right)$*. Let* $\lambda \in (0,1)$*. The sequence* $\{z_k\}$ *generated by damped proximal point method* $z_{k+1} = \lambda \mathrm{Prox}_{sL}(z_k) + (1-\lambda)z_k$*, where* $s \in \left(2\xi, \frac{1}{\rho}\right)$*, converges sublinearly to a saddle point* $z^* \in \mathcal{Z}^*$ *of* $\min_{x\in C}\max_{y\in D} L(x,y)$*, i.e. for any* $K \geq 1$*, we have*

$$\min_{l=1,2,\ldots,K} \|z_{k+1} - z_k\|^2 \leq \frac{\lambda\|z_1 - z^*\|^2}{K(1-\lambda)} \tag{3.21}$$

Proof. Notice the prox operator can be written as $\mathrm{Prox}_{sL}(z_k) = [I + s(F_L + N_{C\times D})]^{-1}(z_k)$ where $F_L + N_{C\times D}$ is $-\xi$ comonotone. Proposition 3.4.1 implies that thus defined prox operator is nonexpansive so that the sequence $\{z_k\}_{k\in\mathbb{N}}$ generated by the damped proximal method satisfies

$$\|z_{k+1} - z^*\|^2 \leq \|z_k - z^*\|^2 - \frac{1-\lambda}{\lambda}\|z_{k+1} - z_k\|^2,$$

so that,

$$\|z_{k+1} - z_k\|^2 \leq \frac{\lambda}{1-\lambda}\left[\|z_k - z^*\|^2 - \|z_{k+1} - z^*\|^2\right]. \tag{3.22}$$

One can thus write,

$$\min_{k=1,2,\ldots,K} \|z_{k+1} - z_k\|^2 \leq \frac{1}{K}\sum_{k=1}^{K} \|z_{k+1} - z_k\|^2 \leq \frac{\lambda}{K(1-\lambda)}\|z_1 - z^*\|^2$$

where the last inequality follows from (Equation 3.22). □

CITED LITERATURE

[Alves et al., 2016] Alves, M. M., Monteiro, R. D. C., , and Svaiter, B. F. (2016). Regularized hpe-type methods for solving monotone inclusions with improved pointwise iteration-complexity bounds. *SIAM Journal on Optimization*, 26(4):2730–2743.

[Applegate et al., 2022] Applegate, D., Hinder, O., Lu, H., and Lubin, M. (2022). Faster first-order primal-dual methods for linear programming using restarts and sharpness. *Mathematical Programming*, (1):1–52.

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *34th International Conference on Machine Learning (ICML)*, pages 214–223.

[Attouch et al., 1986] Attouch, H., Aze, D., , and Wets, R. J.-B. (1986). On continuity properties of the partial legendrefenchel transform: Convergence of sequences of augmented lagrangian functions, moreau-yosida approximates and subdifferential operators. *Fermat Days 85: Mathematics for Optimization*, 129:1–42.

[Başar and Olsder, 1998] Başar, T. and Olsder, G. J. (1998). *Dynamic Noncooperative Game Theory.* Society for Industrial and Applied Mathematics, 2 edition.

[Bauschke and Combettes, 2017] Bauschke, H. H. and Combettes, P. L. (2017). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* CMS Books in Mathematics, Springer.

[Bauschke1 et al., 2021] Bauschke1, H. H., Moursi, W. M., and Wang, X. (2021). Generalized monotone operators and their averaged resolvents. *Mathematical Programming*, (189):55–74.

[Chambolle and Pock, 2011] Chambolle, A. and Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145.

[Clarke, 1990] Clarke, F. H. (1990). *Optimization and Nonsmooth Analysis.* SIAM, Classics in Applied Mathematics.

[Dai et al., 2018] Dai, B., Shaw, A., He, N., Li, L., and Song, L. (2018). Boosting the actor with dual critic. In *International Conference on Learning Representations (ICLR).*

[Daskalakis et al., 2018] Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. (2018). Training gans with optimism. In *International Conference on Learning Representations (ICLR).*

[Daskalakis and Panageas, 2018] Daskalakis, C. and Panageas, I. (2018). The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems 31*, page 9236–9246.

[Davis and Drusvyatskiy, 2019] Davis, D. and Drusvyatskiy, D. (2019). Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239.

[Davis and Grimmer, 2019] Davis, D. and Grimmer, B. (2019). Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Jounral on Optimization*, 29(3):908–1930.

[Diakonikolas et al., 2021] Diakonikolas, J., Daskalakis, C., and Jordan, M. I. (2021). Efficient methods for structured nonconvex-nonconcave min-max optimization. volume 130, pages 2746–2754. PMLR.

[Douglas and Rachford, 1956] Douglas, J. and Rachford, H. H. (1956). On the numerical solution of heat conduction problems in two and three space variables. *Transactions on American Mathematical Society*, 82(2):421–439.

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

[Gidel et al., 2020] Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. (2020). A variational inequality perspective on generative adversarial networks.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, pages 2672–2680.

[Grimmer et al., 2022a] Grimmer, B., Lu, H., Worah, P., and Mirrokni, V. (2022a). The landscape of the proximal point method for nonconvex–nonconcave minimax optimization. *Mathematical Programming*, pages 1–35.

[Grimmer et al., 2022b] Grimmer, B., Lu, H., Worah, P., and Mirrokni, V. (2022b). Limiting behaviors of nonconvex-nonconcave minimax optimization via continuous-time systems. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, pages 465–487.

[Hajizadeh et al., 2023] Hajizadeh, S., Lu, H., and Grimmer, B. (2023). On the linear convergence of extra-gradient methods for nonconvex-nonconcave minimax problems. *To Appear in INFORMS Journal on Optimization.*

[Hast et al., 2013] Hast, M., Åström, K., Bernhardsson, B., and Boyd, S. (2013). Pid design by convex-concave optimization. In *ECCIEEE*.

[Jin et al., 2020] Jin, C., Netrapalli, P., and Jordan, M. I. (2020). What is local optimality in nonconvex-nonconcave minimax optimization? In *Proceedings of the 37th International Conference on Machine Learning*, pages 4880–4889.

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

[Lee and Kim, 2021] Lee, S. and Kim, D. (2021). Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *Advances in Neural Information Processing Systems.*

[Lin et al., 2020a] Lin, T., Jin, C., and Jordan, M. (2020a). On gradient descent ascent for nonconvex-concave minimax problems. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6083–6093.

[Lin et al., 2020b] Lin, T., Jin, C., and Jordan, M. I. (2020b). Near-optimal algorithms for minimax optimization. In *33rd Annual Conference on Learning Theory*, pages 1–42.

[Liu et al., 2021] Liu, M., Rafique, H., Lin, Q., and Yang, T. (2021). First-order convergence theory for weakly-convex-weakly-concave min-max problems. *Journal of Machine Learning Research*, 22:1–34.

[Lu, 2022] Lu, H. (2022). An $\mathsf{o}\left(s^r\right)$-resolution ode framework for understanding discrete-time algorithms and applications to the linear convergence of minimax problems. *Mathematical Programming*, 194:1061–1112.

[Madry et al., 2018] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083v4.

[Malitsky, 2020] Malitsky, Y. (2020). Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184:383–410.

[Martinet, 1970] Martinet, B. (1970). Regularisation d'ineq́uations variationelles par approximations successives. *Rev. Francaise Inf. Rech. Oper.*, pages 154–159.

[Mertikopoulos et al., 2019] Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., , and Piliouras, G. (2019). Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference of Learning Representation.*

[Mokhtari et al., 2020] Mokhtari, A., Ozdaglar, A., and Pattathil, S. (2020). A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics.*

[Monteiro and Svaiter, 2010] Monteiro, R. D. C. and Svaiter, B. F. (2010). On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6):2755–2787.

[Moreau, 1965] Moreau, J. J. (1965). Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, (93):273–299.

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

[Nesterov, 1983] Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate $\mathsf{o}\left(\frac{1}{k^2}\right)$. *Soviet Math. Doklady*, 27:372–376.

[Nesterov, 2005] Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming*, (103):127–152.

[O'Connor and Vandenberghe, 2018] O'Connor, D. and Vandenberghe, L. (2018). On the equivalence of the primal-dual hybrid gradient method and douglas–rachford splitting. *Mathematical Programming*, pages 1–24.

[Pethick et al., 2022] Pethick, T., Latafat, P., Patrinos, P., Fercoq, O., and Cevher, V. (2022). Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *International Conference on Learning Representations (ICLR).*

[Poliquin and Rockafellar, 1996] Poliquin, R. A. and Rockafellar, R. T. (1996). Prox-regular functions in variational analysis. *Transactions of the American Mathematical society*, 348(5):pp. 1805–1838.

[Rafique et al., 2022] Rafique, H., Liu, M., Lin, Q., and Yang, T. (2022). Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121.

[Rockafellar, 1970] Rockafellar, R. T. (1970). On the maximal monotonicity of subdifferential mappings. *Pacific Journal of Mathematics*, 33(1):209–216.

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

[Rockafellar and Wets, 1998] Rockafellar, T. and Wets, R. J.-B. (1998). *Variational Analysis.* Springer.

[Shi et al., 2022] Shi, B., Du, S. S., Jordan, M. I., and Su, W. J. (2022). Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195:79–148.

[Solodov and Svaiter, 1999a] Solodov, M. V. and Svaiter, B. F. (1999a). A hybrid approximate extragradient – proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, pages 323–345.

[Solodov and Svaiter, 1999b] Solodov, M. V. and Svaiter, B. F. (1999b). A hybrid projection-proximal point algorithm. *Journal of Convex Analysis*, pages 59–70.

[Solodov and Svaiter, 2000] Solodov, M. V. and Svaiter, B. F. (2000). An inexact hybrid generalized proximal point algorithmand some new results on the theory of bregman functions. *Mathematics of Operations Research*, 25:214–230.

[Solodov and Svaiter, 2001] Solodov, M. V. and Svaiter, B. F. (2001). A unified framework for some inexact proximal point algorithms. *Numerical Functional Analysis and Optimization*, 20:1013–1035.

[Song et al., 2020] Song, C., Zhou, Z., Zhou, Y., Jiang, Y., and Ma, Y. (2020). Optimistic dual extrapolation for coherent non-monotone variational inequalities. In *34th Conference on Neural Information Processing Systems (NeurIPS).*

[Su et al., 2016] Su, W., Boyd, S., and Candés, E. J. (2016). A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17:1–43.

[Sutton and Barto, 2018] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An Introduction.* MIT press.

[Syrgkanis et al., 2015] Syrgkanis, V., Agarwal, A., Luo, H., and Schapire, R. E. (2015). Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, page 2989–2997.

[Thekumparampil et al., 2019] Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. (2019). Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems 32 (NeurIPS).*

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

[von Stackelberg, 2010] von Stackelberg, H. (2010). *Market Structure and Equilibrium.* Springer Berlin, Heidelberg.

[Yang et al., 2020] Yang, J., Kiyavash, N., and He, N. (2020). Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems.

[Zhang et al., 2022] Zhang, G., Poupart, P., and Yu, Y. (2022). Optimality and stability in non-convex smooth games. *Journal of Machine Learning Research*, (23):1–71.

Copyright Agreement with INFORMS

Saeid Hajizadeh <shajiz2@uic.edu>

RE: INFORMS Journal on Optimization: IJO-2022-01-004.R1 Proof of Permission

Chris Asher <casher@informs.org> Fri, Jun 23, 2023 at 8:27 AM
To: "shajiz2@uic.edu" <shajiz2@uic.edu>
Cc: Jeffrey Elmore <jelmore@informs.org>

Dear Saeid,

Thank you for contacting us with your questions, I'm happy to help.

Our copyright policy permits authors to use all or part of their work in any future projects without permission or fees, provided a full citation to the article's version of record is provided in the references and citations. The only thing you can't do is post the entire formatted/copyedited version of the paper somewhere (though you can post the accepted, pre-formatted version of your paper anywhere you like).

If you have any other questions, please don't hesitate to contact me.

My best,

Chris

Chris Asher

Senior Managing Editor

5521 Research Park Drive, Suite 200, Catonsville, MD 21228 USA
p: 443-757-3583 | e: chris.asher@informs.org

www.informs.org

VITA

NAME Saeid Hajizadeh

EDUCATION Ph.D., Mathematical Computer Science, University of Illinois at Chicago, Chicago, United States
M.Sc., Electrical and Computer Engineering, University of Illinpos at Chicago, Chicago, United States
B.A., Electrical Engineering, Ferdowsi University of Mashhad, Mashhad, Iran, 2011

PUBLICATIONS S. Hajizadeh, Haihao Lu, and Benjamin Grimmer, ***On the Linear Convergence of Extra-Gradient Methods for Nonconvex-Nonconcave Minimax Problems***, arXiv:2201.06167v1, To appear in INFORMS Journal on Optimization

S. Berenjian, S. Hajizadeh, R. Ebrahimi, ***An Incentive Security Model to Provide Fairness for Peer-to-Peer Networks***, *IEEE Conference on Applications, Information and Network Security*, 19-21 Nov. 2019, Penang, Malaysia.

M. Monemizadeh, H. Fehri, G. Abed Hodtani, S. Hajizadeh ***Capacity Bounds and High-SNR Capacity of the Additive Exponential Noise Channel With Additive Exponential Interference***, *Iranian Journal of Electrical and Electronic Engineering*, Aug. 2019.

S. Hajizadeh, N. Devroye ***Dependence Balance Outer Bounds for the Discrete Memoryless Two-way Multiple Access Broadcast Channel***, 52nd *Annual Allerton Conference on Commu- nication, Control, and Computing*, Monticello, IL, Oct. 2014.

S. Hajizadeh, M. Monemizadeh, and E. Bahmani ***State-dependendent Z Channels***, 48th *Annual Conference on Information Sciences and Systems (CISS)*, Princeton University, March 19-21, 2014. More complete version available at ArXiv.

S. Hajizadeh, G. A. Hodtani ***Three-receiver Broadcast Channels with Side Information***, *IEEE Int. Symp. on Inf. Theory*, Boston, MA, July 2012.

S. Hajizadeh, G. A. Hodtani ***Asymmetric Broadcast Channels***, 50th *annual Allerton Conference on Communications, Control, and Computing*, Monticello, IL, Oct. 2012.

S. Hajizadeh, M. Monemizadeh, G. A. Hodtani ***A Coding Theorem for the Discrete Mem- oryless Compound Multiple Access Channels with Common Message and Generalized Feedback***, 50th *annual Allerton Conference on Communications, Control, and Computing*, Monticello, IL, Oct. 2012.

S. Hajizadeh ***Broadcast Channels***, *B.Sc. Thesis*, September 2011, Ferdowsi University of Mashhad, Mashhad, Iran.