OPTIMAL COMPLEXITY FOR PŁ-PŁ MINIMAX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the deterministic first-order oracle complexity of smooth two-sided Polyak-Łojasiewicz (PŁ) minimax optimization, where f is jointly ℓ-smooth, $f(\cdot;\boldsymbol{y})$ is $\mu_{\boldsymbol{x}}$-PŁ for every $\boldsymbol{y}$, and $-f(\boldsymbol{x};\cdot)$ is $\mu_{\boldsymbol{y}}$-PŁ for every $\boldsymbol{x}$. Let $\kappa_{\boldsymbol{x}} := \ell/\mu_{\boldsymbol{x}}$, $\kappa_{\boldsymbol{y}} := \ell/\mu_{\boldsymbol{y}}$, and $\Phi(\boldsymbol{x}) := \max_{\boldsymbol{y}\in\mathbb{R}^n} f(\boldsymbol{x},\boldsymbol{y})$. Assuming $\Phi(\mathbf{0}) - \min_{\boldsymbol{x}\in\mathbb{R}^m}\Phi(\boldsymbol{x}) \le \Delta$, we prove that every deterministic first-order method requires $\Omega(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\log(\Delta/\epsilon))$ oracle queries in the worst case to find a point $\boldsymbol{x}$ satisfying $\Phi(\boldsymbol{x}) - \min_{\boldsymbol{x}\in\mathbb{R}^m}\Phi(\boldsymbol{x}) \le \epsilon$. We further show, through a sharper analysis, that the existing single-loop Smoothed Gradient-Descent-Ascent (S-GDA) method attains a matching upper bound, removing an extra logarithmic factor from previous upper bounds. Together, these results establish the optimal deterministic first-order oracle complexity of smooth two-sided PŁ minimax optimization.

1 INTRODUCTION

Smooth strongly convex-strongly concave minimax problems admit global linear convergence (Kovalev & Gasnikov, 2022), but strong convexity and concavity can be restrictive. The Polyak-Łojasiewicz (PŁ) condition (Polyak, 1963) provides a weaker alternative: it can hold in the absence of convexity while still ensuring global linear convergence of first-order methods (Karimi et al., 2016). This motivates the study of two-sided PŁ minimax problems, a natural nonconvex-nonconcave class for which global linear convergence remains possible (Yang et al., 2020; Chen et al., 2022; Rankawat et al., 2026). Despite these advances, the optimal deterministic first-order oracle complexity of this class remains unclear.

To address this question, we study

$$\min_{\boldsymbol{x}\in\mathbb{R}^m}\max_{\boldsymbol{y}\in\mathbb{R}^n} f(\boldsymbol{x};\boldsymbol{y}), \tag{P}$$

where f is jointly ℓ-smooth, $f(\cdot;\boldsymbol{y})$ satisfies the $\mu_{\boldsymbol{x}}$-PŁ condition for every $\boldsymbol{y}\in\mathbb{R}^n$, and $f(\boldsymbol{x};\cdot)$ satisfies the $\mu_{\boldsymbol{y}}$-PŁ condition for maximization for every $\boldsymbol{x}\in\mathbb{R}^m$. We write $\kappa_{\boldsymbol{x}} := \ell/\mu_{\boldsymbol{x}}$ and $\kappa_{\boldsymbol{y}} := \ell/\mu_{\boldsymbol{y}}$, and define $\Phi(\boldsymbol{x}) := \max_{\boldsymbol{y}\in\mathbb{R}^n} f(\boldsymbol{x};\boldsymbol{y})$ and $\Phi^\star := \inf_{\boldsymbol{x}\in\mathbb{R}^m}\Phi(\boldsymbol{x})$. We call $\boldsymbol{x}$ ϵ-stationary if $\Phi(\boldsymbol{x}) - \Phi^\star \le \epsilon$, and assume $\Phi(\mathbf{0}) - \Phi^\star \le \Delta$. We ask:

What is the optimal deterministic first-order complexity of finding an ϵ-stationary point, and how should the condition numbers $\kappa_{\boldsymbol{x}}$ and $\kappa_{\boldsymbol{y}}$ enter this complexity?

Several upper bounds are known for this problem. The single-loop AGDA method achieves $\mathcal{O}(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}^2\log(1/\epsilon))$ oracle complexity (Yang et al., 2020). More recently, Rankawat et al. (2026) obtained a single-loop complexity of $\mathcal{O}(\kappa_f(\kappa_f+\kappa_{xy}^2)\log^2(1/\epsilon))$, where $\kappa_f := \max\{\kappa_{\boldsymbol{x}},\kappa_{\boldsymbol{y}}\}$ and κ_{xy} denotes the cross-condition number associated with the cross-variable Lipschitz constant. This bound can be substantially sharper when the interaction between the two variables is weak, but does not improve the worst-case complexity over the general two-sided PŁ class. For the general class, multi-loop methods improve the condition-number dependence and currently provide the best-known upper bound, with complexity $\mathcal{O}(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\log^2(1/\epsilon))$ up to logarithmic factors in the condition numbers (Chen et al., 2022). These upper bounds, however, do not determine which dependence on $\kappa_{\boldsymbol{x}}$, $\kappa_{\boldsymbol{y}}$, and ϵ is intrinsic to the problem class. On the lower-bound side, the situation is largely open: to the best of our knowledge, no deterministic first-order oracle lower bound is known for general two-sided PŁ minimax optimization.

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.

We resolve this question by establishing matching lower and upper bounds. First, when $\kappa_{\boldsymbol{x}}, \kappa_{\boldsymbol{y}} \gtrsim 1$ and $0 < \epsilon \lesssim \Delta$, we prove that every deterministic first-order method requires $\Omega(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}} \log(\Delta/\epsilon))$ oracle queries in the worst case. The lower bound applies to arbitrary deterministic first-order methods, rather than only zero-respecting algorithms. Second, we show, through a sharper analysis of the existing single-loop Smoothed Gradient-Descent-Ascent (S-GDA) method of Zhang et al. (2020), that it achieves the matching oracle complexity $\mathcal{O}(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}} \log(\Delta/\epsilon))$, thereby removing one factor of $\log(1/\epsilon)$ from the best-known iteration complexity. Together, these results establish the optimal deterministic first-order oracle complexity for smooth two-sided PŁ minimax optimization. These results further imply a matching $\Theta(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}} \log(1/\epsilon))$ dependence for game stationarity (GS), where a pair $(\boldsymbol{x}, \boldsymbol{y})$ is ϵ-GS if $\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}; \boldsymbol{y})\| \le \epsilon$ and $\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}; \boldsymbol{y})\| \le \epsilon$; see Corollary 3.1.

For the lower bound, the dual chain is the main chain in our hard instance, as illustrated in Figure 1. It consists of $M = \Theta(\kappa_{\boldsymbol{x}} \log(\Delta/\epsilon))$ consecutive blocks of length $N = \Theta(\kappa_{\boldsymbol{y}})$, with the last coordinate of each block connected to a primal variable. For zero-respecting methods, coordinate discovery proceeds sequentially along the dual chain. Reaching the end of a block allows the corresponding primal coordinate to be revealed, so each new primal coordinate requires progress through another N dual coordinates. The primal chain is constructed so that finding an ϵ-stationary point requires revealing $\Omega(M)$ primal coordinates. This yields $\Omega(MN) = \Omega(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}} \log(\Delta/\epsilon))$ oracle queries. A resisting-oracle argument extends the bound to arbitrary deterministic first-order methods.

The main technical challenge is to realize this sequential dependence within a single jointly smooth two-sided PŁ instance. To this end, we adapt the scalar gates and weighted PŁ-chain structure of Pan & Li (2026) to construct a block-structured dual chain and a geometrically weighted primal chain. Unlike the NC-PŁ construction of Pan & Li (2026), where each transition from x_{i-1} to x_i is mediated by a separate length-N dual chain, our dual blocks are connected directly end to end and form a single continuous dual chain. We couple the terminal coordinates of the dual blocks to rescalings of the primal chain so that progress along the dual chain controls the revelation of the corresponding primal coordinates. The coupling preserves both PŁ conditions, while maximizing over the dual variables recovers the weighted primal chain exactly and hence preserves its PŁ and gap properties.

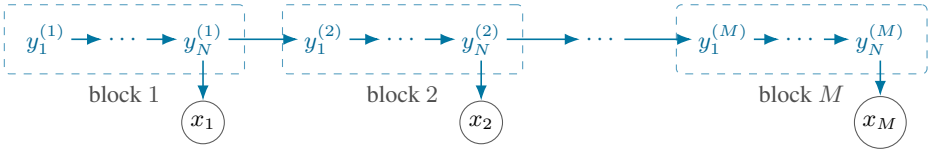


Figure 1: The dual chain governs coordinate discovery. Its blocks are connected end to end, and the terminal gate of block j controls the revelation of the corresponding primal coordinate x_j.

For the matching upper bound, we revisit the Lyapunov analysis of the existing S-GDA method for the NC-PŁ setting (Yang et al., 2022). The additional primal PŁ condition yields an error bound that controls the Lyapunov function by the same quantities appearing in the descent estimate. Combining this error bound with the existing Lyapunov descent gives a global linear contraction with condition-number dependence $\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}$. Together with the initialization procedures, this yields $\mathcal{O}(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}} \log(\Delta/\epsilon))$ first-order oracle complexity for all sufficiently small $\epsilon > 0$. Thus, a sharper analysis of the existing single-loop method attains the matching complexity.

Notation. We write $\kappa_{\boldsymbol{x}} := \ell/\mu_{\boldsymbol{x}}$ and $\kappa_{\boldsymbol{y}} := \ell/\mu_{\boldsymbol{y}}$, and assume $\kappa_{\boldsymbol{x}}, \kappa_{\boldsymbol{y}} \ge 1$ throughout. Bold lowercase letters denote vectors, with scalar coordinates written without boldface. We use $\|\cdot\|$ for the Euclidean norm and $\mathbf{0}$ for the zero vector of the appropriate dimension. Let $\mathbb{N} := \{1, 2, \ldots\}$ and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For $d \in \mathbb{N}$, let $[d] := \{1, \ldots, d\}$, with $[0] := \varnothing$. For $\boldsymbol{z} \in \mathbb{R}^d$, let $\mathrm{supp}(\boldsymbol{z}) := \{j \in [d] \mid z_j \neq 0\}$. For $t \in \mathbb{R}$, define $t^- := \max\{-t, 0\}$, and let $\lfloor t \rfloor$ denote the greatest integer not exceeding t. We adopt the convention $\inf \varnothing = +\infty$. Unless otherwise specified, $\log$ denotes the natural logarithm. The symbols $\mathcal{O}$, Ω, and Θ hide only positive numerical constants.

Organization. Section 2 introduces the function class, stationarity criterion, and first-order oracle model. Section 3 states the main results for lower and upper bounds. Section 4 presents the hard-instance construction and proves the lower bound. Section 5 establishes the matching upper bound for S-GDA. The remaining technical proofs are deferred to the appendix.

2 PRELIMINARIES

Throughout the paper, we consider smooth two-sided PŁ minimax problems. We define the primal function as $\Phi(\boldsymbol{x}) := \max_{\boldsymbol{y}\in\mathbb{R}^n} f(\boldsymbol{x};\boldsymbol{y})$ and let $\Phi^\star := \inf_{\boldsymbol{x}\in\mathbb{R}^m}\Phi(\boldsymbol{x})$.

Definition 2.1 (Smooth two-sided PŁ class)**.** Given $\ell, \mu_{\boldsymbol{x}}, \mu_{\boldsymbol{y}}, \Delta > 0$, let $\mathcal{F}(\ell, \mu_{\boldsymbol{x}}, \mu_{\boldsymbol{y}}, \Delta)$ denote the union over all dimensions $m, n \in \mathbb{N}$ of differentiable functions $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ satisfying the following conditions:

(i) (Smoothness) f is jointly ℓ-smooth, i.e., for all $(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2) \in \mathbb{R}^m \times \mathbb{R}^n$,
$$\|\nabla f(\boldsymbol{x}_1;\boldsymbol{y}_1) - \nabla f(\boldsymbol{x}_2;\boldsymbol{y}_2)\| \le \ell\|(\boldsymbol{x}_1,\boldsymbol{y}_1) - (\boldsymbol{x}_2,\boldsymbol{y}_2)\|.$$

(ii) (Two-sided PŁ) For every $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^n$, the minimum of $f(\cdot;\boldsymbol{y})$ over $\mathbb{R}^m$ and the maximum of $f(\boldsymbol{x};\cdot)$ over $\mathbb{R}^n$ are attained, respectively. Moreover,
$$f(\boldsymbol{x};\boldsymbol{y}) - \min_{\boldsymbol{x}\in\mathbb{R}^m} f(\boldsymbol{x};\boldsymbol{y}) \le \frac{1}{2\mu_{\boldsymbol{x}}}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x};\boldsymbol{y})\|^2,$$
$$\max_{\boldsymbol{y}\in\mathbb{R}^n} f(\boldsymbol{x};\boldsymbol{y}) - f(\boldsymbol{x};\boldsymbol{y}) \le \frac{1}{2\mu_{\boldsymbol{y}}}\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x};\boldsymbol{y})\|^2.$$

(iii) (Initial gap) The initialization satisfies $\Phi(\mathbf{0}) - \inf_{\boldsymbol{x}\in\mathbb{R}^m}\Phi(\boldsymbol{x}) \le \Delta$.

We use the following standard stationarity measure throughout the paper.

Definition 2.2. Given $\epsilon > 0$, a point $\boldsymbol{x} \in \mathbb{R}^m$ is an ϵ-stationary point of Problem (P) if
$$\Phi(\boldsymbol{x}) - \min_{\boldsymbol{x}\in\mathbb{R}^m}\Phi(\boldsymbol{x}) \le \epsilon.$$

Remark 2.1 (Relation to game stationarity)**.** A standard criterion in minimax optimization is game stationarity (GS): a pair $(\boldsymbol{x},\boldsymbol{y})$ is ϵ-GS if $\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x};\boldsymbol{y})\| \le \epsilon$ and $\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x};\boldsymbol{y})\| \le \epsilon$. Under the two-sided PŁ condition,
$$\Phi(\boldsymbol{x}) - \Phi^\star \le \Phi(\boldsymbol{x}) - \inf_{\boldsymbol{u}\in\mathbb{R}^m} f(\boldsymbol{u};\boldsymbol{y}) \le \frac{\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x};\boldsymbol{y})\|^2}{2\mu_{\boldsymbol{y}}} + \frac{\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x};\boldsymbol{y})\|^2}{2\mu_{\boldsymbol{x}}}.$$
Hence, every ϵ-GS point is an $\epsilon^2(\mu_{\boldsymbol{x}}^{-1} + \mu_{\boldsymbol{y}}^{-1})/2$-stationary point in the sense of Definition 2.2.

We next specify the first-order information available to the algorithm.

Definition 2.3 (First-order saddle oracle)**.** For a differentiable function $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, the first-order saddle oracle at $(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^m \times \mathbb{R}^n$ returns
$$\mathbb{O}_f(\boldsymbol{x};\boldsymbol{y}) := \big(f(\boldsymbol{x};\boldsymbol{y}), \nabla_{\boldsymbol{x}} f(\boldsymbol{x};\boldsymbol{y}), \nabla_{\boldsymbol{y}} f(\boldsymbol{x};\boldsymbol{y})\big).$$

We consider arbitrary deterministic first-order methods that access the objective only through the oracle in Definition 2.3. For a first-order algorithm A and a differentiable saddle function $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, let $\mathsf{A}_{\boldsymbol{x}}^{(t)}[f] \in \mathbb{R}^m$ and $\mathsf{A}_{\boldsymbol{y}}^{(t)}[f] \in \mathbb{R}^n$ denote the primal and dual components of its t-th oracle query, respectively.

Definition 2.4 (Deterministic first-order algorithm)**.** A first-order algorithm A is deterministic if there exists a sequence of maps $\{\Gamma_t\}_{t\ge1}$ such that, for every such f, it starts from $(\mathsf{A}_{\boldsymbol{x}}^{(0)}[f], \mathsf{A}_{\boldsymbol{y}}^{(0)}[f]) = \mathbf{0}$ and its query sequence satisfies
$$\big(\mathsf{A}_{\boldsymbol{x}}^{(t)}[f], \mathsf{A}_{\boldsymbol{y}}^{(t)}[f]\big) = \Gamma_t\left(\Big(\mathbb{O}_f\big(\mathsf{A}_{\boldsymbol{x}}^{(s)}[f];\mathsf{A}_{\boldsymbol{y}}^{(s)}[f]\big)\Big)_{s<t}\right), \qquad \forall t \ge 1.$$
We denote the class of such algorithms by $\mathcal{A}_{\mathrm{det}}$.

We measure the worst-case number of oracle queries required to find an ϵ-stationary point as follows.

Definition 2.5 (Oracle complexity)**.** Fix $\epsilon > 0$. Let $\mathcal{A} \subseteq \mathcal{A}_{\mathrm{det}}$ be any nonempty subclass of deterministic first-order algorithms, and let $\mathcal{F}$ be a nonempty class of such saddle functions. The oracle complexity of $\mathcal{A}$ over $\mathcal{F}$ is defined by
$$\mathcal{T}_\epsilon(\mathcal{A},\mathcal{F}) := \inf_{\mathsf{A}\in\mathcal{A}}\sup_{f\in\mathcal{F}}\inf\left\{t\in\mathbb{N}_0 \;\middle|\; \mathsf{A}_{\boldsymbol{x}}^{(t)}[f] \text{ is an } \epsilon\text{-stationary point for } f\right\}.$$

We also consider zero-respecting first-order algorithms that access f through the oracle $\mathbb{O}_f$.

Definition 2.6 (Zero-respecting first-order algorithm)**.** A first-order algorithm A is zero-respecting if, for each such f, it starts from $(\mathsf{A}_{\boldsymbol{x}}^{(0)}[f], \mathsf{A}_{\boldsymbol{y}}^{(0)}[f]) = \mathbf{0}$ and its query sequence satisfies

$$\operatorname{supp}\big(\mathsf{A}_{\boldsymbol{x}}^{(t)}[f], \mathsf{A}_{\boldsymbol{y}}^{(t)}[f]\big) \subseteq \bigcup_{s<t} \operatorname{supp}\Big(\nabla_{\boldsymbol{x}} f\big(\mathsf{A}_{\boldsymbol{x}}^{(s)}[f]; \mathsf{A}_{\boldsymbol{y}}^{(s)}[f]\big), \nabla_{\boldsymbol{y}} f\big(\mathsf{A}_{\boldsymbol{x}}^{(s)}[f]; \mathsf{A}_{\boldsymbol{y}}^{(s)}[f]\big)\Big), \qquad \forall t \geq 1.$$

We denote the class of such algorithms by $\mathcal{A}_{\mathrm{zr}}$.

3 Main Results

We now state our main complexity results. We first establish a lower bound for arbitrary deterministic first-order methods under the smooth two-sided PŁ condition. We then show that the same complexity is attained by the existing single-loop Smoothed Gradient-Descent-Ascent (S-GDA) method through a sharper convergence analysis. Together, these results characterize the deterministic first-order oracle complexity of smooth minimax problems satisfying the two-sided PŁ condition.

3.1 Lower bound

Our first result gives a lower bound for the function class in Definition 2.1 under the first-order saddle-oracle model of Definition 2.3.

Theorem 3.1 (Deterministic two-sided PŁ lower bound)**.** There exist numerical constants $c_0, c_1, c_2 > 0$ such that, whenever $\kappa_{\boldsymbol{x}} := \ell/\mu_{\boldsymbol{x}} \geq c_0$, $\kappa_{\boldsymbol{y}} := \ell/\mu_{\boldsymbol{y}} \geq c_0$, and $0 < \epsilon < c_1\Delta$, we have

$$\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathrm{det}}, \mathcal{F}(\ell, \mu_{\boldsymbol{x}}, \mu_{\boldsymbol{y}}, \Delta)\big) \geq c_2\, \kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}} \log \frac{\Delta}{\epsilon}.$$

Remark 3.1 (Arbitrary deterministic output)**.** Theorem 3.1 also applies when the algorithm returns a deterministic point computed from the final oracle transcript that has not been queried previously. Indeed, one may append this primal point, paired with the dual origin, as the algorithm's next query and apply the lower bound to the resulting augmented deterministic algorithm.

Theorem 3.1 shows that the product dependence on $\kappa_{\boldsymbol{x}}$ and $\kappa_{\boldsymbol{y}}$, together with the logarithmic dependence on Δ/ϵ, is necessary for deterministic first-order methods. Section 4 constructs the hard instance and establishes the zero-respecting lower bound. The extension to arbitrary deterministic first-order methods is obtained through a finite-horizon resisting-oracle argument based on orthogonal rotations; see Appendix B.

3.2 Matching upper bound

We establish the upper bound by analyzing the Smoothed Gradient-Descent-Ascent (S-GDA) method of Zhang et al. (2020). Define

$$F_r(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) := f(\boldsymbol{x}, \boldsymbol{y}) + \frac{r}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2.$$

Given $(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k)$, S-GDA performs

$$\begin{aligned} \boldsymbol{x}^{k+1} &= \boldsymbol{x}^k - \eta_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k), \\ \boldsymbol{y}^{k+1} &= \boldsymbol{y}^k + \eta_{\boldsymbol{y}} \nabla_{\boldsymbol{y}} f(\boldsymbol{x}^{k+1}, \boldsymbol{y}^k), \\ \boldsymbol{z}^{k+1} &= \boldsymbol{z}^k + \beta(\boldsymbol{x}^{k+1} - \boldsymbol{z}^k). \end{aligned} \tag{S-GDA}$$

Under the two-sided PŁ condition, a sharper analysis gives the following result.

Theorem 3.2 (Matching upper bound)**.** Fix $f \in \mathcal{F}(\ell, \mu_{\boldsymbol{x}}, \mu_{\boldsymbol{y}}, \Delta)$. For all sufficiently small $\epsilon > 0$, S-GDA with the parameter choice in Section 5 returns an ϵ-stationary point $\boldsymbol{x}_{\mathrm{out}}$ using

$$\mathcal{O}\left(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}} \log \frac{\Delta}{\epsilon}\right)$$

first-order oracle calls, including initialization procedures. The implicit constant is numerical, while the accuracy threshold may depend on the fixed instance and problem parameters.

The initialization procedures, together with the complete bound for $0 < \epsilon \leq \Delta$, are given in Appendices C.3 and D. Theorem 3.2 removes one logarithmic factor in the target accuracy from the upper bound of Chen et al. (2022). Under the condition-number assumptions of Theorem 3.1, the two results give matching lower and upper bounds of $\Theta(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}} \log(\Delta/\epsilon))$ in this accuracy range.

With a suitable initialization procedure, S-GDA also yields a matching upper bound for game stationarity.

Corollary 3.1 (Game stationarity)**.** Fix $f \in \mathcal{F}(\ell, \mu_{\boldsymbol{x}}, \mu_{\boldsymbol{y}}, \Delta)$. For all sufficiently small $\epsilon > 0$, the deterministic S-GDA procedure in Appendix E returns an ϵ-GS point using

$$\mathcal{O}\left(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}} \log \frac{\ell\Delta}{(\kappa_{\boldsymbol{x}} + \kappa_{\boldsymbol{y}})\epsilon^2}\right)$$

first-order oracle calls, including initialization. The implicit constant is numerical, while the accuracy threshold may depend on the fixed instance and problem parameters. Under the condition-number assumptions of Theorem 3.1, a matching worst-case lower bound holds for deterministic first-order methods.

The algorithm and proof of Corollary 3.1 are given in Appendix E.

4 Lower-Bound Construction and Proof Overview

We now prove Theorem 3.1. We first identify the properties required of the hard instance, then construct an instance satisfying these properties and establish the zero-respecting lower bound. The extension to arbitrary deterministic first-order methods is obtained through a rotation argument.

4.1 Certificates for the hard instance

Call a triple $(K, N, S) \in \mathbb{N}^3$ admissible if $K, N \geq 2$ and $S \geq 1$, and set $M := KS$. For each admissible triple, we construct an unscaled instance $\bar{f} : \mathbb{R}^M \times \mathbb{R}^{MN} \to \mathbb{R}$. Let $\bar{\Phi}(\boldsymbol{x}) := \max_{\boldsymbol{y} \in \mathbb{R}^{MN}} \bar{f}(\boldsymbol{x}; \boldsymbol{y})$ and $\bar{\Phi}^\star := \inf_{\boldsymbol{x} \in \mathbb{R}^M} \bar{\Phi}(\boldsymbol{x})$. The primal variable has M coordinates, and we partition the dual variable into M consecutive blocks of length N: $\boldsymbol{y} = (\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(M)})$, where $\boldsymbol{y}^{(i)} \in \mathbb{R}^N$. We require the following four properties.

Condition 4.1 (Unscaled hard instance properties)**.** There are numerical constants $\ell_0, \mu_1, \mu_2 > 0$, independent of K, N, S, such that, for every admissible triple:

(C1) The function $\bar{f}$ is jointly ℓ_0-smooth. For every $(\boldsymbol{x}, \boldsymbol{y})$, the functions $\bar{f}(\cdot; \boldsymbol{y})$ and $-\bar{f}(\boldsymbol{x}; \cdot)$ attain their minima and are (μ_1/K)- and (μ_2/N)-PŁ, respectively.

(C2) For every $(\boldsymbol{x}, \boldsymbol{y})$ and integer $0 \leq h < MN$,

$$\operatorname{supp}(\boldsymbol{y}) \subseteq [h] \quad \Longrightarrow \quad \begin{cases} \operatorname{supp}(\nabla_{\boldsymbol{y}} \bar{f}(\boldsymbol{x}; \boldsymbol{y})) \subseteq [h+1], \\ \operatorname{supp}(\nabla_{\boldsymbol{x}} \bar{f}(\boldsymbol{x}; \boldsymbol{y})) \subseteq \operatorname{supp}(\boldsymbol{x}) \cup [\lfloor h/N \rfloor]. \end{cases}$$

(C3) For every $\boldsymbol{x} \in \mathbb{R}^M$ and $s \in [S]$,

$$\operatorname{supp}(\boldsymbol{x}) \subseteq [(s-1)K] \quad \Longrightarrow \quad \bar{\Phi}(\boldsymbol{x}) - \bar{\Phi}^\star \geq 2^{-s}.$$

(C4) The initial gap satisfies $\Delta_0 := \bar{\Phi}(\mathbf{0}) - \bar{\Phi}^\star \leq 2$.

Conditions (C2) and (C3) provide the information-theoretic part of the lower bound: (C2) restricts coordinate discovery, while (C3) keeps the primal-function gap bounded away from zero when primal coordinates have not been revealed. Conditions (C1) and (C4) control the analytic parameters of the construction and allow the instance to be scaled into the prescribed function class.

We now construct an instance satisfying Condition 4.1.

4.2 THE DUAL PŁ CHAIN

Following the PŁ chain structure of Pan & Li (2026), we introduce the scalar gate functions

$$q(t) = \begin{cases} 0, & t \le 0, \\ 3t^2 - 2t^3, & 0 < t < 1, \\ 1, & t \ge 1, \end{cases} \qquad p(t) = q\left(\frac{t+1}{2}\right). \tag{4.1}$$

For an input $r \in [0,1]$ and normalized dual coordinates $\boldsymbol{w} \in \mathbb{R}^N$, we define the inner block by

$$H(r, \boldsymbol{w}) := 1 - rp(w_1) + \sum_{h=2}^{N} \{1 - q(w_{h-1})p(w_h)\} + \frac{1}{2}\sum_{h=1}^{N} (w_h^-)^2. \tag{4.2}$$

Here, the scalar r serves as the input gate of the block. In particular, r controls the first coordinate w_1, while for $h \ge 2$, the coordinate w_h is controlled by the gate $q(w_{h-1})$. Since $p'(0) > 0$ whereas $q(0) = q'(0) = 0$, the chain reveals at most one new coordinate at a time. Thus, $H(r, \boldsymbol{w})$ defines a length-N sequential chain.

We next concatenate M inner chains. For each block $i \in [M]$, let

$$\boldsymbol{w}^{(i)} = (w_1^{(i)}, \dots, w_N^{(i)}) := \frac{\boldsymbol{y}^{(i)}}{\sqrt{b_i}}, \qquad b_i := 2^{M-i},$$

where $\boldsymbol{y} = (\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(M)}) \in \mathbb{R}^{MN}$ denotes the dual variable. Thus, $\boldsymbol{w}^{(i)} \in \mathbb{R}^N$ gives the normalized coordinates of block i. We then define the terminal gates by

$$r_i(\boldsymbol{y}) := q\left(w_N^{(i)}\right), \qquad i \in [M],$$

and set $r_0(\boldsymbol{y}) \equiv 1$. The first inner chain therefore has a fixed input, while each subsequent chain receives its input from the terminal gate of its predecessor. The resulting weighted dual chain is

$$D(\boldsymbol{y}) := \sum_{i=1}^{M} b_i H\left(r_{i-1}(\boldsymbol{y}), \boldsymbol{w}^{(i)}\right). \tag{4.3}$$

Since each block contains N coordinates and the block weights decrease geometrically, for every $i \in [M]$, $N\sum_{k=i}^{M} b_k \le 2Nb_i$. This estimate allows the PŁ modulus of the dual chain to remain of order $1/N$, independently of the number M of blocks. The following lemma records the properties of D needed later in the primal-dual coupling.

Lemma 4.1 (Properties of the dual chain)**.** Let D be defined in (4.3). The following statements hold.

(i) D has minimum zero and nonpositive coordinate derivatives.

(ii) D is $1/(260N)$-PŁ and 11-smooth.

(iii) For every $\boldsymbol{y} \in \mathbb{R}^{MN}$, we have

$$\sum_{i=1}^{M} \{1 - r_i(\boldsymbol{y})\} \le 4D(\boldsymbol{y}). \tag{4.4}$$

The derivative signs prevent cancellation in the dual gradient, while (4.4) controls the additional dual gap. The proof is in Appendix A.1.

4.3 THE PRIMAL PŁ CHAIN

We next construct a primal chain of length $M = KS$, partitioned into S consecutive segments of length K. Using the functions in (4.1), we assign the weights

$$\omega_j := \frac{1}{K} 2^{-\lfloor (j-1)/K \rfloor}, \qquad j \in [M]. \tag{4.5}$$

For $\boldsymbol{x} = (x_1, \ldots, x_M) \in \mathbb{R}^M$, define the normalized primal coordinates by

$$\xi_j := \frac{x_j}{\sqrt{\omega_j}}, \qquad j \in [M],$$

and set $\xi_0 := 1$. The primal chain is

$$P(\boldsymbol{x}) := \sum_{j=1}^{M} \omega_j \left\{1 - q(\xi_{j-1})p(\xi_j)\right\} + \frac{1}{2}\sum_{j=1}^{M}(x_j^-)^2. \tag{4.6}$$

Therefore, all coordinates in segment s have the same weight, $2^{-(s-1)}/K$. The tail sums satisfy $\sum_{k=j}^{M} \omega_k \le 2K\omega_j$, so the PŁ constant depends on K independently of S. The following lemma records the properties needed for the lower bound.

Lemma 4.2 (Properties of the primal chain)**.** Let P be defined in (4.6). The following statements hold.

(i) P is $1/(260K)$-PŁ and 11-smooth, with $\min_{\boldsymbol{x}\in\mathbb{R}^M} P(\boldsymbol{x}) = 0$ and $P(\mathbf{0}) \le 2$.

(ii) For every $j \in [M]$,

$$\operatorname{supp}(\boldsymbol{x}) \subseteq [j-1] \quad \Longrightarrow \quad \operatorname{supp}(\nabla P(\boldsymbol{x})) \subseteq [j].$$

(iii) For every $s \in [S]$,

$$\operatorname{supp}(\boldsymbol{x}) \subseteq [(s-1)K] \quad \Longrightarrow \quad P(\boldsymbol{x}) \ge 2^{-s}. \tag{4.7}$$

The last property ensures that $P(\boldsymbol{x}) - \min_{\boldsymbol{x}} P(\boldsymbol{x})$ remains bounded away from zero before the s-th primal segment is revealed. The proof is in Appendix A.2.

Remark 4.1 (An alternative PŁ hard instance)**.** The primal chain P is a geometrically weighted modification of the PŁ chain construction of Pan & Li (2026). This construction is structurally different from the hard instance of Yue et al. (2023) for smooth PŁ minimization. Nevertheless, when considered as a standalone minimization instance and appropriately scaled, Lemma 4.2 together with a standard resisting-oracle argument yields the same $\Omega((\ell/\mu)\log(\Delta/\epsilon))$ deterministic first-order lower bound for ℓ-smooth, μ-PŁ functions with initial gap Δ. Thus, P provides an alternative hard instance for smooth PŁ minimization, while its segmented geometric weighting is used here for the two-sided minimax construction. The details are given in Appendix A.3.

4.4 Primal-dual coupling

We now introduce the primal-dual coupling by rescaling the primal chain using the terminal gates of the dual blocks. The main requirement is that, for every fixed $\boldsymbol{y}$, the coupled primal objective retain a PŁ modulus of order $1/K$ uniformly over all partially completed dual states. To this end, the terminal gates of the dual blocks are propagated through all subsequent primal coordinates.

Using the terminal gates defined in Section 4.2, we set

$$\theta_0(\boldsymbol{y}) := 1, \qquad \theta_j(\boldsymbol{y}) := \prod_{i=1}^{j} r_i(\boldsymbol{y}), \qquad j \in [M]. \tag{4.8}$$

For brevity, we use θ_j to denote $\theta_j(\boldsymbol{y})$. These scales satisfy $1 = \theta_0 \ge \theta_1 \ge \cdots \ge \theta_M \ge 0$. For positive scales, define the normalized primal coordinates by

$$\zeta_j := \frac{x_j}{\theta_j\sqrt{\omega_j}}, \qquad j \in [M],$$

and set $\zeta_0 := 1$. The coupled primal term is

$$C(\boldsymbol{x};\boldsymbol{y}) := \sum_{j=1}^{M} \omega_j \Big[\theta_{j-1}^2\{1 - q(\zeta_{j-1})\} + \theta_j^2 q(\zeta_{j-1})\{1 - p(\zeta_j)\}\Big] + \frac{1}{2}\sum_{j=1}^{M}(x_j^-)^2. \tag{4.9}$$

If some scale vanishes, C is defined by continuous extension from positive nested scales; see Appendix A.1. We finally define the unscaled hard instance by

$$\bar{f}(\boldsymbol{x};\boldsymbol{y}) := C(\boldsymbol{x};\boldsymbol{y}) - D(\boldsymbol{y}). \tag{4.10}$$

The following lemma collects the resulting properties.

Lemma 4.3 (Properties of the unscaled instance)**.** Let $\bar{f}$ be defined in (4.10). There are numerical constants $\ell_0 \geq 1$ and $\mu_1, \mu_2 > 0$, independent of K, N, S, such that:

(i) $\bar{f}$ is jointly ℓ_0-smooth. For every $\boldsymbol{y}$, $\bar{f}(\cdot;\boldsymbol{y})$ attains its minimum and is (μ_1/K)-PŁ; for every $\boldsymbol{x}$, $-\bar{f}(\boldsymbol{x};\cdot)$ attains its minimum and is (μ_2/N)-PŁ.

(ii) The primal function of $\bar{f}$ satisfies $\bar{\Phi}(\boldsymbol{x}) = P(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathbb{R}^M$.

Lemma 4.3 verifies Condition 4.1 (C1). Moreover, since $\bar{\Phi} = P$, Lemma 4.2 yields (C3) and (C4). Condition (C2) is verified separately in Appendix A.6. The proof of Lemma 4.3 is given in Appendix A.4.

4.5 Scaling and the zero-respecting lower bound

We use the following scaling to match the prescribed smoothness, PŁ constants, and initial gap while preserving coordinate supports.

Lemma 4.4 (Scaling lemma)**.** Fix an admissible triple (K, N, S), set $M = KS$, and let $\bar{f}$ be defined in (4.10). Given $\ell, \lambda > 0$, define

$$f(\boldsymbol{x};\boldsymbol{y}) := \frac{\ell\lambda^2}{\ell_0}\bar{f}\Big(\frac{\boldsymbol{x}}{\lambda};\frac{\boldsymbol{y}}{\lambda}\Big), \qquad (\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^M \times \mathbb{R}^{MN}. \tag{4.11}$$

Then the following statements hold:

(i) For every $\boldsymbol{y}$, $f(\cdot;\boldsymbol{y})$ attains its minimum; for every $\boldsymbol{x}$, $f(\boldsymbol{x};\cdot)$ attains its maximum. Moreover,

$$\Phi(\boldsymbol{x}) = \frac{\ell\lambda^2}{\ell_0}P\Big(\frac{\boldsymbol{x}}{\lambda}\Big), \qquad \Phi(\mathbf{0}) - \Phi^\star = \frac{\ell\lambda^2}{\ell_0}\Delta_0.$$

(ii) The function f is jointly ℓ-smooth. For every $\boldsymbol{y}$, $f(\cdot;\boldsymbol{y})$ is $\ell\mu_1/(\ell_0 K)$-PŁ; for every $\boldsymbol{x}$, $-f(\boldsymbol{x};\cdot)$ is $\ell\mu_2/(\ell_0 N)$-PŁ.

(iii) The scaling preserves the coordinate supports of the iterates and oracle gradients at corresponding points.

We choose $K = \Theta(\kappa_{\boldsymbol{x}})$ and $N = \Theta(\kappa_{\boldsymbol{y}})$ so that the scaled instance satisfies the prescribed primal and dual PŁ conditions, and choose $S = \Theta(\log(\Delta/\epsilon))$. Condition 4.1 (C2) then implies that revealing one additional primal coordinate requires progress through a dual block of length N. Hence reaching the s-th primal segment requires on the order of sKN oracle queries. By Condition (C3), before that segment is reached the primal-function gap remains at least the corresponding geometric level. Together with the scaling above, this yields the following result.

Proposition 4.1 (Zero-respecting lower bound)**.** There exist numerical constants $c_0, c_1, c_2 > 0$ such that, whenever $\kappa_{\boldsymbol{x}} := \ell/\mu_{\boldsymbol{x}} \geq c_0$, $\kappa_{\boldsymbol{y}} := \ell/\mu_{\boldsymbol{y}} \geq c_0$, and $0 < \epsilon < c_1\Delta$, there exists $f \in \mathcal{F}(\ell, \mu_{\boldsymbol{x}}, \mu_{\boldsymbol{y}}, \Delta)$ such that

$$\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathrm{zr}} \cap \mathcal{A}_{\mathrm{det}}, \{f\}\big) \geq c_2\,\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}} \log\frac{\Delta}{\epsilon}.$$

The proof, including the precise parameter choices, is given in Appendices A.5 and A.6.

5 Matching Upper Bound via S-GDA

We now show that the lower bound in Theorem 3.1 is attained by the single-loop S-GDA method in (S-GDA). Our analysis builds on the Lyapunov descent developed for S-GDA in the NC-PŁ setting (Yang et al., 2022). The additional primal PŁ condition allows us to establish an error bound that controls the Lyapunov function by the same quantities appearing in the descent estimate. Combining these two ingredients yields a linear contraction with factor $1 - \Theta(1/(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}))$, which leads to the matching iteration complexity.

5.1 LYAPUNOV DESCENT

Following Yang et al. (2022), we consider the Lyapunov function

$$\mathcal{V}_r(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \big(F_r(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) - d_r(\boldsymbol{y},\boldsymbol{z})\big) + \big(P_r(\boldsymbol{z}) - d_r(\boldsymbol{y},\boldsymbol{z})\big) + \big(P_r(\boldsymbol{z}) - P_r^\star\big),$$

where $d_r(\boldsymbol{y},\boldsymbol{z}) := \min_{\boldsymbol{x}} F_r(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$, $P_r(\boldsymbol{z}) := \min_{\boldsymbol{x}} \max_{\boldsymbol{y}} F_r(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \min_{\boldsymbol{x}}\{\Phi(\boldsymbol{x}) + \frac{r}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2\}$, and $P_r^\star := \inf_{\boldsymbol{z}} P_r(\boldsymbol{z}) = \Phi^\star$. Up to the additive constant $P_r^\star$, this is the Lyapunov function used in the NC-PŁ analysis of S-GDA, which also yields the descent estimate stated below.

Lemma 5.1 (Lyapunov descent)**.** Let $r = 2\ell$, $\eta_{\boldsymbol{x}} = 1/(3\ell)$, $\eta_{\boldsymbol{y}} = 1/(144\ell)$, and $\beta = \eta_{\boldsymbol{y}}\mu_{\boldsymbol{y}}/1600$. For each $k \geq 0$, define $\boldsymbol{x}_r(\boldsymbol{y}^k,\boldsymbol{z}^k) := \arg\min_{\boldsymbol{x}} F_r(\boldsymbol{x},\boldsymbol{y}^k,\boldsymbol{z}^k)$. Then

$$\mathcal{V}_r^k - \mathcal{V}_r^{k+1} \geq \frac{\eta_{\boldsymbol{x}}}{4}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k,\boldsymbol{y}^k,\boldsymbol{z}^k)\|^2 + \frac{\eta_{\boldsymbol{y}}}{16}\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k,\boldsymbol{z}^k),\boldsymbol{y}^k)\|^2 + \frac{r\beta}{4}\|\boldsymbol{x}^{k+1} - \boldsymbol{z}^k\|^2,$$

where $\mathcal{V}_r^k := \mathcal{V}_r(\boldsymbol{x}^k,\boldsymbol{y}^k,\boldsymbol{z}^k)$.

Lemma 5.1 follows from the deterministic specialization of Yang et al. (2022, Theorem 4.1), in particular Eq. (36) together with Lemma C.2 therein; the notation correspondence is given in Appendix C.

5.2 LINEAR CONTRACTION UNDER THE PRIMAL PŁ CONDITION

We next relate the Lyapunov function to the quantities appearing in the descent estimate of Lemma 5.1. The first two terms in $\mathcal{V}_r$ are controlled by the regularized strong convexity and the dual PŁ condition, respectively, while the additional primal PŁ condition controls the remaining term $P_r(\boldsymbol{z}) - P_r^\star$. This yields the following bound.

Lemma 5.2 (Lyapunov bound)**.** Under the parameter choice of Lemma 5.1, there exists a numerical constant $c_3 > 0$ such that, for every $k \geq 0$,

$$\mathcal{V}_r^k \leq c_3\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\left(\frac{\eta_{\boldsymbol{x}}}{4}\big\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k,\boldsymbol{y}^k,\boldsymbol{z}^k)\big\|^2 + \frac{\eta_{\boldsymbol{y}}}{16}\big\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k,\boldsymbol{z}^k),\boldsymbol{y}^k)\big\|^2 + \frac{r\beta}{4}\big\|\boldsymbol{x}^{k+1} - \boldsymbol{z}^k\big\|^2\right).$$

The proof of Lemma 5.2 is given in Appendix C. The key additional estimate uses the primal PŁ condition to control the proximal gap $P_r(\boldsymbol{z}^k) - P_r^\star$, which is unavailable in the NC-PŁ analysis.

Combining Lemmas 5.1 and 5.2 immediately gives a linear contraction.

Corollary 5.1. There exists a numerical constant $c_4 > 0$ such that, for every $k \geq 0$,

$$\mathcal{V}_r^{k+1} \leq \left(1 - \frac{c_4}{\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}}\right)\mathcal{V}_r^k.$$

Proof of Corollary 5.1. Combining Lemmas 5.1 and 5.2 gives

$$\mathcal{V}_r^k - \mathcal{V}_r^{k+1} \geq \frac{1}{c_3\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}}\mathcal{V}_r^k.$$

Rearranging proves the claim with $c_4 = 1/c_3$. ■

Combining Corollary 5.1 with the common initialization procedures in Appendix C.3 yields Theorem 3.2.

6 CONCLUSION

We establish matching lower and upper bounds for smooth two-sided PŁ minimax optimization under the primal-function gap criterion. Specifically, we prove that every deterministic first-order method requires $\Omega(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\log(\Delta/\epsilon))$ oracle queries in the worst case, and show that the existing single-loop S-GDA method attains $\mathcal{O}(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\log(\Delta/\epsilon))$ complexity through a sharper Lyapunov analysis. These results show that the product dependence on $\kappa_{\boldsymbol{x}}$ and $\kappa_{\boldsymbol{y}}$ is intrinsic. The same scaling extends to game stationarity, yielding $\Theta(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\log(1/\epsilon))$ in the high-accuracy regime.

Extending the lower bound to randomized first-order methods remains open, as the resisting-oracle argument used here applies only to deterministic algorithms.

AI USE STATEMENT

In this work, we used generative AI tools for developing the lower-bound hard-instance construction and assisting with the corresponding proof development and checking. We did not use generative AI tools to develop the upper-bound result or its mathematical analysis, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for drafting and editing parts of the manuscript. We have reviewed all AI-assisted work. The lower-bound construction was developed through iterative collaboration with GPT-5.6 Sol, while the upper-bound analysis was developed independently by the authors. All AI-assisted mathematical arguments were independently verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is theoretical and does not involve human subjects, personal data, or sensitive datasets. We are not aware of any particular ethical concerns arising from the results presented in this paper.

REPRODUCIBILITY STATEMENT

All assumptions, parameter choices, and complexity measures are stated explicitly in the main text, and complete proofs of the lower-bound and upper-bound results are provided in the appendix. An accompanying Lean formalization, available at `https://anonymous.4open.science/r/plpl-lean-3E4E/README.md`, provides a machine-checked verification of the lower-bound construction and the deterministic first-order oracle lower bound established in this paper. No datasets or empirical results are used in this work.

REFERENCES

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1):71–120, 2020.

Lesi Chen, Boyuan Yao, and Luo Luo. Faster stochastic algorithms for minimax optimization under Polyak-Łojasiewicz condition. *Advances in Neural Information Processing Systems*, 35:13921–13932, 2022.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Dmitry Kovalev and Alexander Gasnikov. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. *Advances in Neural Information Processing Systems*, 35:14691–14703, 2022.

Siyu Pan and Jiajin Li. Lower bounds for nonconvex-PŁ minimax optimization, 2026.

B.T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. ISSN 0041-5553.

Mansi Rankawat, Michael Muehlebach, Simon Lacoste-Julien, and Damien Scieur. From Lyapunov analysis to algorithm design in two-sided PL minimax optimization. In *Forty-third International Conference on Machine Learning*, 2026.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 33:1153–1165, 2020.

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pp. 5485–5517. PMLR, 2022.

Pengyun Yue, Cong Fang, and Zhouchen Lin. On the lower bound of minimizing Polyak-Łojasiewicz functions. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2948–2968. PMLR, 2023.

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

A Proofs for Section 4

We give the proofs deferred from Section 4 in order. All vector norms are Euclidean; $\|\cdot\|$ and $\|\cdot\|_F$ denote the operator and Frobenius norms of matrices, respectively. We use the gates p, q defined in (4.1).

A.1 Proof of Lemma 4.1

We first establish a bound that applies to both chains and to the reweighted primal chain. Let $a_1 \geq \cdots \geq a_d > 0$, and let $Q \geq 1$ be a constant satisfying

$$\sum_{k=j}^{d} a_k \leq Q a_j, \qquad j \in [d], \tag{A.1}$$

and let $1 = \theta_0 \geq \theta_1 \geq \cdots \geq \theta_d \geq 0$. When all scales are positive, define

$$H_{\boldsymbol{a},\boldsymbol{\theta}}(\boldsymbol{x}) := \sum_{j=1}^{d} a_j \Big[\theta_{j-1}^2\{1 - q(\zeta_{j-1})\} + \theta_j^2 q(\zeta_{j-1})\{1 - p(\zeta_j)\}\Big] + \frac{1}{2}\sum_{j=1}^{d}(x_j^-)^2, \tag{A.2}$$

where

$$\zeta_j := \frac{x_j}{\theta_j\sqrt{a_j}}, \qquad \zeta_0 := 1.$$

Here ζ_0 is fixed; no auxiliary coordinate or weight is needed.

Lemma A.1 (Uniform weighted-chain bound)**.** For every nonincreasing scale vector $\boldsymbol{\theta}$, the continuous extension of $H_{\boldsymbol{a},\boldsymbol{\theta}}$ is nonnegative, has nonpositive coordinate derivatives, and is 11-smooth. Moreover,

$$H_{\boldsymbol{a},\boldsymbol{\theta}}(\boldsymbol{x}) \leq 65Q\|\nabla_{\boldsymbol{x}} H_{\boldsymbol{a},\boldsymbol{\theta}}(\boldsymbol{x})\|^2. \tag{A.3}$$

Its minimum is zero, attained whenever $x_j \geq \theta_j\sqrt{a_j}$ for every $j \in [d]$.

Proof of Lemma A.1. The gates satisfy $0 \leq q \leq p \leq 1$, $q', p' \geq 0$, and

$$1 - q(t) \leq 4\big(p'(t)^2 + (t^-)^2\big). \tag{A.4}$$

To verify the last inequality, for $0 \leq t < 1$ we have $(1 - q(t))/p'(t)^2 = 16(1 + 2t)/(9(1 + t)^2) \leq 16/9$. For $-1 \leq t \leq 0$, $p'(t)^2 + t^2 = 9(1 - t^2)^2/16 + t^2 \geq 5/9$. The cases $t \leq -1$ and $t \geq 1$ are immediate.

Suppose first that all scales are positive, and abbreviate $H = H_{\boldsymbol{a},\boldsymbol{\theta}}$. Set

$$\nu_j = a_j\theta_j^2, \qquad \delta_j = 1 - q(\zeta_j), \qquad \eta_j = q(\zeta_{j-1})^2 p'(\zeta_j)^2 + (\zeta_j^-)^2.$$

Differentiating gives

$$-\partial_{x_j} H = \sqrt{a_j}\theta_j \left[q(\zeta_{j-1})p'(\zeta_j) + \zeta_j^- + \mathbf{1}_{\{j<d\}}\frac{a_{j+1}}{a_j} q'(\zeta_j)\left(1 - \frac{\theta_{j+1}^2}{\theta_j^2}\{1 - p(\zeta_{j+1})\}\right)\right], \tag{A.5}$$

where the last term is omitted when $j = d$. All terms in brackets are nonnegative. Hence every coordinate derivative is nonpositive and $\|\nabla H\|^2 \geq \sum_j \nu_j\eta_j$.

Fix k with $\delta_k > 0$, and take the first $j \le k$ for which $\delta_j \ge \delta_k/2$. Since $\delta_0 = 0$, this choice gives $q(\zeta_{j-1}) \ge 1/2$. By (A.4), we have

$$\delta_k \le 2\delta_j \le 8\{p'(\zeta_j)^2 + (\zeta_j^-)^2\} \le 32\eta_j \le 32 \sum_{h \le k} \eta_h.$$

The same bound holds when $\delta_k = 0$. Reindexing the first terms in (A.2) and using $a_{j+1} \le a_j$ and $p \ge q$, we obtain

$$H \le 2 \sum_k \nu_k \delta_k + \frac{1}{2} \sum_k \nu_k (\zeta_k^-)^2.$$

The scales are nonincreasing, so $\sum_{k \ge j} \nu_k \le Q\nu_j$. Therefore

$$H \le 64 \sum_j \eta_j \sum_{k \ge j} \nu_k + \frac{1}{2} \sum_j \nu_j \eta_j \le (64Q + 1/2) \sum_j \nu_j \eta_j \le 65Q \|\nabla H\|^2,$$

where $Q \ge 1$ follows from (A.1).

For smoothness, the gate derivatives satisfy $\|p'\|_\infty \le 3/4$, $\|q'\|_\infty \le 3/2$, $|p''| \le 3/2$, and $|q''| \le 6$ almost everywhere. The Hessian is tridiagonal. Its diagonal entries have absolute value at most $3/2 + 1 + 6 = 17/2$, and its off-diagonal entries at most $9/8$, by (A.5) and the monotonicity of the weights and scales. Thus its absolute row sums are at most $43/4 < 11$. Since the summands are $C^{1,1}$, this proves global 11-smoothness.

It remains to justify zero scales. Let r be the number of positive scales. Keep the bounded summands with indices $j \le r$, the boundary term

$$a_{r+1}\theta_r^2\{1 - q(\zeta_r)\} \qquad \text{if } r < d, \tag{A.6}$$

and all negative-part quadratics. For $r = 0$, the boundary term is zero because $\theta_0 = \zeta_0 = 1$. This defines the continuous extension: the j-th bounded summand is at most $2a_j\theta_{j-1}^2$, and its term involving $p(\zeta_j)$ is at most $a_j\theta_j^2$. The bounded part of its j-th gradient component satisfies

$$\left|\partial_{x_j} H + x_j^-\right| \le \theta_j \sqrt{a_j} \left(\frac{3}{4} + \mathbf{1}_{\{j<d\}} \frac{3a_{j+1}}{2a_j} \right).$$

It vanishes with θ_j. If θ_j stays positive while θ_{j+1} vanishes, the dependence on ζ_{j+1} in (A.5) disappears because it is multiplied by $(\theta_{j+1}/\theta_j)^2$. Thus the function and its gradient with respect to $\boldsymbol{x}$ converge locally uniformly as scales vanish. Approximation by $\theta_j^\tau = (1-\tau)\theta_j + \tau$, $\tau \downarrow 0$, preserves (A.3), the derivative signs, and 11-smoothness. Finally, every summand is nonnegative and vanishes when $x_j \ge \theta_j\sqrt{a_j}$ for all j, proving the minimum claim. ■

Proof of Lemma 4.1. Repeat each weight $b_i = 2^{M-i}$ exactly N times to form a vector $\boldsymbol{d}$. Flattening the dual blocks gives $D = H_{\boldsymbol{d},\mathbf{1}}$. A tail starting in block i has weight at most $N \sum_{k=i}^M b_k \le 2Nb_i$. Lemma A.1 with $Q = 2N$ therefore proves (i) and (ii).

For (iii), the gate formulas give $1 - q(t) \le 4\{1 - p(t)\}$. Indeed, for $0 < t < 1$ the ratio is $4(1+2t)/(2+t) \le 4$; outside this interval the inequality is immediate. Since $N \ge 2$, the last internal term of each block gives

$$D(\boldsymbol{y}) \ge \sum_{i=1}^M b_i \left\{ 1 - p\left(\frac{y_N^{(i)}}{\sqrt{b_i}} \right) \right\} \ge \frac{1}{4} \sum_{i=1}^M \{1 - r_i(\boldsymbol{y})\},$$

where the last step also uses $b_i \ge 1$. ■

A.2 Proof of Lemma 4.2

Proof of Lemma 4.2. We have $P = H_{\boldsymbol{\omega},\mathbf{1}}$. The weights satisfy

$$\sum_{j=1}^M \omega_j = 2(1 - 2^{-S}) < 2, \qquad \sum_{k=j}^M \omega_k \le 2K\omega_j.$$

For the tail bound, the remainder of the current segment contributes at most $K\omega_j$, and the later segments together contribute at most another $K\omega_j$. Lemma A.1 with $Q = 2K$ gives the PŁ and smoothness bounds and $\min P = 0$. Since $P(\mathbf{0}) \le \sum_j \omega_j < 2$, this proves (i).

If $\mathrm{supp}(\boldsymbol{x}) \subseteq [j-1]$, then (A.5) gives $\partial_{x_k} P = 0$ for $k > j$, because $q(0) = q'(0) = 0$. This proves (ii).

If $\mathrm{supp}(\boldsymbol{x}) \subseteq [(s-1)K]$, all coordinates in segment s are zero. Its first term contributes at least half its weight, and the other $K-1$ terms contribute their full weights. Since all remaining terms are nonnegative,

$$P(\boldsymbol{x}) \ge \left(1 - \frac{1}{2K}\right) 2^{-(s-1)} \ge 2^{-s}. \tag{A.7}$$

This proves (iii). ■

A.3 Proof of Remark 4.1

Proof of Remark 4.1. For $\ell/\mu \ge 5720$ and $0 < \epsilon \le \Delta/64$, take $K = \lfloor \ell/(2860\mu) \rfloor$, $S = 2\lfloor \log_2(\Delta/(4\epsilon)) \rfloor$, and $g(\boldsymbol{x}) = (\Delta/2)P(\boldsymbol{x}/\lambda)$ with $\lambda^2 = 11\Delta/(2\ell)$. Lemma 4.2 makes g ℓ-smooth and μ-PŁ, with initial gap at most Δ. A zero-respecting method reveals at most one coordinate per query. For $t \le (S/2-1)K$, its query is supported on $[(S/2-1)K]$, so (A.7) gives a gap at least $(\Delta/2)2^{-S/2} \ge 2\epsilon$. The same finite-horizon rotation argument, specialized to the single-variable setting, extends the bound to arbitrary deterministic first-order methods; see Appendix B. Since $(S/2-1)K = \Omega((\ell/\mu)\log(\Delta/\epsilon))$, the claimed lower bound follows. ■

A.4 Proof of Lemma 4.3

We first treat the terminal gate values as independent variables.

Lemma A.2 (Derivatives with respect to the gate values)**.** Let $a_1 \ge \cdots \ge a_d > 0$ and $\theta_j(\boldsymbol{\rho}) = \prod_{h=1}^{j} \rho_h$ for $\boldsymbol{\rho} \in [0,1]^d$. Let $H(\boldsymbol{x}, \boldsymbol{\rho}) = H_{\boldsymbol{a},\boldsymbol{\theta}(\boldsymbol{\rho})}(\boldsymbol{x})$ and $A_* = \sum_j a_j$. The function H and its first derivatives extend continuously to $\mathbb{R}^d \times [0,1]^d$. There are numerical constants $c_{\rho\rho}, c_{x\rho} > 0$, independent of the dimension, weights, scales, and evaluation points, such that

$$0 \le H_{\rho_i} \le 4\sum_{j=i}^{d} a_j, \qquad |H_{\rho_i\rho_k}| \le c_{\rho\rho} A_*, \qquad \|\nabla_{\boldsymbol{x}} H_{\rho_i}\|^2 \le c_{x\rho}^2 A_*. \tag{A.8}$$

The second-derivative bounds hold almost everywhere on $\mathbb{R}^d \times (0,1)^d$.

Proof of Lemma A.2. Set $v = 1 - p$ and define

$$R_p(t) = 2v(t) + tp'(t), \qquad R_q(t) = 2\{1 - q(t)\} + tq'(t).$$

The gate formulas give $0 \le R_p, R_q \le 2$. Indeed, $R_p(t) = 1 - 3t/4 - t^3/4$ on $(-1,1)$ and $R_q(t) = 2 - 2t^3$ on $(0,1)$; each equals 2 to the left and 0 to the right of its respective interval.

For the bounded summand E_j in (A.2), with $j > 1$, put

$$\tau = \theta_{j-1}, \quad r = \rho_j, \quad s = \frac{x_{j-1}}{\tau\sqrt{a_{j-1}}}, \quad t = \frac{x_j}{\tau r\sqrt{a_j}}.$$

Then $E_j = a_j\tau^2\{1 - q(s) + r^2 q(s)v(t)\}$, and differentiation at positive gate values gives

$$\begin{aligned} \partial_{\rho_i} E_j &= \frac{a_j\tau^2}{\rho_i}\big[2\{1-q(s)\} + sq'(s)\{1 - r^2 v(t)\} + r^2 q(s) R_p(t)\big], \quad i < j, \\ \partial_{\rho_j} E_j &= a_j\tau^2 r q(s) R_p(t). \end{aligned} \tag{A.9}$$

For $i > j$ the derivative is zero. The bracket in the first line lies between zero and $R_q(s) + 2r^2 q(s) \le 4$. The second line also covers $j = 1$ by taking $\tau = 1$ and $q(s) = 1$. Since $\tau^2/\rho_i \le 1$ for $i < j$, summing over the affected terms proves the first bound in (A.8).

Further differentiation introduces factors such as $sq'(s)$, $s^2q''(s)$, $tp'(t)$, and $t^2p''(t)$. They are uniformly bounded because the gate derivatives vanish outside bounded intervals. The remaining prefactors are $\tau^2/(\rho_i\rho_k)$, τ^2/ρ_i^2, $\tau^2 r/\rho_i$, or τ^2, with earlier gate values in the denominators. Each is at most one. Consequently $|\partial^2_{\rho_i\rho_k} E_j| \le c_{\rho\rho} a_j$, which gives the second bound after summation.

For the mixed derivatives, an individual summand has

$$\partial_{x_j} E_j = -\sqrt{a_j}\,\tau r q(s) p'(t), \qquad \partial_{x_{j-1}} E_j = -\frac{a_j}{\sqrt{a_{j-1}}}\,\tau q'(s)\{1 - r^2 v(t)\}.$$

A gate derivative again introduces bounded factors, with $\theta_h/\rho_i \le 1$ whenever ρ_i is a factor of θ_h. Since $a_j \le a_{j-1}$ and each coordinate occurs in at most two summands, $|H_{x_j\rho_i}| \le c_{x\rho}\sqrt{a_j}$. Squaring and summing over j proves the last bound in (A.8). The negative-part quadratics do not depend on $\boldsymbol{\rho}$.

These estimates also justify continuity at zero gate values. If $\tau = \theta_{j-1}$ vanishes, the bounded summand is $O(\tau^2)$ and its derivatives with respect to $\boldsymbol{x}$ are $O(\tau)$. Its gate derivatives vanish as well: every vanishing factor of τ still occurs with a positive exponent in τ^2/ρ_i. If only r vanishes, the term $a_j\tau^2 r^2 q(s) v(t)$ and its first derivatives vanish, leaving $a_j\tau^2\{1 - q(s)\}$ and its derivatives. The factors involving p, q and their derivatives are uniformly bounded, so the same conclusions hold when several gates vanish simultaneously, locally uniformly in $\boldsymbol{x}$. This proves the claimed extension. ■

Proof of Lemma 4.3. For independent gate values, let

$$\mathcal{C}(\boldsymbol{x}; \boldsymbol{\rho}) = H_{\boldsymbol{\omega}, \boldsymbol{\theta}(\boldsymbol{\rho})}(\boldsymbol{x}), \qquad \boldsymbol{\rho}(\boldsymbol{y}) = (r_1(\boldsymbol{y}), \dots, r_M(\boldsymbol{y})).$$

Then $C(\boldsymbol{x}; \boldsymbol{y}) = \mathcal{C}(\boldsymbol{x}; \boldsymbol{\rho}(\boldsymbol{y}))$ and $P(\boldsymbol{x}) = \mathcal{C}(\boldsymbol{x}; \mathbf{1})$. We will use the bounds

$$A_* := \sum_j \omega_j < 2, \qquad B_* := \sum_i b_i^{-1} < 2, \qquad \sum_{k=j}^{M} \omega_k \le 2K\omega_j. \tag{A.10}$$

Primal minima, dual maxima, and the primal function. Lemma A.2 shows that $\mathcal{C}$ is nondecreasing in each gate value. Thus $C(\boldsymbol{x}; \boldsymbol{y}) \le P(\boldsymbol{x})$, and $D \ge 0$ gives $\bar{f}(\boldsymbol{x}; \boldsymbol{y}) \le P(\boldsymbol{x})$. Equality holds when $y_h^{(i)} = \sqrt{b_i}$ for every i, h: all gates then equal one and $D = 0$. For fixed $\boldsymbol{y}$, Lemma A.1 shows that $C(\cdot; \boldsymbol{y})$ attains its minimum value of zero. Therefore

$$\bar{\Phi}(\boldsymbol{x}) = P(\boldsymbol{x}), \qquad \min_{\boldsymbol{x} \in \mathbb{R}^M} \bar{f}(\boldsymbol{x}; \boldsymbol{y}) = -D(\boldsymbol{y}). \tag{A.11}$$

This proves (ii) and the attainment statements in (i).

Primal and dual PŁ conditions. Lemma A.1 with $Q = 2K$ gives

$$C(\boldsymbol{x}; \boldsymbol{y}) \le 130K \|\nabla_{\boldsymbol{x}} C(\boldsymbol{x}; \boldsymbol{y})\|^2. \tag{A.12}$$

By (A.11), $C(\boldsymbol{x}; \boldsymbol{y})$ is the gap between $\bar{f}(\boldsymbol{x}; \boldsymbol{y})$ and its minimum over $\boldsymbol{x}$, so this proves the primal PŁ condition with $\mu_1 = 1/260$.

For the dual gap $\Gamma_{\boldsymbol{x}}(\boldsymbol{y}) := P(\boldsymbol{x}) - \bar{f}(\boldsymbol{x}; \boldsymbol{y})$, integrate $0 \le \mathcal{C}_{\rho_i} \le 4A_*$ from $\boldsymbol{\rho}(\boldsymbol{y})$ to $\mathbf{1}$. Lemma 4.1(iii) yields

$$0 \le P(\boldsymbol{x}) - C(\boldsymbol{x}; \boldsymbol{y}) \le 4A_* \sum_i \{1 - r_i(\boldsymbol{y})\} \le 16 A_* D(\boldsymbol{y}).$$

Moreover, $q' \ge 0$ and $\mathcal{C}_{\rho_i} \ge 0$ imply $\nabla_{\boldsymbol{y}} C \ge 0$, while $\nabla D \le 0$ by Lemma 4.1(i). Hence, for a numerical constant $c_{\mathrm{gap}} > 0$,

$$D(\boldsymbol{y}) \le \Gamma_{\boldsymbol{x}}(\boldsymbol{y}) \le c_{\mathrm{gap}} D(\boldsymbol{y}), \qquad \nabla_{\boldsymbol{y}} \bar{f} \ge -\nabla D \ge 0 \quad \text{coordinatewise.} \tag{A.13}$$

Set $c_{\mathrm{dual}} := 130 c_{\mathrm{gap}}$. The PŁ bound for D gives

$$\Gamma_{\boldsymbol{x}}(\boldsymbol{y}) \le c_{\mathrm{dual}} N \|\nabla D(\boldsymbol{y})\|^2 \le c_{\mathrm{dual}} N \|\nabla_{\boldsymbol{y}} \bar{f}(\boldsymbol{x}; \boldsymbol{y})\|^2. \tag{A.14}$$

This proves the dual PŁ condition with $\mu_2 = 1/(2c_{\mathrm{dual}})$.

Joint smoothness. To handle zero gate values, first replace $\boldsymbol{\rho}(\boldsymbol{y})$ by $\boldsymbol{\rho}^\varepsilon(\boldsymbol{y}) = \varepsilon\mathbf{1} + (1-2\varepsilon)\boldsymbol{\rho}(\boldsymbol{y})$, where $0 < \varepsilon < 1/2$, and set $C_\varepsilon(\boldsymbol{x};\boldsymbol{y}) = \mathcal{C}(\boldsymbol{x};\boldsymbol{\rho}^\varepsilon(\boldsymbol{y}))$. Let $d_i = \partial_{y_N^{(i)}}\rho_i^\varepsilon$ and $e_i = \partial^2_{y_N^{(i)}}\rho_i^\varepsilon$. For numerical constants $c_{q,1}, c_{q,2} > 0$, the gate formulas give $|d_i| \le c_{q,1}/\sqrt{b_i}$ and $|e_i| \le c_{q,2}/b_i$ almost everywhere, uniformly in ε. Only terminal dual coordinates occur in C_ε; on these coordinates, the chain rule gives

$$(C_\varepsilon)_{\boldsymbol{xy}} = [\mathcal{C}_{\boldsymbol{x}\rho_i} d_i]_{i=1}^M, \qquad ((C_\varepsilon)_{\boldsymbol{yy}})_{ik} = \mathcal{C}_{\rho_i\rho_k} d_i d_k + \mathbf{1}_{\{i=k\}}\mathcal{C}_{\rho_i} e_i. \tag{A.15}$$

All derivatives of $\mathcal{C}$ are evaluated at $(\boldsymbol{x};\boldsymbol{\rho}^\varepsilon(\boldsymbol{y}))$. Set $c_{xy} := c_{x\rho}^2 c_{q,1}^2$ and $c_{yy} := c_{\rho\rho}c_{q,1}^2 + 4c_{q,2}$. Lemma A.2 and $(\sum_i b_i^{-2})^{1/2} \le B_*$ imply

$$\|(C_\varepsilon)_{\boldsymbol{xy}}\|_F^2 \le c_{xy}A_*B_*, \qquad \|(C_\varepsilon)_{\boldsymbol{yy}}\|_F \le c_{yy}A_*B_*.$$

The primal Hessian has norm at most 11 by Lemma A.1. Together with (A.10), these estimates bound the full Hessian by a numerical constant independent of K, N, S, ε. The positive-scale formulas are locally $C^{1,1}$, so this gives a uniform global Lipschitz bound for ∇C_ε.

The continuous first-derivative extension in Lemma A.2 gives $C_\varepsilon \to C$ and $\nabla C_\varepsilon \to \nabla C$ locally uniformly as $\varepsilon \downarrow 0$. Passing the uniform Lipschitz bound to the limit proves that C is jointly smooth, including at zero scales. Since D is 11-smooth, $\bar{f} = C - D$ is jointly ℓ_0-smooth for a numerical $\ell_0 \ge 1$. This completes (i). ■

A.5 Proof of Lemma 4.4

Proof of Lemma 4.4. Let $a = \ell\lambda^2/\ell_0$. The coordinate dilations are bijective and $a > 0$, so Lemmas 4.3 and 4.2(i) give attainment of the minimum of $f(\cdot;\boldsymbol{y})$ for every $\boldsymbol{y}$ and the maximum of $f(\boldsymbol{x};\cdot)$ for every $\boldsymbol{x}$, together with

$$\Phi(\boldsymbol{x}) = aP(\boldsymbol{x}/\lambda), \qquad \min_{\boldsymbol{x}\in\mathbb{R}^M}\Phi(\boldsymbol{x}) = 0, \qquad \Phi(\mathbf{0}) = a\Delta_0.$$

The chain rule gives

$$\nabla f(\boldsymbol{x};\boldsymbol{y}) = \frac{a}{\lambda}\nabla\bar{f}(\boldsymbol{x}/\lambda;\boldsymbol{y}/\lambda).$$

Thus the smoothness constant is multiplied by a/λ^2. The primal and dual gaps scale by a, and the corresponding squared partial gradient norms scale by a^2/λ^2. Hence both PŁ constants are multiplied by $a/\lambda^2 = \ell/\ell_0$. This proves (i) and (ii). Positive scalar multiplication preserves the supports of queries and gradients, proving (iii). ■

A.6 Proof of Proposition 4.1

Proof of Proposition 4.1. We first verify Condition 4.1 (C2). Flatten the dual coordinates in block order and suppose $\mathrm{supp}(\boldsymbol{y}) \subseteq [h]$, where $0 \le h < MN$. The weighted-chain formula and $q(0) = q'(0) = 0$ give $\mathrm{supp}(\nabla D(\boldsymbol{y})) \subseteq [h+1]$. The coupling depends on $\boldsymbol{y}$ only through the terminal gates. Their derivatives vanish at zero, and the gate derivatives of $\mathcal{C}$ remain continuous there by Lemma A.2. Thus $\mathrm{supp}(\nabla_{\boldsymbol{y}} C) \subseteq \mathrm{supp}(\boldsymbol{y})$, proving the dual inclusion. For $j > \lfloor h/N \rfloor$, we have $r_j = \theta_j = 0$. The bounded part of C is then independent of x_j, so $\partial_{x_j}\bar{f} = -x_j^-$, which vanishes whenever $x_j = 0$. This proves the primal inclusion.

For a zero-respecting algorithm, let $p_x(t), p_y(t)$ be the lengths of the smallest prefixes containing all primal and dual gradient supports returned before query t, respectively. Both are zero at $t = 0$, and Definition 2.6 places the next query within these prefixes. As long as $p_y(t) < MN$, (C2) gives

$$p_y(t+1) \le p_y(t) + 1, \qquad p_x(t+1) \le \max\{p_x(t), \lfloor p_y(t)/N \rfloor\}.$$

Induction, followed by the trivial dimension bounds once all dual coordinates have been revealed, yields

$$p_y(t) \le t, \qquad p_x(t) \le \lfloor t/N \rfloor. \tag{A.16}$$

Lemma 4.4 preserves these restrictions under scaling.

Fix the constants in Lemma 4.3 and set

$$c_0 := \max\left\{2, \frac{2\ell_0}{\mu_1}, \frac{2\ell_0}{\mu_2}\right\}, \qquad c_1 := \frac{1}{64}, \qquad c_2 := \frac{\mu_1\mu_2}{64\ell_0^2\log 2}.$$

For parameters satisfying the proposition's assumptions, choose

$$K = \left\lfloor \frac{\mu_1 \kappa_{\boldsymbol{x}}}{\ell_0} \right\rfloor, \qquad N = \left\lfloor \frac{\mu_2 \kappa_{\boldsymbol{y}}}{\ell_0} \right\rfloor, \qquad S = 2\left\lfloor \log_2 \frac{\Delta}{4\epsilon} \right\rfloor, \qquad M = KS, \\ \lambda^2 = \frac{\ell_0 \Delta}{2\ell}. \tag{A.17}$$

Then $K, N \geq 2$, and $S \geq 8$ is even. For f defined by (4.11), the scaling lemma and (C4) give

$$\mathrm{Lip}(\nabla f) \leq \ell, \qquad \frac{\ell \mu_1}{\ell_0 K} \geq \mu_{\boldsymbol{x}}, \qquad \frac{\ell \mu_2}{\ell_0 N} \geq \mu_{\boldsymbol{y}}, \qquad \Phi(\mathbf{0}) - \min \Phi = \frac{\Delta}{2}\Delta_0 \leq \Delta.$$

Hence $f \in \mathcal{F}(\ell, \mu_{\boldsymbol{x}}, \mu_{\boldsymbol{y}}, \Delta)$.

Set $T_\star = \lfloor MN/8 \rfloor$. For $0 \leq t \leq T_\star$, (A.16) gives $p_x(t) \leq M/8 \leq (S/2-1)K$. Thus $\mathsf{A}_{\boldsymbol{x}}^{(t)}[f]/\lambda$ is supported on $[(S/2-1)K]$. Using (C3) with $s = S/2$ and $\ell\lambda^2/\ell_0 = \Delta/2$, we obtain

$$\Phi\Big(\mathsf{A}_{\boldsymbol{x}}^{(t)}[f]\Big) - \min_{\boldsymbol{x} \in \mathbb{R}^M} \Phi(\boldsymbol{x}) \geq \frac{\Delta}{2}\, 2^{-S/2} \geq 2\epsilon > \epsilon. \tag{A.18}$$

This holds on the same instance for every zero-respecting algorithm.

Finally, $KNS \geq 32$ gives $T_\star \geq KNS/16$, while

$$K \geq \frac{\mu_1 \kappa_{\boldsymbol{x}}}{2\ell_0}, \qquad N \geq \frac{\mu_2 \kappa_{\boldsymbol{y}}}{2\ell_0}, \qquad S \geq \log_2(\Delta/\epsilon).$$

For the last inequality, put $u = \log_2(\Delta/\epsilon) \geq 6$ and use $2\lfloor u - 2 \rfloor \geq 2u - 6 \geq u$. Consequently,

$$T_\star \geq c_2\, \kappa_{\boldsymbol{x}} \kappa_{\boldsymbol{y}} \log \frac{\Delta}{\epsilon}. \tag{A.19}$$

Together with (A.18), this proves the claimed lower bound on the singleton class $\{f\}$. ■

B Proof of Theorem 3.1

We extend Proposition 4.1 to arbitrary deterministic first-order methods using the finite-horizon rotation argument of Carmon et al. (2020); see also Pan & Li (2026) for its primal-dual formulation.

For $d' \geq d$, let $\mathrm{O}(d', d)$ denote the set of $d' \times d$ matrices with orthonormal columns. Given $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, $\mathbf{U} \in \mathrm{O}(m', m)$, and $\mathbf{V} \in \mathrm{O}(n', n)$, define

$$f_{\mathbf{U},\mathbf{V}}(\boldsymbol{x}; \boldsymbol{y}) := f(\mathbf{U}^\top \boldsymbol{x}; \mathbf{V}^\top \boldsymbol{y}), \qquad \Phi_{\mathbf{U},\mathbf{V}}(\boldsymbol{x}) := \max_{\boldsymbol{y} \in \mathbb{R}^{n'}} f_{\mathbf{U},\mathbf{V}}(\boldsymbol{x}; \boldsymbol{y}).$$

Lemma B.1 (Rotation invariance)**.** If $f \in \mathcal{F}(\ell, \mu_{\boldsymbol{x}}, \mu_{\boldsymbol{y}}, \Delta)$, then $f_{\mathbf{U},\mathbf{V}} \in \mathcal{F}(\ell, \mu_{\boldsymbol{x}}, \mu_{\boldsymbol{y}}, \Delta)$ for every $\mathbf{U}, \mathbf{V}$ as above.

Proof of Lemma B.1. At corresponding points, the chain rule gives

$$\nabla_{\boldsymbol{x}} f_{\mathbf{U},\mathbf{V}} = \mathbf{U} \nabla_{\boldsymbol{x}} f, \qquad \nabla_{\boldsymbol{y}} f_{\mathbf{U},\mathbf{V}} = \mathbf{V} \nabla_{\boldsymbol{y}} f.$$

Since $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns, partial gradient norms are preserved, while $\mathbf{U}^\top$ and $\mathbf{V}^\top$ are contractions. Hence joint ℓ-smoothness is preserved.

The transpose maps are also surjective, so the primal minima and dual maxima agree with the corresponding extrema of f and remain attained. Together with the preservation of partial gradient norms, this preserves both PŁ conditions. Finally,

$$\Phi_{\mathbf{U},\mathbf{V}}(\boldsymbol{x}) = \Phi(\mathbf{U}^\top \boldsymbol{x}), \qquad \Phi^\star_{\mathbf{U},\mathbf{V}} = \Phi^\star,$$

so the initial primal-function gap is unchanged. ■

We use the following finite-horizon resisting-oracle lemma from Pan & Li (2026).

Lemma B.2 (Finite-horizon resisting oracle)**.** For any $T_0 \in \mathbb{N}$ and $\mathsf{A} \in \mathcal{A}_{\mathrm{det}}$, there exists a zero-respecting deterministic algorithm $\mathsf{Z} \in \mathcal{A}_{\mathrm{zr}} \cap \mathcal{A}_{\mathrm{det}}$ such that, for every smooth saddle function $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, one can find $\mathbf{U} \in \mathrm{O}(m + T_0, m)$ and $\mathbf{V} \in \mathrm{O}(n + T_0, n)$ satisfying

$$\big(\mathsf{Z}_{\boldsymbol{x}}^{(t)}[f], \mathsf{Z}_{\boldsymbol{y}}^{(t)}[f]\big) = \big(\mathbf{U}^\top \mathsf{A}_{\boldsymbol{x}}^{(t)}[f_{\mathbf{U},\mathbf{V}}], \mathbf{V}^\top \mathsf{A}_{\boldsymbol{y}}^{(t)}[f_{\mathbf{U},\mathbf{V}}]\big), \qquad t = 0, \dots, T_0 - 1.$$

Proof of Theorem 3.1. Take c_0, c_1, c_2 from Proposition 4.1, and let f and $T_\star$ be the instance and horizon constructed in its proof. For any $\mathsf{A} \in \mathcal{A}_{\mathrm{det}}$, apply Lemma B.2 with $T_0 = T_\star + 1$. It gives $\mathsf{Z} \in \mathcal{A}_{\mathrm{zr}} \cap \mathcal{A}_{\mathrm{det}}$ and isometries $\mathbf{U}, \mathbf{V}$ such that

$$\mathbf{U}^\top \mathsf{A}_{\boldsymbol{x}}^{(t)}[f_{\mathbf{U},\mathbf{V}}] = \mathsf{Z}_{\boldsymbol{x}}^{(t)}[f], \qquad 0 \le t \le T_\star.$$

By Lemma B.1, $f_{\mathbf{U},\mathbf{V}} \in \mathcal{F}(\ell, \mu_{\boldsymbol{x}}, \mu_{\boldsymbol{y}}, \Delta)$. Moreover,

$$\Phi_{\mathbf{U},\mathbf{V}}(\boldsymbol{x}) = \Phi(\mathbf{U}^\top \boldsymbol{x}), \qquad \Phi^\star_{\mathbf{U},\mathbf{V}} = \Phi^\star.$$

Thus (A.18) gives

$$\Phi_{\mathbf{U},\mathbf{V}}\Big(\mathsf{A}_{\boldsymbol{x}}^{(t)}[f_{\mathbf{U},\mathbf{V}}]\Big) - \Phi^\star_{\mathbf{U},\mathbf{V}} = \Phi\Big(\mathsf{Z}_{\boldsymbol{x}}^{(t)}[f]\Big) - \Phi^\star > \epsilon$$

for every $0 \le t \le T_\star$.

The rotated instance may depend on A, as allowed by the order of the infimum and supremum in Definition 2.5. Taking these extrema and using (A.19) proves Theorem 3.1 with the same constants. ■

Proof of Remark 3.1. Append the reported primal point, paired with the dual origin, as the algorithm's next query. An output after at most $T_\star$ oracle calls is then a query with index at most $T_\star$, so the preceding argument rules it out as well.

If the method reports without making any oracle query, first insert a query at the origin and ignore its reply; the reported point can then be appended as the next query, which remains within the finite horizon. ■

C Proofs for Section 5

C.1 Proof of Lemma 5.1

Proof of Lemma 5.1. We use the notation correspondence $p = r$, $\tau_1 = \eta_{\boldsymbol{x}}$, $\tau_2 = \eta_{\boldsymbol{y}}$, $\widehat{f} = F_r$, $\Psi = d_r$, and $P = P_r$. The Lyapunov function of Yang et al. (2022) differs from $\mathcal{V}_r$ only by the additive constant $P_r^\star$.

Under the deterministic specialization, Eq. (36) of Yang et al. (2022) gives

$$\begin{aligned}\mathcal{V}_r^k - \mathcal{V}_r^{k+1} \ge \frac{\eta_{\boldsymbol{x}}}{4}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k)\|^2 + \frac{\eta_{\boldsymbol{y}}}{8}\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}^k)\|^2 \\ + \frac{r}{4\beta}\|\boldsymbol{z}^{k+1} - \boldsymbol{z}^k\|^2 - 24r\beta\|\boldsymbol{x}_r^\star(\boldsymbol{z}^k) - \boldsymbol{x}_r(\boldsymbol{y}_+^k, \boldsymbol{z}^k)\|^2,\end{aligned}$$

where $\boldsymbol{x}_r^\star(\boldsymbol{z}) := \operatorname{argmin}_{\boldsymbol{x}}\{\Phi(\boldsymbol{x}) + \frac{r}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2\}$ and $\boldsymbol{y}_+^k := \boldsymbol{y}^k + \eta_{\boldsymbol{y}} \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}^k)$.

Moreover, Lemma C.2 of Yang et al. (2022) gives

$$\|\boldsymbol{x}_r^\star(\boldsymbol{z}^k) - \boldsymbol{x}_r(\boldsymbol{y}_+^k, \boldsymbol{z}^k)\|^2 \le \frac{1}{(r-\ell)\mu_{\boldsymbol{y}}}\left(1 + \eta_{\boldsymbol{y}}\ell + \frac{\eta_{\boldsymbol{y}}\ell(r+\ell)}{r-\ell}\right)^2 \|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}^k)\|^2.$$

With $r = 2\ell$, $\eta_{\boldsymbol{y}} = 1/(144\ell)$, and $\beta = \eta_{\boldsymbol{y}}\mu_{\boldsymbol{y}}/1600$, this implies

$$24r\beta\|\boldsymbol{x}_r^\star(\boldsymbol{z}^k) - \boldsymbol{x}_r(\boldsymbol{y}_+^k, \boldsymbol{z}^k)\|^2 \le \frac{\eta_{\boldsymbol{y}}}{16}\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}^k)\|^2.$$

Combining the preceding two inequalities yields

$$\mathcal{V}_r^k - \mathcal{V}_r^{k+1} \ge \frac{\eta_{\boldsymbol{x}}}{4}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k)\|^2 + \frac{\eta_{\boldsymbol{y}}}{16}\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}^k)\|^2 + \frac{r}{4\beta}\|\boldsymbol{z}^{k+1} - \boldsymbol{z}^k\|^2.$$

Finally, using $\boldsymbol{z}^{k+1} - \boldsymbol{z}^k = \beta(\boldsymbol{x}^{k+1} - \boldsymbol{z}^k)$ gives Lemma 5.1. ■

C.2 Proof of Lemma 5.2

Recall that

$$\mathcal{V}_r^k = \big(F_r(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k) - d_r(\boldsymbol{y}^k, \boldsymbol{z}^k)\big) + \big(P_r(\boldsymbol{z}^k) - d_r(\boldsymbol{y}^k, \boldsymbol{z}^k)\big) + \big(P_r(\boldsymbol{z}^k) - P_r^\star\big).$$

We bound the three terms above separately.

Proof of Lemma 5.2. Since $f(\cdot, \boldsymbol{y})$ is ℓ-smooth, it is ℓ-weakly convex. Hence, for $r = 2\ell$, $F_r(\cdot, \boldsymbol{y}, \boldsymbol{z})$ is ℓ-strongly convex. Therefore,

$$F_r(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k) - d_r(\boldsymbol{y}^k, \boldsymbol{z}^k) \le \frac{1}{2\ell}\big\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k)\big\|^2.$$

For the second term in $\mathcal{V}_r^k$, by the definitions of P_r and $\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k)$,

$$\begin{aligned} P_r(\boldsymbol{z}^k) - d_r(\boldsymbol{y}^k, \boldsymbol{z}^k) &\le \max_{\boldsymbol{y}} F_r(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}, \boldsymbol{z}^k) - F_r(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}^k, \boldsymbol{z}^k) \\ &= \max_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}) - f(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}^k) \\ &\le \frac{1}{2\mu_{\boldsymbol{y}}}\big\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}^k)\big\|^2, \end{aligned}$$

where the last inequality follows from the dual PŁ condition.

It remains to control $P_r(\boldsymbol{z}^k) - P_r^\star$. Let

$$\boldsymbol{x}_r^\star(\boldsymbol{z}^k) := \arg\min_{\boldsymbol{x}} \left\{\Phi(\boldsymbol{x}) + \frac{r}{2}\|\boldsymbol{x} - \boldsymbol{z}^k\|^2\right\}.$$

Since Φ is ℓ-weakly convex, the above minimizer is unique for $r = 2\ell$. Its optimality condition gives $\nabla\Phi(\boldsymbol{x}_r^\star(\boldsymbol{z}^k)) + r(\boldsymbol{x}_r^\star(\boldsymbol{z}^k) - \boldsymbol{z}^k) = 0$. Using the PŁ condition for Φ (Yang et al., 2020, Lemma A.3),

$$\begin{aligned} P_r(\boldsymbol{z}^k) - P_r^\star &= \Phi(\boldsymbol{x}_r^\star(\boldsymbol{z}^k)) - \Phi^\star + \frac{r}{2}\|\boldsymbol{x}_r^\star(\boldsymbol{z}^k) - \boldsymbol{z}^k\|^2 \\ &\le \frac{1}{2\mu_{\boldsymbol{x}}}\|\nabla\Phi(\boldsymbol{x}_r^\star(\boldsymbol{z}^k))\|^2 + \frac{r}{2}\|\boldsymbol{x}_r^\star(\boldsymbol{z}^k) - \boldsymbol{z}^k\|^2 \\ &= \frac{r(r+\mu_{\boldsymbol{x}})}{2\mu_{\boldsymbol{x}}}\|\boldsymbol{x}_r^\star(\boldsymbol{z}^k) - \boldsymbol{z}^k\|^2 \\ &\le 3\ell\kappa_{\boldsymbol{x}}\|\boldsymbol{x}_r^\star(\boldsymbol{z}^k) - \boldsymbol{z}^k\|^2, \end{aligned}$$

where the last inequality uses $r = 2\ell$ and $\mu_{\boldsymbol{x}} \le \ell$.

We next relate $\boldsymbol{x}_r^\star(\boldsymbol{z}^k)$ to $\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k)$. Since $\Phi(\cdot) + \frac{r}{2}\|\cdot - \boldsymbol{z}^k\|^2$ is ℓ-strongly convex,

$$\begin{aligned} \frac{\ell}{2}\big\|\boldsymbol{x}_r^\star(\boldsymbol{z}^k) - \boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k)\big\|^2 &\le \max_{\boldsymbol{y}} F_r(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}, \boldsymbol{z}^k) - P_r(\boldsymbol{z}^k) \\ &\le \max_{\boldsymbol{y}} F_r(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}, \boldsymbol{z}^k) - F_r(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}^k, \boldsymbol{z}^k) \\ &\le \frac{1}{2\mu_{\boldsymbol{y}}}\big\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}^k)\big\|^2. \end{aligned}$$

Thus,

$$\big\|\boldsymbol{x}_r^\star(\boldsymbol{z}^k) - \boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k)\big\|^2 \le \frac{1}{\ell\mu_{\boldsymbol{y}}}\big\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k), \boldsymbol{y}^k)\big\|^2.$$

The ℓ-strong convexity of $F_r(\cdot, \boldsymbol{y}^k, \boldsymbol{z}^k)$ also gives

$$\|\boldsymbol{x}^k - \boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k)\| \le \frac{1}{\ell}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k)\|.$$

Since $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \eta_{\boldsymbol{x}}\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k)$ and $\eta_{\boldsymbol{x}} = 1/(3\ell)$,

$$\|\boldsymbol{x}^{k+1} - \boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k)\|^2 \le \frac{16}{9\ell^2}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k)\|^2.$$

Consequently,

$$\begin{aligned}\|\boldsymbol{x}_r^\star(\boldsymbol{z}^k)-\boldsymbol{z}^k\|^2 &\le 3\|\boldsymbol{x}_r^\star(\boldsymbol{z}^k)-\boldsymbol{x}_r(\boldsymbol{y}^k,\boldsymbol{z}^k)\|^2+3\|\boldsymbol{x}_r(\boldsymbol{y}^k,\boldsymbol{z}^k)-\boldsymbol{x}^{k+1}\|^2\\ &\quad+3\|\boldsymbol{x}^{k+1}-\boldsymbol{z}^k\|^2\\ &\le \frac{3}{\ell\mu_{\boldsymbol{y}}}\left\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k,\boldsymbol{z}^k),\boldsymbol{y}^k)\right\|^2+\frac{16}{3\ell^2}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k,\boldsymbol{y}^k,\boldsymbol{z}^k)\|^2\\ &\quad+3\|\boldsymbol{x}^{k+1}-\boldsymbol{z}^k\|^2.\end{aligned}$$

Substituting this estimate into the preceding bound on $P_r(\boldsymbol{z}^k)-P_r^\star$ gives

$$\begin{aligned}P_r(\boldsymbol{z}^k)-P_r^\star &\le \frac{16\kappa_{\boldsymbol{x}}}{\ell}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k,\boldsymbol{y}^k,\boldsymbol{z}^k)\|^2+\frac{9\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}}{\ell}\left\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k,\boldsymbol{z}^k),\boldsymbol{y}^k)\right\|^2\\ &\quad+9\ell\kappa_{\boldsymbol{x}}\|\boldsymbol{x}^{k+1}-\boldsymbol{z}^k\|^2.\end{aligned}$$

Combining the bounds for the three terms of $\mathcal{V}_r^k$ and using $\kappa_{\boldsymbol{x}},\kappa_{\boldsymbol{y}}\ge 1$, we obtain

$$\mathcal{V}_r^k \le \frac{17\kappa_{\boldsymbol{x}}}{\ell}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k,\boldsymbol{y}^k,\boldsymbol{z}^k)\|^2+\frac{10\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}}{\ell}\left\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k,\boldsymbol{z}^k),\boldsymbol{y}^k)\right\|^2+9\ell\kappa_{\boldsymbol{x}}\|\boldsymbol{x}^{k+1}-\boldsymbol{z}^k\|^2.$$

Finally, under the parameter choice of Lemma 5.1,

$$\frac{\eta_{\boldsymbol{x}}}{4}=\frac{1}{12\ell},\qquad \frac{\eta_{\boldsymbol{y}}}{16}=\frac{1}{2304\ell},\qquad \frac{r\beta}{4}=\frac{\ell}{460800\kappa_{\boldsymbol{y}}}.$$

Hence the preceding inequality implies, for some numerical constant $c_3>0$,

$$\mathcal{V}_r^k \le c_3\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\left(\frac{\eta_{\boldsymbol{x}}}{4}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k,\boldsymbol{y}^k,\boldsymbol{z}^k)\|^2+\frac{\eta_{\boldsymbol{y}}}{16}\left\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k,\boldsymbol{z}^k),\boldsymbol{y}^k)\right\|^2+\frac{r\beta}{4}\|\boldsymbol{x}^{k+1}-\boldsymbol{z}^k\|^2\right),$$

which finishes the proof. ■

C.3 Initialization

The initialization below is shared by the primal-function gap guarantee in Theorem 3.2 and the game-stationarity guarantee in Corollary 3.1. We first establish a fixed-anchor contraction that will be used to initialize the Lyapunov function.

For fixed $\boldsymbol{z}$, define

$$\mathcal{W}_r(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) := F_r(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})-d_r(\boldsymbol{y},\boldsymbol{z})+P_r(\boldsymbol{z})-d_r(\boldsymbol{y},\boldsymbol{z}).$$

Thus, $\mathcal{V}_r(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})=\mathcal{W}_r(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})+P_r(\boldsymbol{z})-P_r^\star$. We first show that, when the anchor $\boldsymbol{z}$ is fixed, the first two S-GDA updates contract $\mathcal{W}_r$ at rate $1-\Theta(1/\kappa_{\boldsymbol{y}})$.

Lemma C.1 (S-GDA with a fixed proximal variable)**.** Fix $\boldsymbol{z}$ and consider

$$\begin{aligned}\boldsymbol{x}^{t+1} &= \boldsymbol{x}^t-\eta_{\boldsymbol{x}}\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}),\\ \boldsymbol{y}^{t+1} &= \boldsymbol{y}^t+\eta_{\boldsymbol{y}}\nabla_{\boldsymbol{y}} f(\boldsymbol{x}^{t+1},\boldsymbol{y}^t),\end{aligned}$$

with $r=2\ell$, $\eta_{\boldsymbol{x}}=1/(3\ell)$, and $\eta_{\boldsymbol{y}}=1/(144\ell)$. Then, for every $t\ge 0$,

$$\mathcal{W}_r(\boldsymbol{x}^{t+1},\boldsymbol{y}^{t+1},\boldsymbol{z})\le\left(1-\frac{1}{144\kappa_{\boldsymbol{y}}}\right)\mathcal{W}_r(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}).$$

Proof of Lemma C.1. Since $f(\cdot,\boldsymbol{y})$ is ℓ-smooth, it is ℓ-weakly convex. Hence, for $r=2\ell$, $F_r(\cdot,\boldsymbol{y},\boldsymbol{z})$ is ℓ-strongly convex and 3ℓ-smooth. Moreover, by Lemma B.2 of Zhang et al. (2020), we have

$$\|\boldsymbol{x}_r(\boldsymbol{y}_1,\boldsymbol{z})-\boldsymbol{x}_r(\boldsymbol{y}_2,\boldsymbol{z})\|\le\|\boldsymbol{y}_1-\boldsymbol{y}_2\|.$$

By Danskin's theorem, $\nabla_{\boldsymbol{y}} d_r(\boldsymbol{y},\boldsymbol{z})=\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y},\boldsymbol{z}),\boldsymbol{y})$, and therefore $d_r(\cdot,\boldsymbol{z})$ is 2ℓ-smooth (Zhang et al., 2020, Lemma B.3).

The $\boldsymbol{x}$-update and the 3ℓ-smoothness of $F_r(\cdot, \boldsymbol{y}^t, \boldsymbol{z})$ yield

$$F_r(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}) \le F_r(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}) - \frac{\eta_{\boldsymbol{x}}}{2}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z})\|^2.$$

For the $\boldsymbol{y}$-update, the ℓ-smoothness of $f(\boldsymbol{x}^{t+1}, \cdot)$ and the 2ℓ-smoothness of $d_r(\cdot, \boldsymbol{z})$ imply

$$\begin{aligned}&\mathcal{W}_r(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}) - \mathcal{W}_r(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z})\\ \le\ & -\eta_{\boldsymbol{y}}\|\nabla_{\boldsymbol{y}} d_r(\boldsymbol{y}^t, \boldsymbol{z})\|^2 + \eta_{\boldsymbol{y}}\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t) - \nabla_{\boldsymbol{y}} d_r(\boldsymbol{y}^t, \boldsymbol{z})\|^2 + \frac{5\ell}{2}\eta_{\boldsymbol{y}}^2\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t)\|^2.\end{aligned}$$

Since $\nabla_{\boldsymbol{y}} d_r(\boldsymbol{y}^t, \boldsymbol{z}) = \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^t, \boldsymbol{z}), \boldsymbol{y}^t)$,

$$\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t) - \nabla_{\boldsymbol{y}} d_r(\boldsymbol{y}^t, \boldsymbol{z})\| \le \ell\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_r(\boldsymbol{y}^t, \boldsymbol{z})\| \le \frac{4}{3}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z})\|,$$

where the last inequality follows from Lemma B.2 of Zhang et al. (2020) and $\eta_{\boldsymbol{x}} = 1/(3\ell)$. Moreover, by the AM-GM inequality,

$$\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t)\|^2 \le 2\|\nabla_{\boldsymbol{y}} d_r(\boldsymbol{y}^t, \boldsymbol{z})\|^2 + 2\ell^2\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_r(\boldsymbol{y}^t, \boldsymbol{z})\|^2.$$

Using $\eta_{\boldsymbol{y}} = 1/(144\ell)$, we obtain

$$\mathcal{W}_r(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}) - \mathcal{W}_r(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}) \ge \frac{1}{8\ell}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z})\|^2 + \frac{\eta_{\boldsymbol{y}}}{2}\|\nabla_{\boldsymbol{y}} d_r(\boldsymbol{y}^t, \boldsymbol{z})\|^2.$$

Finally, the ℓ-strong convexity of $F_r(\cdot, \boldsymbol{y}^t, \boldsymbol{z})$ and the dual PŁ condition give

$$F_r(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}) - d_r(\boldsymbol{y}^t, \boldsymbol{z}) \le \frac{1}{2\ell}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z})\|^2$$

and

$$P_r(\boldsymbol{z}) - d_r(\boldsymbol{y}^t, \boldsymbol{z}) \le \frac{1}{2\mu_{\boldsymbol{y}}}\|\nabla_{\boldsymbol{y}} d_r(\boldsymbol{y}^t, \boldsymbol{z})\|^2.$$

Since $\eta_{\boldsymbol{y}} = 1/(144\ell)$ and $\kappa_{\boldsymbol{y}} = \ell/\mu_{\boldsymbol{y}}$, comparison of the preceding inequalities gives

$$\mathcal{W}_r(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}) - \mathcal{W}_r(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}) \ge \frac{1}{144\kappa_{\boldsymbol{y}}}\mathcal{W}_r(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}),$$

which proves the claim. ■

Define

$$Q_{\text{in}} := \frac{\|\nabla_{\boldsymbol{x}} f(\boldsymbol{0}, \boldsymbol{0})\|^2}{6\ell} + \frac{\left(\|\nabla_{\boldsymbol{y}} f(\boldsymbol{0}, \boldsymbol{0})\| + \|\nabla_{\boldsymbol{x}} f(\boldsymbol{0}, \boldsymbol{0})\|/3\right)^2}{2\mu_{\boldsymbol{y}}}.$$

Since $F_r(\cdot, \boldsymbol{0}, \boldsymbol{0})$ is ℓ-strongly convex,

$$F_r(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0}) - d_r(\boldsymbol{0}, \boldsymbol{0}) \le \frac{\|\nabla_{\boldsymbol{x}} f(\boldsymbol{0}, \boldsymbol{0})\|^2}{2\ell}.$$

Moreover, $\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}_r(\boldsymbol{0}, \boldsymbol{0}), \boldsymbol{0}, \boldsymbol{0}) = 0$ and strong monotonicity give $\|\boldsymbol{x}_r(\boldsymbol{0}, \boldsymbol{0})\| \le \|\nabla_{\boldsymbol{x}} f(\boldsymbol{0}, \boldsymbol{0})\|/\ell$. The dual PŁ condition therefore yields

$$P_r(\boldsymbol{0}) - d_r(\boldsymbol{0}, \boldsymbol{0}) \le \frac{\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{0}, \boldsymbol{0}), \boldsymbol{0})\|^2}{2\mu_{\boldsymbol{y}}},$$

while joint smoothness gives

$$\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{0}, \boldsymbol{0}), \boldsymbol{0})\| \le \|\nabla_{\boldsymbol{y}} f(\boldsymbol{0}, \boldsymbol{0})\| + \|\nabla_{\boldsymbol{x}} f(\boldsymbol{0}, \boldsymbol{0})\|.$$

Consequently, $\mathcal{W}_r(\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0}) \le 9Q_{\text{in}}$.

Set $\boldsymbol{z}^0 = \boldsymbol{0}$ and apply the fixed-anchor updates in Lemma C.1, initialized at $(\boldsymbol{0}, \boldsymbol{0})$. After

$$\mathcal{O}\left(\kappa_{\boldsymbol{y}} \log\left(1 + \frac{Q_{\text{in}}}{\Delta}\right)\right)$$

oracle calls, we obtain a pair, relabeled as $(\boldsymbol{x}^0, \boldsymbol{y}^0)$, satisfying $\mathcal{W}_r(\boldsymbol{x}^0, \boldsymbol{y}^0, \boldsymbol{z}^0) \le \Delta$. Since $P_r(\boldsymbol{z}^0) - P_r^\star \le \Phi(\boldsymbol{z}^0) - \Phi^\star \le \Delta$, we have

$$\mathcal{V}_r^0 \le 2\Delta. \tag{C.1}$$

D PROOF OF THEOREM 3.2

We now convert the Lyapunov contraction into a guarantee for the primal-function gap.

Lemma D.1 (Primal-function gap bound)**.** For every $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$,

$$\Phi(\boldsymbol{x}) - \Phi^\star \le P_r(\boldsymbol{z}) - P_r^\star + 4\kappa_{\boldsymbol{y}} \mathcal{W}_r(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}).$$

Proof of Lemma D.1. For fixed $\boldsymbol{z}$, $d_r(\cdot, \boldsymbol{z})$ is 2ℓ-smooth and $\nabla_{\boldsymbol{y}} d_r(\boldsymbol{y}, \boldsymbol{z}) = \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}, \boldsymbol{z}), \boldsymbol{y})$. Since $d_r(\boldsymbol{y}', \boldsymbol{z}) \le P_r(\boldsymbol{z})$ for every $\boldsymbol{y}'$, smoothness gives

$$\|\nabla_{\boldsymbol{y}} d_r(\boldsymbol{y}, \boldsymbol{z})\|^2 \le 4\ell\big(P_r(\boldsymbol{z}) - d_r(\boldsymbol{y}, \boldsymbol{z})\big).$$

The ℓ-strong convexity of $F_r(\cdot, \boldsymbol{y}, \boldsymbol{z})$ gives

$$\|\boldsymbol{x} - \boldsymbol{x}_r(\boldsymbol{y}, \boldsymbol{z})\|^2 \le \frac{2}{\ell}\big(F_r(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) - d_r(\boldsymbol{y}, \boldsymbol{z})\big).$$

Hence,

$$\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})\|^2 \le 8\ell\big(P_r(\boldsymbol{z}) - d_r(\boldsymbol{y}, \boldsymbol{z})\big) + 4\ell\big(F_r(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) - d_r(\boldsymbol{y}, \boldsymbol{z})\big).$$

By the dual PŁ condition,

$$\max_{\boldsymbol{y}'} f(\boldsymbol{x}, \boldsymbol{y}') - f(\boldsymbol{x}, \boldsymbol{y}) \le 4\kappa_{\boldsymbol{y}}\big(P_r(\boldsymbol{z}) - d_r(\boldsymbol{y}, \boldsymbol{z})\big) + 2\kappa_{\boldsymbol{y}}\big(F_r(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) - d_r(\boldsymbol{y}, \boldsymbol{z})\big).$$

Using

$$F_r(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) - P_r(\boldsymbol{z}) = \big(F_r(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) - d_r(\boldsymbol{y}, \boldsymbol{z})\big) - \big(P_r(\boldsymbol{z}) - d_r(\boldsymbol{y}, \boldsymbol{z})\big)$$

and $\kappa_{\boldsymbol{y}} \ge 1$ gives

$$\max_{\boldsymbol{y}'} F_r(\boldsymbol{x}, \boldsymbol{y}', \boldsymbol{z}) - P_r(\boldsymbol{z}) \le 4\kappa_{\boldsymbol{y}} \mathcal{W}_r(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}).$$

Finally, $\Phi(\boldsymbol{x}) \le \max_{\boldsymbol{y}'} F_r(\boldsymbol{x}, \boldsymbol{y}', \boldsymbol{z})$ and $P_r^\star = \Phi^\star$. ■

Corollary D.1 (Output guarantee)**.** Suppose that the main S-GDA phase produces $\mathcal{V}_r^K \le \epsilon/2$. Hold $\boldsymbol{z}^K$ fixed and apply the first two S-GDA updates starting from $(\boldsymbol{x}^K, \boldsymbol{y}^K)$. After $\mathcal{O}(\kappa_{\boldsymbol{y}} \log(1 + \kappa_{\boldsymbol{y}}))$ additional oracle calls, the resulting point $\boldsymbol{x}_{\mathrm{out}}$ satisfies $\Phi(\boldsymbol{x}_{\mathrm{out}}) - \Phi^\star \le \epsilon$.

Proof of Corollary D.1. Since $\mathcal{V}_r^K \le \epsilon/2$, $P_r(\boldsymbol{z}^K) - P_r^\star \le \epsilon/2$ and $\mathcal{W}_r(\boldsymbol{x}^K, \boldsymbol{y}^K, \boldsymbol{z}^K) \le \epsilon/2$. Lemma C.1 therefore implies that after $\mathcal{O}(\kappa_{\boldsymbol{y}} \log(1 + \kappa_{\boldsymbol{y}}))$ additional calls,

$$\mathcal{W}_r(\boldsymbol{x}_{\mathrm{out}}, \boldsymbol{y}_{\mathrm{out}}, \boldsymbol{z}^K) \le \frac{\epsilon}{8\kappa_{\boldsymbol{y}}}.$$

Lemma D.1 then gives

$$\Phi(\boldsymbol{x}_{\mathrm{out}}) - \Phi^\star \le \frac{\epsilon}{2} + 4\kappa_{\boldsymbol{y}} \frac{\epsilon}{8\kappa_{\boldsymbol{y}}} = \epsilon.$$

■

Proof of Theorem 3.2. By (C.1), the initialization yields $\mathcal{V}_r^0 \le 2\Delta$. Corollary 5.1 therefore gives $\mathcal{V}_r^K \le \epsilon/2$ after

$$\mathcal{O}\left(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\left[1 + \log\left(\frac{\Delta}{\epsilon}\right)\right]\right)$$

main S-GDA iterations. Corollary D.1 then requires $\mathcal{O}(\kappa_{\boldsymbol{y}} \log(1 + \kappa_{\boldsymbol{y}}))$ additional calls. Including the common initialization cost, the complete oracle complexity is

$$\mathcal{O}\left(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\left[1 + \log\left(\frac{\Delta}{\epsilon}\right)\right] + \kappa_{\boldsymbol{y}} \log\left(\left(1 + \frac{Q_{\mathrm{in}}}{\Delta}\right)(1 + \kappa_{\boldsymbol{y}})\right)\right).$$

In particular, suppose that

$$\log\left(\frac{\Delta}{\epsilon}\right) \ge 1, \qquad \text{and} \qquad \log\left(\left(1 + \frac{Q_{\mathrm{in}}}{\Delta}\right)(1 + \kappa_{\boldsymbol{y}})\right) \le \kappa_{\boldsymbol{x}} \log\left(\frac{\Delta}{\epsilon}\right).$$

Then both the initialization and output costs are absorbed by the main accuracy-dependent term, and the complete oracle complexity becomes

$$\mathcal{O}\left(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\log\left(\frac{\Delta}{\epsilon}\right)\right).$$

This proves the theorem.

■

We next verify that the lower-bound instances satisfy the high-accuracy condition above. For the scaled instance used in the proof of Proposition 4.1, the gate identities $p'(0) = 3/4$ and $q(0) = q'(0) = 0$ give

$$\nabla_{\boldsymbol{x}} f(\mathbf{0};\mathbf{0}) = \mathbf{0}, \qquad \frac{Q_{\mathrm{in}}}{\Delta} = \frac{9\kappa_{\boldsymbol{y}}}{64\ell_0}\,2^{M-1}.$$

Using $\ell_0 \geq 1$ and the choices of K and S, we obtain

$$\begin{aligned}\log\left[\left(1+\frac{Q_{\mathrm{in}}}{\Delta}\right)(1+\kappa_{\boldsymbol{y}})\right] &\leq M\log 2 + 2\log(1+\kappa_{\boldsymbol{y}}) \\ &\leq \frac{2\mu_1}{\ell_0}\,\kappa_{\boldsymbol{x}}\log\left(\frac{\Delta}{\epsilon}\right) + 2\log(1+\kappa_{\boldsymbol{y}}).\end{aligned}$$

Consequently, whenever $\log(1+\kappa_{\boldsymbol{y}}) \leq \kappa_{\boldsymbol{x}}\log(\Delta/\epsilon)$, the initialization cost is $\mathcal{O}(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\log(\Delta/\epsilon))$. This condition holds at sufficiently high accuracy for any fixed condition numbers. The isometries used to obtain the deterministic lower bound preserve both partial gradient norms at the origin, and hence preserve Q_{in}. Thus, the lower-bound instances remain in the regime where the complete upper bound has the claimed order.

For the GS lower bound, the same calculation applies with the primal accuracy replaced by $\delta = (\kappa_{\boldsymbol{x}} + \kappa_{\boldsymbol{y}})\epsilon^2/(2\ell)$, as in the proof of Corollary 3.1.

E Proof of Corollary 3.1

The lower bound follows from Remark 2.1. For the upper bound, we use the descent accumulated over a short window to select a queried point with controlled error, and then restart S-GDA with the smaller of the two PŁ constants on the minimization side.

Lemma E.1 (Window selection)**.** Under the parameter choice of Lemma 5.1, define

$$E_k := \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^k,\boldsymbol{y}^k)\|^2 + \kappa_{\boldsymbol{y}}\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}^k,\boldsymbol{y}^k)\|^2.$$

For every $t \in \mathbb{N}_0$ and $H \in \mathbb{N}$, let

$$j \in \arg\min_{t\leq k<t+H} E_k.$$

Then there exists a numerical constant $c_5 > 0$ such that

$$\Phi(\boldsymbol{x}^j) - \inf_{\boldsymbol{u}\in\mathbb{R}^m} f(\boldsymbol{u},\boldsymbol{y}^j) \leq \frac{c_5\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}}{H}\mathcal{V}_r^t, \tag{E.1}$$

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^j,\boldsymbol{y}^j)\|^2 + \|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}^j,\boldsymbol{y}^j)\|^2 \leq \frac{c_5\ell\kappa_{\boldsymbol{y}}}{H}\mathcal{V}_r^t. \tag{E.2}$$

In particular, the index j can be selected using only gradients returned by the first-order oracle.

Proof of Lemma E.1. Let

$$\mathcal{D}_k := \frac{\eta_{\boldsymbol{x}}}{4}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k,\boldsymbol{y}^k,\boldsymbol{z}^k)\|^2 + \frac{\eta_{\boldsymbol{y}}}{16}\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}_r(\boldsymbol{y}^k,\boldsymbol{z}^k),\boldsymbol{y}^k)\|^2 + \frac{r\beta}{4}\|\boldsymbol{x}^{k+1}-\boldsymbol{z}^k\|^2.$$

Lemma 5.1 gives $\mathcal{D}_k \leq \mathcal{V}_r^k - \mathcal{V}_r^{k+1}$. Summing over the window and using the nonnegativity of $\mathcal{V}_r^{t+H}$ gives

$$\sum_{k=t}^{t+H-1}\mathcal{D}_k \leq \mathcal{V}_r^t.$$

Since $F_r(\cdot, \boldsymbol{y}^k, \boldsymbol{z}^k)$ is ℓ-strongly convex,

$$\|\boldsymbol{x}^k - \boldsymbol{x}_r(\boldsymbol{y}^k, \boldsymbol{z}^k)\| \le \frac{1}{\ell}\|\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k)\|.$$

Joint smoothness and the parameter choices therefore imply

$$\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}^k, \boldsymbol{y}^k)\|^2 \le c_{y,\mathrm{grad}}\ell\,\mathcal{D}_k.$$

Here $c_{y,\mathrm{grad}} > 0$ is a numerical constant.

Moreover, the $\boldsymbol{x}$-update gives $\boldsymbol{x}^k - \boldsymbol{z}^k = \boldsymbol{x}^{k+1} - \boldsymbol{z}^k + \eta_{\boldsymbol{x}}\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k)$, and hence

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^k, \boldsymbol{y}^k) = (1 - r\eta_{\boldsymbol{x}})\nabla_{\boldsymbol{x}} F_r(\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{z}^k) - r(\boldsymbol{x}^{k+1} - \boldsymbol{z}^k).$$

Using $r = 2\ell$, $\eta_{\boldsymbol{x}} = 1/(3\ell)$, $\eta_{\boldsymbol{y}} = 1/(144\ell)$, and $\beta = \eta_{\boldsymbol{y}}\mu_{\boldsymbol{y}}/1600$ gives

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^k, \boldsymbol{y}^k)\|^2 \le c_{x,\mathrm{grad}}\ell\kappa_{\boldsymbol{y}}\mathcal{D}_k.$$

Here $c_{x,\mathrm{grad}} > 0$ is another numerical constant.

Set $c_5 := c_{x,\mathrm{grad}} + c_{y,\mathrm{grad}}$. Consequently, $E_k \le c_5\ell\kappa_{\boldsymbol{y}}\mathcal{D}_k$. Since j minimizes E_k over the window,

$$E_j \le \frac{1}{H}\sum_{k=t}^{t+H-1} E_k \le \frac{c_5\ell\kappa_{\boldsymbol{y}}}{H}\sum_{k=t}^{t+H-1}\mathcal{D}_k \le \frac{c_5\ell\kappa_{\boldsymbol{y}}}{H}\mathcal{V}_r^t.$$

Since $\kappa_{\boldsymbol{y}} \ge 1$,

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^j, \boldsymbol{y}^j)\|^2 + \|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}^j, \boldsymbol{y}^j)\|^2 \le E_j,$$

which proves (E.2).

Moreover, by the two-sided PŁ condition,

$$\begin{aligned}\Phi(\boldsymbol{x}^j) - \inf_{\boldsymbol{u}} f(\boldsymbol{u}, \boldsymbol{y}^j) &\le \frac{\kappa_{\boldsymbol{x}}}{2\ell}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^j, \boldsymbol{y}^j)\|^2 + \frac{\kappa_{\boldsymbol{y}}}{2\ell}\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}^j, \boldsymbol{y}^j)\|^2 \\ &\le \frac{\kappa_{\boldsymbol{x}}}{2\ell}E_j \le \frac{c_5\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}}{H}\mathcal{V}_r^t,\end{aligned}$$

where we used $\kappa_{\boldsymbol{x}} \ge 1$ in the second inequality. This proves (E.1). ■

Proof of Corollary 3.1. For the lower bound, let

$$\delta := \frac{\epsilon^2}{2}\left(\frac{1}{\mu_{\boldsymbol{x}}} + \frac{1}{\mu_{\boldsymbol{y}}}\right) = \frac{(\kappa_{\boldsymbol{x}} + \kappa_{\boldsymbol{y}})\epsilon^2}{2\ell}.$$

By Remark 2.1, every ϵ-GS point is δ-stationary in the sense of Definition 2.2. Hence, whenever δ lies in the accuracy range of Theorem 3.1, that theorem gives

$$\Omega\left(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\log\frac{2\ell\Delta}{(\kappa_{\boldsymbol{x}} + \kappa_{\boldsymbol{y}})\epsilon^2}\right)$$

oracle queries.

We next prove the upper bound. Start from the common initialization (C.1), so that $\mathcal{V}_r^0 \le 2\Delta$. Apply Lemma E.1 over $H = \lceil\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\rceil$ iterations and select the queried iterate minimizing the observable quantity E_k over this window. It yields a queried pair $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})$ satisfying

$$\Phi(\widehat{\boldsymbol{x}}) - \inf_{\boldsymbol{u}} f(\boldsymbol{u}, \widehat{\boldsymbol{y}}) \le 2c_5\Delta.$$

Set $\alpha := \min\{\mu_{\boldsymbol{x}}, \mu_{\boldsymbol{y}}\}$. If $\mu_{\boldsymbol{x}} \le \mu_{\boldsymbol{y}}$, keep the original problem. Otherwise, consider

$$\widehat{f}(\boldsymbol{u}, \boldsymbol{v}) := -f(\boldsymbol{v}, \boldsymbol{u}),$$

which interchanges the primal and dual PŁ constants. In either case, denote the resulting problem by $\widehat{f}$, and use hats for its associated quantities. The minimization-side PŁ constant is then α, while the product of the two condition numbers remains $\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}$. The selected pair is relabeled as $(\widehat{\boldsymbol{x}}^0, \widehat{\boldsymbol{y}}^0)$, and we set $\widehat{\boldsymbol{z}}^0 = \widehat{\boldsymbol{x}}^0$.

The primal-dual gap of the selected pair is unchanged by the possible interchange of variables. Since $\widehat{\boldsymbol{z}}^0 = \widehat{\boldsymbol{x}}^0$, the nonnegativity of the regularization gives $\widehat{d}_r(\widehat{\boldsymbol{y}}^0, \widehat{\boldsymbol{z}}^0) \geq \inf_{\boldsymbol{u}} \widehat{f}(\boldsymbol{u}, \widehat{\boldsymbol{y}}^0)$, while $\widehat{P}_r(\widehat{\boldsymbol{z}}^0) \leq \widehat{\Phi}(\widehat{\boldsymbol{x}}^0)$ and $\widehat{P}_r^\star = \widehat{\Phi}^\star \geq \inf_{\boldsymbol{u}} \widehat{f}(\boldsymbol{u}, \widehat{\boldsymbol{y}}^0)$. Consequently, we have

$$\begin{aligned}
\widehat{F}_r(\widehat{\boldsymbol{x}}^0, \widehat{\boldsymbol{y}}^0, \widehat{\boldsymbol{z}}^0) - \widehat{d}_r(\widehat{\boldsymbol{y}}^0, \widehat{\boldsymbol{z}}^0) &\leq \widehat{\Phi}(\widehat{\boldsymbol{x}}^0) - \inf_{\boldsymbol{u}} \widehat{f}(\boldsymbol{u}, \widehat{\boldsymbol{y}}^0), \\
\widehat{P}_r(\widehat{\boldsymbol{z}}^0) - \widehat{d}_r(\widehat{\boldsymbol{y}}^0, \widehat{\boldsymbol{z}}^0) &\leq \widehat{\Phi}(\widehat{\boldsymbol{x}}^0) - \inf_{\boldsymbol{u}} \widehat{f}(\boldsymbol{u}, \widehat{\boldsymbol{y}}^0), \\
\widehat{P}_r(\widehat{\boldsymbol{z}}^0) - \widehat{P}_r^\star &\leq \widehat{\Phi}(\widehat{\boldsymbol{x}}^0) - \inf_{\boldsymbol{u}} \widehat{f}(\boldsymbol{u}, \widehat{\boldsymbol{y}}^0).
\end{aligned}$$

Therefore, we have $\widehat{\mathcal{V}}_r^0 \leq 6c_5\Delta$.

Run S-GDA on $\widehat{f}$ from this restarted state. Corollary 5.1 gives

$$\widehat{\mathcal{V}}_r^t \leq 6c_5\Delta \exp\left(-\frac{c_4 t}{\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}}\right).$$

Let $\widehat{\kappa}_{\boldsymbol{y}}$ denote the dual condition number after the possible interchange. Applying Lemma E.1 over another window of length $H = \lceil \kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}} \rceil$ and selecting the iterate minimizing the corresponding observable quantity E_k yields a queried pair satisfying

$$\|\nabla_{\boldsymbol{x}} \widehat{f}\|^2 + \|\nabla_{\boldsymbol{y}} \widehat{f}\|^2 \leq c_5 \, \frac{\ell \widehat{\kappa}_{\boldsymbol{y}}}{\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}} \widehat{\mathcal{V}}_r^t = c_5 \alpha \, \widehat{\mathcal{V}}_r^t.$$

Thus an ϵ-GS point is obtained after

$$\mathcal{O}\left(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\left[1 + \log\left(1 + \frac{\alpha\Delta}{\epsilon^2}\right)\right]\right)$$

oracle calls after the common initialization. Here, we have $\alpha = \frac{\ell}{\max\{\kappa_{\boldsymbol{x}}, \kappa_{\boldsymbol{y}}\}}$, $\alpha \leq \frac{2\ell}{\kappa_{\boldsymbol{x}} + \kappa_{\boldsymbol{y}}} \leq 2\alpha$. Hence the upper bound can equivalently be written as

$$\mathcal{O}\left(\kappa_{\boldsymbol{x}}\kappa_{\boldsymbol{y}}\left[1 + \log\left(1 + \frac{2\ell\Delta}{(\kappa_{\boldsymbol{x}} + \kappa_{\boldsymbol{y}})\epsilon^2}\right)\right]\right),$$

in addition to the common initialization cost. In the high-accuracy regime, this matches the lower bound above up to numerical constants. ■