OBTAINING GAME-STATIONARY POINTS FOR SMOOTH NONCONVEX–NONCONCAVE MINIMAX PROBLEMS VIA FIRST-ORDER METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

Minimax optimization is a fundamental framework in machine learning, robust optimization, and game theory, yet finding first-order stationary points of general nonconvex–nonconcave minimax problems remains challenging without additional structural assumptions. Existing guarantees often rely on global PL- or KL-type conditions that connect max-player stationarity to global inner optimality, or on Minty-type conditions that impose a global relation on the game gradient field relative to a reference solution; local KL variants relax the former requirement but typically require initialization and tracking within a near-optimal region. Such conditions may be difficult to satisfy in many applications. In contrast, we develop a first-order method that finds an ϵ-stationary point within $\widetilde{\mathcal{O}}(\epsilon^{-2})$ first-order iterations under a local inverse-Lipschitz regularity condition around approximate max-player stationary points, together with a compactness condition on a penalty sublevel set. Our condition places no optimality requirement on stationary points of the inner maximization problem: they need not be globally, or even locally, maximizing. We further provide sufficient conditions for the required regularity. In the unconstrained setting, it follows from uniform nonsingularity of the maximization-variable Hessian near stationary points; for constrained upper Moreau envelopes, it follows from standard KKT regularity conditions. These results establish first-order complexity guarantees for classes of nonconvex–nonconcave minimax problems not covered by the above PL-, KL-, or Minty-type frameworks.

1 INTRODUCTION

Minimax optimization is a basic modeling framework for machine learning and engineering, including adversarial training, robust optimization, generative modeling, multi-agent systems, fair representation learning, reinforcement learning and robust control (Goodfellow et al., 2014; Madry et al., 2018; Sinha et al., 2018; Madras et al., 2018; Dai et al., 2018; Nachum et al., 2019; Başar & Bernhard, 2008; Ozdaglar et al., 2023).

A minimax problem takes the following form:

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^m} F(x,y). \tag{1}$$

Convex–concave instances of (1) admit a mature theory based on monotone variational inequalities (Nemirovski, 2004; Mokhtari et al., 2020; Yazdandoost Hamedani & Aybat, 2021; Zhang et al., 2024; Zheng et al., 2026).When only the minimizing component is nonconvex, existing approaches include gradient descent–ascent methods, two-timescale algorithms, proximal-point-type methods, and smoothing techniques (Lin et al., 2020; Rafique et al., 2022; Nouiehed et al., 2019; Kong & Monteiro, 2021; Zhang et al., 2020; Yang et al., 2022). However, these methods do not readily extend to the *nonconvex–nonconcave* (NC-NC) regime, where the absence of convexity and concavity presents substantial additional challenges. Indeed, several existing works (Jin et al., 2020; Grimmer et al., 2023; Daskalakis et al., 2021; Diakonikolas et al., 2021) have highlighted the difficulties inherent in this setting. For instance, gradient descent–ascent methods may exhibit cyclic behavior or diverge, the choice of an appropriate local solution concept requires careful consideration and solving the general nonconvex–nonconcave problem is computationally challenging.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Motivated by these challenges, we study the nonconvex-nonconcave setting and our goal is to find a *first-order game stationary point*, namely, a pair (x, y) satisfying

$$\nabla_x F(x, y) = 0, \qquad \nabla_y F(x, y) = 0. \tag{2}$$

Accordingly, we consider an ϵ-approximate version of (2) defined by $\|\nabla_x F(x, y)\| + \|\nabla_y F(x, y)\| \leq \epsilon$. Finding an ϵ-stationary point under nonconvex-nonconcave setting remains a standard computational target (Zheng et al., 2023; Li et al., 2025; Tsaknakis & Hong, 2021). Although (2) is equivalent to $\nabla F = 0$, applying gradient descent to both variables can easily fail. Standard stationarity guarantees rely on F being bounded below, whereas descending y in minimax problem can drive F to $-\infty$.

1.1 Literature review

Many convergence guarantees for nonconvex–nonconcave problems relate the max-player stationarity residual to the global value $F^\star(x) := \sup_y F(x, y)$. A broad class of results assumes that, on a feasible set $\mathcal{Y}$, $\operatorname{dist}\big(0, -\nabla_y F(x, y) + N_{\mathcal{Y}}(y)\big) \geq c\big(F^\star(x) - F(x, y)\big)^\theta$, for some $c > 0$ and $\theta > 0$. This condition includes the unconstrained one-sided Polyak–Łojasiewicz (PL) condition with $\theta = \frac{1}{2}$ (Nouiehed et al., 2019; Yang et al., 2022), as well as the global Kurdyka–Łojasiewicz (KL) condition with more general values of θ (Zheng et al., 2023; Li et al., 2025). In particular, these assumptions imply that every max-player stationary point in their respective regions of validity is globally optimal.

Recent work relaxes this requirement by restricting the region on which the inequality is assumed to hold. Specifically, Lu & Wang (2026) impose a local KL condition on x-dependent near-optimal level sets, thereby allowing suboptimal stationary points outside these regions. Their guarantee, however, requires an initial y^0 that is already near-optimal for the inner problem at x^0 and uses successive inner solutions to initialize subsequent subproblems. Obtaining the first such solution requires either an initialization within the local KL region or additional favorable structure. Consequently, their guarantee does not cover arbitrary joint initialization.

A different line of work assumes a Minty variational inequality (MVI). In the unconstrained setting, it requires the existence of a point $z^\star$ such that $\langle G(z), z - z^\star\rangle \geq 0$ for all z, where $G = (\nabla_x F, -\nabla_y F)$. This condition implies that $z^\star$ is a global saddle point and, consequently, a game-stationary point, restricting its applicability to general nonconvex–nonconcave objectives, which limits its applicability to broad classes of nonconvex–nonconcave objectives (Zheng et al., 2023). The weak MVI relaxes this requirement, for example, to $\langle G(z), z - z^\star\rangle \geq -\eta\|G(z)\|^2$ for a game stationary reference point $z^\star$. Although subsequent work has expanded the admissible range of $\eta > 0$, the assumption still imposes a global condition on the gradient field relative to a single solution (Diakonikolas et al., 2021; Pethick et al., 2022; Fan et al., 2024; Alacaoglu et al., 2024).

Smooth merit functions provide another approach by reformulating game stationarity as a minimization problem (Raghunathan et al., 2019; Tsaknakis & Hong, 2021). Their global minimizers characterize the stationary points of the original game. However, these merit functions are generally nonconvex, so first-order methods typically provide guarantees only for approximate stationarity of the merit objective. Without additional structure, such guarantees do not certify game stationarity of the original problem. The challenge is therefore to construct a tractable reformulation whose approximate stationary points yield the desired certificate.

These limitations of existing approaches motivate the following question:

Can we achieve global convergence to game stationarity without requiring max-player stationarity to imply global optimality or imposing global conditions relative to a reference solution?

We answer this question affirmatively. We develop a penalty method that enforces y to be nearly stationary and prove that approximate stationarity of the penalty function can be transferred to the approximate game stationarity of the original problem under suitable regularity and growth assumptions. The penalty objective is smooth and admits an efficiently computable gradient. Building on these properties, our method computes an ϵ-approximate game stationary point using $\widetilde{O}(\epsilon^{-2})$ total gradient evaluations.

1.2 Our contributions

Summary of contributions. We address the preceding question through a smooth minimization reformulation. First, under local regularity and growth assumptions, we construct a function whose approximate stationary points are approximate game stationary points of the original objective, using a fixed finite penalty parameter. Second, we develop an inexact gradient descent method with $\widetilde{O}(\varepsilon^{-2})$ first-order oracle complexity. Third, we provide sufficient conditions for the required regularity, covering nonsingular Hessians at stationary points for max problem and a class of constrained problems satisfying strict complementarity and reduced-Hessian nondegeneracy.

Our formulation. For an L-smooth objective F and fix $p > L$, define the upper proximal point of $F(x, \cdot)$ at y and the associated stationarity gap by

$$w(x, y) := \arg\max_{v \in \mathbb{R}^m} \left\{ F(x, v) - \frac{p}{2}\|v - y\|^2 \right\}, \tag{3}$$

$$\mathcal{G}(x, y) := F(x, w(x, y)) - \frac{p}{2}\|w(x, y) - y\|^2 - F(x, y). \tag{4}$$

The inner problem in (3) is strongly concave, and its gap $\mathcal{G}$ is nonnegative and vanishes exactly when $\nabla_y F(x, y) = 0$ (Lemma 2.1). We therefore minimize

$$P_\rho(x, y) := F(x, y) + \rho\mathcal{G}(x, y). \tag{5}$$

This penalty function has a Lipschitz gradient that can be evaluated using first-order information and a strongly concave response solve.

Stationary exactness and complexity. Our key result transfers approximate stationarity of P_ρ to stationarity of F. Under the stated local regularity, growth, and compact-sublevel assumptions, a sufficiently large finite ρ ensures

$$\|\nabla P_\rho(x, y)\| \le \varepsilon \quad \Longrightarrow \quad \|\nabla_x F(x, y)\| + \|\nabla_y F(x, y)\| \le C_\rho \varepsilon, \tag{6}$$

on the relevant sublevel set whenever the merit gradient is sufficiently small. Here C_ρ is independent of the target accuracy. Thus, stationary points of P_ρ are game stationary points of F (Corollary 3.2). This transfer yields an inexact gradient descent method requiring $\widetilde{O}(\varepsilon^{-2})$ gradient evaluations of F to find a ε-game stationary point (Theorem 3.5). This is the canonical rate of gradient descent for finding an ε-stationary point of a smooth nonconvex function, now obtained for game stationarity of a nonconvex–nonconcave minimax problem.

Local regularity beyond global gradient domination. Our local regularity condition compares the proximal displacement $w(x, y) - y$ with the corresponding change in $\nabla_y F$. This condition is imposed locally near max-player stationary points and holds uniformly over the relevant compact sublevel set (Assumption 3). For C^2 objectives, Proposition 2.3 shows that the condition is satisfied whenever $\nabla^2_{yy} F$ is nonsingular at all max-player stationary points on this compact set. In particular, it allows for local minima, maxima and saddle points on the max-player side. (Proposition 2.3). When strongly convex constraints are introduced, we further show that the corresponding upper Moreau envelope satisfies this condition under strict complementarity and nondegeneracy of the reduced Lagrangian Hessian (Proposition 2.4).

2 Setup and the regularized gap penalty

Throughout, vector norms denote Euclidean norms, and matrix norms denote operator norms. For the minimax problem (1), we first make the following assumption.

Assumption 1 (Smoothness)**.** The function $F : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ is continuously differentiable, and its gradient is L-Lipschitz continuous; that is, for any $x_1, x_2 \in \mathbb{R}^d$ and $y_1, y_2 \in \mathbb{R}^m$,

$$\|\nabla F(x_1, y_1) - \nabla F(x_2, y_2)\| \le L(\|x_1 - x_2\| + \|y_1 - y_2\|). \tag{7}$$

Recall that fix $p > L$, the proximal point $w(x, y)$, the stationarity gap $\mathcal{G}(x, y)$, and the penalty function $P_\rho(x, y)$ for $\rho > 0$, are defined in (3), (4), and (5), respectively. Then we have the following lemma.

Lemma 2.1. *Under Assumption 1, the following statements hold.*

(i) The maximization problem in (3) *is* $(p-L)$*-strongly concave and has a unique maximizer* $w(x,y)$*. Moreover,*

$$\nabla_y F(x, w(x,y)) = p(w(x,y) - y). \tag{8}$$

(ii) The map w *is globally Lipschitz with constant* $\kappa_w := \frac{\sqrt{L^2+p^2}}{p-L}$*. More precisely, for* $z_i = (x_i, y_i)$ *and* $w_i = w(x_i, y_i)$ *with* $i = 1, 2$*,*

$$\|w_1 - w_2\| \le \frac{L\|x_1 - x_2\| + p\|y_1 - y_2\|}{p-L} \le \frac{\sqrt{L^2+p^2}}{p-L}\|z_1 - z_2\|. \tag{9}$$

(iii) With $w = w(x,y)$*, we have*

$$\mathcal{G}(x,y) \ge \frac{p-L}{2}\|w - y\|^2. \tag{10}$$

In addition,

$$\frac{1}{2(p+L)}\|\nabla_y F(x,y)\|^2 \le \mathcal{G}(x,y) \le \frac{1}{2(p-L)}\|\nabla_y F(x,y)\|^2. \tag{11}$$

Consequently,

$$\mathcal{G}(x,y) = 0 \quad \Longleftrightarrow \quad \nabla_y F(x,y) = 0. \tag{12}$$

(iv) The gap $\mathcal{G}(x,y)$ *is continuously differentiable and*

$$\nabla_x \mathcal{G}(x,y) = \nabla_x F(x,w) - \nabla_x F(x,y), \tag{13}$$
$$\nabla_y \mathcal{G}(x,y) = \nabla_y F(x,w) - \nabla_y F(x,y) = p(w-y) - \nabla_y F(x,y). \tag{14}$$

Consequently,

$$\nabla_x P_\rho(x,y) = \nabla_x F(x,y) + \rho\big[\nabla_x F(x,w) - \nabla_x F(x,y)\big], \tag{15}$$
$$\nabla_y P_\rho(x,y) = \nabla_y F(x,y) + \rho\big[\nabla_y F(x,w) - \nabla_y F(x,y)\big]. \tag{16}$$

(v) The function P_ρ *has a globally Lipschitz gradient with constant*

$$L_\rho = L + \rho L(1 + \sqrt{1+\kappa_w^2}). \tag{17}$$

The proof relies primarily on the $(p-L)$-strong concavity of the problem in (3) and Danskin's theorem. The detailed proof is provided in Appendix A.1.

Remark 2.2 (Only the max-player curvature constrains p). Let $L_y \le L$ be a Lipschitz constant for $\nabla_y F(x,\cdot)$, uniformly in x. The condition $p > L$ is used only to ensure the strong concavity of the inner problem in (3). Thus, all results remain valid under the weaker condition $p > L_y$, with every occurrence of $p - L$ replaced by $p - L_y$ in Lemma 2.1 and Section 3.

Consider $C_0 \ge \max_y F(x^0, y) + 1$ for some initial x^0. We make the following assumptions.

Assumption 2. There exists a constant $\rho_C \ge 1$ and a compact set $\mathcal{K}$ such that

$$\{z := (x,y) \in \mathbb{R}^{d+m} : P_\rho(z) \le C_0\} \subseteq \mathcal{K}, \quad \forall \rho \ge \rho_C. \tag{18}$$

Assumption 3 (Local inverse–Lipschitz regularity)**.** There exist constants $\eta_s > 0$, $\delta_s > 0$, and $\kappa_s > 0$ such that the following holds. For every $(x,y) \in \mathcal{K}$, let $w = w(x,y)$. If

$$\max\{\|\nabla_y F(x,y)\|, \|\nabla_y F(x,w)\|\} \le \eta_s, \qquad \|w - y\| \le \delta_s, \tag{19}$$

then

$$\|w - y\| \le \kappa_s \|\nabla_y F(x,w) - \nabla_y F(x,y)\|. \tag{20}$$

We next present two settings under which Assumption 3 is satisfied. The first setting concerns the nondegeneracy of the max-player Hessian and is formalized in the following proposition.

Proposition 2.3. *Suppose F is twice continuously differentiable on a neighborhood of $\mathcal{K}$, and there exists $\mu_0 > 0$ such that*

$$\sigma_{\min}\big(\nabla^2_{yy} F(x,y)\big) \geq \mu_0, \tag{21}$$

for every $(x,y) \in \mathcal{K}$ satisfying $\nabla_y F(x,y) = 0$. Then Assumption 3 holds.

Note that (21) imposes no sign condition on $\nabla^2_{yy} F$ and it excludes only zero eigenvalues at max-player critical points in $\mathcal{K}$. Detailed proof is provided in Appendix A.2.1.

Although problem (1) is unconstrained, practical applications often restrict the max-player to a feasible set. A representative example is budget-constrained adversarial training (Madry et al., 2018), where inner perturbations u must satisfy $c(u) \leq 0$. By leveraging the upper Moreau envelope technique, such problems fit seamlessly into our framework. Moreover, when a perturbation exhausts its budget while the loss continues to rise along the outward normal direction, the constraint prevents further ascent (Kanai et al., 2023) – yielding a strictly positive KKT multiplier at an active max-player stationary point. Motivated by this intuition, we give our second sufficient setting for Assumption 3.

Proposition 2.4. *Let $\lambda > 0$, and suppose that F admits the representation*

$$F(x,y) = \max_{u:\, c(u) \leq 0} \left\{ g(x,u) - \frac{\lambda}{2}\|u-y\|^2 \right\}. \tag{22}$$

Assume the following conditions.

(i) The functions g and c are twice continuously differentiable, $\nabla_u g(x,\cdot)$ is L_g-Lipschitz uniformly in x, and $\lambda > L_g$.

(ii) The function c is μ_c-strongly convex for some $\mu_c > 0$ and satisfies the Slater condition: there exists u^{s} such that $c(u^{\mathrm{s}}) < 0$.

(iii) Define

$$\mathcal{K}^+ := \left\{ (x,y) : \inf_{(\bar{x},\bar{y}) \in \mathcal{K}} \|(x,y) - (\bar{x},\bar{y})\| \leq 1 \right\}.$$

There exists $\eta_M > 0$ such that strict complementarity holds for every KKT point associated with $(x,y) \in \mathcal{K}^+$ satisfying $\|\nabla_y F(x,y)\| \leq \eta_M$. Namely, if $u = u(x,y)$ is the unique maximizer in (22) and $\alpha = \alpha(x,y)$ is its multiplier, then

$$c(u) = 0 \quad \Longrightarrow \quad \alpha > 0. \tag{23}$$

(iv) At every KKT point described in part (iii), define

$$H(x,u,\alpha) := -\nabla^2_{uu} g(x,u) + \alpha \nabla^2 c(u).$$

If $c(u) < 0$, then $H(x,u,0)$ is nonsingular. If $c(u) = 0$, let $Z(u)$ have orthonormal columns spanning $\{d : \nabla c(u)^\top d = 0\}$. Then the reduced Lagrangian Hessian $Z(u)^\top H(x,u,\alpha) Z(u)$ is nonsingular.

Then Assumption 3 holds.

Note that this result does not require $g(x,\cdot)$ to be convex or concave. The proof is provided in Appendix A.2.2.

We next provide a concrete example satisfying Assumption 3.

Example 2.5. Consider a label-adversarial regression with $g(x,u) := \frac{1}{2}(x-u)^2 + r(x)$, $r(x) := 10x^2 + 23(\cos x - 1)$, and $|u| \leq 5$. Its upper-Moreau reformulation is

$$F(x,y) := \max_{u^2 - 25 \leq 0} \left\{ g(x,u) - (u-y)^2 \right\}, \tag{24}$$

for which $\max_y F(x,y) = \max_{u^2-25\leq 0} g(x,u)$. There is a stationary point $(0,0)$ with $F(0,0) = 0$ while $\max_y F(0,y) = 25/2$. Hence, any global PL or gap-based KL inequality imposed at $(0,0)$ fails. Our condition nonetheless holds, with $\kappa_s = 1/2$. Details are shown in Appendix B.

3 STATIONARITY TRANSFER AND GRADIENT-DESCENT COMPLEXITY

3.1 APPROXIMATE STATIONARITY TRANSFER

Define

$$\underline{F}_{\mathcal{K}} := \min_{z\in\mathcal{K}} F(z), \qquad B := C_0 - \underline{F}_{\mathcal{K}}, \qquad D := \sqrt{\frac{2B}{p-L}}, \qquad C_y := (p+L)D.$$

Note that B is finite and nonnegative. With constants $\eta_s, \delta_s, \kappa_s$ from Assumption 3, we set

$$\bar{\rho} := \max\left\{\rho_c, \frac{C_y^2}{\eta_s^2}, \frac{p^2D^2}{\eta_s^2}, \frac{D^2}{\delta_s^2}, 2(1+p\kappa_s)\right\}. \tag{25}$$

We are now ready to establish the relationship between the stationary points of P_ρ and F in the following theorem and corollary.

Theorem 3.1. *Suppose Assumptions 1–3 hold. Let $\rho \geq \bar{\rho}$, and suppose that (x, y) satisfies*

$$P_\rho(x,y) \leq C_0, \qquad \|\nabla P_\rho(x,y)\| \leq \varepsilon. \tag{26}$$

Let $w = w(x,y)$ and let $\Delta_y = \nabla_y F(x,w) - \nabla_y F(x,y)$ be the gradient difference. Then

$$\|\Delta_y\| \leq \frac{\varepsilon}{\rho - 1 - p\kappa_s}, \qquad \|w - y\| \leq \frac{\kappa_s \varepsilon}{\rho - 1 - p\kappa_s}, \tag{27}$$

$$\|\nabla_y F(x,y)\| \leq \frac{(1+p\kappa_s)\varepsilon}{\rho - 1 - p\kappa_s}, \qquad \|\nabla_x F(x,y)\| \leq \left(1 + \frac{\rho L \kappa_s}{\rho - 1 - p\kappa_s}\right)\varepsilon. \tag{28}$$

Consequently,

$$\|\nabla_x F(x,y)\| + \|\nabla_y F(x,y)\| \leq C_\rho \varepsilon, \qquad C_\rho := 1 + \frac{\rho L\kappa_s + 1 + p\kappa_s}{\rho - 1 - p\kappa_s}. \tag{29}$$

In particular, every stationary point of P_ρ in the common low-level region is a stationary point of F.

The proof is provided in Appendix A.3

Corollary 3.2. *Under the assumptions of Theorem 3.1, let $\rho \geq \bar{\rho}$. If*

$$P_\rho(x,y) \leq C_0, \qquad \nabla P_\rho(x,y) = 0, \tag{30}$$

then

$$\nabla_x F(x,y) = 0, \qquad \nabla_y F(x,y) = 0. \tag{31}$$

Proof. Apply Theorem 3.1 with $\varepsilon = 0$. Then (27) gives $w(x,y) = y$, and (29) gives the conclusion. □

3.2 EXACT GRADIENT DESCENT ON THE PENALTY OBJECTIVE

We first show how to obtain an initial point $z^0 = (x^0, y^0)$ in the common low-level set, that is, $P_\rho(z^0) \leq C_0$ for a prescribed $\rho \geq \bar{\rho}$.

Proposition 3.3. *Suppose Assumption 1 holds and fix $\rho > 0$. Starting from any $u^0 \in \mathbb{R}^m$, consider the gradient ascent iteration $u^{t+1} = u^t + \alpha \nabla_y F(x^0, u^t)$ within $T := \left\lceil \frac{\rho\left(\overline{F}(x^0) - F(x^0,u^0)\right)}{\alpha(2-\alpha L)(p-L)} \right\rceil$ steps. Then there exists $0 \leq t \leq T$ such that $\|\nabla_y F(x^0,u^t)\|^2 \leq \frac{2(p-L)}{\rho}$, and consequently $P_\rho(x^0,u^t) \leq C_0$. In particular, $y^0 := u^t$ yields $P_\rho(x^0,y^0) \leq C_0$.*

Proof. Since $F(x^0,\cdot)$ is L-smooth, bounded above by $\overline{F}(x^0)$, and $\alpha < \frac{1}{L}$, the standard convergence analysis of gradient ascent with step size $\alpha \leq \frac{1}{L}$ gives $\min_{0\leq t\leq T} \|\nabla_y F(x^0,u^t)\|^2 \leq \frac{2(p-L)}{\rho}$. For such t, the upper bound in (11) yields

$$P_\rho(x^0,u^t) = F(x^0,u^t) + \rho\,\mathcal{G}(x^0,u^t) \leq \overline{F}(x^0) + \frac{\rho}{2(p-L)}\|\nabla_y F(x^0,u^t)\|^2 \leq \overline{F}(x^0) + 1 \leq C_0.$$

□

Since ρ is fixed throughout, T is a constant independent of the target accuracy ϵ, and this initialization does not affect the complexity bounds below. For a fixed $\rho \geq \bar{\rho}$, let $z^0 := (x^0, y^0)$ be an initial point satisfying Assumption 2, and consider

$$z^{k+1} = z^k - \frac{1}{L_\rho} \nabla P_\rho(z^k). \tag{32}$$

Here L_ρ is given in (17). Equivalently, if $w^k = w(x^k, y^k)$, the gradient in (32) is computed from

$$\nabla P_\rho(x^k, y^k) = \nabla F(x^k, y^k) + \rho\big[\nabla F(x^k, w^k) - \nabla F(x^k, y^k)\big].$$

Given w^k, the outer gradient requires evaluations of ∇F at (x^k, y^k) and (x^k, w^k). The point w^k is the unique solution of the strongly concave problem in (3). We next establish the convergence complexity of the exact gradient descent method.

Theorem 3.4. *Suppose Assumptions 1–3 hold, and fix $\rho \geq \bar{\rho}$. Let*

$$P_c^* := \min_{z \in \mathcal{K}} P_{\rho_c}(z), \qquad \Delta_c := C_0 - P_c^*. \tag{33}$$

Then the iterates in (32) *satisfy*

$$P_\rho(z^{k+1}) \leq P_\rho(z^k) - \frac{1}{2L_\rho} \|\nabla P_\rho(z^k)\|^2, \tag{34}$$

and remain in the common compact set $\mathcal{K}$. Moreover, for every integer $N \geq 1$,

$$\min_{0 \leq k < N} \|\nabla P_\rho(z^k)\|^2 \leq \frac{2L_\rho \Delta_c}{N}. \tag{35}$$

Consequently, if

$$N \geq \frac{2L_\rho \Delta_c C_\rho^2}{\epsilon^2}, \tag{36}$$

then at least one iterate $z^k = (x^k, y^k)$ satisfies

$$\|\nabla_x F(x^k, y^k)\| + \|\nabla_y F(x^k, y^k)\| \leq \epsilon. \tag{37}$$

For every fixed $\rho \geq \bar{\rho}$, the outer iteration complexity is $O(\epsilon^{-2})$.

The proof in Appendix A.4.1 applies the descent lemma to P_ρ, uses $P_\rho \geq P_{\rho_c}$ to ensure that all iterates remain in $\mathcal{K}$, and invokes Theorem 3.1 at the iterate with the smallest gradient norm.

3.3 Inexact gradient descent on the penalty objective

In practice, exact optimal solutions are often not available, and we instead obtain an approximate best response $\widetilde{w}^k$ satisfying

$$\|\widetilde{w}^k - w(x^k, y^k)\| \leq \delta, \tag{38}$$

for some $\delta > 0$. We then replace the exact penalty gradient by an inexact gradient $\widetilde{g}^k$ and perform the update

$$z^{k+1} = z^k - \frac{1}{L_\rho} \widetilde{g}^k, \qquad \widetilde{g}^k := \nabla F(x^k, y^k) + \rho\big[\nabla F(x^k, \widetilde{w}^k) - \nabla F(x^k, y^k)\big]. \tag{39}$$

Given (38), the inexact gradient error satisfies

$$\|\widetilde{g}^k - \nabla P_\rho(z^k)\| = \rho \|\nabla F(x^k, \widetilde{w}^k) - \nabla F(x^k, w^k)\| \leq \rho L \|\widetilde{w}^k - w^k\|. \tag{40}$$

We next establish the convergence complexity of the inexact gradient descent method.

Theorem 3.5. *Suppose Assumptions 1–3 hold, and fix $\rho \geq \bar{\rho}$. Let P_c^* and Δ_c be defined as in* (33)*, and suppose that* (38) *holds at every iteration. Then the iterates in* (39) *satisfy*

$$P_\rho(z^{k+1}) \leq P_\rho(z^k) - \frac{1}{2L_\rho} \|\nabla P_\rho(z^k)\|^2 + \frac{\rho^2 L^2 \delta^2}{2L_\rho}. \tag{41}$$

Moreover, for every integer $N \geq 1$ there exists an index k with $0 \leq k < N$ such that

$$P_\rho(z^j) \leq C_0 \quad \textit{for all } 0 \leq j \leq k, \qquad \textit{in particular } P_\rho(z^k) \leq C_0 \textit{ and } z^k \in \mathcal{K}, \tag{42}$$

and

$$\|\nabla P_\rho(z^k)\|^2 \leq \frac{2L_\rho \Delta_c}{N} + \rho^2 L^2 \delta^2. \tag{43}$$

For this iterate,

$$\|\nabla_x F(x^k, y^k)\| + \|\nabla_y F(x^k, y^k)\| \leq C_\rho \sqrt{\frac{2L_\rho \Delta_c}{N} + \rho^2 L^2 \delta^2}. \tag{44}$$

In particular, if (36) *holds and $\delta \leq \epsilon$, then at least one iterate $z^k = (x^k, y^k)$ satisfies*

$$\|\nabla_x F(x^k, y^k)\| + \|\nabla_y F(x^k, y^k)\| \leq (1 + C_\rho \rho L)\epsilon. \tag{45}$$

For every fixed $\rho \geq \bar{\rho}$, choosing $\delta = O(\epsilon)$ therefore preserves the $O(\epsilon^{-2})$ outer iteration complexity.

The proof, in Appendix A.4.2, follows that of Theorem 3.4: the gradient error (40) contributes the additive term, and descent holds while $\|\nabla P_\rho(z^k)\| \geq \rho L \delta$, which keeps the selected iterate in $\mathcal{K}$.

Corollary 3.6 (Total gradient complexity)**.** *The best-response problem in* (3) *is strongly concave and smooth, so gradient ascent initialized at y^k returns $\widetilde{w}^k$ satisfying* (38) *after $\mathcal{O}(\log(1/\delta))$ gradient evaluations. Taking $\delta \leq \frac{\epsilon}{\sqrt{2}C_\rho \rho L}$, the total number of gradient evaluations of F needed to reach an ϵ-stationary point is $\widetilde{\mathcal{O}}(\epsilon^{-2})$.*

4 Numerical experiments

In this section we evaluate our method through quadratic examples and compare its performance with other relevant algorithms. We consider the quadratic family

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^{2d}} F_d(x, y) := -\frac{1}{2}\|x\|_2^2 + x^\top \begin{bmatrix} I_d & 2I_d \end{bmatrix} Q_d^\top y + \frac{1}{2} y^\top Q_d \begin{bmatrix} I_d & 0 \\ 0 & -I_d \end{bmatrix} Q_d^\top y, \tag{46}$$

where $Q_d \in \mathbb{R}^{2d \times 2d}$ is orthogonal. We test $d = 1$ with $Q_1 = I_2$ and $d = 10$ with a fixed random orthogonal Q_{10}, giving problems of total dimension 3 and 30, respectively. In the coordinates $\widetilde{y} = Q_d^\top y$, the objective decomposes into d identical three-variable quadratic blocks. Consequently, both problems have the unique game stationary point $z^\star = (0, 0)$, while $\nabla^2_{yy} F_d$ has eigenvalues $+1$ and -1, each with multiplicity d. Thus $y^\star$ is a saddle point of $F_d(x^\star, \cdot)$. The rotation mixes the displayed y-coordinates but preserves the Euclidean geometry and the dynamics of the gradient methods under the corresponding change of variables.

In both experiments, we compare inexact gradient descent on P_ρ with joint GD on F_d using step size 0.05 and simultaneous GDA with $(\tau_x, \tau_y) \in \{(0.05, 0.05), (0.01, 0.2), (0.1, 0.01)\}$. All methods use the same initial point within each experiment. Each regularized response is computed by gradient ascent initialized at the current y^k. We report the distance $\|z^k - z^\star\|_2$ and the stationarity residual

$$r_k := \|\nabla_x F_d(z^k)\|_2 + \|\nabla_y F_d(z^k)\|_2,$$

both against outer updates and, for the residual, against gradient evaluations including all inner iterations. Runs are terminated upon reaching $r_k \leq 10^{-6}$ or $\|z^k\|_2 \geq 10^6$.

For the three-dimensional experiment, we take $z^0 = (1, -1, 2)$ and $\rho = 16$, with outer step size $1/22$ and inner ascent step size $1/6$. For the thirty-dimensional experiment, we use a random unit-norm z^0, $p = 3.75$, and $\rho = 9.50$, with outer step size 0.076 and inner ascent step size 0.21. In the latter experiment, the inner solve stops when the norm of the gradient of the regularized inner objective is at most 10^{-10}.

The proposed method reaches $r_k \leq 10^{-6}$ in both experiments: 1,120 outer updates and 14,586 gradient evaluations in dimension 3, and 654 outer updates and 9,910 gradient evaluations in dimension 30. In contrast, GD and all three GDA configurations reach $\|z^k\|_2 \geq 10^6$ within 93–501 updates in dimension 3 and 81–536 updates in dimension 30 (Figures 1 and 2). These results illustrate convergence of the proposed method to game stationarity despite indefinite max-player curvature. Since the rotated problem preserves the block spectrum and uses different numerical parameters, the iteration counts are not intended as a comparison of convergence speed across dimensions.

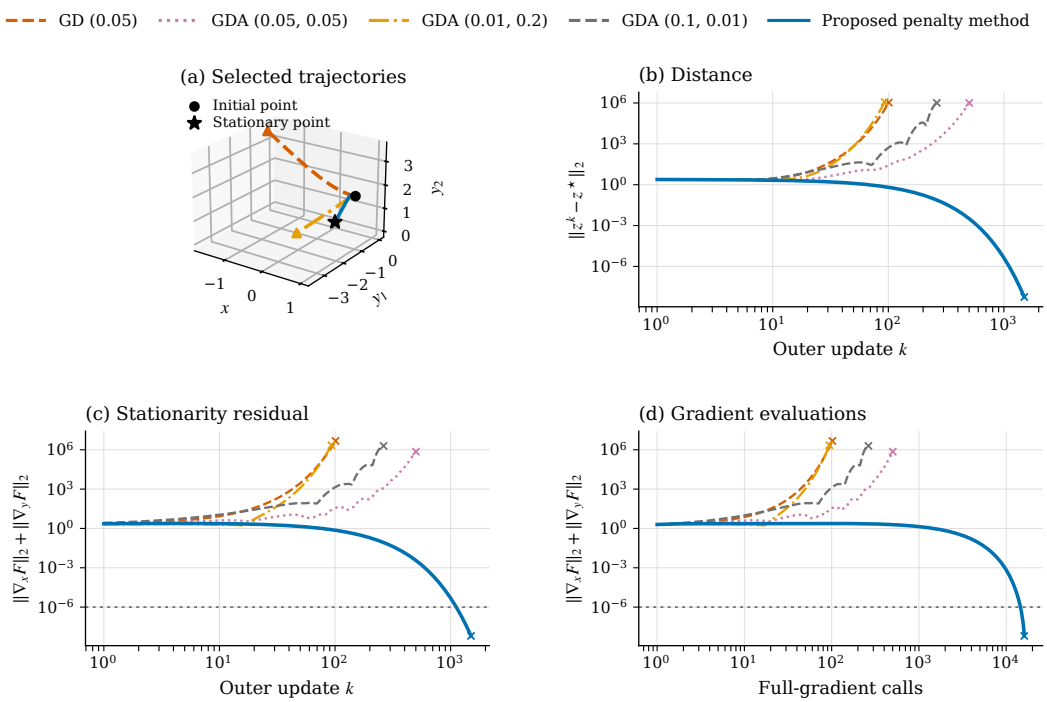


Figure 1: Three-dimensional comparison for (46) with $d = 1$, $Q_1 = I_2$, and $z^0 = (1, -1, 2)$. (a) Trajectories up to their first exit from $\|z\|_2 \le 4$. (b) Distance to $z^\star$. (c) Stationarity residual versus outer update. (d) Residual versus gradient evaluations, including inner solves.

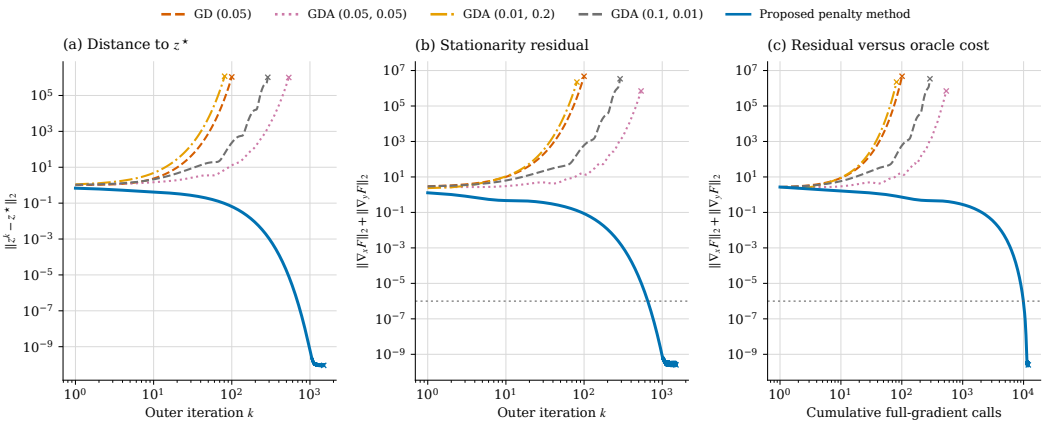


Figure 2: Thirty-dimensional comparison for (46) with $d = 10$, a fixed random orthogonal Q_{10}, and a common random unit-norm initialization. (a) Distance to $z^\star$. (b) Stationarity residual versus outer update. (c) Residual versus gradient evaluations, including inner solves.

AI DISCLOSURE STATEMENT

GPT-5.6 was utilized solely for editing and improving the clarity and readability of the writing in this paper.

REFERENCES

Ahmet Alacaoglu, Donghwan Kim, and Stephen J Wright. Revisiting inexact fixed-point iterations for min-max problems: Stochasticity and structured nonconvexity. *arXiv preprint arXiv:2402.05071*, 2024.

Tamer Başar and Pierre Bernhard. *H^∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Birkhäuser, 2 edition, 2008. doi: 10.1007/978-0-8176-4757-5.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1125–1134. PMLR, 2018.

Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 1466–1478. ACM, 2021. doi: 10.1145/3406325.3451125.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2746–2754, 2021.

Yifeng Fan, Yongqiang Li, and Bo Chen. Weaker MVI condition: Extragradient methods with multi-step exploration. In *International Conference on Learning Representations*, 2024.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2672–2680, 2014.

Benjamin Grimmer, Haihao Lu, Pratik Worah, and Vahab Mirrokni. The landscape of the proximal point method for nonconvex–nonconcave minimax optimization. *Mathematical Programming*, 2023. doi: 10.1007/s10107-022-01910-8.

Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. What is local optimality in nonconvex–nonconcave minimax optimization? In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4880–4889. PMLR, 2020.

Sekitoshi Kanai, Masanori Yamada, Hiroshi Takahashi, Yuki Yamanaka, and Yasutoshi Ida. Relationship between nonsmoothness in adversarial training, constraints of attacks, and flatness in the input space. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10817–10831, 2023.

Weiwei Kong and Renato D. C. Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *SIAM Journal on Optimization*, 31(4):2558–2585, 2021. doi: 10.1137/20M1313222.

Jiajin Li, Linglingzhi Zhu, and Anthony Man-Cho So. Nonsmooth nonconvex–nonconcave minimax optimization: Primal–dual balancing and iteration complexity analysis. *Mathematical Programming*, 214(1–2):591–641, 2025. doi: 10.1007/s10107-025-02197-1.

Tianyi Lin, Chi Jin, and Michael I. Jordan. On gradient descent ascent for nonconvex–concave minimax problems. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6083–6093. PMLR, 2020.

Zhaosong Lu and Xiangyuan Wang. A first-order method for constrained nonconvex-nonconcave minimax optimization. *Mathematical Programming*, 2026. doi: 10.1007/s10107-026-02415-4.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3384–3393. PMLR, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1497–1507. PMLR, 2020.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

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. doi: 10.1137/S1052623403425629.

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*, volume 32, pp. 14905–14916, 2019.

Asuman E. Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. Revisiting the linear-programming framework for offline RL with general function approximation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26769–26791. PMLR, 23–29 Jul 2023.

Thomas Pethick, Puya Latafat, Panos Patrinos, Olivier Fercoq, and Volkan Cevher. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *International Conference on Learning Representations*, 2022.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: Provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022. doi: 10.1080/10556788.2021.1895152.

Arvind Raghunathan, Anoop Cherian, and Devesh Jha. Game theoretic optimization via gradient-based nikaido-isoda function. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5291–5300. PMLR, 09–15 Jun 2019.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Ioannis Tsaknakis and Mingyi Hong. Finding first-order nash equilibria of zero-sum games with the regularized nikaido-isoda function. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1189–1197. PMLR, 13–15 Apr 2021.

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 5485–5517. PMLR, 2022.

Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021. doi: 10.1137/18M1213488.

Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhi-Quan Luo. A single-loop smoothed gradient descent–ascent algorithm for nonconvex–concave min–max problems. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7377–7389, 2020.

Xuan Zhang, Necdet Serhat Aybat, and Mert Gürbüzbalaban. Robust accelerated primal-dual methods for computing saddle points. *SIAM Journal on Optimization*, 34(1):1097–1130, 2024. doi: 10.1137/21M1462775.

Taoli Zheng, Linglingzhi Zhu, Anthony Man-Cho So, Jose Blanchet, and Jiajin Li. Universal gradient descent ascent method for nonconvex–nonconcave minimax optimization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 54075–54110, 2023.

Taoli Zheng, Jiajin Li, and Anthony Man-Cho So. Last-iterate convergence of single-loop stochastic methods for constrained convex-concave minimax problems. *arXiv preprint arXiv:2607.11056*, 2026.

A PROOFS

A.1 PROOF OF LEMMA 2.1

Proof. (i) For fixed (x, y), write $w = w(x, y)$ and define

$$q_{x,y}(v) := F(x, v) - \frac{p}{2}\|v - y\|^2.$$

Since $F(x, \cdot)$ is L-smooth, it is L-weakly concave. Hence $q_{x,y}$ is $(p - L)$-strongly concave, proving uniqueness of $w(x, y)$. Moreover, the first-order condition of $q_{x,y}$ is exactly (8).

(ii) Write the first-order conditions (8) at z_1 and z_2:

$$p(w_i - y_i) = \nabla_y F(x_i, w_i), \qquad i = 1, 2. \tag{47}$$

After subtraction, rearranging the terms and applying the smoothness of F in Assumption 1,

$$\begin{aligned} p\|w_1 - w_2\| &\le p\|y_1 - y_2\| + \|\nabla_y F(x_1, w_1) - \nabla_y F(x_2, w_2)\| \\ &\le p\|y_1 - y_2\| + L\|x_1 - x_2\| + L\|w_1 - w_2\|. \end{aligned}$$

Rearranging gives

$$\|w_1 - w_2\| \le \frac{L\|x_1 - x_2\| + p\|y_1 - y_2\|}{p - L} \le \frac{\sqrt{L^2 + p^2}}{p - L}\|z_1 - z_2\|,$$

where the second inequality follows from the Cauchy–Schwarz inequality.

(iii) Strong concavity and $\nabla q_{x,y}(w) = 0$ imply

$$q_{x,y}(w) - q_{x,y}(y) \ge \frac{p - L}{2}\|w - y\|^2.$$

Note that the left-hand side is $\mathcal{G}(x, y)$ since $q_{x,y}(y) = F(x, y)$, which proves (10).

To prove (11), let $g := \nabla_y F(x, y)$ and substitute $v = y + s$ in (3)–(4), so that

$$\mathcal{G}(x, y) = \max_{s \in \mathbb{R}^m} \phi(s), \qquad \phi(s) := F(x, y + s) - F(x, y) - \frac{p}{2}\|s\|^2. \tag{48}$$

Applying the smoothness of $F(x, \cdot)$ and subtracting $\frac{p}{2}\|s\|^2$ from each side, for all s, we have

$$\langle g, s\rangle - \frac{p + L}{2}\|s\|^2 \;\le\; \phi(s) \;\le\; \langle g, s\rangle - \frac{p - L}{2}\|s\|^2. \tag{49}$$

For any $b > 0$ one has $\max_s\{\langle g, s\rangle - \frac{b}{2}\|s\|^2\} = \|g\|^2/(2b)$ attained at $s = g/b$. Taking the maximum over s in the right-hand inequality of (49) and using $b = p - L > 0$ proves the upper bound in (11).

For the lower bound, note that the maximum in (48) dominates the value of ϕ at any single point. Evaluating the left-hand inequality of (49) at the choice $s = g/(p + L)$, which maximizes that lower bound, gives

$$\mathcal{G}(x, y) \ge \phi\Big(\frac{g}{p + L}\Big) \ge \frac{\|g\|^2}{2(p + L)}.$$

The equivalence (12) follows immediately.

(iv) Because the maximizer is unique, Danskin's theorem gives (13) and the first equality in (14). The second equality in (14) follows from (8). Equations (15) and (16) follow from $P_\rho = F + \rho\mathcal{G}$.

(v) Define $\widehat{z} := (x, w(x, y))$. Since the norm on (x, y) is Euclidean, $\|x_1 - x_2\| \le \|z_1 - z_2\|$, and the bound $\|w_1 - w_2\| \le \kappa_w\|z_1 - z_2\|$ from (9) yields

$$\|\widehat{z}_1 - \widehat{z}_2\|^2 = \|x_1 - x_2\|^2 + \|w_1 - w_2\|^2 \le (1 + \kappa_w^2)\|z_1 - z_2\|^2.$$

Hence, the map $z \mapsto \widehat{z}$ is $\sqrt{1+\kappa_w^2}$-Lipschitz. Equations (13)–(14) imply $\nabla\mathcal{G}(z) = \nabla F(\widehat{z}) - \nabla F(z)$. Hence, by the triangle inequality and the Lipschitz continuity of ∇F, we have

$$\begin{aligned}\|\nabla\mathcal{G}(z_1) - \nabla\mathcal{G}(z_2)\| &\le \|\nabla F(\widehat{z}_1) - \nabla F(\widehat{z}_2)\| + \|\nabla F(z_1) - \nabla F(z_2)\| \\ &\le L\|\widehat{z}_1 - \widehat{z}_2\| + L\|z_1 - z_2\| \le L(1+\sqrt{1+\kappa_w^2})\|z_1 - z_2\|.\end{aligned}$$

Together with $P_\rho = F + \rho\mathcal{G}$, we have

$$\begin{aligned}\|\nabla P_\rho(z_1) - \nabla P_\rho(z_2)\| &\le \|\nabla F(z_1) - \nabla F(z_2)\| + \rho\|\nabla\mathcal{G}(z_1) - \nabla\mathcal{G}(z_2)\| \\ &\le L\|z_1 - z_2\| + \rho L(1+\sqrt{1+\kappa_w^2})\|z_1 - z_2\| \\ &= (L + \rho L(1+\sqrt{1+\kappa_w^2}))\|z_1 - z_2\|,\end{aligned}$$

where we are applying the Lipschitz continuity of ∇F and $\nabla\mathcal{G}$ in the second inequality. This proves that P_ρ has a globally Lipschitz gradient with constant $L_\rho = L + \rho L(1+\sqrt{1+\kappa_w^2})$. □

A.2 Proofs of the sufficient conditions for Assumption 3

A.2.1 Proof of Proposition 2.3

Proof. We first claim that there exists $\eta_0 > 0$ such that

$$\sigma_{\min}\big(\nabla_{yy}^2 F(x,y)\big) \ge \frac{3\mu_0}{4} \text{ whenever } (x,y) \in \mathcal{K} \text{ and } \|\nabla_y F(x,y)\| \le \eta_0. \tag{50}$$

Suppose not, then there would exist a sequence $z_j = (x_j, y_j) \in \mathcal{K}$ such that

$$\|\nabla_y F(x_j,y_j)\| \to 0, \qquad \sigma_{\min}\big(\nabla_{yy}^2 F(x_j,y_j)\big) < \frac{3\mu_0}{4}.$$

Compactness of $\mathcal{K}$ gives a convergent subsequence with limit $z_\star = (x_\star, y_\star) \in \mathcal{K}$. By continuity,

$$\nabla_y F(x_\star,y_\star) = 0, \qquad \sigma_{\min}\big(\nabla_{yy}^2 F(x_\star,y_\star)\big) \le \frac{3\mu_0}{4},$$

which contradicts (21). Let U be an open neighborhood of $\mathcal{K}$ on which F is twice continuously differentiable. Since $\mathcal{K}$ is compact, there exists $r_1 > 0$ such that the closed enlargement $\mathcal{K}_{r_1} := \{z' \in \mathbb{R}^{d+m} : \inf_{z\in\mathcal{K}} \|z' - z\| \le r_1\}$ is contained in U. As $\mathcal{K}_{r_1}$ is compact, $\nabla_{yy}^2 F$ is uniformly continuous on $\mathcal{K}_{r_1}$. Therefore, there exists $r_0 \in (0, r_1]$ such that

$$\|\nabla_{yy}^2 F(x,v) - \nabla_{yy}^2 F(x,y)\| \le \frac{\mu_0}{4}, \text{ for all } (x,y) \in \mathcal{K} \text{ with } \|v - y\| \le r_0. \tag{51}$$

Set $\eta_s := \min\{\eta_0, pr_0\}$ and suppose that $(x,y) \in \mathcal{K}$, $w = w(x,y)$, and $\max\{\|\nabla_y F(x,y)\|, \|\nabla_y F(x,w)\|\} \le \eta_s$. The first-order condition (8) then gives $\|w - y\| = \|\nabla_y F(x,w)\|/p \le r_0$. Define

$$A(x,y) := \int_0^1 \nabla_{yy}^2 F(x, y + t(w-y))\, dt.$$

Then $\|A(x,y) - \nabla_{yy}^2 F(x,y)\| \le \int_0^1 \|\nabla_{yy}^2 F\big(x, y+t(w-y)\big) - \nabla_{yy}^2 F(x,y)\|\, dt \le \frac{\mu_0}{4}$. Combining this estimate with (50) and the singular-value perturbation inequality proves

$$\sigma_{\min}\big(A(x,y)\big) \ge \sigma_{\min}\big(\nabla_{yy}^2 F(x,y)\big) - \|A(x,y) - \nabla_{yy}^2 F(x,y)\| \ge \frac{3\mu_0}{4} - \frac{\mu_0}{4} = \frac{\mu_0}{2}.$$

The fundamental theorem of calculus gives

$$\nabla_y F(x,w) - \nabla_y F(x,y) = A(x,y)(w-y).$$

Consequently,

$$\|w - y\| \le \frac{2}{\mu_0}\|\nabla_y F(x,w) - \nabla_y F(x,y)\|.$$

Thus Assumption 3 holds with $\eta_s = \min\{\eta_0, pr_0\}$, $\delta_s = r_0$, and $\kappa_s = 2/\mu_0$. We note that the proximal structure of w entered only through (8), which was used to bound $\|w - y\|$; the argument shows more generally that $\|v - y\| \le \frac{2}{\mu_0}\|\nabla_y F(x,v) - \nabla_y F(x,y)\|$ for every $(x,y) \in \mathcal{K}$ with $\|\nabla_y F(x,y)\| \le \eta_0$ and every v with $\|v - y\| \le r_0$. □

A.2.2 Proof of Proposition 2.4

Proof. Throughout the proof we write

$$\phi_{x,y}(u) := g(x,u) - \frac{\lambda}{2}\|u-y\|^2, \qquad \mathcal{C} := \{u : c(u) \le 0\},$$

so that $F(x,y) = \max_{u\in\mathcal{C}} \phi_{x,y}(u)$. The proof proceeds in five steps.

Step 1: Well-posedness, continuity, and the residual map. Since $\nabla_u g(x,\cdot)$ is L_g-Lipschitz and $\lambda > L_g$, $\phi_{x,y}$ is $(\lambda - L_g)$-strongly concave. The set $\mathcal{C}$ is closed and convex because c is continuous and convex, nonempty by the Slater condition, and bounded because the strongly convex function c is coercive; thus $\mathcal{C}$ is compact. Consequently the maximizer $u(x,y)$ of $\phi_{x,y}$ over $\mathcal{C}$ exists and is unique, and it is characterized by the variational inequality

$$\langle \nabla\phi_{x,y}(u(x,y)),\, v - u(x,y)\rangle \le 0 \qquad \text{for all } v \in \mathcal{C}. \tag{52}$$

Fix two parameter pairs (x,y) and (x',y') and abbreviate $u = u(x,y)$, $u' = u(x',y')$. Applying (52) at (x,y) with $v = u'$ and at (x',y') with $v = u$, and adding the two inequalities, gives $\langle \nabla\phi_{x,y}(u) - \nabla\phi_{x',y'}(u'), u - u'\rangle \ge 0$. Inserting $\pm\nabla\phi_{x,y}(u')$ and using strong concavity of $\phi_{x,y}$ yields

$$(\lambda - L_g)\|u-u'\|^2 \le -\langle \nabla\phi_{x,y}(u) - \nabla\phi_{x,y}(u'), u-u'\rangle \le \langle \nabla\phi_{x,y}(u') - \nabla\phi_{x',y'}(u'), u-u'\rangle.$$

Since $\nabla\phi_{x,y}(u') - \nabla\phi_{x',y'}(u') = \nabla_u g(x,u') - \nabla_u g(x',u') + \lambda(y-y')$, applying the Cauchy–Schwarz inequality to the right-hand side, we conclude that

$$\|u(x,y) - u(x',y')\| \le \frac{1}{\lambda - L_g}\Big(\|\nabla_u g(x,u(x',y')) - \nabla_u g(x',u(x',y'))\| + \lambda\|y-y'\|\Big). \tag{53}$$

Since $\nabla_u g$ is continuous, (53) shows that $(x,y) \mapsto u(x,y)$ is continuous on $\mathbb{R}^d \times \mathbb{R}^m$, and with $x = x'$ it gives the Lipschitz bound

$$\|u(x,y) - u(x,y')\| \le \frac{\lambda}{\lambda - L_g}\|y-y'\|. \tag{54}$$

Since $\mathcal{C}$ is compact, $\phi_{x,y}$ has a unique maximizer, and $\nabla_y\phi_{x,y}(u) = \lambda(u-y)$ is continuous, Danskin's theorem shows that $F(x,\cdot)$ is continuously differentiable with

$$R_x(y) := \nabla_y F(x,y) = \lambda(u(x,y) - y). \tag{55}$$

In particular $R : (x,y) \mapsto R_x(y)$ is continuous, and by (54),

$$\|R_x(y) - R_x(y')\| \le \lambda\big(\|u(x,y) - u(x,y')\| + \|y-y'\|\big) \le L_F\|y-y'\|, \qquad L_F := \frac{\lambda(2\lambda - L_g)}{\lambda - L_g}.$$

Hence $\nabla_y F(x,\cdot)$ is L_F-Lipschitz continuous. This is the only property needed in the proof of Lemma 2.1(i): for any $p > L$, the objective in (3) is $(p-L)$-strongly concave, so the regularized best response $w(x,y)$ is well defined and satisfies (8).

Step 2: The KKT system and the multiplier. Because $\phi_{x,y}$ is concave, c is convex, and the Slater condition holds, the KKT conditions are necessary and sufficient for optimality: $u \in \mathcal{C}$ maximizes $\phi_{x,y}$ over $\mathcal{C}$ if and only if there exists $\alpha \ge 0$ with $\nabla\phi_{x,y}(u) = \alpha\nabla c(u)$ and $\alpha c(u) = 0$. At $u = u(x,y)$ we have $\nabla\phi_{x,y}(u) = \nabla_u g(x,u) - \lambda(u-y) = \nabla_u g(x,u) - r$ with $r := R_x(y)$, so the KKT conditions read

$$-\nabla_u g(x,u) + r + \alpha\nabla c(u) = 0, \tag{56}$$

$$c(u) \le 0, \qquad \alpha \ge 0, \qquad \alpha c(u) = 0. \tag{57}$$

Conversely, if (u,α) satisfies (56)–(57) for some vector r, then (u,α) is a KKT pair of (22) with center $y := u - r/\lambda$, and therefore $u = u(x,y)$ and $r = R_x(y)$ by (55). This observation is the basis of the local inversion below.

We next show that $\nabla c(u) \ne 0$ whenever $c(u) = 0$. Indeed, if $\nabla c(\bar u) = 0$ at some $\bar u$ with $c(\bar u) = 0$, then strong convexity gives $c(v) \ge c(\bar u) + \frac{\mu_c}{2}\|v-\bar u\|^2 \ge 0$ for every v, contradicting the Slater condition. Consequently the multiplier is unique: $\alpha = 0$ if $c(u) < 0$, while if $c(u) = 0$, taking the inner product of (56) with $\nabla c(u)$ gives

$$\alpha = \frac{\langle \nabla_u g(x,u) - r, \nabla c(u)\rangle}{\|\nabla c(u)\|^2}. \tag{58}$$

We write $\alpha(x,y)$ for this multiplier, so that $(u(x,y), \alpha(x,y))$ is the unique KKT pair at (x,y).

Step 3: Compactness of the small-residual set and stability of the active set. Define

$$\mathcal{D}_{\eta_M} := \left\{(x,y) \in \mathcal{K}^+ : \|R_x(y)\| \le \eta_M\right\}.$$

The set $\mathcal{K}$ is compact by Assumption 2, so its closed unit enlargement $\mathcal{K}^+$ is compact, and $\mathcal{D}_{\eta_M}$ is a closed subset of $\mathcal{K}^+$ because R is continuous. Hence $\mathcal{D}_{\eta_M}$ is compact.

Fix $\bar{z} := (\bar{x}, \bar{y}) \in \mathcal{D}_{\eta_M}$ and write $\bar{u} := u(\bar{x}, \bar{y})$, $\bar{\alpha} := \alpha(\bar{x}, \bar{y})$, and $\bar{r} := R_{\bar{x}}(\bar{y})$. We claim that there is an open neighborhood N^0 of $\bar{z}$ in $\mathbb{R}^d \times \mathbb{R}^m$ on which the active set is constant and α is continuous.

If $c(\bar{u}) < 0$, continuity of u gives an open neighborhood N^0 on which $c(u(x,y)) < 0$, hence $\alpha \equiv 0$ on N^0.

If $c(\bar{u}) = 0$, then $\bar{\alpha} > 0$ by strict complementarity (iii), because $(\bar{x}, \bar{y}) \in \mathcal{K}^+$ and $\|\bar{r}\| \le \eta_M$. Suppose, for contradiction, that there were points $(x_k, y_k) \to (\bar{x}, \bar{y})$ with $c(u_k) < 0$, where $u_k := u(x_k, y_k)$ and $r_k := R_{x_k}(y_k)$. Then $\alpha(x_k, y_k) = 0$, and (56) reduces to $\nabla_u g(x_k, u_k) = r_k$. By continuity of u, R, and $\nabla_u g$, letting $k \to \infty$ gives $\nabla_u g(\bar{x}, \bar{u}) = \bar{r}$. Comparing with (56) at $(\bar{x}, \bar{y})$ yields $\bar{\alpha}\nabla c(\bar{u}) = 0$, which is impossible since $\bar{\alpha} > 0$ and $\nabla c(\bar{u}) \neq 0$. Hence there is an open neighborhood N^0 of $(\bar{x}, \bar{y})$ on which $c(u(x,y)) = 0$. On N^0 the multiplier is given by (58) with $u = u(x,y)$ and $r = R_x(y)$, and $\nabla c(u(x,y)) \neq 0$ there, so α is continuous on N^0.

In both cases, the map $(x,y) \mapsto (u(x,y), \alpha(x,y))$ is continuous on N^0.

Step 4: Local inversion of R_x near $\bar{z}$. We keep $\bar{z} = (\bar{x}, \bar{y}) \in \mathcal{D}_{\eta_M}$ and the notation of Step 3, and apply the implicit function theorem to the KKT system in the variables (x, r). Below, $B(a, \epsilon)$ denotes the open Euclidean ball of radius ϵ centered at a.

Inactive case, $c(\bar{u}) < 0$. Set $\Phi(x, r, u) := -\nabla_u g(x, u) + r$. Then $\Phi(\bar{x}, \bar{r}, \bar{u}) = 0$ and $\partial_u \Phi(\bar{x}, \bar{r}, \bar{u}) = -\nabla^2_{uu} g(\bar{x}, \bar{u}) = H(\bar{x}, \bar{u}, 0)$, which is nonsingular by (iv).

Active case, $c(\bar{u}) = 0$. Set $\Psi(x, r, u, \alpha) := \begin{pmatrix} -\nabla_u g(x,u) + r + \alpha \nabla c(u) \\ c(u) \end{pmatrix}$. Then $\Psi(\bar{x}, \bar{r}, \bar{u}, \bar{\alpha}) = 0$, and the Jacobian of Ψ with respect to (u, α) at this point is

$$M = \begin{bmatrix} H(\bar{x}, \bar{u}, \bar{\alpha}) & \nabla c(\bar{u}) \\ \nabla c(\bar{u})^\top & 0 \end{bmatrix}.$$

To see that M is nonsingular, let $M \begin{bmatrix} d \\ \beta \end{bmatrix} = 0$, i.e.,

$$H(\bar{x}, \bar{u}, \bar{\alpha}) d + \beta \nabla c(\bar{u}) = 0, \qquad \nabla c(\bar{u})^\top d = 0.$$

The second equation places d in the range of $Z(\bar{u})$, so $d = Z(\bar{u}) q$ for some q. Multiplying the first equation by $Z(\bar{u})^\top$ and using $Z(\bar{u})^\top \nabla c(\bar{u}) = 0$ gives $Z(\bar{u})^\top H(\bar{x}, \bar{u}, \bar{\alpha}) Z(\bar{u}) q = 0$. The reduced-Hessian assumption (iv) yields $q = 0$, hence $d = 0$, and then $\beta \nabla c(\bar{u}) = 0$ with $\nabla c(\bar{u}) \neq 0$ gives $\beta = 0$. Thus M is nonsingular.

Implicit function chart. In both cases, the implicit function theorem provides $\epsilon_{\bar{z}} > 0$, an open neighborhood $W_{\bar{z}}$ of $(\bar{u}, \bar{\alpha})$, and a continuously differentiable map

$$(x, r) \mapsto \big(\hat{u}_{\bar{z}}(x, r), \hat{\alpha}_{\bar{z}}(x, r)\big) \in W_{\bar{z}} \qquad \text{on} \qquad V_{\bar{z}} := B(\bar{x}, \epsilon_{\bar{z}}) \times B(\bar{r}, \epsilon_{\bar{z}}),$$

with $(\hat{u}_{\bar{z}}, \hat{\alpha}_{\bar{z}})(\bar{x}, \bar{r}) = (\bar{u}, \bar{\alpha})$, such that for every $(x, r) \in V_{\bar{z}}$ the pair $(\hat{u}_{\bar{z}}(x, r), \hat{\alpha}_{\bar{z}}(x, r))$ is the unique $(u, \alpha) \in W_{\bar{z}}$ satisfying (56) together with the active-set condition of $\bar{z}$, namely $\alpha = 0$ in the inactive case and $c(u) = 0$ in the active case. In the active case this is the theorem applied to Ψ. In the inactive case it is the theorem applied to Φ, which provides a neighborhood W' of $\bar{u}$; we then set $\hat{\alpha}_{\bar{z}} \equiv 0$ and $W_{\bar{z}} := W' \times \mathbb{R}$. Shrinking $\epsilon_{\bar{z}}$, we may assume that the continuity of $\partial_r \hat{u}_{\bar{z}}$ gives a constant $\kappa_{\bar{z}}$ with $\|\partial_r \hat{u}_{\bar{z}}\| \le \kappa_{\bar{z}}$ on $V_{\bar{z}}$. Each slice $\{x\} \times B(\bar{r}, \epsilon_{\bar{z}})$ of the product $V_{\bar{z}}$ is convex, so integrating along the segment joining (x, r_2) and (x, r_1) gives

$$\|\hat{u}_{\bar{z}}(x, r_1) - \hat{u}_{\bar{z}}(x, r_2)\| \le \kappa_{\bar{z}} \|r_1 - r_2\| \qquad \text{whenever } (x, r_1), (x, r_2) \in V_{\bar{z}}. \tag{59}$$

Right inverse. Define

$$Y_{\bar{z}}(x, r) := \hat{u}_{\bar{z}}(x, r) - \frac{r}{\lambda}, \qquad (x, r) \in V_{\bar{z}}. \tag{60}$$

Since $(\hat{u}_{\bar{z}}(x,r), \hat{\alpha}_{\bar{z}}(x,r))$ satisfies (56)–(57) at (x,r), the converse statement in Step 2 gives $\hat{u}_{\bar{z}}(x,r) = u(x, Y_{\bar{z}}(x,r))$ and

$$R_x(Y_{\bar{z}}(x,r)) = r \qquad \text{for all } (x,r) \in V_{\bar{z}},$$

and (59) yields

$$\|Y_{\bar{z}}(x,r_1) - Y_{\bar{z}}(x,r_2)\| \le \Big(\kappa_{\bar{z}} + \frac{1}{\lambda}\Big)\|r_1 - r_2\| \qquad \text{whenever } (x,r_1),(x,r_2) \in V_{\bar{z}}. \tag{61}$$

Left inverse on a chart in (x,y)-space. The map $(x,y) \mapsto (x, R_x(y))$ is continuous and sends $\bar{z}$ to $(\bar{x},\bar{r}) \in V_{\bar{z}}$, and by Step 3 the map $(x,y) \mapsto (u(x,y), \alpha(x,y))$ is continuous on N^0 and sends $\bar{z}$ to $(\bar{u},\bar{\alpha}) \in W_{\bar{z}}$. Hence there is an open neighborhood $N_{\bar{z}} \subseteq N^0$ of $\bar{z}$ such that

$$(x, R_x(y)) \in V_{\bar{z}} \qquad \text{and} \qquad (u(x,y), \alpha(x,y)) \in W_{\bar{z}} \qquad \text{for every } (x,y) \in N_{\bar{z}}.$$

Let $(x,y) \in N_{\bar{z}}$ and $r := R_x(y)$. By Step 2, $(u(x,y), \alpha(x,y))$ satisfies (56) at (x,r), and by Step 3 it satisfies the active-set condition of $\bar{z}$. The uniqueness property of the chart gives $u(x,y) = \hat{u}_{\bar{z}}(x,r)$, and hence, by (55),

$$Y_{\bar{z}}(x, R_x(y)) = u(x,y) - \frac{1}{\lambda} R_x(y) = u(x,y) - \big(u(x,y) - y\big) = y.$$

Thus R_x and $Y_{\bar{z}}(x,\cdot)$ are mutually inverse between $\{y : (x,y) \in N_{\bar{z}}\}$ and its image. Combining this with (61), and using that $(x, R_x(y_i)) \in V_{\bar{z}}$ for $(x,y_i) \in N_{\bar{z}}$, we obtain

$$\|y_1 - y_2\| = \|Y_{\bar{z}}(x, R_x(y_1)) - Y_{\bar{z}}(x, R_x(y_2))\| \le \Big(\kappa_{\bar{z}} + \frac{1}{\lambda}\Big)\|R_x(y_1) - R_x(y_2)\| \quad \text{whenever } (x,y_1),(x,y_2) \in N_{\bar{z}}. \tag{62}$$

The constant here depends on $\bar{z}$ but not on x.

Step 5: Uniform constants. The family $\{N_{\bar{z}}\}_{\bar{z} \in \mathcal{D}_{\eta_M}}$ is an open cover of the compact set $\mathcal{D}_{\eta_M}$. Extract a finite subcover $N_1, \dots, N_J$ with associated constants $\kappa_1, \dots, \kappa_J$, and let $\ell > 0$ be a Lebesgue number of the cover $\{N_j \cap \mathcal{D}_{\eta_M}\}_{j=1}^J$ of the compact metric space $\mathcal{D}_{\eta_M}$: every subset of $\mathcal{D}_{\eta_M}$ of diameter less than ℓ is contained in a single N_j.

Set

$$\eta_s := \eta_M, \qquad \delta_s := \min\Big\{1, \frac{\ell}{2}, \frac{\eta_M}{p}\Big\}, \qquad \kappa_s := \max_{1 \le j \le J}\Big(\kappa_j + \frac{1}{\lambda}\Big). \tag{63}$$

Let $(x,y) \in \mathcal{K}$ and $w = w(x,y)$ satisfy (19) with these constants. Then $\|w - y\| \le \delta_s \le 1$, so $(x,w) \in \mathcal{K}^+$, while $(x,y) \in \mathcal{K} \subseteq \mathcal{K}^+$. Moreover, $\|R_x(y)\| \le \eta_s = \eta_M$ by (19), and the first-order condition (8) gives $\|R_x(w)\| = \|\nabla_y F(x,w)\| = p\|w - y\| \le p\,\delta_s \le \eta_M$. Hence both (x,y) and (x,w) belong to $\mathcal{D}_{\eta_M}$. They share the same x, so

$$\|(x,y) - (x,w)\| = \|w - y\| \le \delta_s \le \frac{\ell}{2} < \ell,$$

and the Lebesgue property places both points in a common chart N_j. Applying (62) on N_j yields

$$\|w - y\| \le \Big(\kappa_j + \frac{1}{\lambda}\Big)\|R_x(w) - R_x(y)\| \le \kappa_s \|\nabla_y F(x,w) - \nabla_y F(x,y)\|,$$

which is (20). Therefore Assumption 3 holds with the constants (63). □

Remark A.1 (Where each hypothesis is used). Hypothesis (i) makes the inner problem strongly concave, which gives uniqueness and continuity of $u(x,y)$, Danskin's formula (55), and the sufficiency of the KKT conditions used in Steps 2 and 4. Hypothesis (ii) makes the feasible set compact and convex, and gives $\nabla c(u) \neq 0$ at active points, hence the uniqueness and the explicit formula (58) of the multiplier in Step 2. Hypothesis (iii) rules out the degenerate case $\alpha = 0$ with $c(u) = 0$, in which the KKT system is not locally a square smooth system in (u,α) and the active set can change under arbitrarily small perturbations of r; it is what keeps the active set locally constant in Step 3. Hypothesis (iv) is the nondegeneracy that makes the implicit function theorem applicable in Step 4; it is the exact analogue, for the constrained problem, of (21) in Proposition 2.3.

A.3 PROOFS OF STATIONARITY TRANSFER THEOREM

Proof. Since $\rho \geq \rho_c$ and $\mathcal{G} \geq 0$, $P_{\rho_c}(x, y) \leq P_\rho(x, y) \leq C_0$. Hence $(x, y) \in \mathcal{K}$. Also,

$$\rho\mathcal{G}(x, y) = P_\rho(x, y) - F(x, y) \leq C_0 - \underline{F}_{\mathcal{K}} = B.$$

Combining above with (10) yields the coarse localization bound

$$\|w - y\| \leq \sqrt{\frac{2\mathcal{G}(x, y)}{p - L}} \leq \frac{D}{\sqrt{\rho}}.$$

By the first-order condition (8) and L-smoothness give

$$\begin{aligned}\|\nabla_y F(x, y)\| &\leq \|\nabla_y F(x, y) - \nabla_y F(x, w)\| + \|\nabla_y F(x, w)\| \\ &\leq (L + p)\|w - y\| \leq \frac{C_y}{\sqrt{\rho}}.\end{aligned}$$

ALso, $\|\nabla_y F(x, w)\| = p\|w - y\| \leq \frac{pD}{\sqrt{\rho}}$. Since $\rho \geq \max\{C_y^2/\eta_s^2, p^2D^2/\eta_s^2, D^2/\delta_s^2\}$, the displayed bounds give $\|\nabla_y F(x, y)\| \leq \eta_s$, $\|\nabla_y F(x, w)\| \leq \eta_s$, and $\|w - y\| \leq \delta_s$, so (19) holds. Hence Assumption 3 gives

$$\|w - y\| \leq \kappa_s\|\Delta_y\|. \tag{64}$$

By (8), $\nabla_y F(x, y) = p(w - y) - \Delta_y$. Together with (64), this yields

$$\|\nabla_y F(x, y)\| \leq (1 + p\kappa_s)\|\Delta_y\|. \tag{65}$$

Applying (26), we obtain

$$\varepsilon \geq \|\nabla_y P_\rho(x, y)\| = \|\nabla_y F(x, y) + \rho\Delta_y\| \geq \rho\|\Delta_y\| - \|\nabla_y F(x, y)\| \geq (\rho - 1 - p\kappa_s)\|\Delta_y\|,$$

where the equality follows from (16), the second inequality is by the triangle inequality, and the last inequality is by (65). Since $\rho \geq 2(1 + p\kappa_s)$, the denominator is positive and proves the first bound in (27). The second bound in (27) and the first bound in (28) follow from (64) and (65), respectively.

Finally, using (15) first, applying the Lipschitz continuity of $\nabla_x F$ and (26) in the second inequality, and applying second bound of (27) in the third, we have

$$\begin{aligned}\|\nabla_x F(x, y)\| &\leq \|\nabla_x P_\rho(x, y)\| + \rho\|\nabla_x F(x, w) - \nabla_x F(x, y)\| \\ &\leq \varepsilon + \rho L\|w - y\| \leq \left(1 + \frac{\rho L\kappa_s}{\rho - 1 - p\kappa_s}\right)\varepsilon.\end{aligned}$$

This proves the second bound in (28). Combined with the first, this proves (29). Setting $\varepsilon = 0$ proves the last assertion. □

A.4 PROOFS OF THE COMPLEXITY THEOREMS

A.4.1 PROOF OF THEOREM 3.4

Proof. By Lemma 2.1, P_ρ is L_ρ-smooth. The standard descent lemma applied to (32) gives (34). Since $P_\rho(z^0) \leq C_0$ by the construction of initialization, (34) gives

$$P_\rho(z^k) \leq C_0 \qquad \text{for all } k.$$

Since $\rho \geq \rho_c$ and $\mathcal{G} \geq 0$,

$$P_{\rho_c}(z^k) \leq P_\rho(z^k) \leq C_0 \implies z^k \in \mathcal{K} \qquad \text{for all } k.$$

Summing (34) from $k = 0$ to $N - 1$ yields

$$\frac{1}{2L_\rho}\sum_{k=0}^{N-1}\|\nabla P_\rho(z^k)\|^2 \leq P_\rho(z^0) - P_\rho(z^N).$$

Because $P_\rho \geq P_{\rho_c}$, $P_\rho(z^N) \geq P_c^*$. Therefore, $\sum_{k=0}^{N-1}\|\nabla P_\rho(z^k)\|^2 \leq 2L_\rho\Delta_c$, which proves (35).

Under (36), some iterate satisfies $\|\nabla P_\rho(z^k)\| \leq \epsilon/C_\rho$. This iterate also satisfies $P_\rho(z^k) \leq C_0$, so Theorem 3.1 gives (37). □

A.4.2 Proof of Theorem 3.5

Proof. Since P_ρ is L_ρ-smooth, the update (39) gives

$$\begin{aligned} P_\rho(z^{k+1}) &\le P_\rho(z^k) - \frac{1}{L_\rho}\langle \nabla P_\rho(z^k), \widetilde{g}^k\rangle + \frac{1}{2L_\rho}\|\widetilde{g}^k\|^2 \\ &= P_\rho(z^k) - \frac{1}{2L_\rho}\|\nabla P_\rho(z^k)\|^2 + \frac{1}{2L_\rho}\|\widetilde{g}^k - \nabla P_\rho(z^k)\|^2 \\ &\le P_\rho(z^k) - \frac{1}{2L_\rho}\|\nabla P_\rho(z^k)\|^2 + \frac{\rho^2 L^2\delta^2}{2L_\rho}, \end{aligned}$$

which proves (41).

For the second assertion, observe first that (40) and (41) imply the conditional monotonicity

$$\|\nabla P_\rho(z^j)\| \ge \rho L\delta \quad\Longrightarrow\quad P_\rho(z^{j+1}) \le P_\rho(z^j). \tag{66}$$

Consider the case in which there exists an index $k < N$ such that

$$\|\nabla P_\rho(z^k)\| < \rho L\delta, \tag{67}$$

and let k be the first such index. For every $j < k$, $\|\nabla P_\rho(z^j)\| \ge \rho L\delta$, so (66) gives $P_\rho(z^k) \le P_\rho(z^0) \le C_0$, and $z^k \in \mathcal{K}$ follows from $P_{\rho_c} \le P_\rho$. The bound (43) holds because $\|\nabla P_\rho(z^k)\|^2 < \rho^2 L^2\delta^2$.

Otherwise, $\|\nabla P_\rho(z^k)\| \ge \rho L\delta$ for every $k < N$, and (66) gives $P_\rho(z^k) \le P_\rho(z^0) \le C_0$, and hence $z^k \in \mathcal{K}$, for all $k \le N$. Summing (41) from $k = 0$ to $N-1$ gives

$$\frac{1}{2L_\rho}\sum_{k=0}^{N-1}\|\nabla P_\rho(z^k)\|^2 \le P_\rho(z^0) - P_\rho(z^N) + \frac{N\rho^2 L^2\delta^2}{2L_\rho}.$$

Since $z^N \in \mathcal{K}$ and $P_{\rho_c} \le P_\rho$, we have $P_\rho(z^N) \ge P_{\rho_c}(z^N) \ge P_c^*$, whence

$$\sum_{j=0}^{N-1}\|\nabla P_\rho(z^j)\|^2 \le 2L_\rho\Delta_c + N\rho^2 L^2\delta^2.$$

Dividing by N, some iterate satisfies (43). In both cases the selected iterate satisfies $P_\rho(z^k) \le C_0$ and (43), so applying Theorem 3.1 gives (44).

Finally, under (36) and $\delta \le \epsilon$, some iterate satisfies

$$\|\nabla_x F(x^k, y^k)\| + \|\nabla_y F(x^k, y^k)\| \le \sqrt{1 + C_\rho^2\rho^2 L^2}\,\epsilon \le (1 + C_\rho\rho L)\epsilon. \tag{68}$$

□

B Robust regression example illustrating Assumption 3

Consider the one-sample response-adversarial regression problem

$$\min_{x\in\mathbb{R}}\max_{|u|\le 5} g(x,u), \qquad g(x,u) := \frac{1}{2}(x-u)^2 + r(x), \qquad r(x) := 10x^2 + 23(\cos x - 1). \tag{69}$$

The regularizer is smooth and coercive but nonconvex, since $r''(x) = 20 - 23\cos x$ and $r''(0) = -3$. With $c(u) := u^2 - 25$ and $\lambda = 2$, the upper Moreau envelope is

$$F(x,y) := \max_{c(u)\le 0}\Big\{g(x,u) - \frac{\lambda}{2}(u-y)^2\Big\} = \max_{u^2-25\le 0}\Big\{\frac{1}{2}(x-u)^2 + r(x) - (u-y)^2\Big\}. \tag{70}$$

Since $\nabla_u g(x,u) = u - x$, we have $L_g = 1 < \lambda$, so the inner objective is 1-strongly concave in u with unique maximizer $u(x,y) = \Pi_{[-5,5]}(2y - x)$. Danskin's theorem gives $\nabla_x F(x,y) = x - u(x,y) + r'(x)$ and

$$\nabla_y F(x,y) = 2\big(u(x,y) - y\big) = \begin{cases} -2(y+5), & 2y - x \le -5, \\ 2(y-x), & |2y-x| < 5, \\ -2(y-5), & 2y - x \ge 5. \end{cases} \tag{71}$$

Because u is Lipschitz and $r'(x) = 20x - 23\sin x$ is 43-Lipschitz, ∇F is globally Lipschitz; moreover, $\nabla_y F(x,y)$ is continuous and piecewise affine with slopes ± 2, so Assumption 1 holds with $L_y = 2$. We take $p = 4$.

Proposition B.1. *For the example above with $p = 4$, the following hold.*

(i) F is nonconvex in x and nonconcave in y.

(ii) Assumption 2 holds with $\rho_c = 7$, $C_0 = 15$, and $z_\rho^0 = (0,0)$ for every $\rho \geq 7$.

(iii) Assumption 3 holds.

(iv) No global max-player PL inequality, global-gap KL inequality, or global error bound to the set of maximizers holds.

Consequently, this example lies within the scope of Theorems 3.1 and 3.4.

Proof. We verify (i)–(iv) in turn.

(i) On the region $|2y - x| < 5$, we have $u(x,y) = 2y - x$, and (70) becomes $F(x,y) = r(x) + (x-y)^2$. Hence $\nabla^2_{xx} F(0,0) = r''(0) + 2 = -1 < 0$ and $\nabla^2_{yy} F(0,0) = 2 > 0$. In particular, the max-player stationary point $y = 0$ is a strict local minimum of $F(0,\cdot)$.

(ii) The first bound below follows from taking $u = 0$ in (70) and $23(\cos x - 1) \geq -46$; the second from $|u(x,y)| \leq 5$ and $y^2 \leq 2(u-y)^2 + 2u^2$:

$$F(x,y) \geq \frac{21}{2}x^2 - y^2 - 46, \qquad |\nabla_y F(x,y)|^2 = 4\big(u(x,y) - y\big)^2 \geq 2y^2 - 100.$$

By Lemma 2.1, $\mathcal{G}(x,y) \geq |\nabla_y F(x,y)|^2/(2(p + L_y))$. Combining these bounds gives

$$\begin{aligned} P_7(x,y) &\geq F(x,y) + \frac{7}{12}|\nabla_y F(x,y)|^2 \\ &\geq \frac{21}{2}x^2 + \frac{1}{6}y^2 - \frac{313}{3}. \end{aligned}$$

Hence,

$$\mathcal{K} = \{P_7(x,y) \leq 15\} \subseteq \{\frac{21}{2}x^2 + \frac{1}{6}y^2 - \frac{313}{3} \leq 15\}$$

is compact. At the origin, $u(0,0) = 0$, $F(0,0) = 0$, and $\nabla_y F(0,0) = 0$, so $\mathcal{G}(0,0) = 0$ by (12) and $P_\rho(0,0) = 0 = C_0$ for every ρ. Finally, ((ii)) gives $|x| \leq \sqrt{(313/3 + 15) \times 2/21} \approx 3.371$ on $\mathcal{K}$; hence, on the unit enlargement $\mathcal{K}^+$ of Proposition 2.4,

$$|x| \leq \bar{x} := 1 + \sqrt{(313/3 + 15) \times 2/21} \approx 4.371 < 5. \tag{72}$$

(iii) We check the hypotheses of Proposition 2.4. Hypothesis (i) holds with $L_g = 1 < \lambda = 2$; (ii) holds with Slater point $u = 0$. (iii) holds vacuously for any $\eta_M > 0$ because $|u(x,y) = |2y - x| < 5$ on $\mathcal{K}^+$, so the constraint is never active; and (iv) holds since $H(x,u,0) = -\nabla^2_{uu} g(x,u) = -1$. Hence Assumption 3 holds.

(iv) The proximal term in (70) is nonpositive and vanishes at $y = u$, so $F^\star(x) := \max_y F(x,y) = \max_{|u| \leq 5} g(x,u)$. At $x = 0$, the value $F^\star(0) = \frac{25}{2}$ is attained exactly on $Y^\star(0) = \{-5, 5\}$, whereas $F(0,0) = 0$ and $\nabla_y F(0,0) = 0$. Hence, for all $\mu, C, \theta, \kappa > 0$, each of

$$\begin{aligned} \frac{1}{2}|\nabla_y F(x,y)|^2 &\geq \mu\big(F^\star(x) - F(x,y)\big), \\ C\big(F^\star(x) - F(x,y)\big)^\theta &\leq |\nabla_y F(x,y)|, \\ \operatorname{dist}\big(y, Y^\star(x)\big) &\leq \kappa |\nabla_y F(x,y)| \end{aligned} \tag{73}$$

fails at $(x,y) = (0,0)$, since the gradient vanishes there while $F^\star(0) - F(0,0) = \frac{25}{2}$ and $\operatorname{dist}(0, Y^\star(0)) = 5$. The same obstruction is already present in the original objective: $\nabla_u g(0,0) = 0$, while $\max_{|u| \leq 5} g(0,u) - g(0,0) = \frac{25}{2}$.

□