The Landscape of the Proximal Point Method for Nonconvex-Nonconcave Minimax Optimization

Benjamin Grimmer[*], Haihao Lu[†], Pratik Worah[‡], Vahab Mirrokni[§]

Abstract

Minimax optimization has become a central tool for modern machine learning with applications in generative adversarial networks, robust optimization, reinforcement learning, etc. These applications are often nonconvex-nonconcave, but the existing theory is unable to identify and deal with the fundamental difficulties posed by nonconvex-nonconcave structures. In this paper, we study the classic proximal point method (PPM) for solving nonconvex-nonconcave minimax problems. We develop a new analytic tool, the saddle envelope, generalizing the Moreau envelope. The saddle envelope not only smooths the objective but can convexify and concavify it based on the level of interaction present between the minimizing and maximizing variables. From this, we identify three distinct regions of nonconvex-nonconcave minimax problems. For problems where interaction is sufficiently strong, we derive global linear convergence guarantees. Conversely when the interaction is fairly weak, we derive local linear convergence guarantees with a proper initialization. Between these two settings, we show that PPM may diverge or converge to a limit cycle and present a "Lyapunov"-type function limiting how quickly PPM can diverge.

1 Introduction

Minimax optimization has become a central tool for modern machine learning, recently receiving increasing attention in optimization and machine learning communities. The problem of interest is the following saddle point optimization problem:

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m} L(x,y) \ , \tag{1}$$

where $L(x,y)$ is a differentiable function in x and y. Many important problems in modern machine learning can be formulated as a minimax optimization problem with the form (1), and often the objective $L(x,y)$ is neither convex in x nor concave in y. For example,

- **(GANs).** Generative adversarial networks (GANs) [1] learn the distribution of observed samples through a two-player zero-sum game. While the generative network (parameterized by G) generates samples minimizing their difference from the true data distribution, the

[*] bdg79@cornell.edu; Google Research, New York NY and Cornell University, Ithaca NY

[†] Haihao.Lu@chicagobooth.edu; Google Research, New York NY and University of Chicago, Chicago IL

[‡] pworah@google.com; Google Research, New York NY

[§] mirrokni@google.com;Google Research, New York NY

discriminative network (parameterized by D) maximizes its ability to distinguish between these distributions. This gives rise to the minimax formulation

$$\min_G \max_D \mathbb{E}_{s\sim p_{data}} [\log D(s)] + \mathbb{E}_{e\sim p_{latent}} [\log(1 - D(G(e)))] \ ,$$

where p_{data} is the data distribution, and p_{latent} is the latent distribution.

- **(Robust Training).** Minimax optimization has a long history in robust optimization. Recently, it has found usage with neural networks, which have shown great success in machine learning tasks but are vulnerable to adversarial attack. Robust training [2] aims to overcome such issues by solving the minimax problem

  $$\min_x \mathbb{E}_{(u,v)} \left[\max_{y\in S} \ell(u + y, v, x) \right] \ ,$$

  where u is a feature vector, v is its label, x is the model parameters being trained, y is an adversarial modification, and S is the set of possible corruptions.

- **(Reinforcement Learning).** In reinforcement learning, the solution to Bellman equations can be obtained by solving a primal-dual minimax formulation. Such an approach can be viewed as having a dual critic seeking a solution satisfying the Bellman equation and a primal actor seeking state-action pairs to break this satisfaction [3, 4].

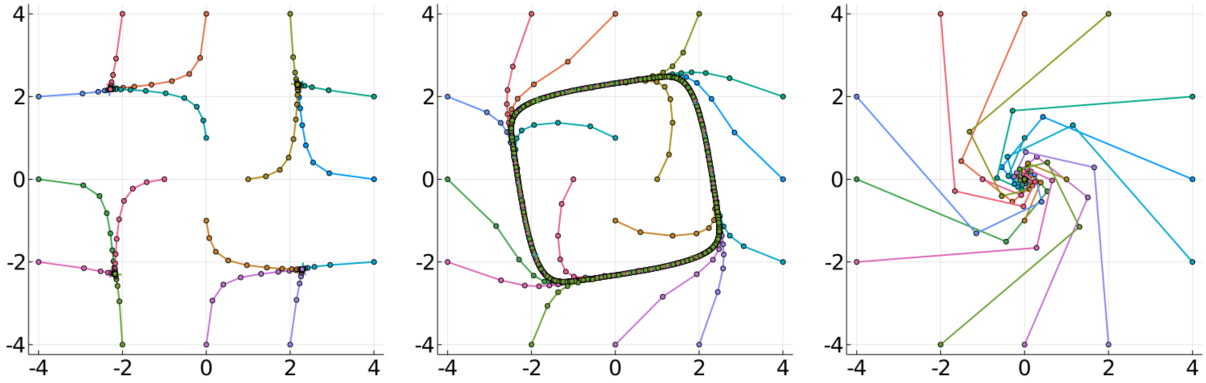
The Proximal Point Method (PPM) may be the most classic first-order method for solving minimax problems. It was first studied in the seminal work by Rockafellar in [5], and many practical algorithms for minimax optimization developed later on turn out to be approximations of PPM, such as Extragradient Method (EGM) [6, 7] and Optimistic Gradient Descent Ascent [8]. The update rule of PPM with step-size η is given by the proximal operator:

$$(x_{k+1}, y_{k+1}) = \text{prox}_\eta(x_k, y_k) := \arg \min_{u\in\mathbb{R}^n} \max_{v\in\mathbb{R}^m} L(u, v) + \frac{\eta}{2}\|u - x_k\|^2 - \frac{\eta}{2}\|v - y_k\|^2. \tag{2}$$

For convex-concave minimax problems, PPM is guaranteed to converge to an optimal solution. However, the dynamics of PPM for nonconvex-nonconcave minimax problem are much more complicated. For example, consider the special case of minimax optimization problem with bilinear interaction defined as

$$\min_x \max_y L(x, y) = f(x) + x^T Ay - g(y). \tag{3}$$

Figure 1 presents the sample paths of PPM from different initial solutions solving a simple two-dimensional nonconvex-nonconcave minimax problem (3) with $f(x) = g(x) = (x - 3)(x - 1)(x + 1)(x + 3)$ and different interaction terms A. This example may be the simplest non-trivial example of a nonconvex-nonconcave minimax problem. It turns out the behaviors of PPM heavily relies on the scale of the interaction term A: when the interaction term is small, PPM converges to local stationary solutions, as the interaction term increases, PPM may fall into a limit cycle indefinitely, and eventually when the interaction term is large enough, PPM converges globally to stationary solution. Similar behaviors also happen in other classic algorithms for nonconvex-nonconcave minimax problems, in particular, EGM, which is known as one of the most effective algorithms for minimax problem. See Figure 2 in Appendix A for their trajectories for solving this simple two-dimension examples (the study of these other algorithms is beyond the scope of this paper).



(a) $A = 1$ (b) $A = 10$ (c) $A = 100$

Figure 1: Sample paths of PPM from different initial solutions applied to (3) with $f(x) = (x+3)(x+1)(x-1)(x-3)$ and $g(y) = (y+3)(y+1)(y-1)(y-3)$ and different scalars A. As $A \geq 0$ increases, the solution path transitions from having four locally attractive stationary points, to a globally attractive cycle, and finally to a globally attractive stationary point.

In practice, it is also well-known that classic first-order methods may fail to converge to a stable solution for minimax problems, such as GANs [9].

The goal of this paper is to understand these varied behaviors of PPM when solving nonconvex-nonconcave minimax problems. To do so, we develop a new analytic tool, the saddle envelope (see Section 2), which is a natural extension of Moreau envelope [10] to a minimax problem. Though sharing many similarities, the saddle envelope has fundamentally different properties compared to Moreau envelope. Most outstandingly, the saddle envelope not only smooths the objective but also can convexify and concavify nonconvex-nonconcave problems based on the level of the interaction between x and y. Understanding this structure turns out to be the cornerstone of explaining the above varied behaviors of PPM. Utilizing this machinery, we find that the three regions shown in the simple two-dimensional example (Figure 1) happen with generality for solving (1). Informally speaking,

1. When the interaction between x and y is dominate, PPM has global linear convergence to a stationary point of $L(x, y)$ (Figure 1 (c)). This argument utilizes the fact that, in this case, the closely related saddle envelope becomes convex-concave thanks to the high interaction terms, even though the original function $L(x, y)$ is nonconvex-nonconcave.

2. When the interaction between x and y is weak, properly initializing PPM yields local linear convergence to a nearby stationary point of $L(x, y)$ (Figure 1 (a)). The intuition is that due to the low interaction we do not lose much by ignoring the interaction and decomposing the minimax problems to a nonconvex minimization problem and a nonconcave maximization problem (where the local convergence of PPM is typical).

3. Between these interaction dominate and weak regimes, PPM may fail to converge at all and fall into cycling (Figure 1 (b)) or divergence (see the example in Section 5.1). In this scenario, we construct a "Lyapunov"-type function that characterizes how fast PPM may diverge and show that the resulting diverging bound is tight for PPM by constructing a worst-case example.

Furthermore, we believe this saddle envelope will be broadly impactful outside its use herein analyzing the proximal point method. We give it a full development in Section 2 that we believe will be of independent interest. As a byproduct of our analysis of the saddle envelope, we clearly see that the interaction term helps the convergence of PPM for minimax problems. This may not be the case for other algorithms, such as gradient descent ascent (GDA) and alternating gradient descent ascent (AGDA) (see Figure 2 in Appendix A for some examples and [11] for theoretical analysis).

We comment on the meaning of stationary points $\nabla L(z) = 0$ for nonconvex-nonconcave problems. By viewing the problem (1) as a simultaneous zero-sum game between a player selecting x and a player selecting y, a stationary point can be thought of as a first-order Nash Equilibrium. That is, neither player tends to deviate from their position based on their first-order information. One can instead view the minimax problem as a sequential zero-sum game (where the minimizing player selects x and then the maximizing player exploits that choice in choosing y). Unlike the convex-concave case, the solutions between these two types of games no longer coincide and the optimal (sequential) minimax solution need not be a stationary point. In this case, a different asymmetric measure of optimality may be called for [8, 12, 9]. However such approaches are beyond the scope of this paper as the limit points of the proximal point method are all stationary points.

In the rest of this section, we discuss the assumptions, related literature, and preliminaries that will be used later on. In Section 2, we introduce the saddle envelope and develop its general theory. In particular, we introduce the interaction dominance condition (Definition 2.9) that naturally comes out as a condition for convexity-concavity of the saddle envelope. In Section 3, we present the global linear convergence of PPM for solving interaction dominate minimax problems. In Section 4, we discuss the behaviors of PPM in the interaction weak case and show that with a good initialization, PPM converges to a local stationary point. In Section 5, we show that PPM may diverge when our interaction dominance condition is slightly violated, showing the tightness of our global convergence theory. Further, we propose a natural "Lyapunov"-type function that applies to generic minimax problems, providing an upper bound on how quickly problems can divergence in the difficult interaction moderate setting. This bound on PPM's divergence is tight under our basic assumptions as we provide a worst-case example.

1.1 Assumptions and Algorithms

Basic definitions and assumptions. We say a function $M(x, y)$ is β-smooth if its gradient is uniformly β-Lipschitz

$$\|\nabla M(z) - \nabla M(z')\| \leq \beta \|z - z'\|$$

or equivalently for twice differentiable functions, if $\|\nabla^2 M(z)\| \leq \beta$. Further, we say a twice differentiable $M(x, y)$ is μ-strongly convex-strongly concave for some $\mu \geq 0$ if

$$\nabla^2_{xx} M(z) \succeq \mu I \ , \quad -\nabla^2_{yy} M(z) \succeq \mu I \ .$$

When $\mu = 0$, this corresponds to M being convex with respect to x and concave with respect to y.

Throughout this paper, we are primarily interested in the weakening of this convexity condition to allow negative curvature given by ρ-weak convexity and ρ-weak concavity: we assume that L is twice differentiable, and for any $z = (x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ that

$$\nabla^2_{xx} L(z) \succeq -\rho I \ , \quad -\nabla^2_{yy} L(z) \succeq -\rho I \ . \tag{4}$$

Notice that the objective $L(x, y)$ is convex-concave when $\rho = 0$, and strongly convex-strongly concave when $\rho < 0$. Here our primary interest is in the regime where $\rho > 0$ is positive, quantifying how nonconvex-nonconcave the given problem instance is.

Algorithms for minimax problems. Besides PPM, Gradient Descent Ascent (GDA) is another classic algorithm for minimax problem (1). The update rule is given by

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - s \begin{bmatrix} \nabla_x L(x_k, y_k) \\ -\nabla_y L(x_k, y_k) \end{bmatrix} , \tag{5}$$

with stepsize parameter $s > 0$. However, GDA is known to work only for strongly convex-strongly concave minimax problems, and it may diverge even for simple convex-concave problems [8, 13].

In this paper, we study a more generalized algorithm, damped PPM, with damping parameter $\lambda \in (0, 1]$ and proximal parameter $\eta > 0$. The damped proximal point method updates by

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = (1 - \lambda) \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \lambda \operatorname{prox}_\eta(x_k, y_k) . \tag{6}$$

In particular, when $\lambda = 1$, we recover the traditional PPM (2). Interestingly, we find through our theory that some nonconvex-nonconcave problems only have the proximal point method converge when damping is employed (that is, $\lambda < 1$ strictly).

1.2 Related Literature.

There is a long history of research into convex-concave minimax optimization. Rockafellar [5] studies PPM for solving monotone variational inequalities, and shows that, as a special case, PPM converges to the stationary point linearly when $L(x, y)$ is strongly convex-strongly concave or when $L(x, y)$ is bilinear. Later on, Tseng [6] shows that EGM converges linearly to a stationary point under similar conditions. Nemirovski [7] shows that EGM approximates PPM and presents the sublinear rate of EGM. Recently, minimax problems have gained the attention of the machine learning community, perhaps due to the thriving of research on GANs. Daskalakis and Panageas [8] present an Optimistic Gradient Descent Ascent algorithm (OGDA) and shows that it converges linearly to the saddle-point when $L(x, y)$ is bilinear. Mokhtari et al. [14] show that OGDA is a different approximation to PPM. Lu [13] presents an ODE approach, which leads to unified conditions under which each algorithm converges, including a class of nonconvex-nonconcave problems.

There are also extensive studies on convex-concave minimax problems when the interaction term is bilinear (similar to our setting (1)). Some influential algorithms include Nesterov's smoothing [15], Douglas-Rachford splitting (a special case is Alternating Direction Method of Multipliers (ADMM)) [16, 17] and Primal-Dual Hybrid Gradient Method (PDHG) [18].

Recently, a number of works have been undertaken considering nonconvex-concave minimax problems. The basic technique is to first turn the minimax problem (1) to a minimization problem on $\Phi(x) = \max_y L(x, y)$, which is well-defined since $L(x, y)$ is concave in y, and then utilize the recent developments in nonconvex optimization [19, 20, 21, 22].

Unfortunately, the above technique cannot be extended to nonconvex-nonconcave setting, because $\Phi(x)$ is now longer tractable to compute (even approximately) as it is a nonconcave maximization problem itself. Indeed, the current understanding of nonconvex-nonconcave minimax problems is fairly limited, in particular compared with the growing literature on nonconvex optimization.

The recent research on nonconvex-nonconcave minimax problems mostly relies on some form of convex-concave-like assumptions, such as Minty's Variational Inequality [23] and Polyak-Lojasiewicz conditions [24, 25], which are strong in general and successfully bypass the inherent difficulty in the nonconvex-nonconcave setting. Such theory, unfortunately, presupposes the existence of a globally attractive solution. As such, fundamental nonconvex-nonconcave structures like local solutions and cycling are prohibited.

In an early version of this work [26], we presented preliminary results for analyzing nonconvex-nonconcave bilinear problem (3). Simultaneous to (or after) the early version, [27] presents examples of nonconvex-nonconcave minimax problems where a reasonably large class of algorithms do not converge; [28] presents an ODE analysis for the limiting behaviors of different algorithms with shrinking step-size (equivalently it studies the ODE when step-size of an algorithm goes to 0) and shows the possibility to converge to an attractive circle; [29] utilizes tools from discrete-time dynamic systems to study the behaviors of algorithms around a local stationary solution, which involves the non-transparent complex eigenvalues of the Jacobian matrix at a stationary solution; [11] studies higher-order resolution ODEs of different algorithms for nonconvex-nonconcave minimax problems, which presents more transparent conditions for when a stationary solution is locally attractive, and characterizes the threshold of phase transitions between limit cycles and limit points. Compared to these recent works, we introduce new theoretical machinery, the saddle envelope, to directly analyze nonconvex-nonconcave minimax problems, which enable us to obtain a global understanding of the trajectories, as well as more transparent conditions under which the methods converges/diverges.

The idea to utilize a generalization of the Moreau envelope for nonconvex-nonconcave minimax problems is well motivated by the nonconvex optimization literature. In recent years, the Moreau envelope has found great success as an analysis tool in nonsmooth nonconvex optimization [30, 31, 32] and in nonconvex-concave optimization [21]. There, the Moreau envelope provides an angle of attack for describing stationarity in settings where gradients need not converge to zero even as first-order methods converge. Although we identify a different primary barrier (PPM may not converge at all as cycling and divergence arise from reasonable instances), we still find the Moreau envelope provides the key insight. In our setting, the critical finding is that the saddle envelope can convexify and concavify nonconvex-nonconcave problems, which has no parallel in the classic Moreau envelope.

1.3 Preliminaries

Review of convex-concave saddle point optimization. Strongly convex-strongly concave minimax optimization problems $\min_x \max_y M(x,y)$ are well understood. The following lemma is key to the convergence of gradient descent ascent on these problems. In the language of monotone operators, this lemma corresponds to showing $F(x,y) := (\nabla_x M(x,y), -\nabla_y M(x,y))$ is locally strongly monotone (or coercive). From this, the subsequent theorem below shows that gradient descent ascent contracts towards a stationary point when strong convexity-strong concavity and smoothness hold in a region around it. Proofs of these two standard results are given in Appendix B for completeness.

Lemma 1.1. *Suppose $M(x,y)$ is μ-strongly convex-strongly concave on a convex set $S = S_x \times S_y$, then it holds for any $(x,y),(x',y') \in S$ that*

$$\mu \left\| \begin{bmatrix} x - x' \\ y - y' \end{bmatrix} \right\|^2 \le \left(\begin{bmatrix} \nabla_x M(x,y) \\ -\nabla_y M(x,y) \end{bmatrix} - \begin{bmatrix} \nabla_x M(x',y') \\ -\nabla_y M(x',y') \end{bmatrix} \right)^T \begin{bmatrix} x - x' \\ y - y' \end{bmatrix}.$$

In particular, when $\nabla M(x', y') = 0$, the distance to this stationary point is bounded by

$$\left\| \begin{bmatrix} x - x' \\ y - y' \end{bmatrix} \right\| \leq \frac{\|\nabla M(x, y)\|}{\mu} .$$

Theorem 1.2. *Consider any minimax problem $\min_{x\in\mathbb{R}^n} \max_{y\in\mathbb{R}^m} M(x, y)$ where $M(x, y)$ is β-smooth and μ-strongly convex-strongly concave on a set $B(x_0, r)\times B(y_0, r)$ with $r \geq 2\|\nabla M(x_0, y_0)\|/\mu$. Then GDA (5) with initial solution (x_0, y_0) and step-size $s \in (0, 2\mu/\beta^2)$ linearly converges to a stationary point $(x^*, y^*) \in B((x_0, y_0), r/2)$ with*

$$\left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 \leq \left(1 - 2\mu s + \beta^2 s^2\right)^k \left\| \begin{bmatrix} x_0 - x^* \\ y_0 - y^* \end{bmatrix} \right\|^2 .$$

Review of the Moreau envelope's properties. Denote the Moreau envelope of a function f with proximal parameter $\eta > 0$ by

$$e_\eta \left\{f\right\} (x) = \min_u \ f(u) + \frac{\eta}{2}\|u - x\|^2 . \tag{7}$$

The Moreau envelope of a function provides a lower bound on it everywhere as all $x \in \mathbb{R}^n$ have

$$e_\eta\{f\}(x) \leq f(x) \tag{8}$$

and if f is ρ-weakly convex and $\eta > \rho$, these functions are equal at the stationary points of f

$$e_\eta\{f\}(x^*) = f(x^*) \quad \iff \quad \nabla f(x^*) = 0 . \tag{9}$$

Moreover, for ρ-weakly convex functions, there is a nice calculus for the Moreau envelope. Its gradient at some $x \in \mathbb{R}^n$ is determined by the proximal step $x_+ = \operatorname{argmin}_u f(x) + \frac{\eta}{2}\|u - x\|^2$ having

$$\nabla e_\eta \left\{f\right\} (x) = \eta(x - x_+) = \nabla f(x_+) . \tag{10}$$

For twice differentiable f, the Moreau envelope is twice differentiable as well with Hessian

$$\nabla^2 e_\eta\{f\}(x) = \eta I - (\eta I + \nabla^2 f(x_+))^{-1} . \tag{11}$$

From this formula, we can extract the following bounds related to smoothness and convexity

$$(\eta^{-1} - \rho^{-1})^{-1} I \preceq \nabla^2 e_\eta\{f\}(x) \preceq \eta I . \tag{12}$$

These bounds ensure the Moreau envelope is has a $\max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\}$-Lipschitz gradient, which simplifies for convex f (that is, $\rho \leq 0$) to have an η-Lipschitz gradient. Noting that $(\eta^{-1} - \rho^{-1})^{-1}$ always has the same sign as $-\rho$, we see that the Moreau envelope is (strongly/weakly) convex exactly when the given function f is (strongly/weakly) convex.

2 The Saddle Envelope

In this section, we develop a broad theory characterizing our core proof tool and innovation, the saddle envelope. Formally, the saddle envelope with proximal parameter $\eta > 0$ is defined as

$$L_\eta(x,y) := \min_{u\in\mathbb{R}^n} \max_{v\in\mathbb{R}^m} L(u,v) + \frac{\eta}{2}\|u-x\|^2 - \frac{\eta}{2}\|v-y\|^2 \ . \tag{13}$$

Throughout our theory, we require that the parameter η is selected with $\eta > \rho$, which ensures the minimax problem in (13) is strongly convex-strongly concave. As a result, the saddle envelope is well-defined (as its subproblem has a unique minimax point) and often can be efficiently approximated.

The saddle envelope generalizes the Moreau envelope from the minimization literature to minimax problems. To see this reduction, taking any objective $L(x,y) = g(x)$ (that is, one constant with respect to y) immediately recovers the Moreau envelope

$$L_\eta(x,y) = \min_u \max_v \ g(u) + \frac{\eta}{2}\|u-x\|^2 - \frac{\eta}{2}\|v-y\|^2 = e_\eta\{g\}(x) \ .$$

Throughout this section, we develop theory characterizing how the saddle envelope relates to L (in Section 2.1), the calculus of the saddle envelope (in Section 2.2), and the smoothing and convexifying effects of this operation (in Section 2.3). We take careful note throughout our theory of similarities and differences from the simpler case of Moreau envelopes.

2.1 Relationship between the Saddle Envelope L_η and the Objective L

We begin by considering how the value of the Saddle Envelope L_η relates to the origin objective L. Recall from (8), the Moreau envelope $e_\eta\{f\}$ provides a lower bound on f everywhere

$$e_\eta\{f\}(x) \le f(x) \ , \quad \text{for all } x \in \mathbb{R}^n$$

and from (9), we know they must agree at their shared set of minimizers

$$e_\eta\{f\}(x^*) = f(x^*) \ , \quad \text{for all } x^* \in \operatorname{argmin} f = \operatorname{argmin} e_\eta\{f\} \ .$$

Sadly, neither of these properties carry over to the saddle envelope in such generality.

The saddle envelope fails to provide a lower bound like (8). Intuitively, this is due to minimax optimization having a mixture of increasing and decreasing forces. If the objective function is constant with respect to y, having $L(x,y) = g(x)$, the saddle envelope becomes a Moreau envelope and provides a lower bound for every (x,y),

$$L_\eta(x,y) = \min_u \max_v g(u) + \frac{\eta}{2}\|u-x\|^2 - \frac{\eta}{2}\|v-y\|^2 = e_\eta\{g\}(x) \le g(x) = L(x,y) \ .$$

Conversely, if the objective is constant in x, having $L(x,y) = h(y)$, then it provides an upper bound

$$L_\eta(x,y) = \min_u \max_v h(v) + \frac{\eta}{2}\|u-x\|^2 - \frac{\eta}{2}\|v-y\|^2 = -e_\eta\{-h\}(y) \ge h(y) = L(x,y) \ .$$

In generic settings between these extremes, the saddle envelope L_η need not provide any kind of bound on L. The only generic relationship we can establish between the values of L and L_η everywhere is that as $\eta \to \infty$, they approach each other. If the objective L is uniformly Lipschitz, we can establish a constant bound on how far above or below L the saddle envelope L_η can be.

Proposition 2.1. *The saddle envelope has*

$$\lim_{\eta\to\infty} L_\eta(z) - L(z) = 0 \ , \quad \textit{for all } z \in \mathbb{R}^n \times \mathbb{R}^m$$

and if the objective function L is l-Lipschitz continuous, this difference is at most $O(1/\eta)$

$$|L_\eta(z) - L(z)| \leq \frac{(3\eta - 2\rho)l^2}{2(\eta-\rho)^2} \ , \quad \textit{for all } z \in \mathbb{R}^n \times \mathbb{R}^m \ .$$

Proof. Consider the $(\eta-\rho)$-strongly convex-strongly concave proximal subproblem $M(u,v) = L(u,v) + \frac{\eta}{2}\|u-x\|^2 - \frac{\eta}{2}\|v-y\|^2$ at some $z=(x,y)$. Noting that $\nabla M(z) = \nabla L(z)$, Lemma 1.1 bounds the distance from z to $z_+ = \mathrm{prox}_\eta(z)$ as

$$\|z_+ - z\| \leq \|\nabla L(z)\|/(\eta-\rho) \ .$$

Then continuity of the objective function gives the claimed limit as

$$\lim_{\eta\to\infty} L_\eta(z) = \lim_{\eta\to\infty} L(z_+) + \frac{\eta}{2}\|x_+ - x\|^2 - \frac{\eta}{2}\|y_+ - y\|^2 = L(z) + \lim_{\eta\to\infty} \frac{\eta\|\nabla L(z)\|}{(\eta-\rho)^2} = L(z) \ .$$

If L is l-Lipschitz continuous, then $\|\nabla L(z)\| \leq l$ everywhere. Applying the constant bound $\|z_+ - z\| \leq l/(\eta-\rho)$ lets us bound the difference between L and L_η as decreasing with $O(1/\eta)$

$$\begin{aligned} |L(z) - L_\eta(z)| &= |L(z) - L(z_+) - \frac{\eta}{2}\|x_+ - x\|^2 + \frac{\eta}{2}\|y_+ - y\|^2| \\ &\leq |L(z) - L(z_+)| + \left|\frac{\eta}{2}\|x_+ - x\|^2 - \frac{\eta}{2}\|y_+ - y\|^2\right| \\ &\leq l\|z - z_+\| + \frac{\eta}{2}\|z_+ - z\|^2 \\ &\leq \frac{l^2}{\eta-\rho} + \frac{\eta l^2}{2(\eta-\rho)^2} = \frac{(3\eta-2\rho)l^2}{2(\eta-\rho)^2} \ . \end{aligned}$$

□

Unlike the lower bound relationship (8), the saddle envelope does satisfy a version of the optimality relationship (9), albeit somewhat weakened. For generic L, the following proposition shows the saddle envelope provides a lower bound on the original problem's minimax solution. When L is convex-concave, we can improve this result to recover (9), showing both functions have the same minimax solutions and minimax objective values.

Proposition 2.2. *The saddle envelope has optimal objective value bounded by*

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m} L(x,y) \geq \min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m} L_\eta(x,y) \geq \max_{y\in\mathbb{R}^m}\min_{x\in\mathbb{R}^n} L_\eta(x,y) \geq \max_{y\in\mathbb{R}^m}\min_{x\in\mathbb{R}^n} L(x,y) \ .$$

Proof. Consider an augmented version of (1) with the addition of two dummy variables u and v as

$$\min_x \max_y L(x,y) = \min_x \min_u \max_y \max_v L(u,v) + \frac{\eta}{2}\|u-x\|^2 - \frac{\eta}{2}\|v-y\|^2 \ .$$

Equality holds here since the minimum value over x always occurs at $x=u$ and the maximum value over y always occurs at $y=v$. Interchanging the middle minimization and maximization operations can only decrease the objective value. Hence we have the claimed inequality

$$\begin{aligned}\min_x \max_y L(x,y) &= \min_x \min_u \max_y \max_v L(u,v) + \frac{\eta}{2}\|u-x\|^2 - \frac{\eta}{2}\|v-y\|^2 \\ &\geq \min_x \max_y \min_u \max_v L(u,v) + \frac{\eta}{2}\|u-x\|^2 - \frac{\eta}{2}\|v-y\|^2 \\ &= \min_x \max_y L_\eta(x,y) \ .\end{aligned}$$

Exchanging the remaining minimum and maximum can only further decrease the optimal value

$$\min_x \max_y L_\eta(x,y) \geq \max_y \min_x L_\eta(x,y) \ .$$

Reexpanding the definition of L_η and again interchanging the middle minimum and maximum gives our final claimed inequality

$$\begin{aligned}\max_y \min_x L_\eta(x,y) &= \max_y \min_x \max_v \min_u L(u,v) + \frac{\eta}{2}\|u-x\|^2 - \frac{\eta}{2}\|v-y\|^2 \\ &\geq \max_y \max_v \min_x \min_u L(u,v) + \frac{\eta}{2}\|u-x\|^2 - \frac{\eta}{2}\|v-y\|^2 \\ &= \max_y \min_x L(x,y) \ .\end{aligned}$$

□

Remark 2.3. *The inequalities in Proposition 2.2 all arise from exchanging the order of a minimum and a maximum. Famously, the minimax theorem tells us minimums and maximums can be exchanged without changing the solution set or optimal value whenever the given objective is convex-concave and satisfies a modest regularity condition (typically that the domain of one of the variables is compact [33], but in our unconstrained setting of* (1)*, strong convexity or strong concavity in one of the variables would suffice). Thus the four optimization problems considered in Proposition 2.2 are, in fact, all equivalent in more classic settings than our nonconvex-nonconcave focus here.*

Remark 2.4. *When viewed as a sequential game, the saddle envelope can be seen as adding a quadratically penalized recourse phase:*

$$\min_x \max_y L_\eta(x,y) = \min_x \max_y \min_u \max_v L(u,v) + \frac{\eta}{2}\|u-x\|^2 - \frac{\eta}{2}\|v-y\|^2 \ .$$

After the minimizing agent declares x and then the maximizing agent declares y, the players can change their strategies to any u and v, paying a penalty based on the square of the distance moved. Adding this recourse naturally favors the minimizing agent, who previously had to select their strategy with no information about the maximizing agent. This explains the direction of our inequality

$$\min_{x\in\mathbb{R}^n} \max_{y\in\mathbb{R}^m} L(x,y) \geq \min_{x\in\mathbb{R}^n} \max_{y\in\mathbb{R}^m} L_\eta(x,y) \ .$$

2.2 Calculus for the Saddle Envelope L_η

Here we develop a calculus for the saddle envelope L_η, giving formulas for its gradient and Hessian in terms of the original objective L and the proximal operator. These results immediately give algorithmic insights into the behavior of the proximal point method. First, we show that a direct generalization of the Moreau envelope gradient formula (10) holds, establishing that the gradient of the saddle envelope is entirely determined by the proximal operator.

Lemma 2.5. *The gradient of the saddle envelope $L_\eta(x,y)$ at $z=(x,y)$ is given by*

$$\begin{bmatrix}\nabla_x L_\eta(z)\\ \nabla_y L_\eta(z)\end{bmatrix} = \begin{bmatrix}\eta(x-x_+)\\ \eta(y_+-y)\end{bmatrix} = \begin{bmatrix}\nabla_x L(z_+)\\ \nabla_y L(z_+)\end{bmatrix}$$

where $z_+=(x_+,y_+)=\mathrm{prox}_\eta(z)$ is given by the proximal operator.

Proof. Notice that the saddle envelope can be described as the composition of Moreau envelopes

$$\begin{aligned} L_\eta(x,y) &= \min_u \left(\max_v L(u,v) - \frac{\eta}{2}\|v-y\|^2\right) + \frac{\eta}{2}\|u-x\|^2 \\ &= \min_u \ -e_\eta\{-L(u,\cdot)\}(y) + \frac{\eta}{2}\|u-x\|^2 \\ &= e_\eta\{g(\cdot,y)\}(x) \end{aligned}$$

where $g(u,y) = -e_\eta\{-L(u,\cdot)\}(y)$. Applying the gradient formula for Moreau envelopes (10) gives our first claimed gradient formula in x of $\nabla_x L_\eta(x,y) = \eta(x-x_+)$ since x_+ is the unique minimizer of $u \mapsto g(u,y) + \frac{\eta}{2}\|u-x\|^2$. Likewise, exchanging the minimum and maximum defining the saddle envelope (since $\eta > \rho$ ensures the problem is convex-concave), gives the composition

$$\begin{aligned} L_\eta(x,y) &= \max_v \left(\min_u L(u,v) + \frac{\eta}{2}\|u-x\|^2\right) - \frac{\eta}{2}\|v-y\|^2 \\ &= \max_v \ e_\eta\{L(\cdot,v)\}(x) - \frac{\eta}{2}\|v-y\|^2 \\ &= -e_\eta\{-h(x,\cdot)\}(y) \end{aligned}$$

where $h(x,v) = e_\eta\{L(\cdot,v)\}(x)$. Then the Moreau envelope gradient formula (10) gives our first claimed gradient formula in y of $\nabla_y L_\eta(x,y) = \eta(y_+-y)$ since y_+ is the unique minimizer of $v \mapsto -h(x,v) + \frac{\eta}{2}\|v-y\|^2$. The second claimed equality is precisely the first-order optimality condition for (2). That is,

$$\begin{aligned} \nabla_x L(x_+,y_+) + \eta(x_+-x) = 0 \ , \\ -\nabla_y L(x_+,y_+) + \eta(y_+-y) = 0 \ . \end{aligned}$$

□

Corollary 2.6. *The stationary points of L_η are exactly the same as the stationary points of L.*

Proof. First consider any stationary point $z=(x,y)$ of L. Denote $z_+ = \mathrm{prox}_\eta(z)$ and the objective function defining the proximal operator (2) as $M(u,v) = L(u,v) + \frac{\eta}{2}\|u-x\|^2 - \frac{\eta}{2}\|v-y\|^2$. Then observing that $\nabla M(z)=0$, z must be the unique minimax point of M (that is, $z = z_+ = \mathrm{prox}_\eta(z)$). Hence z must be a stationary point of L_η as well since $\nabla L_\eta(z) = \nabla L(z_+) = \nabla L(z) = 0$.

Conversely consider a stationary point $z=(x,y)$ of L_η. Then $\eta(x-x_+) = \nabla_x L_\eta(z) = 0$ and $\eta(y_+-y) = \nabla_y L_\eta(z) = 0$. Hence we again find that $z = z_+ = \mathrm{prox}_\eta(z)$ and consequently, this point must be a stationary point of L as well since $\nabla L(z) = \nabla L(z_+) = \nabla L_\eta(z) = 0$. □

Corollary 2.7. *One step of the (potentially damped) PPM (6) on the original objective L is equivalent to one step of GDA (5) on the saddle envelope L_η with $s=\lambda/\eta$.*

Proof. Let $(x_k^+, y_k^+) = \mathrm{prox}_\eta(x_k,y_k)$ and let (x_{k+1},y_{k+1}) be a step of GDA on $L_\eta(x,y)$ from (x_k,y_k) with step-size $s=\lambda/\eta$. Then it follows from Lemma 2.5 that

$$\begin{bmatrix}x_{k+1}\\ y_{k+1}\end{bmatrix} = \begin{bmatrix}x_k\\ y_k\end{bmatrix} - s\begin{bmatrix}\nabla_x L_\eta(z_k)\\ -\nabla_y L_\eta(z_k)\end{bmatrix} = \begin{bmatrix}x_k\\ y_k\end{bmatrix} - \frac{\lambda}{\eta}\begin{bmatrix}\eta(x_k-x_k^+)\\ -\eta(y_k^+-y_k)\end{bmatrix} = (1-\lambda)z_k + \lambda\ \mathrm{prox}_\eta(z_k)\ .$$

□

Similar to our previous lemma, our next result shows that the Hessian of the saddle envelope at some z is determined by the Hessian of L at $z_+ = \mathrm{prox}_\eta(z)$. This formula generalizes the Moreau envelope's formula (11) by considering an example that is constant with respect to y (i.e., setting $\nabla^2_{xy}L(z) = 0$ and $\nabla^2_{yy}L(z) = 0$).

Lemma 2.8. *The Hessian of the saddle envelope $L_\eta(z)$ is given by*

$$\begin{bmatrix} \nabla^2_{xx}L_\eta(z) & \nabla^2_{xy}L_\eta(z) \\ -\nabla^2_{yx}L_\eta(z) & -\nabla^2_{yy}L_\eta(z) \end{bmatrix} = \eta I - \eta^2 \left(\eta I + \begin{bmatrix} \nabla^2_{xx}L(z_+) & \nabla^2_{xy}L(z_+) \\ -\nabla^2_{yx}L(z_+) & -\nabla^2_{yy}L(z_+) \end{bmatrix}\right)^{-1}$$

where $z_+ = \mathrm{prox}_\eta(z)$ is given by the proximal operator. Since $\eta > \rho$, we have

$$\nabla^2_{xx}L_\eta(z) = \eta I - \eta^2 \left(\eta I + \nabla_{xx}L(z_+) + \nabla^2_{xy}L(z_+)(\eta I - \nabla^2_{yy}L(z_+))^{-1}\nabla^2_{yx}L(z_+)\right)^{-1},$$

$$\nabla^2_{yy}L_\eta(z) = -\eta I + \eta^2 \left(\eta I + \nabla_{yy}L(z_+) + \nabla^2_{yx}L(z_+)(\eta I + \nabla^2_{xx}L(z_+))^{-1}\nabla^2_{xy}L(z_+)\right)^{-1}.$$

Proof. Consider some $z = (x, y)$ and a nearby point $z^\Delta = z + \Delta$. Denote one proximal step from each of these points by $z_+ = (x_+, y_+) = \mathrm{prox}_\eta(z)$ and $z^\Delta_+ = (x^\Delta_+, y^\Delta_+) = \mathrm{prox}_\eta(z^\Delta)$. Then our claimed Hessian formula amounts to showing

$$\begin{bmatrix} \nabla_x L_\eta(z^\Delta) \\ -\nabla_y L_\eta(z^\Delta) \end{bmatrix} = \begin{bmatrix} \nabla_x L_\eta(z) \\ -\nabla_y L_\eta(z) \end{bmatrix} + \left(\eta I - \eta^2 \left(\eta I + \begin{bmatrix} \nabla^2_{xx}L(z_+) & \nabla^2_{xy}L(z_+) \\ -\nabla^2_{yx}L(z_+) & -\nabla^2_{yy}L(z_+) \end{bmatrix}\right)^{-1}\right) \begin{bmatrix} \Delta_x \\ \Delta_y \end{bmatrix} + o(\|\Delta\|).$$

Recall Lemma 2.5 showed the gradient of the saddle envelope is given by $\nabla_x L_\eta(z) = \eta(x - x_+)$ and $\nabla_y L_\eta(z) = \eta(y_+ - y)$. Applying this at z and z_+ and dividing η, our claimed Hessian formula becomes

$$\begin{bmatrix} x^\Delta_+ - x_+ \\ y^\Delta_+ - y_+ \end{bmatrix} = \eta \left(\eta I + \begin{bmatrix} \nabla^2_{xx}L(z_+) & \nabla^2_{xy}L(z_+) \\ -\nabla^2_{yx}L(z_+) & -\nabla^2_{yy}L(z_+) \end{bmatrix}\right)^{-1} \begin{bmatrix} \Delta_x \\ \Delta_y \end{bmatrix} + o(\|\Delta\|) . \tag{14}$$

Our proof shows this in two steps: first considering a proximal step on the second-order Taylor approximation of L at z_+ and then showing this closely matches the result of a proximal step on L.

First, consider the following quadratic model of the objective around z_+:

$$\widetilde{L}(z) = L(z_+) + \nabla L(z_+)^T(z - z_+) + \frac{1}{2}(z - z_+)^T\nabla^2 L(z_+)(z - z_+) .$$

Denote the result of one proximal step on $\widetilde{L}$ from z^Δ by $\widetilde{z}^\Delta_+ = (\widetilde{x}^\Delta_+, \widetilde{y}^\Delta_+)$. Since the proximal subproblem is strongly convex-strongly concave, this solution is uniquely determined by

$$\begin{bmatrix} \nabla_x \widetilde{L}(\widetilde{x}^\Delta_+, \widetilde{y}^\Delta_+) \\ -\nabla_y \widetilde{L}(\widetilde{x}^\Delta_+, \widetilde{y}^\Delta_+) \end{bmatrix} + \begin{bmatrix} \eta(\widetilde{x}^\Delta_+ - x^\Delta) \\ \eta(\widetilde{y}^\Delta_+ - y^\Delta) \end{bmatrix} = 0 .$$

Plugging in the definition of our quadratic model $\widetilde{L}$ yields

$$\begin{aligned}
&\begin{bmatrix}\nabla_x L(z_+)\\ -\nabla_y L(z_+)\end{bmatrix} + \begin{bmatrix}\nabla^2_{xx}L(z_+) & \nabla^2_{xy}L(z_+)\\ -\nabla^2_{yx}L(z_+) & -\nabla^2_{yy}L(z_+)\end{bmatrix}\begin{bmatrix}\widetilde{x}^\Delta_+ - x_+\\ \widetilde{y}^\Delta_+ - y_+\end{bmatrix} + \begin{bmatrix}\eta(\widetilde{x}^\Delta_+ - x^\Delta)\\ \eta(\widetilde{y}^\Delta_+ - y^\Delta)\end{bmatrix} = 0 ,\\
\implies &\left(\eta I + \begin{bmatrix}\nabla^2_{xx}L(z_+) & \nabla^2_{xy}L(z_+)\\ -\nabla^2_{yx}L(z_+) & -\nabla^2_{yy}L(z_+)\end{bmatrix}\right)\begin{bmatrix}\widetilde{x}^\Delta_+ - x_+\\ \widetilde{y}^\Delta_+ - y_+\end{bmatrix} = \eta\begin{bmatrix}x^\Delta - x_+ - \eta^{-1}\nabla_x L(z_+)\\ y^\Delta - y_+ + \eta^{-1}\nabla_y L(z_+)\end{bmatrix},\\
\implies &\left(\eta I + \begin{bmatrix}\nabla^2_{xx}L(z_+) & \nabla^2_{xy}L(z_+)\\ -\nabla^2_{yx}L(z_+) & -\nabla^2_{yy}L(z_+)\end{bmatrix}\right)\begin{bmatrix}\widetilde{x}^\Delta_+ - x_+\\ \widetilde{y}^\Delta_+ - y_+\end{bmatrix} = \eta\begin{bmatrix}\Delta_x\\ \Delta_y\end{bmatrix},\\
\implies &\begin{bmatrix}\widetilde{x}^\Delta_+ - x_+\\ \widetilde{y}^\Delta_+ - y_+\end{bmatrix} = \eta\left(\eta I + \begin{bmatrix}\nabla^2_{xx}L(z_+) & \nabla^2_{xy}L(z_+)\\ -\nabla^2_{yx}L(z_+) & -\nabla^2_{yy}L(z_+)\end{bmatrix}\right)^{-1}\begin{bmatrix}\Delta_x\\ \Delta_y\end{bmatrix}.
\end{aligned}$$

This is nearly our target condition (14). All that remains is to show our second-order approximation satisfies $\|z^\Delta_+ - \widetilde{z}^\Delta_+\| = o(\|\Delta\|)$. Denote the proximal subproblem objective by $M^\Delta(u,v) = L(u,v) + \frac{\eta}{2}\|u - x^\Delta\|^2 - \frac{\eta}{2}\|v - y^\Delta\|^2$ and its approximation by $\widetilde{M}^\Delta(u,v) = \widetilde{L}(u,v) + \frac{\eta}{2}\|u - x^\Delta\|^2 - \frac{\eta}{2}\|v - y^\Delta\|^2$. Noting that $\|\nabla\widetilde{M}^\Delta(x_+,y_+)\| = \eta\|\Delta\|$, we can apply Lemma 1.1 to the $(\eta-\rho)$-strongly convex-strongly concave function $\widetilde{M}^\Delta$ to bound the distance to its minimax point as

$$\|z_+ - \widetilde{z}^\Delta_+\| \le \frac{\eta}{\eta-\rho}\|\Delta\| \ .$$

Consequently, we can bound difference in gradients between L and its quadratic model $\widetilde{L}$ at $\widetilde{z}^\Delta_+$ by $\|\nabla L(\widetilde{z}^\Delta_+) - \nabla\widetilde{L}(\widetilde{z}^\Delta_+)\| = o(\|\Delta\|)$. Therefore $\|\nabla M^\Delta(\widetilde{z}^\Delta_+)\| = o(\|\Delta\|)$ and so applying Lemma 1.1 to the strongly convex-strongly concave function M^Δ bounds the distance to its minimax point as $\|z^\Delta_+ - \widetilde{z}^\Delta_+\| = o(\|\Delta\|)$, which completes our proof. □

A careful understanding of the saddle envelope's Hessian allows us to describe its smoothness and when it is convex-concave. This is carried out in the following section and forms the crucial step in enabling our convergence analysis for nonconvex-nonconcave problems.

2.3 Smoothing and Convexifing from the Saddle Envelope

Recall that the Moreau envelope $e_\eta\{f\}$ serves as a smoothing of any ρ-weakly convex function since has Hessian uniform bounds above and below (12). The lower bound on the Moreau envelope's Hessian guarantees it is convex exactly when the given function f is convex (that is, $\rho = 0$), and strongly convex if and only if f is strongly convex.

Our Hessian formula in Lemma 2.8 allows us to generalize this result to the saddle envelope. Outstandingly, we find that the minimax extension of this result is much more powerful than its Moreau counterpart. The saddle envelope will be convex-concave not just when L is convex-concave, but whenever the following interaction dominance condition holds with a nonnegative parameter α. This relationship is formalized in our next proposition.

Definition 2.9. *A function L is α-interaction dominate with respect to x if*

$$\nabla^2_{xx}L(z) + \nabla^2_{xy}L(z)(\eta I - \nabla^2_{yy}L(z))^{-1}\nabla^2_{yx}L(z) \succeq \alpha I \tag{15}$$

and α-interaction dominate with respect to y if

$$-\nabla^2_{yy}L(z) + \nabla^2_{yx}L(z)(\eta I + \nabla^2_{xx}L(z))^{-1}\nabla^2_{xy}L(z) \succeq \alpha I \ . \tag{16}$$

For any ρ-weakly convex-weakly concave function L, interaction dominance holds with $\alpha = -\rho$ since the second term in these definitions is always positive semidefinite. As a consequence, any convex-concave function is $\alpha \geq 0$-interaction dominate with respect to both x and y. Further, nonconvex-nonconcave functions are interaction dominate with $\alpha \geq 0$ when the second term above is sufficiently positive definite (hence the name "interaction dominate" as the interaction term of the Hessian $\nabla^2_{xy}L(z)$ is dominating any negative curvature in Hessians $\nabla^2_{xx}L(z)$ and $-\nabla^2_{yy}L(z)$). For example, any problem with β-Lipschitz gradient in y has interaction dominance in x hold with non-negative parameter whenever

$$\frac{\nabla^2_{xy}L(z)\nabla^2_{yx}L(z)}{\eta+\beta} \succeq -\nabla^2_{xx}L(z)$$

since $\eta I - \nabla^2_{yy}L(z) \preceq (\eta+\beta)I$. Similarly, any problem with β-Lipschitz gradient in x has interaction dominance in y with a non-negative parameter whenever

$$\frac{\nabla^2_{yx}L(z)\nabla^2_{xy}L(z)}{\eta+\beta} \succeq \nabla^2_{yy}L(z) \ .$$

The following proposition derives bounds on the Hessian of the saddle envelope showing it is convex in x (concave in y) whenever $\alpha \geq 0$-interaction dominance holds in x (in y). Further, its Hessian lower bounds ensure that L_η is $(\eta^{-1}+\alpha^{-1})^{-1}$-strongly convex in x (strongly concave in y) whenever $\alpha > 0$-interaction dominance holds in x (in y).

Proposition 2.10. *If the x-interaction dominance* (15) *holds with $\alpha \in \mathbb{R}$, the saddle envelope is smooth and weakly convex with respect to x*

$$(\eta^{-1}+\alpha^{-1})^{-1}I \preceq \nabla^2_{xx}L_\eta(z) \preceq \eta I \ ,$$

and if the y-interaction dominance condition (16) *holds with $\alpha \in \mathbb{R}$, the saddle envelope is smooth and weakly concave with respect to y*

$$(\eta^{-1}+\alpha^{-1})^{-1}I \preceq -\nabla^2_{yy}L_\eta(z) \preceq \eta I \ .$$

Proof. Recall the formula for the x component of the saddle envelope's Hessian given by Lemma 2.8. Then applying the interaction dominance condition (15) lets us lower bound this by

$$\begin{aligned}
\nabla^2_{xx}L_\eta(z) &= \eta I - \eta^2\left(\eta I + \nabla_{xx}L(z_+) + \nabla^2_{xy}L(z_+)(\eta I - \nabla^2_{yy}L(z_+))^{-1}\nabla^2_{yx}L(z_+)\right)^{-1} \\
&\succeq \eta I - \eta^2\left(\eta I + \alpha I\right)^{-1} \\
&= (\eta - \eta^2/(\eta+\alpha))I \\
&= (\eta^{-1}+\alpha^{-1})^{-1}I \ .
\end{aligned}$$

Note that $\eta I + \nabla_{xx}L(z_+)$ is positive definite (since $\eta > \rho$) and $\nabla^2_{xy}L(z_+)(\eta I + \nabla^2_{yy}L(z_+))^{-1}\nabla^2_{yx}L(z_+)$ is positive semidefinite (since its written as a square). Then the inverse of their sum must also be positive definite and consequently

$$\nabla^2_{xx}L_\eta(z) = \eta I - \eta^2\left(\eta I + \nabla_{xx}L(z_+) + \nabla^2_{xy}L(z_+)(\eta I + \nabla^2_{yy}L(z_+))^{-1}\nabla^2_{yx}L(z_+)\right)^{-1} \preceq \eta I \ .$$

Symmetric reasoning applies to give the upper and lower bounds on $\nabla^2_{yy}L_\eta(z)$. □

Remark 2.11. *Note that our definition of interaction dominance depends on the choice of the proximal parameter $\eta > \rho$. In our convergence theory, we will show that interaction dominance with nonnegative $\alpha > 0$ captures when the proximal point method with the same parameter η converges.*

Remark 2.12. *Proposition 2.10 generalizes the Hessian bounds for the Moreau envelope* (12) *since for any $L(x, y)$ that is constant in y, the α-interaction dominance condition in x simplifies to simply be ρ-weak convexity $\nabla^2_{xx}L(z) + \nabla^2_{xy}L(z)(\eta I - \nabla^2_{yy}L(z))^{-1}\nabla^2_{yx}L(z) = \nabla^2_{xx}L(z) \succeq \alpha I$. Hence this special case has $\alpha = -\rho$ and so our theory recovers the Moreau envelope's bounds.*

In addition to bounding the Hessians of the x and y variables separately, we can also bound the overall smoothness of the saddle envelope. In the case of Moreau envelopes, the Hessian formula (12) ensures the envelope is $\max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\}$-smooth. Our next result shows that the saddle envelope maintains the smoothing nature of the Moreau envelope, possessing a uniformly $\max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\}$-Lipschitz gradient regardless of the smoothness of the original function or lack thereof.

Proposition 2.13. *The saddle envelope has a uniformly Lipschitz gradient with constant*

$$\max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\} \ .$$

Proof. Consider two points $z = (x, y)$ and $\bar{z} = (\bar{x}, \bar{y})$ and denote one proximal step from each of them by $z_+ = (x_+, y_+) = \mathrm{prox}_\eta(z)$ and $\bar{z}_+ = (\bar{x}_+, \bar{y}_+) = \mathrm{prox}_\eta(\bar{z})$. Define the $(\eta - \rho)$-strongly convex-strongly concave function underlying the computation of the saddle envelope at z as

$$M(u, v) = L(u, v) + \frac{\eta}{2}\|u - x\|^2 - \frac{\eta}{2}\|v - y\|^2.$$

First we compute the gradient of M at $\bar{z}_+$ which is given by

$$\begin{bmatrix} \nabla_x M(\bar{z}_+) \\ \nabla_y M(\bar{z}_+) \end{bmatrix} = \begin{bmatrix} \nabla_x L(\bar{z}_+) + \eta(\bar{x}_+ - x) \\ \nabla_y L(\bar{z}_+) - \eta(\bar{y}_+ - y) \end{bmatrix} = \begin{bmatrix} \nabla_x L(\bar{z}_+) + \eta(\bar{x}_+ - \bar{x} + \bar{x} - x) \\ \nabla_y L(\bar{z}_+) - \eta(\bar{y}_+ - \bar{y} + \bar{y} - y) \end{bmatrix} = \eta \begin{bmatrix} \bar{x} - x \\ y - \bar{y} \end{bmatrix}.$$

Then applying Lemma 1.1, and noting that $z_+ = \mathrm{prox}_\eta(z)$ has $\nabla M(z_+) = 0$, we conclude that

$$\left\| \begin{bmatrix} \bar{x}_+ - x_+ \\ \bar{y}_+ - y_+ \end{bmatrix} \right\|^2 \leq \frac{\eta}{\eta - \rho} \begin{bmatrix} \bar{x} - x \\ \bar{y} - y \end{bmatrix}^T \begin{bmatrix} \bar{x}_+ - x_+ \\ \bar{y}_+ - y_+ \end{bmatrix}.$$

Recall from Lemma 2.5, the gradients of the saddle envelope are given by

$$\nabla L_\eta(z) = \eta \begin{bmatrix} x - x_+ \\ y_+ - y \end{bmatrix} \qquad \text{and} \qquad \nabla L_\eta(\bar{z}) = \eta \begin{bmatrix} \bar{x} - \bar{x}_+ \\ \bar{y}_+ - \bar{y} \end{bmatrix}.$$

Then we can upper bound the difference in gradients of the saddle envelope by

$$\begin{aligned} \frac{1}{\eta^2}\|\nabla L_\eta(z) - \nabla L_\eta(\bar{z})\|^2 &= \left\| \begin{bmatrix} x - x_+ \\ y_+ - y \end{bmatrix} - \begin{bmatrix} \bar{x} - \bar{x}_+ \\ \bar{y}_+ - \bar{y} \end{bmatrix} \right\|^2 \\ &= \left\| \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix} \right\|^2 + 2 \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix}^T \begin{bmatrix} \bar{x}_+ - x_+ \\ \bar{y}_+ - y_+ \end{bmatrix} + \left\| \begin{bmatrix} \bar{x}_+ - x_+ \\ \bar{y}_+ - y_+ \end{bmatrix} \right\|^2 \\ &\leq \left\| \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix} \right\|^2 + \left(\frac{\eta}{\eta - \rho} - 2 \right) \begin{bmatrix} \bar{x} - x \\ \bar{y} - y \end{bmatrix}^T \begin{bmatrix} \bar{x}_+ - x_+ \\ \bar{y}_+ - y_+ \end{bmatrix}. \end{aligned}$$

Notice that $\begin{bmatrix}\bar{x}-x\\ \bar{y}-y\end{bmatrix}^T \begin{bmatrix}\bar{x}_+-x_+\\ \bar{y}_+-y_+\end{bmatrix}$ is non-negative but the sign of $\left(\frac{\eta}{\eta-\rho}-2\right)$ may be positive or negative. If this coefficient is negative, we can upperbound the second term above by zero, giving a smoothness constant of η as

$$\|\nabla L_\eta(z)-\nabla L_\eta(\bar{z})\|^2 \leq \eta^2 \left\|\begin{bmatrix}x-\bar{x}\\ y-\bar{y}\end{bmatrix}\right\|^2 .$$

If instead, $\left(\frac{\eta}{\eta-\rho}-2\right)$ is non-negative, then we arrive at a smoothness constant of $|\eta^{-1}-\rho^{-1}|^{-1}$ by the Cauchy–Schwarz inequality and using that $\left\|\begin{bmatrix}\bar{x}_+-x_+\\ \bar{y}_+-y_+\end{bmatrix}\right\| \leq \frac{\eta}{\eta-\rho}\left\|\begin{bmatrix}\bar{x}-x\\ \bar{y}-y\end{bmatrix}\right\|$:

$$\begin{aligned}\|\nabla L_\eta(z)-\nabla L_\eta(\bar{z})\|^2 &\leq \eta^2\left(1+\left(\frac{\eta}{\eta-\rho}-2\right)\frac{\eta}{\eta-\rho}\right)\left\|\begin{bmatrix}x-\bar{x}\\ y-\bar{y}\end{bmatrix}\right\|^2\\ &= \eta^2\left(\frac{\eta}{\eta-\rho}-1\right)^2\left\|\begin{bmatrix}x-\bar{x}\\ y-\bar{y}\end{bmatrix}\right\|^2\\ &= \left(\frac{\eta\rho}{\eta-\rho}\right)^2\left\|\begin{bmatrix}x-\bar{x}\\ y-\bar{y}\end{bmatrix}\right\|^2 .\end{aligned}$$

□

The setting of taking the Moreau envelope of a convex function gives a simpler smoothness bound of η since having $\rho \leq 0$ implies $\eta = \max\{\eta, |\eta^{-1}-\rho^{-1}|^{-1}\}$. The same simplification holds when applying our saddle envelope machinery to convex-concave problems: the saddle envelope of any convex-concave L is η-smooth.

3 Interaction Dominate Regime

Our theory for the saddle envelope $L_\eta(z)$ shows it is much more structured than the original objective function $L(z)$. Proposition 2.10 established that for x and y interaction dominant problems, the saddle envelope is strongly convex-strongly concave. Proposition 2.13 established that the saddle envelope is always smooth (has a uniformly Lipschitz gradient). Both of these results hold despite us not assuming convexity, concavity, or smoothness of the original objective. Historically these two conditions are the key to linear convergence (see Theorem 1.2) and indeed we find interaction dominance causes the proximal point method to linearly converge. The proof of this result is deferred to the end of the section.

Theorem 3.1. *For any objective L that is ρ-weakly convex-weakly concave and $\alpha > 0$-interaction dominate in both x and y, the damped PPM (6) with η and λ satisfying*

$$\lambda \leq 2\frac{\min\{1, (\eta/\rho-1)^2\}}{\eta/\alpha+1}$$

linearly converges to the unique stationary point (x^*, y^*) of (1) with

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

For example, setting $\eta = 2\rho$ and $\lambda = \frac{1}{1+\eta/\alpha}$, our convergence rate simplifies to

$$\left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 \leq \left(1 - \frac{1}{(2\rho/\alpha+1)^2}\right)^k \left\| \begin{bmatrix} x_0 - x^* \\ y_0 - y^* \end{bmatrix} \right\|^2 .$$

Remark 3.2. *Theorem 3.1 is valid even if $\alpha > 0$-interaction dominance only holds locally. That is, as long as α-interaction dominance holds within an l_2-ball around a local stationary point, and the initial point is sufficiently within this ball, then PPM converges linearly to this local stationary point. Of course in this case there can be many local stationary points.*

Remark 3.3. *For μ-strongly convex-strongly concave problems, this theorem recovers the standard proximal point convergence rate for any choice of $\eta > 0$. In this case, we have $\rho = -\mu$, $\alpha = \mu$, and can set $\lambda = \frac{1}{\eta/\mu+1}$, giving a $O(\eta^2/\mu^2 \log(1/\varepsilon))$ convergence rate matching [5].*

Remark 3.4. *The $\alpha > 0$-interaction dominance condition is tight for obtaining global linear convergence. A nonconvex-nonconcave quadratic example illustrating the sharpness of this boundary is presented is Section 5.1. Moreover, our example shows that it is necessary to utilize the damping parameter (that is, selecting $\lambda < 1$) for the proximal point method to converge for some $\alpha > 0$-interaction dominant problems.*

If we only have $\alpha > 0$-interaction dominance with y, then the saddle envelope L_η is still much more structured than the original objective L. In this case, $L_\eta(x, y)$ may still be nonconvex in x, but Proposition 2.13 ensures it is strongly concave in y. Then our theory allows us to extend existing convergence guarantees for nonconvex-concave problems to this larger class of y interaction dominate problems.

Nonconvex-concave problems have been considered by numerous recent works giving first-order methods that converge to stationary points [19, 20, 21, 22]. For example, Lin et al. [19] recently showed that GDA with different, carefully chosen stepsize parameters for x and y will converge to a stationary point at a rate of $O(\varepsilon^{-2})$. We find that running the following damped proximal point method is equivalent to running their variant of GDA on the saddle envelope

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} \lambda x_k^+ + (1-\lambda)x_k \\ \gamma y_k^+ + (1-\gamma)y_k \end{bmatrix} \text{ where } \begin{bmatrix} x_k^+ \\ y_k^+ \end{bmatrix} = \mathrm{prox}_\eta(x_k, y_k) \tag{17}$$

for proper choice of the parameters $\lambda, \gamma \in [0, 1]$. From this, we derive the following sublinear convergence rate for nonconvex-nonconcave problems whenever y interaction dominance holds. The proof of this result is deferred to the end of the section.

Theorem 3.5. *For any objective L that is ρ-weakly convex-weakly concave and $\alpha > 0$-interaction dominate in y, consider the PPM variant (17) with damping constants $\lambda = \Theta\left(\frac{\min\{1, |\eta/\rho - 1|^3\}}{(1+\eta/\alpha)^2}\right)$ and $\gamma = \Theta(\min\{1, |\eta/\rho - 1|\})$. If the sequence y_k is bounded*[1]*, then a stationary point $\|\nabla L(x_T^+, y_T^+)\| \leq \varepsilon$ will be found by iteration $T \leq O(\varepsilon^{-2})$.*

Remark 3.6. *Symmetrically, we can guarantee sublinear convergence assuming only x-interaction dominance. Considering the maximin problem of $\max_y \min_x L(x, y) = -\min_y \max_x -L(x, y)$, which is now interaction dominate with respect to the inner maximization variable, we can apply Theorem 3.5. This reduction works since although the original minimax problem and this maximin problem need not have the same solutions, they always have the same stationary points.*

[1] We do not believe this boundedness condition is fundamentally needed, but we make it to leverage the results of [19] which utilize compactness.

3.1 Proof of Theorem 3.1

Propositions 2.10 and 2.13 show that L_η is $\mu = (\eta^{-1} + \alpha^{-1})^{-1}$-strongly convex-strongly concave and has a $\beta = \max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\}$-Lipschitz gradient. Having strong convexity and strong concavity ensures L_η has a unique stationary point (x^*, y^*), which in turn must be the unique stationary point of L by Corollary 2.6. Recall Corollary 2.7 showed that the damped PPM (6) on L is equivalent to GDA (5) with $s = \lambda/\eta$ on L_η. Then provided

$$\lambda \le 2\frac{\min\left\{1, (\eta/\rho - 1)^2\right\}}{\eta/\alpha + 1} = \frac{2(\eta^{-1} + \alpha^{-1})^{-1}}{\max\{\eta^2, (\eta^{-1} - \rho^{-1})^{-2}\}} ,$$

we have $s = \lambda/\eta \in (0, 2\mu/\beta^2)$. Hence applying Theorem 1.2 shows the iterations of GDA (and consequently PPM) linearly converge to this unique stationary point as

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

3.2 Proof of Theorem 3.5

Proposition 2.10 shows that whenever interaction dominance holds for y the saddle envelope is $\mu = (\eta^{-1} + \alpha^{-1})^{-1}$-strongly concave in y and Proposition 2.13 ensures the saddle envelope has a $\beta = \max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\}$-Lipschitz gradient. Recently, Lin et al. [19] considered such nonconvex-strongly concave problems with a compact constraint $y \in D$. They analyzed the following variant of GDA

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \mathrm{proj}_{\mathbb{R}^n \times D}\left(\begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} -\nabla_x L(x_k, y_k)/\eta_x \\ \nabla_y L(x_k, y_k)/\eta_y \end{bmatrix}\right) \tag{18}$$

which projects onto the feasible region $\mathbb{R}^n \times D$ each iteration and has different stepsize parameters η_x and η_y for x and y. Lin et al. prove the following theorem showing a sublinear guarantee.

Theorem 3.7 (Theorem 4.4 of [19]). *For any β-smooth, nonconvex-μ-strongly concave L, let $\kappa = \beta/\mu$ be the condition number for y. Then for any $\varepsilon > 0$, GDA with stepsizes $\eta_x^{-1} = \Theta(1/\kappa^2\beta)$ and $\eta_y^{-1} = \Theta(1/\beta)$ will find a point satisfying $\|\nabla L(x_T, y_T)\| \le \varepsilon$ by iteration*

$$T \le O\left(\frac{\kappa^2\beta + \kappa\beta^2}{\varepsilon^2}\right).$$

Assuming that the sequence y_k above stays bounded, this projected gradient method is equivalent to running GDA on our unconstrained problem by setting the domain of y as a sufficiently large compact set to contain all the iterates. Consider setting the averaging parameters as $\lambda = \Theta(\eta/\kappa^2\beta) = \Theta\left(\frac{\min\left\{1, |\eta/\rho - 1|^3\right\}}{(1+\eta/\alpha)^2}\right)$ and $\gamma = \Theta(\eta/\beta) = \Theta\left(\min\left\{1, |\eta/\rho - 1|\right\}\right)$. Then using the gradient formula from Lemma 2.5, we see that the damped proximal point method (17) is equivalent to running GDA on the saddle envelope with $\eta_x = \eta/\lambda$ and $\eta_y = \eta/\gamma$:

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} -\nabla_x L_\eta(x_k, y_k)/\eta_x \\ \nabla_y L_\eta(x_k, y_k)/\eta_y \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} -\lambda(x_k - x_k^+) \\ \gamma(y_k^+ - y_k) \end{bmatrix} = \begin{bmatrix} \lambda x_k^+ + (1-\lambda)x_k \\ \gamma y_k^+ + (1-\gamma)y_k \end{bmatrix}.$$

Then the above theorem guarantees that running this variant of the proximal point method on L (or equivalently, applying the GDA variant (18) to the saddle envelope) will converge to a stationary point with $\|\nabla L_\eta(z_T)\| \le \varepsilon$ within $T \le O(\varepsilon^{-2})$ iterations. It immediate follows from the gradient formula that $z_T^+ = \mathrm{prox}_\eta(z_T)$ is approximately stationary for L as $\|\nabla L(z_T^+)\| = \|\nabla L_\eta(z_T)\| \le \varepsilon$. $\square$

4 Interaction Weak Regime

Our previous theory showed that when the interaction between x and y is sufficiently strong, global linear convergence occurs. Now we consider when there is limited interaction between x and y. At the extreme of having no interaction, nonconvex-nonconcave minimax optimization separates into nonconvex minimization and nonconcave maximization. On these separate problems, local convergence of the proximal point method is well-understood[2]. Here we show that under reasonable smoothness and initialization assumptions, this local convergence behavior extends to minimax problems with weak, but nonzero, interaction between x and y.

To formalize this, we make the following regularity assumptions describing how smooth L is

$$\|\nabla^2 L(z)\| \leq \beta \ , \quad \text{for all } z \in \mathbb{R}^n \times \mathbb{R}^m \tag{19}$$

$$\|\nabla^2 L(z) - \nabla^2 L(\bar{z})\| \leq H\|z - \bar{z}\| \ , \quad \text{for all } z, \bar{z} \in \mathbb{R}^n \times \mathbb{R}^m \tag{20}$$

and quantify how weak the interaction between the minimizing and maximizing agents is by

$$\|\nabla^2_{xy} L(z)\| \leq \delta \ , \quad \text{for all } z \in \mathbb{R}^n \times \mathbb{R}^m \tag{21}$$

$$\begin{cases} \|\nabla^2_{xx} L(x,y) - \nabla^2_{xx} L(x,\bar{y})\| \leq \xi\|y - \bar{y}\| \\ \|\nabla^2_{yy} L(x,y) - \nabla^2_{yy} L(\bar{x},y)\| \leq \xi\|x - \bar{x}\| \end{cases} , \quad \text{for all } (x,y),(\bar{x},\bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m \tag{22}$$

for some constants $\beta, H, \delta, \xi \geq 0$. Here we are particularly interested in problems where δ and ξ are sufficiently small. For example, the bilinear setting of (3) satisfies this with $(\delta, \xi) = (\lambda_{max}(A), 0)$ and so we are considering small interaction matrices A.

For such problems, we consider an initialization for the proximal point method based on our motivating intuition that when there is no interaction, we can find local minimizers and maximizers with respect to x and y. For a fixed point $z' = (x', y')$, we compute our PPM initialization $z_0 = (x_0, y_0)$ as

$$\begin{cases} x_0 = \text{a local minimizer of } \min_u L(u, y') \ , \\ y_0 = \text{a local maximizer of } \max_v L(x', v) \ . \end{cases} \tag{23}$$

These subproblems amount to smooth nonconvex minimization, which is well-studied (see for example [34]), and so we take them as a blackbox.

The critical observation explaining why this is a good initialization is that provided δ and ξ are small enough, we have (i) that the interaction dominance conditions (15) and (16) hold at z_0 with a nearly positive $\alpha = \alpha_0$, often with $\alpha_0 > 0$ and (ii) that z_0 is a nearly stationary point of L. Below we formalize each of these properties and arrive at conditions quantifying how small we need ξ and δ to be for our local convergence theory to apply.

[2] For example, considering the decrease in objective value for nonconvex minimization problem $\min_x f(x)$. Each iteration of the proximal point method has $f(x_{k+1}) \leq f(x) - \frac{\eta}{2}\|x_{k+1} - x_k\|^2$. Then inductively applying this and using the gradient formula $\nabla e_\eta\{f\}(x_k) = \eta(x_k - x_{k+1}) = \nabla f(x_{k+1})$, we see the average gradient must converge to zero as

$$\frac{1}{T}\sum_{k=1}^{T} \|\nabla f(x_k)\|^2 \leq \frac{2\eta(f(x_0) - \inf_x f(x))}{T} \ .$$

(i) First, we observe that the interaction dominance conditions (15) and (16) hold at z_0 with a nearly positive coefficient α_0. Since x_0 and y_0 are local optimum, for some $\mu \geq 0$, we must have

$$\nabla^2_{xx} L(x_0, y') \succeq \mu I \quad \text{and} \quad -\nabla^2_{yy} L(x', y_0) \succeq \mu I \ .$$

Then the Hessians at z_0 must be similarly bounded since the amount they can change is limited by (22). Hence

$$\nabla^2_{xx} L(z_0) \succeq (\mu - \xi\|y_0 - y'\|)I \quad \text{and} \quad -\nabla^2_{yy} L(z_0) \succeq (\mu - \xi\|x_0 - x'\|)I \ .$$

Adding a positive semidefinite term onto these (as is done in the definition of interaction dominance) can only increase the righthand-side above. In particular, we can bound the second term added in the interaction dominance conditions (15) and (16) as

$$\begin{aligned}
\nabla^2_{xy} L(z_0)(\eta I - \nabla^2_{yy} L(z_0))^{-1} \nabla^2_{yx} L(z_0) &\succeq \frac{\nabla^2_{xy} L(z_0) \nabla^2_{yx} L(z_0)}{\eta + \beta} \\
&\succeq \frac{\lambda_{min}(\nabla^2_{xy} L(z_0) \nabla^2_{yx} L(z_0))}{\eta + \beta} I \geq 0 \ , \\
\nabla^2_{yx} L(z_0)(\eta I + \nabla^2_{xx} L(z_0))^{-1} \nabla^2_{xy} L(z_0) &\succeq \frac{\nabla^2_{yx} L(z_0) \nabla^2_{xy} L(z_0)}{\eta + \beta} \\
&\succeq \frac{\lambda_{min}(\nabla^2_{yx} L(z_0) \nabla^2_{xy} L(z_0))}{\eta + \beta} I \geq 0 \ .
\end{aligned}$$

Hence interaction dominance holds at z_0 in both x and y with the following constants

$$\begin{aligned}
&\nabla^2_{xx} L(z_0) + \nabla^2_{xy} L(z_0)(\eta I - \nabla^2_{yy} L(z_0))^{-1} \nabla^2_{yx} L(z_0) \\
&\qquad \succeq \left(\mu + \frac{\lambda_{min}(\nabla^2_{xy} L(z_0) \nabla^2_{yx} L(z_0))}{\eta + \beta} - \xi\|y_0 - y'\|\right) I \ , \\
&-\nabla^2_{yy} L(z_0) + \nabla^2_{yx} L(z_0)(\eta I + \nabla^2_{xx} L(z_0))^{-1} \nabla^2_{xy} L(z_0) \\
&\qquad \succeq \left(\mu + \frac{\lambda_{min}(\nabla^2_{yx} L(z_0) \nabla^2_{xy} L(z_0))}{\eta + \beta} - \xi\|x_0 - x'\|\right) I \ .
\end{aligned}$$

For our local linear convergence theory to apply, we need this to hold with non-negative coefficient. Then it suffices to have ξ sufficiently small, satisfying

$$\begin{cases} \xi\|y_0 - y'\| \leq \mu + \dfrac{\lambda_{min}(\nabla^2_{xy} L(z_0) \nabla^2_{yx} L(z_0))}{\eta + \beta} \\ \xi\|x_0 - x'\| \leq \mu + \dfrac{\lambda_{min}(\nabla^2_{yx} L(z_0) \nabla^2_{xy} L(z_0))}{\eta + \beta} \end{cases} . \tag{24}$$

Note this is trivially the case for problems with bilinear interaction (3) as $\xi = 0$. It is also worth noting that even if $\mu = 0$, the right-hand-sides above are still strictly positive if $\nabla_{xy} L(z_0)$ is full rank and the variable dimensions n and m of x and y are equal[3].

[3] This works since having full rank square $\nabla^2_{xy} L(z_0)$ implies that both of its squares $\nabla^2_{xy} L(z_0) \nabla^2_{yx} L(z_0)$ and $\nabla^2_{yx} L(z_0) \nabla^2_{xy} L(z_0)$ are full rank as well. Hence these squares must be strictly positive definite and as a result, have strictly positive minimum eigenvalues.

(ii) Next, we observe that z_0 is nearly stationary by applying (21) and using the first-order optimality conditions of the subproblems (23):

$$\|\nabla L(z_0)\| \leq \left\| \begin{bmatrix} \nabla_x L(x_0, y') \\ \nabla_y L(x', y_0) \end{bmatrix} \right\| + \delta\|z_0 - z'\| = \delta\|z_0 - z'\|.$$

For our convergence theory to apply, we need this gradient to be sufficiently small, quantified as having

$$\delta\|z_0 - z'\| \leq \frac{\alpha_0(\eta - \rho)}{2\left(1 + \frac{4\sqrt{2}(\eta+\alpha_0/2)}{\alpha_0} + \frac{4\sqrt{2}\beta(\eta+\alpha_0/2)}{\alpha_0(\eta-\rho)}\right) H \left(1 + \frac{2\delta}{\eta-\rho} + \frac{\delta^2}{(\eta-\rho)^2}\right)} \ . \tag{25}$$

Under these conditions, we have the following local linear convergence guarantee.

Theorem 4.1. *For any objective L satisfying weak convexity-concavity (4), the smoothness conditions (19) and (20), and the interaction bounds (21) and (22), consider the damped PPM (6) with initialization (x_0, y_0) given by (23) and η and λ satisfying*

$$\lambda \leq 2\frac{\min\left\{1, (\eta/\rho - 1)^2\right\}}{2\eta/\alpha_0 + 1} \ .$$

Then PPM linearly converges to a nearby stationary point (x^*, y^*) of (1) with

$$\left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 \leq \left(1 - \frac{2\lambda}{2\eta/\alpha_0 + 1} + \frac{\lambda^2}{\min\left\{1, (\eta/\rho - 1)^2\right\}}\right)^k \left\| \begin{bmatrix} x_0 - x^* \\ y_0 - y^* \end{bmatrix} \right\|^2$$

provided δ and ξ are small enough to satisfy (24) and (25).

4.1 Proof of Theorem 4.1

Our proof of this local convergence guarantee considers two sets centered at initial point (x_0, y_0): An inner region $B_{\text{inner}} = \{(x, y) | x \in B(x_0, r), y \in B(y_0, r)\}$ with radius r given by

$$r := \frac{4(\eta + \alpha_0/2)}{\alpha_0} \frac{\|\nabla L(z_0)\|}{\eta - \rho}$$

and an outer ball $B_{\text{outer}} = B((x_0, y_0), R)$ with radius R given by

$$R = \left(1 + \frac{4\sqrt{2}(\eta + \alpha_0/2)}{\alpha_0} + \frac{4\sqrt{2}\beta(\eta + \alpha_0/2)}{\alpha_0(\eta - \rho)}\right) \frac{\|\nabla L(z_0)\|}{\eta - \rho} \geq \sqrt{2}r \ .$$

Thus $B_{\text{inner}} \subseteq B_{\text{outer}}$. We are able to conclude in the following lemma that the $\alpha_0 \geq 0$-interaction dominance at z_0 (which we arrived at from our choice of initialization procedure and interaction being sufficiently weak) extends to give $\alpha_0/2$-interaction dominance on the whole outer ball B_{outer}.

Lemma 4.2. *On the outer ball B_{outer}, $\alpha_0/2$-iteration dominance holds in both x and y.*

Proof. First, observe that the functions defining the interaction dominance conditions (15) and (16)

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

are both uniformly Lipschitz with constant

$$H\left(1 + \frac{2\delta}{\eta - \rho} + \frac{\delta^2}{(\eta-\rho)^2}\right).$$

We compute this constant by applying the "product rule"-style formula that the product of two mappings $A(z)B(z)$ is uniformly $(a'b + ab')$-Lipschitz provided $A(z)$ is bounded in norm by a and a'-Lipschitz and $B(z)$ is bounded in norm by b and b'-Lipschitz[4]. Then our Lipschitz constant follows by observing the component functions defining it satisfy the following: $\nabla^2_{xx}L(z)$ and $\nabla^2_{yy}L(z)$ are H-Lipschitz, $\nabla^2_{xy}L(z)$ and its transpose $\nabla^2_{yx}L(z)$ are both H-Lipschitz and bounded in norm by δ, and $(\eta I + \nabla^2_{xx}L(z))^{-1}$ and $(\eta I - \nabla^2_{yy}L(z))^{-1}$ are both $H/(\eta-\rho)^2$-Lipschitz and bounded in norm by $(\eta-\rho)^{-1}$.

Then it immediately follows that every $z \in B_{\text{outer}}$ has $\alpha_0/2$-interaction dominance in x as

$$\begin{aligned}
&\nabla^2_{xx}L(z) + \nabla^2_{xy}L(z)(\eta I - \nabla^2_{yy}L(z))^{-1}\nabla^2_{yx}L(z) \\
&\quad\succeq \nabla^2_{xx}L(z_0) + \nabla^2_{xy}L(z_0)(\eta I - \nabla^2_{yy}L(z_0))^{-1}\nabla^2_{yx}L(z_0) - H\left(1 + \frac{2\delta}{\eta-\rho} + \frac{\delta^2}{(\eta-\rho)^2}\right)RI \\
&\quad\succeq \nabla^2_{xx}L(z_0) + \nabla^2_{xy}L(z_0)(\eta I - \nabla^2_{yy}L(z_0))^{-1}\nabla^2_{yx}L(z_0) - \alpha_0/2I \\
&\quad\succeq \alpha_0 I - \alpha_0/2I = \alpha_0/2I
\end{aligned}$$

where the first inequality uses Lipschitz continuity, the second inequality uses our assumed condition (25) of $H\left(1 + \frac{2\delta}{\eta-\rho} + \frac{\delta^2}{(\eta-\rho)^2}\right)R \le \alpha_0/2$, and the third inequality uses the α_0-interaction dominance at z_0. Symmetrical reasoning shows $\alpha_0/2$-interaction dominance in y holds for each $z \in B_{\text{outer}}$ as well. □

From this, we find that interaction dominance on the outer ball suffices to ensure the saddle envelope is strongly convex-strongly concave on the inner square.

Lemma 4.3. *The saddle envelope is $(\eta^{-1} + (\alpha_0/2)^{-1})^{-1}$-strongly convex-strongly concave on B_{inner}.*

Proof. Given $\alpha_0/2$-interaction dominance holds on B_{outer}, it suffices to show that for any $z = (x, y) \in B_{\text{inner}}$, the proximal step $z_+ = \text{prox}_\eta(z) \in B_{\text{outer}}$ as we can then apply the Hessian bounds from Proposition 2.10 to show strong convexity and strong concavity.

[4] A proof of this is straightforward. For any two points z, z', we have

$$\begin{aligned}
\|A(z)B(z) - A(z')B(z')\| &= \|A(z)B(z) - A(z')B(z) + A(z')B(z) - A(z')B(z')\| \\
&\le \|A(z)B(z) - A(z')B(z)\| + \|A(z')B(z) - A(z')B(z')\| \\
&\le \|A(z) - A(z')\|\|B(z)\| + \|A(z')\|\|B(z) - B(z')\| \le (a'b + b'a)\|z - z'\|\ .
\end{aligned}$$

Define the function underlying the computation of the proximal step at (x, y) as

$$M(u, v) = L(u, v) + \frac{\eta}{2}\|u - x\|^2 - \frac{\eta}{2}\|v - y\|^2.$$

Our choice of $\eta > \rho$ ensures that M is $(\eta - \rho)$-strongly convex-strongly concave. Thus applying Lemma 1.1 and then the β-Lipschitz continuity of $\nabla L(z)$ implies

$$\left\| \begin{bmatrix} x - x_+ \\ y - y_+ \end{bmatrix} \right\| \leq \frac{\|\nabla M(x, y)\|}{\eta - \rho} = \frac{\|\nabla L(x, y)\|}{\eta - \rho} \leq \frac{\|\nabla L(x_0, y_0)\| + \beta\sqrt{2}r}{\eta - \rho} .$$

Hence (x_+, y_+) must lie in the outer ball as

$$\left\| \begin{bmatrix} x_0 - x_+ \\ y_0 - y_+ \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} x_0 - x \\ y_0 - y \end{bmatrix} \right\| + \left\| \begin{bmatrix} x - x_+ \\ y - y_+ \end{bmatrix} \right\| \leq \sqrt{2}r + \frac{\|\nabla L(x_0, y_0)\| + \beta\sqrt{2}r}{\eta - \rho} = R . \qquad \square$$

Armed with the knowledge that interaction dominance holds on B_{inner}, we return to the proof of Theorem 4.1. Observe that the gradient of the saddle envelope at $z_0 = (x_0, y_0)$ is bounded by Lemma 2.5 and Lemma 1.1 as

$$\|\nabla L_\eta(z_0)\| = \|\eta(z_0 - z_0^+)\| \leq \frac{\eta}{\eta - \rho}\|\nabla M_0(z_0)\| = \frac{\eta}{\eta - \rho}\|\nabla L(z_0)\|$$

where $z_0^+ = \text{prox}_\eta(z_0)$ and $M_0(u, v) = L(u, v) + \frac{\eta}{2}\|u - x_0\|^2 - \frac{\eta}{2}\|v - y_0\|^2$ is the $\eta - \rho$-strongly convex-strongly concave function defining it. Now we have shown all of the conditions necessary to apply Theorem 1.2 on the square $B(x_0, r) \times B(y_0, r)$ with

$$r = \frac{4(\eta + \alpha_0/2)\|\nabla L(z_0)\|}{\alpha_0(\eta - \rho)} = \frac{2\|\nabla L_\eta(z_0)\|}{\mu}$$

upon which the saddle envelope is $\mu = (\eta^{-1} + (\alpha_0/2)^{-1})^{-1}$-strongly convex-strongly concave and $\beta = \max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\}$-smooth. Hence applying GDA with $s = \lambda/\eta$ to the saddle envelope produces iterates (x_k, y_k) converging to a stationary point (x^*, y^*) with

$$\left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 \leq \left(1 - \frac{2\lambda}{\eta(\eta^{-1} + (\alpha_0/2)^{-1})} + \frac{\lambda^2}{\eta^2(\eta^{-1} - \rho^{-1})^2}\right)^k \left\| \begin{bmatrix} x_0 - x^* \\ y_0 - y^* \end{bmatrix} \right\|^2 .$$

By Corollary 2.6, (x^*, y^*) must also be a stationary point of L. Further, by Corollary 2.7, this sequence (x_k, y_k) is the same as the sequence generated by running the damped PPM on (1). $\square$

5 Interaction Moderate Regime

Between the interaction dominate and interaction weak regimes, the proximal point method may diverge or cycle indefinitely (recall our introductory example in Figure 1 where convergence fails in this middle regime). We begin by considering the behavior of the proximal point method when applied to a nonconvex-nonconcave quadratic example. From this, we find that our interaction dominance condition is tight in that it exactly describes when our example converges or diverges. Motivated by this, we propose a Lyapunov-type function that identifies stationary points of generic minimax problems. Since PPM may cycle indefinitely for problems with moderate amounts of interaction, one cannot guarantee any Lyapunov monotonically decreases. However, analyzing our proposed Lyapunov enables us to bound how quickly PPM can diverge based on how far the given problem instance is from the interaction dominant regime or convexity and concavity.

5.1 Divergence and Tightness of the Interaction Dominance Regime

Consider the following nonconvex-nonconcave quadratic minimax problem of

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^n} L(x,y) = \frac{-\rho}{2}\|x\|^2 + ax^Ty - \frac{-\rho}{2}\|y\|^2 \tag{26}$$

where $a \in \mathbb{R}$ controls the size of the interaction between x and y and $\rho \geq 0$ controls how weakly convex-weakly concave the problem is. Notice this problem has a stationary point at the origin. Even though this problem is nonconvex-nonconcave, the proximal point method will still converge to the origin for some selections of a, ρ, and η. Examining our interaction dominance conditions (15) and (16), we see that this example is $\alpha = -\rho + a^2/(\eta-\rho)$-interaction dominant in both x and y as $\nabla^2_{xx}L(z)L(z) = -\nabla^2_{yy}L(z) = -\rho$ and $\nabla^2_{xy}L(z) = \nabla^2_{yx}L(z) = a$.

For quadratic problems, the proximal point method always corresponds to the matrix multiplication. In the case of (26), the damped PPM iteration is given by

$$\begin{aligned}
\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} &= (1-\lambda)\begin{bmatrix} x_k \\ y_k \end{bmatrix} + \lambda \begin{bmatrix} (1-\rho/\eta)I & aI/\eta \\ -aI/\eta & (1-\rho/\eta)I \end{bmatrix}^{-1} \begin{bmatrix} x_k \\ y_k \end{bmatrix} \\
&= (1-\lambda)\begin{bmatrix} x_k \\ y_k \end{bmatrix} + \frac{\lambda\eta}{\eta-\rho}\left(\begin{bmatrix} I & aI/(\eta-\rho) \\ -aI/(\eta-\rho) & I \end{bmatrix}\right)^{-1} \begin{bmatrix} x_k \\ y_k \end{bmatrix} \\
&= (1-\lambda)\begin{bmatrix} x_k \\ y_k \end{bmatrix} + \frac{\lambda\eta}{a^2/(\eta-\rho)+\eta-\rho}\begin{bmatrix} I & -aI/(\eta-\rho) \\ aI/(\eta-\rho) & I \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} \\
&= \begin{bmatrix} CI & -DI \\ DI & CI \end{bmatrix}\begin{bmatrix} x_k \\ y_k \end{bmatrix}
\end{aligned}$$

for constants $C = 1 - \dfrac{\lambda\alpha}{\eta+\alpha}$ and $D = \dfrac{\lambda\eta a}{(\eta+\alpha)(\eta-\rho)}$. Notice that these constants are well-defined since $\eta-\rho > 0$ and $\eta+\alpha > 0$ (even if α is negative) since $\eta > \rho$ and $\alpha \geq -\rho$. Matrix multiplication of this special final form has the following nice property for any z,

$$\left\|\begin{bmatrix} CI & -DI \\ DI & CI \end{bmatrix} z\right\|^2 = (C^2+D^2)\|z\|^2. \tag{27}$$

Hence this iteration will globally converge to the origin exactly when

$$\left(1-\frac{\lambda\alpha}{\eta+\alpha}\right)^2 + \left(\frac{\lambda\eta a}{(\eta+\alpha)(\eta-\rho)}\right)^2 < 1 \ .$$

Likewise, the damped proximal point method will cycle indefinitely when this holds with equality and diverges when it is strictly violated. As a result, violating $\alpha > 0$-interaction dominance (that is, having $\alpha \leq 0$) leads to divergence in (26) for any choice of the averaging parameter $\lambda \in (0,1]$ since this forces $C \geq 1$ (and so $C^2 + D^2 > 1$). Hence our interaction dominance boundary is essentially tight.

Further, this example shows that considering the damped proximal point method (as opposed to fixing $\lambda = 1$) is necessary to fully capture the convergence for interaction dominant problems. For example, setting $\rho = 1, a = 2, \eta = 3$ has $\alpha = 1$-interaction dominance in x and y and converges exactly when

$$(1-\lambda/4)^2 + (3\lambda/4)^2 < 1$$

which is satisfied when $\lambda \in (0, 0.8)$, but not by the undamped proximal point method with $\lambda = 1$. Our theory from Theorem 3.1 is slightly more conservative, guaranteeing convergence whenever $\lambda \leq 0.5 = 2\min\left\{1, (\eta/\rho - 1)^2\right\}/(\eta/\alpha + 1)$.

5.2 A Candidate Lyapunov for Interaction Moderate Problems

The standard analysis of gradient descent on nonconvex optimization relies on the fact that the function value monotonically decays at the level of gradient norm square every iteration. Thus as long as the gradient is large, the function value has sufficient decay. Consequently, the iterates of gradient descent either have gradient norm converge to 0 or objective value approach $-\infty$. However, such arguments no longer hold in the nonconvex-nonconcave minimax setting: the objective is neither monotonically decreasing nor increasing while PPM runs. Worse yet, since we know the proximal point method may cycle indefinitely with gradients bounded away from zero (for example, recall the interaction moderate regime trajectories shown in Figure 1), no "Lyapunov"-type quantity can exist that monotonically decreases along the iterates of the proximal point method.

In order to obtain a similar analysis as the standard nonconvex optimization approach, we propose to study the following "Lyapunov" function, which captures the difference between smoothing over y and smoothing over x using the classic Moreau envelope,

$$\mathcal{L}(x,y) := -\, e_\eta\{-L(x,\cdot)\}(y) - e_\eta\{L(\cdot,y)\}(x) \ . \tag{28}$$

The following proposition establishes structural properties supporting our consideration of $\mathcal{L}(x,y)$.

Theorem 5.1. *The Lyapnuov function $\mathcal{L}(x,y)$ has the following structural properties:*

1. *$\mathcal{L}(x,y) \geq 0$,*
2. *When $\eta > \rho$, $\mathcal{L}(x,y) = 0$ if and only if (x,y) is a stationary point to $L(x,y)$,*
3. *When $\eta = 0$, $\mathcal{L}(x,y)$ recovers the well-known primal-dual gap of the function $L(x,y)$*

$$\mathcal{L}(x,y) = \max_v L(x,v) - \min_u L(u,y).$$

Proof. Recall that a Moreau envelope $e_\eta\{f(\cdot)\}(x)$ provides a lower bound (8) on f everywhere. Hence $e_\eta\{-L(x,\cdot)\}(y) \leq -L(x,y)$ and $e_\eta\{L(\cdot,y)\}(x) \leq L(x,y)$, and so our proposed Lyapunov is always nonnegative since

$$\mathcal{L}(x,y) = -e_\eta\{-L(x,\cdot)\}(y) - e_\eta\{L(\cdot,y)\}(x) \geq L(x,y) - L(x,y) = 0 \ .$$

Further, recall (9) stated that for any ρ-weakly convex function f, selecting $\eta > \rho$ ensures the Moreau envelope equals the given function precisely at its stationary point. Then the preceding nonnegativity argument holds with equality if and only if

$$\nabla_y - L(x,\cdot)(y) = 0 \quad \text{and} \quad \nabla_x L(\cdot,y)(x) = 0 \ .$$

Hence we have $\mathcal{L}(x,y) = 0 \iff \nabla L(x,y) = 0$. Lastly, when $\eta = 0$, observe that this Lyapunov simplifies to be the primal-dual gap for $L(x,y)$ as

$$\begin{aligned}\mathcal{L}(x,y) &= -\min_v \left\{-L(x,v) + \frac{\eta}{2}\|v-y\|^2\right\} - \min_u \left\{-L(u,y) + \frac{\eta}{2}\|u-x\|^2\right\} \\ &= \max_v L(x,v) - \min_u L(u,y) \ .\end{aligned}$$

□

Computing the Moreau envelopes defining $\mathcal{L}(z)$ for the quadratic example (26) gives

$$e_\eta\{L(\cdot,y)\}(x) = \frac{1}{2}(\eta^{-1}-\rho^{-1})^{-1}\|x\|^2 + \frac{\eta a}{\eta-\rho}x^Ty - \frac{\alpha}{2}\|y\|^2 \tag{29}$$

$$e_\eta\{-L(x,\cdot)\}(y) = -\frac{\alpha}{2}\|x\|^2 - \frac{\eta a}{\eta-\rho}x^Ty + \frac{1}{2}(\eta^{-1}-\rho^{-1})^{-1}\|y\|^2 \tag{30}$$

where $\alpha = -\rho + a^2/(\eta-\rho)$ is the interaction dominance parameter for this problem. Hence

$$\mathcal{L}(z) = \frac{1}{2}\left(\alpha - (\eta^{-1}-\rho^{-1})^{-1}\right)\|z\|^2 \ .$$

Noting that $\alpha \geq -\rho$ and $-(\eta^{-1}-\rho^{-1})^{-1} > -\rho$, we see that the origin is the unique minimizer of $\mathcal{L}(z)$ and consequently the unique stationary point of L. In this case, minimizing $\mathcal{L}(z)$ is simple convex optimization.

Future works could identify further regions of tractable nonconvex-nonconcave problems where algorithms can minimize $\mathcal{L}(x,y)$ instead as all of its global minimums are stationary points of the original objective. Since this problem is purely one of minimization, cycling can be ruled out directly. As previously observed, the proximal point method is not such an algorithm since it may fall into a cycle and fail to monotonically decrease $\mathcal{L}(z)$. Instead, we find the following weakened descent condition for $\mathcal{L}(z)$, relating its change to our α-interaction dominance conditions. Note that this result holds regardless of whether the interaction dominance parameter α is positive or negative.

Theorem 5.2. *For any ρ-weakly convex-weakly concave, $\alpha \in \mathbb{R}$-interaction dominant in x and y problem, taking a proximal step from any $z \in \mathbb{R}^n \times \mathbb{R}^m$ to $z_+ = \mathrm{prox}_\eta(z)$ has*

$$\mathcal{L}(z_+) \leq \mathcal{L}(z) - \frac{1}{2}\left(\alpha + (\eta^{-1}-\rho^{-1})^{-1}\right)\|z_+ - z\|^2 \ .$$

Remark 5.3. *This upper bound is attained by our example diverging problem* (26)*. Hence the worst case increase of this Lyapunov is given by simple quadratic. To see why this is example attains our bound, note that the proof of Theorem 5.2 only introduces inequalities by using the following four Hessian bounds for every (u,v)*

$$\nabla^2_{xx} - e_\eta\{-L(u,\cdot)\}(v) \succeq \alpha I \ , \quad \nabla^2_{yy} - e_\eta\{-L(u,\cdot)\}(v) \preceq -(\eta^{-1}-\rho^{-1})^{-1}I \ ,$$
$$\nabla^2_{yy} - e_\eta\{L(\cdot,v)\}(u) \succeq \alpha I \ , \quad \nabla^2_{xx} - e_\eta\{L(\cdot,v)\}(u) \preceq -(\eta^{-1}-\rho^{-1})^{-1}I \ .$$

Observing that all four of these bounds hold with equality everywhere in (29) *and* (30) *shows our recurrence holds with equality.*

Remark 5.4. *For reasonably interaction dominate problems, $\alpha > -(\eta^{-1}-\rho^{-1})^{-1}$, we indeed have a decrease in the Lyapunov*[5] *on the order of $O(\|z_+ - z\|^2) = O(\|\nabla L(z_+)\|^2/\eta^2)$. This gives another insight into why PPM converges to stationary points for interaction dominate problems: inductively applying this descent condition (and using that $\mathcal{L}(x,y) \geq 0$) bounds the average gradient by*

$$\frac{1}{T}\sum_{k=1}^{T}\|\nabla L(z_k)\|^2 \leq \frac{2\eta^2\mathcal{L}(x_0,y_0)}{T\left(\alpha + (\eta^{-1}-\rho^{-1})^{-1}\right)} \ .$$

[5] Noting that $-(\eta^{-1}-\rho^{-1})^{-1} > 0$ whenever $\rho > 0$ shows this setting is a restriction on $\alpha > 0$-interaction dominant problems. Some form of restriction is needed here since we have seen a quadratic example where the proximal point method without damping diverges despite having $\alpha > 0$-interaction dominance.

Remark 5.5. *For generic minimax problems, Theorem 5.2 bounds how quickly the proximal point method can diverge. For example, consider an objective L that is l-Lipschitz and nearly convex-concave, satisfying weak convexity-weak concavity* (4) *with some $\rho = \epsilon$. Then noting interaction dominance holds with coefficient $\alpha = -\rho = -\epsilon$, the increase in the Lyapanov is bounded by a constant $O(\epsilon)$ as*

$$\mathcal{L}(z_+) - \mathcal{L}(z) \leq -\frac{1}{2}\left(\alpha - \frac{\eta\rho}{\eta-\rho}\right)\|\nabla L(z_+)/\eta\|^2 \leq \frac{\epsilon l^2}{2\eta^2}\left(1 + \frac{\eta}{\eta-\epsilon}\right) \approx \frac{\epsilon l^2}{\eta^2} \ .$$

Lastly, we remark the average gradient norm seen when the proximal point method falls into cycling can be lower bounded by the isoperimetric inequality. It says that, if PPM "coarsely" converges to a cyclic attractor, that winds around a large ball, then the average of the squared gradient will be large. Details and definitions in the appendix.

Theorem 5.6. *For any problem $L(x, y)$, if the limiting behaviour of PPM* (6)*, converges uniformly to a cyclic attractor $\mathcal{C}$, such that there exists a point on the minimal surface bounded by $\mathcal{C}$ that is at (geodesic) distance at least R from every point on $\mathcal{C}$ then, small enough choice of step size s and a large enough choice of T and S (where $T \gg S$), we have*

$$\frac{1}{T-S}\sum_{k=S}^{T}\|\nabla L(z_k)\|^2 \geq \frac{C \cdot R^2}{s^2 N^2}$$

where C is a constant depending on the properties of the minimal surface $\mathcal{S}$, and N measures the "coarseness" of the algorithm.

5.3 Proof of Theorem 5.2

First we derive the following bounds on the Hessians of the functions defining our Lyapunov $\mathcal{L}(z)$.

Lemma 5.7. *If the x-interaction dominance* (15) *holds with $\alpha \in \mathbb{R}$, the function $e_\eta\{L(\cdot, y)\}(x)$ has Hessians in x and y bounded by*

$$(\eta^{-1} - \rho^{-1})^{-1} I \preceq \nabla^2_{xx} e_\eta\{L(\cdot, y)\}(x) \preceq \eta I \quad \textit{and} \quad \nabla^2_{yy} e_\eta\{L(\cdot, y)\}(x) \preceq -\alpha I \ .$$

Symmetrically, if the y-interaction dominance (16) *holds with $\alpha \in \mathbb{R}$, the function $e_\eta\{-L(x, \cdot)\}(y)$ has Hessians in x and y bounded by*

$$\nabla^2_{xx} e_\eta\{-L(x, \cdot)\}(y) \preceq -\alpha I \quad \textit{and} \quad (\eta^{-1} - \rho^{-1})^{-1} I \preceq \nabla^2_{yy} e_\eta\{-L(x, \cdot)\}(y) \preceq \eta I.$$

Proof. For the Hessian bound in the x variable, this follows directly from the Moreau envelope Hessian bounds (12). Considering $e_\eta\{L(\cdot, y)\}(x)$ as a function of y, we find that its the gradient and Hessian are given by

$$\begin{aligned}
&\nabla_y e_\eta\{L(\cdot, y)\}(x) = \nabla_y L(x_+, y) \\
&\nabla^2_{yy} e_\eta\{L(\cdot, y)\}(x) = \nabla^2_{yy} L(x_+, y) - \nabla^2_{yx} L(x_+, y)(\eta I + \nabla^2_{xx} L(x_+, y))^{-1} \nabla^2_{xy} L(x_+, y)
\end{aligned}$$

evaluated at (x_+, y) where $x_+ = \text{argmin}_u L(u, y) + \frac{\eta}{2}\|u - x\|^2$. Noting that this Hessian formula is the matrix from the α-interaction dominance condition (16) gives our bound on $-\nabla^2_{yy} e_\eta\{L(\cdot, y)\}(x)$.

All that remains is to derive our claimed gradient and Hessian formulas in y. Consider a nearby point $y^\Delta = y+\Delta$ and denote $x_+^\Delta = \mathrm{argmin}_u L(u,y^\Delta) + \frac{\eta}{2}\|u-x\|^2$. Consider the second-order Taylor model of the objective L around (x_+,y) denoted by $\widetilde{L}(u,v)$ with value

$$L(x_+,y) + \begin{bmatrix}\nabla_x L(x_+,y)\\ \nabla_y L(x_+,y)\end{bmatrix}^T \begin{bmatrix}u-x_+\\ v-y\end{bmatrix} + \frac{1}{2}\begin{bmatrix}u-x_+\\ v-y\end{bmatrix}^T \begin{bmatrix}\nabla^2_{xx}L(x_+,y) & \nabla^2_{xy}L(x_+,y)\\ \nabla^2_{yx}L(x_+,y) & \nabla^2_{yy}L(x_+,y)\end{bmatrix}\begin{bmatrix}u-x_+\\ v-y\end{bmatrix} .$$

Denote the $\widetilde{x}_+^\Delta = \mathrm{argmin}_u \widetilde{L}(u,y^\Delta) + \frac{\eta}{2}\|u-x\|^2$. Noting this point is uniquely defined by its first-order optimality conditions, we have

$$\begin{aligned}&\nabla_x L(x_+,y) + \nabla^2_{xx}L(x_+,y)(\widetilde{x}_+^\Delta - x_+) + \nabla^2_{xy}L(x_+,y)\Delta + \eta(\widetilde{x}_+^\Delta - x) = 0 ,\\ \implies & (\eta I + \nabla^2_{xx}L(x_+,y))(\widetilde{x}_+^\Delta - x_+) = -\nabla^2_{xy}L(x_+,y)\Delta ,\\ \implies & \widetilde{x}_+^\Delta - x_+ = -(\eta I + \nabla^2_{xx}L(x_+,y))^{-1}\nabla^2_{xy}L(x_+,y)\Delta .\end{aligned}$$

Denote the proximal subproblem objective by $M^\Delta(u,v) = L(u,y^\Delta)+\frac{\eta}{2}\|u-x\|^2$ and its approximation by $\widetilde{M}^\Delta(u,v) = \widetilde{L}(u,y^\Delta)+\frac{\eta}{2}\|u-x\|^2$. Noting that $\|\nabla_x\widetilde{M}^\Delta(x_+,y^\Delta)\| = \|\nabla^2_{xy}L(x_+,y)\Delta\|$, the $(\eta-\rho)$-strongly convexity of $\widetilde{M}^\Delta$ bounds the distance to its minimizer by

$$\|x_+ - \widetilde{x}_+^\Delta\| \leq \frac{\|\nabla^2_{xy}L(x_+,y)\Delta\|}{\eta-\rho} = O(\|\Delta\|) .$$

Consequently, we can bound difference in gradients between L and its quadratic model $\widetilde{L}$ at $\widetilde{x}_+^\Delta$ by $\|\nabla L(\widetilde{x}_+^\Delta, y^\Delta) - \nabla\widetilde{L}(\widetilde{x}_+^\Delta,y^\Delta)\| = o(\|\Delta\|)$. Therefore $\|\nabla M^\Delta(\widetilde{x}_+^\Delta,y^\Delta)\| = o(\|\Delta\|)$. Then using the strongly convexity of M^Δ with this gradient bound, we conclude the distance from $\widetilde{x}_+^\Delta$ to the minimizer x_+^Δ is bounded by

$$\|\widetilde{x}_+^\Delta - x_+^\Delta\| = o(\|\Delta\|) .$$

Then our claimed gradient formula follows as

$$\begin{aligned}e_\eta\{L(\cdot,y^\Delta)\}(x) - e_\eta\{L(\cdot,y)\}(x) &= L(x_+^\Delta,y^\Delta) + \frac{\eta}{2}\|x_+^\Delta - x\|^2 - L(x_+,y) - \frac{\eta}{2}\|x_+-x\|^2\\ &= \begin{bmatrix}\nabla_x L(x_+,y) + \eta(x_+-x)\\ \nabla_y L(x_+,y)\end{bmatrix}^T\begin{bmatrix}x_+^\Delta - x_+\\ \Delta\end{bmatrix} + o(\|\Delta\|)\\ &= \nabla_y L(x_+,y)^T\Delta + o(\|\Delta\|) .\end{aligned}$$

Moreover, our claimed Hessian formula follows as

$$\begin{aligned}&\nabla_y e_\eta\{L(\cdot,y^\Delta)\}(x) - \nabla_y e_\eta\{L(\cdot,y)\}(x)\\ &\quad = \nabla_y L(x_+^\Delta,y^\Delta) - \nabla_y L(x_+,y)\\ &\quad = \nabla_y \widetilde{L}(\widetilde{x}_+^\Delta,y^\Delta) - \nabla_y L(x_+,y) + o(\|\Delta\|)\\ &\quad = \begin{bmatrix}\nabla^2_{xy}L(x_+,y)\\ \nabla^2_{yy}L(x_+,y)\end{bmatrix}^T\begin{bmatrix}-(\eta I + \nabla^2_{xx}L(x_+,y))^{-1}\nabla^2_{xy}L(x_+,y)\Delta\\ \Delta\end{bmatrix} + o(\|\Delta\|) .\end{aligned}$$

□

Notice that $-e_\eta\{-L(u,\cdot)\}(y)$ has gradient at x_+ of $\nabla_x - e_\eta\{-L(x_+,\cdot)\}(y) = \nabla_x L(z_+) = \eta(x-x_+)$ and from Lemma 5.7 that its Hessian in x is uniformly lower bounded by αI. As a result, we have the following decrease in $-e_\eta\{-L(u,\cdot)\}(y)$ when moving from x to x_+

$$\begin{aligned}-e_\eta\{-L(x_+,\cdot)\}(y) &\leq -e_\eta\{-L(x,\cdot)\}(y) + \nabla_x L(z_+)^T(x_+-x) - \frac{\alpha}{2}\|x_+-x\|^2\\ &= -e_\eta\{-L(x,\cdot)\}(y) - \left(\eta+\frac{\alpha}{2}\right)\|x_+-x\|^2\ .\end{aligned}$$

From the gradient formula (10), we know that $\nabla_y - e_\eta\{-L(x_+,\cdot)\}(y) = \nabla_y L(z_+) = \eta(y_+-y)$ and from Lemma 5.7 that its Hessian in y is uniformly bounded above by $-(\eta^{-1}-\rho^{-1})^{-1}I$. Then we can upper bound the change in $-e_\eta\{-L(x_+,\cdot)\}(v)$ when moving from y to y_+ as

$$\begin{aligned}-e_\eta\{-L(x_+,\cdot)\}(y_+) + e_\eta\{-L(x_+,\cdot)\}(y) &\leq \nabla_y L(z_+)^T(y_+-y) + \frac{-(\eta^{-1}-\rho^{-1})^{-1}}{2}\|y_+-y\|^2\\ &= \left(\eta - \frac{(\eta^{-1}-\rho^{-1})^{-1}}{2}\right)\|y_+-y\|^2\ .\end{aligned}$$

Summing these two inequalities yields

$$-e_\eta\{-L(x_+,\cdot)\}(y_+) + e_\eta\{-L(x,\cdot)\}(y) \leq \left(\eta - \frac{(\eta^{-1}-\rho^{-1})^{-1}}{2}\right)\|y_+-y\|^2 - \left(\eta+\frac{\alpha}{2}\right)\|x_+-x\|^2\ .$$

Symmetrically, the change in $-e_\eta\{L(\cdot,y)\}(x)$ from z to z_+ is

$$-e_\eta\{L(\cdot,y_+)\}(x_+) + e_\eta\{L(\cdot,y)\}(x) \leq \left(\eta - \frac{(\eta^{-1}-\rho^{-1})^{-1}}{2}\right)\|x_+-x\|^2 - \left(\eta+\frac{\alpha}{2}\right)\|y_+-y\|^2\ .$$

Summing these two results gives the claimed bound

$$\mathcal{L}(z_+) \leq \mathcal{L}(z) - \frac{1}{2}\left(\alpha + (\eta^{-1}-\rho^{-1})^{-1}\right)\|z_+-z\|^2\ . \qquad \square$$

Acknowledgements. Benjamin Grimmer was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650441.

References

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

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

[3] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* MIT press, 2018.

[4] Bo Dai, Albert Shaw, Niao He, Lihong Li, and Le Song. Boosting the actor with dual critic. In *ICLR 2018.*

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

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

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

[8] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9236–9246. Curran Associates, Inc., 2018.

[9] Farzan Farnia and Asuman Ozdaglar. Do gans always have nash equilibria? In *International Conference on Machine Learning*, pages 3029–3039. PMLR, 2020.

[10] Jean Jacques Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 255:2897–2899, 1962.

[11] Benjamin Grimmer, Haihao Lu, Pratik Worah, and Vahab Mirrokni. Limiting behaviors of nonconvex-nonconcave minimax optimization via continuous-time systems. *arXiv preprint arXiv:2010.10628*, 2020.

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

[13] Haihao Lu. An $o(s^r)$-resolution ode framework for discrete-time optimization algorithms and applications to convex-concave saddle-point problems. *arXiv preprint arXiv:2001.08826*, 2020.

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

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

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

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

[18] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

[19] Tianyi Lin, Chi Jin, and Michael I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv preprint arXiv:1906.00331*, 2019.

[20] Tianyi Lin, Chi Jin, and Michael I. Jordan. Near-optimal algorithms for minimax optimization. *arXiv preprint arXiv:2002.02417*, 2020.

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

[22] Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems 32*, pages 12680–12691. Curran Associates, Inc., 2019.

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

[24] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems 32*, pages 14934–14942. Curran Associates, Inc., 2019.

[25] Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020.

[26] Benjamin Grimmer, Haihao Lu, Pratik Worah, and Vahab Mirrokni. The landscape of nonconvex-nonconcave minimax optimization. *arXiv preprint arXiv:2006.08667*, 2020.

[27] Alistair Letcher. On the impossibility of global convergence in multi-loss optimization. *arXiv preprint arXiv:2005.12649*, 2020.

[28] Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher. The limits of min-max optimization algorithms: convergence to spurious non-critical sets. *arXiv preprint arXiv:2006.09065*, 2020.

[29] Guojun Zhang, Pascal Poupart, and Yaoliang Yu. Optimality and stability in non-convex-non-concave min-max optimization. *arXiv preprint arXiv:2002.11875*, 2020.

[30] Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

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

[32] Siqi Zhang and Niao He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.

[33] Maurice Sion. On general minimax theorems. *Pacific J. Math.*, 8(1):171–176, 1958.

[34] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016.

[35] Robert Osserman. The isoperimetric inequality. *Bull. Amer. Math. Soc.*, 84(6):1182–1238, 1978.

A Sample Paths From Other First-Order Methods

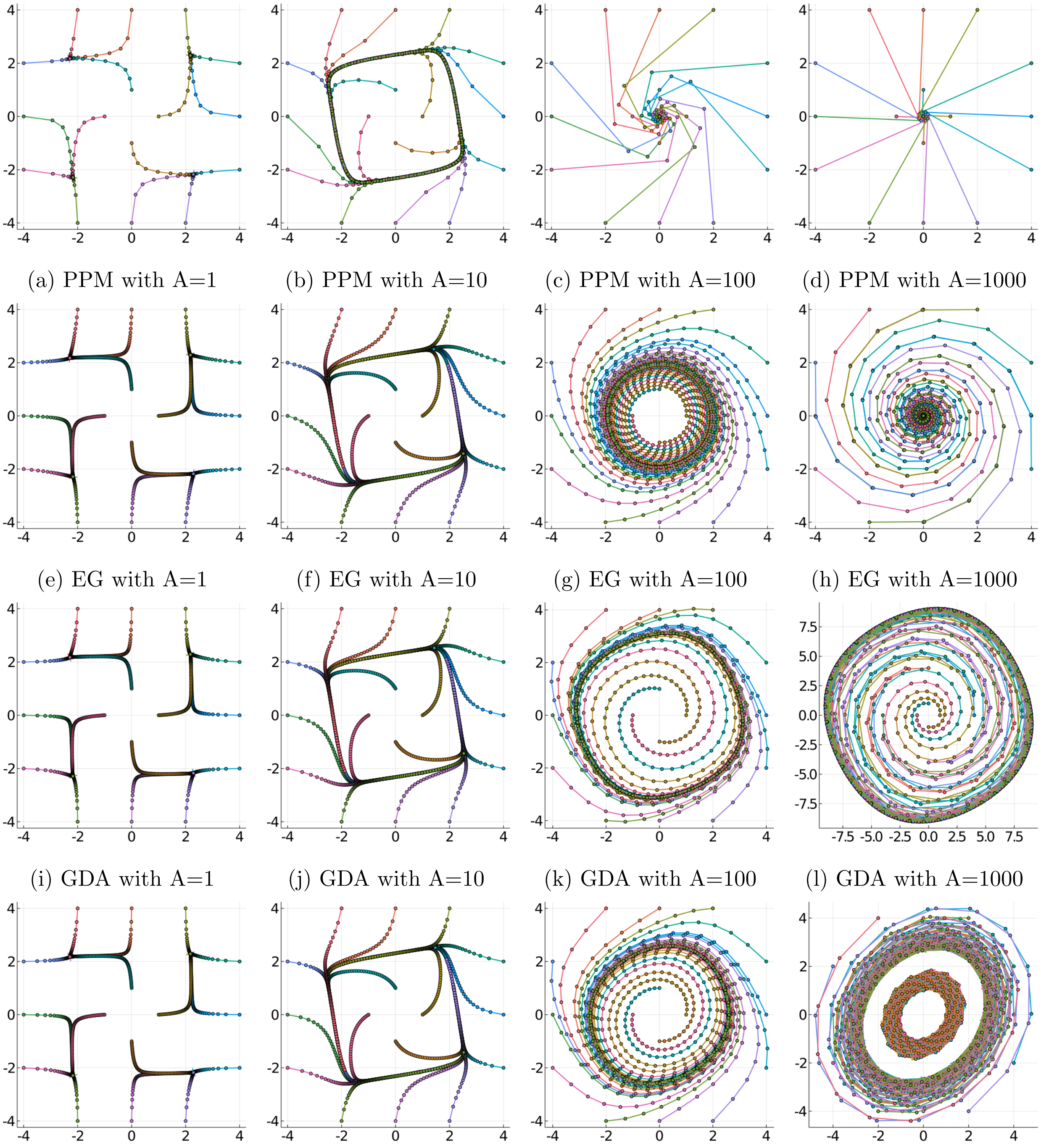
Figure 2 plots the solution paths of four common first-order methods for minimax problem for solving a two-dimensional nonconvex-nonconcave minimax problem:

$$\min_x \max_y L(x,y) = (x+3)(x+1)(x-1)(x-3) + Axy - (y+3)(y+1)(y-1)(y-3), \tag{31}$$

with four different levels of interaction term, $A = 1, 10, 100, 1000$. This problem is globally $\rho = 20$-weakly convex and $\beta = 172$-smooth on the box $[-4,4] \times [-4,4]$.

Each plot in Figure 2 shows the sample paths generated by running 100 iterations of the given method from the twelve different initial solutions around the boundary of the plot $(4,0)$, $(0,4)$, $(-4,0)$, $(0,-4)$, $(4,2)$, $(2,4)$, $(4,-2)$, $(2,-4)$, $(-4,2)$, $(-2,4)$, $(-4,-2)$, $(-2,-4)$ and four initial solutions towards the center of the plot $(1,0)$, $(0,1)$, $(-1,0)$, $(0,-1)$.

Plots (a)-(d) show the behavior of the Proximal Point Method (PPM) (6) with $\eta = 2\rho = 40$ and $\lambda = 1$. These figures match the landscape described by our theory: $A = 1$ is small enough to have local convergence to four different stationary points (each around $\{\pm 2\} \times \{\pm 2\}$), $A = 10$ has moderate size and every sample path is attracted into a limit cycle, and finally $A = 100$ and $A = 1000$ give a large enough interaction term to create a globally attractive stationary point (moreover, comparing plots (c) and (d) shows as A becomes larger the rate of convergence increases).



(a) PPM with A=1 (b) PPM with A=10 (c) PPM with A=100 (d) PPM with A=1000

(e) EG with A=1 (f) EG with A=10 (g) EG with A=100 (h) EG with A=1000

(i) GDA with A=1 (j) GDA with A=10 (k) GDA with A=100 (l) GDA with A=1000

(m) AGDA with A=1 (n) AGDA with A=10 (o) AGDA with A=100 (p) AGDA with A=1000

Figure 2: Sample paths of 100 iterations of four common first-order methods for minimax optimization, Proximal Point Method (PPM), Extragradient Method (EG), Gradient Descent Ascent (GDA) and Alternating Gradient Descent Ascent (AGDA), for solving (31) with different levels of interaction term $A = 1, 10, 100, 1000$.

Plots (e)-(h) show the behavior of the Extragradient Method (EG), which is defined by

$$\begin{bmatrix}\tilde{x}\\ \tilde{y}\end{bmatrix} = \begin{bmatrix}x_k\\ y_k\end{bmatrix} + s\begin{bmatrix}-\nabla_x L(x_k, y_k)\\ \nabla_y L(x_k, y_k)\end{bmatrix}$$
$$\begin{bmatrix}x_{k+1}\\ y_{k+1}\end{bmatrix} = \begin{bmatrix}x_k\\ y_k\end{bmatrix} + s\begin{bmatrix}-\nabla_x L(\tilde{x}, \tilde{y})\\ \nabla_y L(\tilde{x}, \tilde{y})\end{bmatrix} \tag{32}$$

with stepsize chosen as $s = 1/2(\beta + A) = 1/(344 + 2A)$. This stepsize was chosen since the objective function has a $\beta + A$-Lipschitz gradient. These figures show that the extragradient method follows the same general trajectory as described by our theory for the proximal point method. For small $A = 1$, local convergence occurs. For moderate sized $A = 10$ and $A = 100$, the algorithm falls into an attractive limit cycle, never converging. For large enough $A = 1000$, the method globally converges to a stationary point. The extragradient method only differs from the proximal point method's landscape in that it requires a larger A to transition into the interaction dominate regime.

Plots (i)-(l) show the behavior of Gradient Descent Ascent (GDA) (5) with $s = 1/2(\beta+A) = 1/(344+2A)$. This method is know to be unstable and diverge even for convex-concave problems. The same behavior carries over to our nonconvex-nonconcave example. For small A, we still see local convergence. However for $A = 10, 100, 1000$, we find that GDA falls into a limit cycle with increasingly large radius as A grows.

Lastly, plots (m)-(p) show the behavior of Alternating Gradient Descent Ascent (AGDA), defined by

$$x_{k+1} = x_k - s\nabla_x L(x_k, y_k)$$
$$y_{k+1} = y_k + s\nabla_y L(x_{k+1}, y_k) \tag{33}$$

with $s = 1/2(\beta + A) = 1/(344 + 2A)$. Again for small A, we still see local convergence, but for larger $A = 10, 100, 1000$, AGDA always falls into a limit cycle.

B Convex-Concave Optimization Analysis

B.1 Proof of Lemma 1.1

Observe that

$$\begin{aligned} M(x', y') &\le M(x, y') - \nabla_x M(x', y')^T(x - x') - \frac{\mu}{2}\|x - x'\|^2 \\ &\le M(x, y) + \nabla_y M(x, y)^T(y' - y) - \nabla_x M(x', y')^T(x - x') - \frac{\mu}{2}\|y - y'\|^2 - \frac{\mu}{2}\|x - x'\|^2 \end{aligned}$$

where the first inequality uses strong convexity of M in x and the second uses strong concavity in y. Symmetrically,

$$\begin{aligned} M(x', y') &\ge M(x', y) - \nabla_y M(x', y')^T(y - y') + \frac{\mu}{2}\|y - y^*\|^2 \\ &\ge M(x, y) + \nabla_x M(x, y)^T(x' - x) - \nabla_y M(x', y')^T(y - y') + \frac{\mu}{2}\|x - x'\|^2 + \frac{\mu}{2}\|y - y'\|^2. \end{aligned}$$

Combining the above two inequalities gives the first claimed inequality

$$\mu\left\|\begin{bmatrix}x - x'\\ y - y'\end{bmatrix}\right\|^2 \le \left(\begin{bmatrix}\nabla_x M(x, y)\\ -\nabla_y M(x, y)\end{bmatrix} - \begin{bmatrix}\nabla_x M(x', y')\\ -\nabla_y M(x', y')\end{bmatrix}\right)^T\begin{bmatrix}x - x'\\ y - y'\end{bmatrix}.$$

Furthermore, when $\nabla M(x', y') = 0$, we have

$$\mu\left\|\begin{bmatrix}x - x'\\ y - y'\end{bmatrix}\right\|^2 \le \|\nabla M(x, y)\|\left\|\begin{bmatrix}x - x'\\ y - y'\end{bmatrix}\right\|,$$

which finishes the proof of the second inequality. □

B.2 Proof of Theorem 1.2

First we use Lemma 1.1 to conclude that if the set S is large enough, M must have a stationary point in S. Now define $B(z,r) = \{z' | \|z - z'\| \le r\}$ as the closed Euclidean ball centered as a with radius r.

Lemma B.1. *Suppose M is μ-strongly convex-strongly concave in a set $B(x,r) \times B(y,r)$ for some fixed (x,y) and $r \ge 2\|\nabla M(x,y)\|/\mu$, then there exists a stationary point of M in $B((x,y), r/2)$.*

Proof. Consider the following constrained minimax problem $\min_{x' \in B(x,r)} \max_{y' \in B(y,r)} M(x,y)$. Since $M(x,y)$ is strongly convex-strongly concave, it must have a unique solution (x^*, y^*). The first-order optimality condition for (x^*, y^*) ensures

$$\nabla_x M(x^*, y^*) = -\lambda(x^* - x) \quad \text{and} \quad -\nabla_y M(x^*, y^*) = -\gamma(y^* - y)$$

for some constants $\lambda, \gamma \ge 0$ that are nonzero only if x^* or y^* are on the boundary of $B(x,r)$ and $B(y,r)$ respectively. Taking an inner product with $(x^* - x, y^* - y)$ gives

$$\begin{bmatrix} \nabla_x M(x^*, y^*) \\ -\nabla_y M(x^*, y^*) \end{bmatrix}^T \begin{bmatrix} x^* - x \\ y^* - y \end{bmatrix} = -\left\| \begin{bmatrix} \sqrt{\lambda}(x^* - x) \\ \sqrt{\gamma}(y^* - y) \end{bmatrix} \right\|^2 \le 0. \tag{34}$$

Applying Lemma 1.1 and utilizing (34), we conclude that

$$\mu \left\| \begin{bmatrix} x^* - x \\ y^* - y \end{bmatrix} \right\|^2 + \begin{bmatrix} \nabla_x M(x,y) \\ -\nabla_y M(x,y) \end{bmatrix}^T \begin{bmatrix} x^* - x \\ y^* - y \end{bmatrix} \le 0. \tag{35}$$

Hence

$$\left\| \begin{bmatrix} x^* - x \\ y^* - y \end{bmatrix} \right\|^2 \le \frac{1}{\mu} \left\| \begin{bmatrix} \nabla_x M(x,y) \\ -\nabla_y M(x,y) \end{bmatrix} \right\| \left\| \begin{bmatrix} x^* - x \\ y^* - y \end{bmatrix} \right\| ,$$

whereby

$$\left\| \begin{bmatrix} x^* - x \\ y^* - y \end{bmatrix} \right\| \le \frac{1}{\mu} \left\| \begin{bmatrix} \nabla_x M(x,y) \\ -\nabla_y M(x,y) \end{bmatrix} \right\| < r/2 ,$$

where the last inequality utilize the condition on r. Since (x^*, y^*) lies strictly inside the ball $B((x,y), r/2)$, the first-order optimality condition implies (x^*, y^*) is a stationary point of M. □

Lemma B.1 ensures the existence of a nearby stationary point (x^*, y^*). Then the standard proof of strongly monotone (from Lemma 1.1) and Lipschitz operators gives a contraction whenever $s \in (0, 2\mu/\beta^2)$:

$$\begin{aligned}
\left\| \begin{bmatrix} x_{k+1} - x^* \\ y_{k+1} - y^* \end{bmatrix} \right\|^2 &= \left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 - 2s \begin{bmatrix} \nabla_x M(x_k, y_k) \\ -\nabla_y M(x_k, y_k) \end{bmatrix}^T \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} + s^2 \left\| \begin{bmatrix} \nabla_x M(x_k, y_k) \\ -\nabla_y M(x_k, y_k) \end{bmatrix} \right\|^2 \\
&\le \left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 - 2\mu s \left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 + \beta^2 s^2 \left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 \\
&= \left(1 - 2\mu s + \beta^2 s^2\right) \left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 ,
\end{aligned}$$

where the inequality utilizes (35) at $(x,y) = (x_k, y_k)$ and the smoothness of $M(x,y)$. □

C Analysis for Interaction Moderate Region

C.1 Proof of Theorem 5.6

We introduce necessary terminology for the proof. Let $\mathcal{C}$ be a rectifiable Jordan curve in $\mathbb{R}^n$, and let $\mathcal{S}$ be the solution to Plateau's problem with respect to $\mathcal{C}$, i.e., $\mathcal{S}$ is a simply connected *minimal surface* with boundary $\mathcal{C}$ (for details, see for example [35]).

While our proof applies to the discrete time setting, cycling behavior is best defined with respect to a continuous time setting. Therefore, we need to take the limit of the dynamical system corresponding to our PPM algorithm (Equation 6), as step size goes to zero. The limiting dynamical system is a system of ODEs with the property that the solution (assuming the same initial condition as the discrete time system) will be a rectifiable closed curve in $\mathbb{R}^n$. Moreover, for any given positive ε, if we choose a small enough step-size $s(\epsilon)$ then the path, denoted $\mathcal{C}'$, taken by the discrete dynamical will lie within a tube of radius ε around the path $\mathcal{C}$.

Note that [13] already studies such limits of dynamical systems. There it is shown, that in general, for any loss function, and any positive ε, there exists a step-size $s(\varepsilon, T)$, such that PPM will remain with a tube of radius ε around $\mathcal{C}$ for at least T iterations.

We note that for our purposes, it's better if the choice of step-size s does not depend upon T – uniform convergence. The lower bound below is non-trivial if s does not depend upon T, therefore we will assume that from here on. The above discussion explains what we mean by *uniform convergence to a cyclic attarctor*.

Finally, suppose the PPM in Equation 6 runs for T iterations and traverses the curve $\mathcal{C}'$. Observe that there is a natural map $\pi : \mathcal{C} \mapsto \mathcal{C}'$ which maps points in $\mathcal{C}'$ to the corresponding closest point in $\mathcal{C}$. Therefore, for a small enough choice of $s(\varepsilon)$, one can think of $\mathcal{C}'$ as a curve which stays within ε of $\mathcal{C}$, and the former winds around the latter. Assume that $\mathcal{C}'$ winds around $\mathcal{C}$ for κ times, where κ can be a fraction. Define N to be the average number of iterations in a single traversal of $\mathcal{C}$. In other words, N is essentially measures the "coarseness" of our PPM algorithm.

Theorem C.1. *For any problem $L(x, y)$, if the limiting behaviour of PPM* (6)*, converges uniformly to a cyclic attractor $\mathcal{C}$, such that there exists a point on the minimal surface bounded by $\mathcal{C}$ that is at (geodesic) distance at least R from every point on $\mathcal{C}$ then, small enough choice of step size s and a large enough choice of T and S (where $T \gg S$), we have*

$$\frac{1}{T-S}\sum_{k=S}^{T}\|\nabla L(x_{k+1}, y_{k+1})\|^2 \geq \frac{C \cdot R^2}{s^2 N^2} \tag{36}$$

where C is a constant depending on the properties of the minimal surface $\mathcal{S}$.

Proof. Let ℓ and ℓ' be the arc lengths of $\mathcal{C}$ and $\mathcal{C}'$, respectively. Then $\ell' \simeq \kappa\ell$, where $\simeq$ denotes asymptotic equivalence as $\varepsilon \to 0$. It suffices to choose $\varepsilon \ll R$. Additionally, if $T - S$ is large enough, for a given choice of ε, then $\frac{\lfloor\kappa\rfloor}{\kappa} \to 1$, and we have $\ell' \simeq \lfloor\kappa\rfloor\ell$ for a small enough ε and large enough $T - S$.[6]

We know from the properties of PPM that

$$\ell' = s\sum_{k=S}^{T}\|\nabla L(x_{k+1}, y_{k+1})\|_2. \tag{37}$$

Moreover, by Cauchy-Schwarz inequality,

$$\sum_{k=S}^{T}\|\nabla L(x_{k+1}, y_{k+1})\|_2^2 \geq \frac{1}{T-S}\left(\sum_{k=S}^{T}\|\nabla L(x_{k+1}, y_{k+1})\|_2\right)^2. \tag{38}$$

Therefore, we have for a small enough ε,

$$\frac{1}{T-S}\sum_{k=S}^{T}\|\nabla L(x_{k+1}, y_{k+1})\|_2^2 \geq \frac{1}{(T-S)^2 s^2}\cdot(\lfloor\kappa\rfloor\ell)^2 \to \frac{\ell^2}{s^2N^2}. \tag{39}$$

[6] Here it is helpful that we have *uniform* convergence. Since the choice of s does not depend on the number of iterations $T - S$, otherwise we need to ensure the simultaneous existence of a small enough ε and large enough $T - S$ such that RHS of (36) is greater than zero, i.e., the result remains non-trivial.

where the last simplification above uses that $T - S$ is large enough so that $\frac{T-S}{\lfloor \kappa \rfloor} \to \frac{T-S}{\kappa} = N$.

However, the isoperimetric inequality for minimal surfaces, see for example Theorem 4.2 in [35], allows us to lower bound ℓ^2 in terms of the area of the minimal surface that is bounded by $\mathcal{C}$. Furthermore, by our assumption, there exists points on the minimal surface that is at distance R from each point on $\mathcal{C}$, we can lower bound the area of $\mathcal{S}$ enclosed by $\mathcal{C}$ by $C \cdot R^2$, for some constant C depending on the curvature of $\mathcal{S}$. Hence the proof follows.

□