ZOFOSGDA: A HYBRID METHOD FOR NONSMOOTH NONCONVEX-NONCONCAVE MINIMAX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Minimax optimization arises in many machine-learning problems, including robust learning, adversarial training, and AUC maximization. Some of these formulations are nonsmooth and may exhibit nonconvex–nonconcave geometry.However, most existing optimization methods are designed primarily for smooth objectives, and their convergence guarantees typically apply only to weakly convex–(strongly) concave or nonconvex–(strongly) concave settings. To address this gap, we study stochastic noncomposite minimax optimization with a nonsmooth, nonconvex outer variable and a smooth, possibly nonconcave inner variable that satisfies a uniform projected Polyak-Łojasiewicz condition. We propose ZOFOSGDA, a hybrid method that combines randomized two-point zeroth-order estimates for the outer variable, stochastic first-order gradients for the inner variable, and SPIDER variance reduction. Our stationarity criterion pairs a projected Goldstein residual in the outer block with a projected-gradient residual in the inner block. Under a uniformly bounded initial potential gap and the canonical smoothing radius $\delta = c_\delta \epsilon = \Theta(\epsilon)$, ZOFOSGDA returns both an expected blockwise generalized Goldstein stationary point and an expected projected Goldstein stationary point of the value function. Its oracle costs are $\widetilde{\mathcal{O}}(d_1^{3/2}\epsilon^{-4})$ stochastic function queries and $\widetilde{\mathcal{O}}(d_1^{1/2}\epsilon^{-4})$ stochastic inner-gradient queries, where d_1 is the outer dimension. An outer-gradient tracking bound enables the value-function guarantee. A controlled classification benchmark verifies the nonconcave projected-PŁ geometry, while a supplementary AUC-maximization task provides an additional finite-sample stress test.

1 INTRODUCTION

Minimax optimization is a standard framework for modeling competing objectives in modern machine learning. Representative applications include robust classification (Fauß et al., 2021; Wang et al., 2023b; Mazuelas et al., 2023), adversarial training (Goodfellow et al., 2015), AUC maximization (Yang et al., 2022), federated learning (Wu et al., 2023; Sharma et al., 2022; Sun & Wei, 2022), and few-shot meta-learning (Wang et al., 2023a; Khodak et al., 2023; Yang et al., 2024). These examples motivate finite-time analyses under explicit structural assumptions.

For smooth minimax objectives, gradient descent–ascent (GDA) and its variants provide the standard algorithmic framework. Existing methods include multi-step and alternating updates (Nouiehed et al., 2019; Lin et al., 2020), accelerated and adaptive schemes (Zhao et al., 2025; Yan et al., 2020; Yang et al., 2020b; Huang et al., 2023), and stochastic variance-reduced algorithms (Rafique et al., 2022; Luo et al., 2020). Zeroth-order counterparts have also been developed for smooth minimax problems when gradient access is restricted (Huang et al., 2022; Wang et al., 2023c; Xu et al., 2024). However, these approaches rely on differentiability or Lipschitz smoothness in the updated variables and therefore do not directly apply when the outer objective is intrinsically nonsmooth.

Recent work has begun to address noncomposite nonsmooth minimax problems. Lin et al. (2025) study weakly convex–(strongly) concave objectives using a first-order two-timescale GDA method, denoted by TTSGDA. Their guarantees concern stationarity of a Moreau envelope of the value function and require weak convexity in the first variable together with concavity or strong concavity in the second variable. Shi & Luo (2026) subsequently allow the first variable to be nonconvex and nonsmooth. Their PGFDA method uses gradient-free oracle information and targets a full-space

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.

Table 1: Representative nonsmooth minimax methods and their expected guarantees. WC/NC and C/SC/PŁ denote outer weak convexity/nonconvexity and inner concavity/strong concavity/projected-PŁ; SP, GSSP, and GGSP denote the three stationarity criteria. Here ρ is the TTSGDA weak-convexity constant. Complexities use the cited works' notation and are not direct dominance comparisons; $\widetilde{\mathcal{O}}$ also hides logarithmic factors. For our method, $0 < \delta \le c_\delta \epsilon$; setting $\delta = c_\delta \epsilon$ gives SZO/SFO costs $\widetilde{\mathcal{O}}(d_1^{3/2}\epsilon^{-4})$ and $\widetilde{\mathcal{O}}(d_1^{1/2}\epsilon^{-4})$.

Method	Ref.	Oracle	Objective	Stationarity criterion	Oracle complexity
TTSGDA	Lin et al. (2025)	1st	WC-SC	SP of $\Phi_{1/(2\rho)}$	$\widetilde{\mathcal{O}}(\epsilon^{-6})$
TTSGDA	Lin et al. (2025)	1st	WC-C	SP of $\Phi_{1/(2\rho)}$	$\mathcal{O}(\epsilon^{-8})$
PGFDA	Shi & Luo (2026)	0th	NC-SC	GSSP of f	$\mathcal{O}((d_1+d_2)^3\delta^{-4}\epsilon^{-3})$
PGFDA	Shi & Luo (2026)	0th	NC-C	GSSP of f	$\mathcal{O}((d_1+d_2)^3\delta^{-4}\epsilon^{-6})$
NL-PGFDA	Shi & Luo (2026)	0th	NC-SC	GGSP of Φ	$\mathcal{O}(d_1^{3.5}d_2\delta^{-3}\epsilon^{-6})$
NL-PGFDA	Shi & Luo (2026)	0th	NC-C	GGSP of Φ	$\mathcal{O}(d_1^{5.5}d_2\delta^{-5}\epsilon^{-8})$
ZOFOSGDA	Ours	0th+1st	NC-PŁ	blockwise GGSP of f projected GGSP of Φ	SZO : $\widetilde{\mathcal{O}}(d_1^{3/2}\delta^{-1}\epsilon^{-3})$ SFO : $\widetilde{\mathcal{O}}(d_1^{1/2}\delta^{-1}\epsilon^{-3})$ regime : $0 < \delta \le c_\delta\epsilon$

Goldstein saddle stationary point of the objective function, while NL-PGFDA targets generalized Goldstein stationarity of the value function. Their finite-time guarantees primarily rely on concavity or strong concavity in the inner variable. Table 1 summarizes these differences in objective geometry, oracle access, stationarity criterion, and oracle complexity. Consequently, none of the methods in the table directly accommodates the combination of a nonsmooth nonconvex outer objective and a nonconcave inner objective considered here.

To bridge this gap, we study stochastic noncomposite minimax problems in which the outer variable is nonsmooth and nonconvex, while the inner variable is smooth and may be nonconcave. We impose a uniform projected-PŁ condition (Li & Li, 2018), which controls the inner optimality gap through the projected-gradient mapping without requiring inner concavity. This asymmetric geometry motivates a hybrid oracle model: randomized two-point function evaluations are used for the outer variable, while stochastic first-order gradients are retained for the inner variable. We develop ZOFOSGDA, which applies SPIDER variance reduction to both estimators. The main theorem directly controls the smoothed outer-gradient mapping and the inner projected-gradient mapping. These bounds imply blockwise GGSP, while an outer-gradient tracking argument additionally yields projected GGSP of the value function Φ. We also illustrate finite-sample behavior on a controlled nonconvex-PŁ classification objective and include a supplementary AUC-maximization stress test for which projected-PŁ is not verified. Our contributions are summarized as follows.

1. We develop ZOFOSGDA for nonsmooth nonconvex minimax problems with a smooth, potentially nonconcave inner objective satisfying a uniform projected-PŁ condition. Its analysis establishes an outer-gradient tracking bound from $\nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y})$ to $\nabla \Phi_\delta(\mathbf{x})$.

2. Under a uniformly bounded initial potential gap, the canonical choice $\delta = c_\delta\epsilon = \Theta(\epsilon)$, for sufficiently small problem-dependent $c_\delta > 0$, yields an expected blockwise GGSP with $\widetilde{\mathcal{O}}(d_1^{3/2}\epsilon^{-4})$ stochastic zeroth-order queries and $\widetilde{\mathcal{O}}(d_1^{1/2}\epsilon^{-4})$ stochastic y-gradient queries. More generally, for $0 < \delta \le c_\delta\epsilon$, the radius-explicit costs are $\widetilde{\mathcal{O}}(d_1^{3/2}\delta^{-1}\epsilon^{-3})$ and $\widetilde{\mathcal{O}}(d_1^{1/2}\delta^{-1}\epsilon^{-3})$, respectively. The same oracle orders yield an expected projected GGSP of the value function. We also derive simultaneous high-probability guarantees for both criteria with explicit confidence dependence.

2 Related Work

We review the two lines of work most closely related to our results: Goldstein-stationarity methods for nonsmooth nonconvex optimization and algorithms for nonconvex minimax optimization.

2.1 NONSMOOTH OPTIMIZATION

For general Lipschitz nonconvex objectives, finding an approximate Clarke stationary point is computationally intractable in the local-oracle model studied by Zhang et al. (2020), which motivates Goldstein stationarity as a finite-time target. Zhang et al. (2020) established nonasymptotic guarantees for an interpolated normalized-gradient method under a specialized subgradient oracle. Random-perturbation schemes subsequently obtained guarantees using ordinary subgradient access (Davis et al., 2022; Tian et al., 2022). Randomized smoothing provides another implementable route to Goldstein stationarity using only function values. Lin et al. (2022) give a gradient-free method that finds a (δ, ϵ)-GGSP with $\mathcal{O}(d^{3/2}\delta^{-1}\epsilon^{-4})$ function queries, and Chen et al. (2023) improve the dependence on accuracy to $\mathcal{O}(d^{3/2}\delta^{-1}\epsilon^{-3})$ through variance reduction. Further refinements use online-to-nonconvex reductions (Cutkosky et al., 2023; Kornowski & Shamir, 2024), while constrained and compositional extensions were developed by Liu et al. (2024a) and Liu et al. (2024b), respectively. These works provide key tools for our treatment of the nonsmooth variable, but they concern single-objective minimization and do not address the coupled descent–ascent dynamics or inner optimality gap of a minimax problem.

2.2 MINIMAX OPTIMIZATION

For smooth minimax objectives, a large body of work analyzes first-order GDA and its multi-step, accelerated, and stochastic variants (Nouiehed et al., 2019; Lin et al., 2020; Zhao et al., 2025; Yan et al., 2020; Yang et al., 2020b). Variance reduction further improves stochastic oracle complexity (Rafique et al., 2022; Luo et al., 2020; Huang et al., 2023), and zeroth-order counterparts have been developed for settings without gradient access Huang et al. (2022); Wang et al. (2023c); Xu et al. (2024). Smooth nonconvex–nonconcave problems have also been studied under PŁ-type inner geometry (Nouiehed et al., 2019; Yang et al., 2020a; Xu et al., 2023). In particular, Xu et al. (2023) proposed deterministic ZO-AGDA and stochastic variance-reduced ZO-VRAGDA for smooth NC-PL minimax problems, with function-value query complexities of $\mathcal{O}(\epsilon^{-2})$ and $\widetilde{\mathcal{O}}(\epsilon^{-3})$, respectively. Their algorithms use zeroth-order estimators for both variables and assume smoothness in the outer variable. For nonsmooth objectives, mirror and proximal GDA schemes primarily exploit structured composite formulations (Huang et al., 2021; Chen et al., 2021; Li et al., 2022). The closest noncomposite results are more recent. Lin et al. (2025) analyzed two-timescale GDA for nonsmooth weakly convex–concave and weakly convex–strongly-concave objectives, targeting stationarity of a Moreau envelope of the value function. Shi & Luo (2026) further allowed nonsmooth nonconvexity in the first variable while retaining concavity or strong concavity in the second. Their PGFDA method targets a full-space Goldstein saddle stationary point of f, whereas NL-PGFDA targets a generalized Goldstein stationary point of the value function Φ.

3 PRELIMINARIES

3.1 NOTATIONS

We use $\|\cdot\|$ for the Euclidean norm of vectors and the spectral norm of matrices; I_d denotes the identity matrix. For $f(\mathbf{x}, \mathbf{y})$, $f(\mathbf{x}, \cdot)$ and $f(\cdot, \mathbf{y})$ denote the corresponding variable slices. Let $\mathbb{B}_\delta(\mathbf{x}) := \{\boldsymbol{z} \in \mathbb{R}^d : \|\boldsymbol{z} - \mathbf{x}\| \le \delta\}$, $\mathbb{B}^d := \mathbb{B}_1(\mathbf{0})$, $\mathbb{S}^{d-1} := \{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\| = 1\}$, and let $\mathrm{conv}(A)$ denote the convex hull of A. All asymptotic notation hides problem-dependent constants, and $\widetilde{\mathcal{O}}(\cdot)$ may additionally hide logarithmic factors. An SZO query evaluates one stochastic function value $f(\mathbf{x}, \mathbf{y}; \xi)$, while an SFO query returns one stochastic partial gradient $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}; \xi)$; their counts are reported separately. For $r > 0$, define the augmented set (Minkowski expansion) $\mathcal{X}_r := \mathcal{X} + r\mathbb{B}^{d_1} = \{\mathbf{x} + \boldsymbol{u} : \mathbf{x} \in \mathcal{X},\ \boldsymbol{u} \in \mathbb{R}^{d_1},\ \|\boldsymbol{u}\| \le r\}$. For candidate directions $\boldsymbol{h} \in \mathbb{R}^{d_1}$ and $\boldsymbol{g} \in \mathbb{R}^{d_2}$, respectively, we use the projected descent and ascent mappings $\mathcal{G}_{\mathbf{x},\gamma}(\mathbf{x}, \mathbf{y}, \boldsymbol{h}) := \frac{\mathbf{x} - \mathcal{P}_{\mathcal{X}}(\mathbf{x} - \gamma \boldsymbol{h})}{\gamma}$ and $\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}, \mathbf{y}, \boldsymbol{g}) := \frac{\mathcal{P}_{\mathcal{Y}}(\mathbf{y} + \lambda \boldsymbol{g}) - \mathbf{y}}{\lambda}$. Here $\mathcal{P}_C$ denotes Euclidean projection onto a nonempty closed convex set C. Thus, $\mathcal{G}$ always denotes a population projected-gradient mapping; computed stochastic mappings will be marked by a tilde.

3.2 PROBLEM SETTING AND ASSUMPTIONS

We consider the following constrained stochastic minimax problem:

$$\min_{\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^{d_1}} \max_{\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^{d_2}} f(\mathbf{x}, \mathbf{y}) := \mathbb{E}_\xi [f(\mathbf{x}, \mathbf{y}; \xi)], \tag{1}$$

where ξ is a random variable following an unknown distribution. The objective may be nonsmooth and nonconvex in the outer variable $\mathbf{x}$. In the inner variable $\mathbf{y}$, it is smooth and satisfies a projected-gradient-mapping Polyak–Łojasiewicz (projected-PŁ) condition (Li & Li, 2018). We impose the following common assumptions.

Assumption 3.1. Fix an ambient radius $\bar{\delta} > 0$. The assumptions below hold on the common domain $\mathcal{X}_{\bar{\delta}} \times \mathcal{Y}$, and the smoothing radius used by the algorithm satisfies $0 < \delta < \bar{\delta}$:

(1) The stochastic sample map $(\mathbf{x}, \mathbf{y}, \xi) \mapsto f(\mathbf{x}, \mathbf{y}; \xi)$ is jointly measurable and is integrable in ξ at every $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}_{\bar{\delta}} \times \mathcal{Y}$. The population objective is well-defined and jointly measurable on this augmented domain, and it satisfies $f(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\xi}[f(\mathbf{x}, \mathbf{y}; \xi)]$. For every $\mathbf{y} \in \mathcal{Y}$, the outer slice $f(\cdot, \mathbf{y})$ admits a locally Lipschitz extension to an open neighborhood of $\mathcal{X}_{\bar{\delta}}$.

(2) For every $\mathbf{x} \in \mathcal{X}_{\bar{\delta}}$, the inner slice $f(\mathbf{x}, \cdot)$ admits a continuously differentiable extension to an open neighborhood of $\mathcal{Y}$.

(3) There is $L > 0$ such that, for all $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2) \in \mathcal{X}_{\bar{\delta}} \times \mathcal{Y}$, $|f(\mathbf{x}_1, \mathbf{y}_1) - f(\mathbf{x}_2, \mathbf{y}_2)| \leq L\|(\mathbf{x}_1, \mathbf{y}_1) - (\mathbf{x}_2, \mathbf{y}_2)\|$.

(4) There is $L_{fy} > 0$ such that, for all $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2) \in \mathcal{X}_{\bar{\delta}} \times \mathcal{Y}$, $\|\nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}_2)\| \leq L_{fy}\|(\mathbf{x}_1, \mathbf{y}_1) - (\mathbf{x}_2, \mathbf{y}_2)\|$.

(5) The feasible sets $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_2}$ are nonempty, closed, and convex, and $\mathcal{Y}$ is compact.

Assumption 3.2 (Projected-PŁ condition)**.** There exist constants $\mu \geq 0$ and $\bar{\lambda} > 0$, independent of the outer variable, such that, for every $\mathbf{x} \in \mathcal{X}_{\bar{\delta}}$, $\mathbf{y} \in \mathcal{Y}$, and $0 < \lambda \leq \bar{\lambda}$, $\frac{1}{2}\|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}, \mathbf{y}, \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))\|^2 \geq \mu\left[\max_{\mathbf{z} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{z}) - f(\mathbf{x}, \mathbf{y})\right]$.

Writing $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ for a joint outer-inner point, we impose the following stochastic-oracle assumptions.

Assumption 3.3 (Stochastic y-gradient oracle)**.** For every $\mathbf{z} \in \mathcal{X} \times \mathcal{Y}$, the stochastic gradient is unbiased, $\mathbb{E}_{\xi}[\nabla_{\mathbf{y}} f(\mathbf{z}; \xi)] = \nabla_{\mathbf{y}} f(\mathbf{z})$, and has bounded variance $\mathbb{E}_{\xi}\|\nabla_{\mathbf{y}} f(\mathbf{z}; \xi) - \nabla_{\mathbf{y}} f(\mathbf{z})\|^2 \leq \bar{\sigma}_y^2$. Moreover, for some $L_{y,\mathrm{ms}} > 0$ and all $\mathbf{z}, \mathbf{z}' \in \mathcal{X} \times \mathcal{Y}$, $\mathbb{E}_{\xi}\|\nabla_{\mathbf{y}} f(\mathbf{z}; \xi) - \nabla_{\mathbf{y}} f(\mathbf{z}'; \xi)\|^2 \leq L_{y,\mathrm{ms}}^2\|\mathbf{z} - \mathbf{z}'\|^2$.

Assumption 3.4 (Mean-square Lipschitz SZO oracle)**.** For almost every ξ, the sample function is defined on $\mathcal{X}_{\bar{\delta}} \times \mathcal{Y}$. There is a measurable random modulus $L(\xi)$ such that, for all $\mathbf{z}, \mathbf{z}' \in \mathcal{X}_{\bar{\delta}} \times \mathcal{Y}$, $|f(\mathbf{z}; \xi) - f(\mathbf{z}'; \xi)| \leq L(\xi)\|\mathbf{z} - \mathbf{z}'\|$ and $\mathbb{E}[L(\xi)^2] \leq L^2$.

All above assumptions are standard in stochastic minimax optimization (Lin et al., 2020; Huang et al., 2021; Lin et al., 2025; Shi & Luo, 2026). Assumption 3.4 is the mean-square Lipschitz model commonly used for stochastic zeroth-order nonsmooth optimization (Lin et al., 2022; Chen et al., 2023; Kornowski & Shamir, 2024; Shi & Luo, 2026). For Assumption 3.2, we give an example of a nonconcave inner objective satisfying the projected-PŁ condition in Appendix F.

Based on Assumptions 3.1–3.4, we have the following lemma, which is useful for the analysis of our method.

Lemma 3.5 (Common almost everywhere x-differentiability)**.** *Under Assumption 3.1, there exist a measurable set $\mathcal{X}_{\mathrm{diff}} \subseteq \mathrm{int}(\mathcal{X}_{\bar{\delta}})$ whose complement in $\mathcal{X}_{\bar{\delta}}$ has zero d_1-dimensional Lebesgue measure and a constant $L_{xy} > 0$ such that $f(\cdot, \mathbf{y})$ is differentiable at every $\mathbf{x} \in \mathcal{X}_{\mathrm{diff}}$ simultaneously for all $\mathbf{y} \in \mathcal{Y}$, and*

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_1) - \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_2)\| \leq L_{xy}\|\mathbf{y}_1 - \mathbf{y}_2\|, \qquad \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}.$$

The common set allows the analysis to compare the x-gradients at the current inner point and at an active maximizer evaluated at the same randomized outer point. Appendix C.1 proves the lemma.

3.3 Background for Nonsmooth Analysis

Recall the generalized derivative of a locally Lipschitz function (Clarke, 1990).

Definition 3.6. Let $\mathcal{U} \subseteq \mathbb{R}^d$ be open. For a locally Lipschitz function $f : \mathcal{U} \to \mathbb{R}$, a point $\mathbf{x} \in \mathcal{U}$, and a direction $\boldsymbol{v}$, the Clarke directional derivative is $f^{\circ}(\mathbf{x}; \boldsymbol{v}) := \limsup_{\boldsymbol{s} \to \mathbf{x},\, t \downarrow 0}[f(\boldsymbol{s} + t\boldsymbol{v}) - f(\boldsymbol{s})]/t$. Its Clarke subdifferential is $\partial f(\mathbf{x}) := \{\boldsymbol{h} \in \mathbb{R}^d : \boldsymbol{h}^{\top}\boldsymbol{v} \leq f^{\circ}(\mathbf{x}; \boldsymbol{v}),\ \forall \boldsymbol{v} \in \mathbb{R}^d\}$.

Then, an ϵ-Clarke stationary point of unconstrained minimization satisfies

$$\min\{\|\boldsymbol{h}\| : \boldsymbol{h} \in \partial f(\mathbf{x})\} \leq \epsilon. \tag{2}$$

However, Zhang et al. (2020) pointed out that finding an ϵ-Clarke stationary point of a general Lipschitz nonconvex function is intractable in finite time in the worst case. This motivates the following neighborhood relaxation.

Definition 3.7. Let $\mathcal{U} \subseteq \mathbb{R}^d$ be open. For a locally Lipschitz function $f : \mathcal{U} \to \mathbb{R}$, a point $\mathbf{x} \in \mathcal{U}$, and $\delta > 0$ such that $\mathbb{B}_\delta(\mathbf{x}) \subset \mathcal{U}$, the Goldstein δ-subdifferential is $\partial_\delta f(\mathbf{x}) := \operatorname{conv}\left(\bigcup_{\boldsymbol{s}\in\mathbb{B}_\delta(\mathbf{x})} \partial f(\boldsymbol{s})\right)$, the convex hull of Clarke subgradients at points in the δ-neighborhood of $\mathbf{x}$ (Clarke, 1990).

At $\delta = 0$, the Goldstein subdifferential reduces to the Clarke subdifferential, i.e., $\partial_0 f(\mathbf{x}) = \partial f(\mathbf{x})$ (Lin et al., 2022; Liu et al., 2024a; Chen et al., 2023). The corresponding approximate stationarity criterion is (Zhang et al., 2020; Lin et al., 2022):

Definition 3.8. A point $\mathbf{x}$ is a (δ, ϵ)-generalized Goldstein stationary point of a Lipschitz function f if $\min\{\|\boldsymbol{h}\| : \boldsymbol{h} \in \partial_\delta f(\mathbf{x})\} \leq \epsilon$.

Finite-time methods therefore target $\delta > 0$ (Zhang et al., 2020; Lin et al., 2022; Chen et al., 2023; Liu et al., 2024a). These unconstrained definitions supply the generalized derivatives used below.

4 Proposed Method

4.1 Generalized Goldstein Stationarity of Minimax Problems

As discussed above, for the nonsmooth $\mathbf{x}$-variable, exact Clarke stationarity is generally too strong for finite-time guarantees. We therefore use a projected Goldstein criterion for $\mathbf{x}$ together with the standard projected-gradient mapping for the smooth $\mathbf{y}$-variable.

For a fixed $\mathbf{y} \in \mathcal{Y}$, let $\partial_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y})$ denote the Clarke subdifferential. Fix $0 < \delta < \bar{\delta}$. Then, for every $\mathbf{x} \in \mathcal{X}$, $\mathbb{B}_\delta(\mathbf{x}) \subset \operatorname{int}(\mathcal{X}_{\bar{\delta}})$. We can therefore define the outer Goldstein subdifferential intrinsically on the common regularity domain:

$$\partial_{\delta,\mathbf{x}} f(\mathbf{x}, \mathbf{y}) := \operatorname{conv}\left(\bigcup_{\boldsymbol{s}\in\mathbb{B}_\delta(\mathbf{x})} \partial_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y})\right). \tag{3}$$

Then, we can define the stationarity criterion for the problem in Eqn. (1).

Definition 4.1 (Generalized Goldstein stationary point)**.** A pair $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ is a $(\gamma, \lambda, \delta, \epsilon)$-generalized Goldstein stationary point (GGSP) if

$$\min_{\boldsymbol{h}\in\partial_{\delta,\mathbf{x}} f(\mathbf{x},\mathbf{y})} \|\mathcal{G}_{\mathbf{x},\gamma}(\mathbf{x}, \mathbf{y}, \boldsymbol{h})\| \leq \epsilon, \quad \|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}, \mathbf{y}, \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))\| \leq \epsilon, \tag{4}$$

At $\delta = 0$, the x-block condition in Definition 4.1 is projected Clarke stationarity, while the smooth y-block condition is unchanged. For a random output, we use the corresponding expected criterion.

Definition 4.2 (Expected generalized Goldstein stationary point)**.** A random pair $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}})$ is an expected $(\gamma, \lambda, \delta, \epsilon)$-GGSP if the two variable-wise residuals satisfy separately

$$\mathbb{E}\left[\min_{\boldsymbol{h}\in\partial_{\delta,\mathbf{x}} f(\widehat{\mathbf{x}},\widehat{\mathbf{y}})} \|\mathcal{G}_{\mathbf{x},\gamma}(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}, \boldsymbol{h})\|\right] \leq \epsilon, \quad \mathbb{E}\left[\|\mathcal{G}_{\mathbf{y},\lambda}(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}, \nabla_{\mathbf{y}} f(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}))\|\right] \leq \epsilon. \tag{5}$$

Definitions 4.1 and 4.2 characterize the stationarity of the algorithmic pair $(\mathbf{x}, \mathbf{y})$. Since the minimax problem in Eqn. (1) is equivalently the minimization of its worst-case value function, we also introduce a stationarity criterion stated directly for this function, independently of the auxiliary inner variable.

Let $\Phi(\mathbf{x}) := \max_{\boldsymbol{s}\in\mathcal{Y}} f(\mathbf{x}, \boldsymbol{s})$ denote the value function. It is L-Lipschitz on $\mathcal{X}_{\bar{\delta}}$ under Assumption 3.1. Since $0 < \delta < \bar{\delta}$ and $\mathbb{B}_\delta(\mathbf{x}) \subset \operatorname{int}(\mathcal{X}_{\bar{\delta}})$ for every $\mathbf{x} \in \mathcal{X}$, $\Phi_\delta(\mathbf{x})$ and $\partial_\delta \Phi(\mathbf{x})$ are defined locally from Φ on this common domain. In particular, $\partial_\delta \Phi$ is the Goldstein subdifferential of the locally Lipschitz restriction of Φ to $\operatorname{int}(\mathcal{X}_{\bar{\delta}})$. Define $\mathfrak{r}_\Phi(\mathbf{x}) := \min_{\boldsymbol{h}\in\partial_\delta\Phi(\mathbf{x})} \left\|\frac{\mathbf{x}-\mathcal{P}_{\mathcal{X}}(\mathbf{x}-\gamma\boldsymbol{h})}{\gamma}\right\|$, we have the following value-function stationarity criterion.

Definition 4.3 (Projected value-function GGSP)**.** A point $\mathbf{x} \in \mathcal{X}$ is a projected $(\gamma, \delta, \epsilon)$-GGSP of Φ if $\mathfrak{r}_\Phi(\mathbf{x}) \le \epsilon$. A random point $\widehat{\mathbf{x}}$ is an expected projected $(\gamma, \delta, \epsilon)$-GGSP of Φ if $\mathbb{E}[\mathfrak{r}_\Phi(\widehat{\mathbf{x}})] \le \epsilon$. When $\mathcal{X} = \mathbb{R}^{d_1}$, this reduces to the usual condition $\min_{\boldsymbol{h} \in \partial_\delta \Phi(\mathbf{x})} \|\boldsymbol{h}\| \le \epsilon$.

4.2 Randomized Smoothing

Let $\boldsymbol{u}$ be uniformly distributed on $\mathbb{B}^{d_1}$ and define

$$f_\delta(\mathbf{x}, \mathbf{y}) := \mathbb{E}_{\boldsymbol{u}}[f(\mathbf{x} + \delta \boldsymbol{u}, \mathbf{y})] = \mathbb{E}_{\boldsymbol{u}, \xi}[f(\mathbf{x} + \delta \boldsymbol{u}, \mathbf{y}; \xi)]. \tag{6}$$

The following lemma connects this surrogate to the original nonsmooth objective; its proof is in Appendix C.3.

Lemma 4.4 (Properties of randomized smoothing)**.** *Under Assumption 3.1, the following statements hold for $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}$, as applicable: (1) $|f_\delta(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}, \mathbf{y})| \le \delta L$; (2) f_δ is differentiable and L-Lipschitz, and $\nabla_{\mathbf{x}} f_\delta(\cdot, \mathbf{y})$ is $L_{x,\delta} := L\sqrt{d_1}/\delta$-Lipschitz; (3) By Lemma 3.5, $\nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \cdot)$ is L_{xy}-Lipschitz; (4) $\nabla_{\mathbf{y}} f_\delta$ is L_{fy}-Lipschitz in $(\mathbf{x}, \mathbf{y})$ and $\|\nabla_{\mathbf{y}} f_\delta(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\| \le L_{fy}\delta$; (5) $\nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y}) \in \partial_{\delta, \mathbf{x}} f(\mathbf{x}, \mathbf{y})$.*

To describe the inner geometry after smoothing, let $\mathcal{Y}^\star(\mathbf{x}) := \arg\max_{\mathbf{z} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{z})$ and $\Delta(\mathbf{x}, \mathbf{y}) := \Phi(\mathbf{x}) - f(\mathbf{x}, \mathbf{y})$. Define the averaged value and gap by $\Phi_\delta(\mathbf{x}) := \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^{d_1})}[\Phi(\mathbf{x} + \delta \boldsymbol{u})]$ and $\Delta_\delta(\mathbf{x}, \mathbf{y}) := \Phi_\delta(\mathbf{x}) - f_\delta(\mathbf{x}, \mathbf{y})$. The projected-PŁ condition yields quadratic growth for each original inner problem and controls the averaged gap up to the smoothing error.

Lemma 4.5 (Quadratic growth from projected-PŁ)**.** *Fix an analysis stepsize $0 < \lambda_{\mathrm{qg}} \le \min\left\{\bar{\lambda}, \frac{1}{L_{fy}}, \frac{1}{2\mu}\right\}$. Under Assumptions 3.1 and 3.2, $f(\mathbf{x}, \cdot)$ satisfies, for every $\mathbf{x} \in \mathcal{X}_{\bar{\delta}}$, the quadratic-growth condition $\mathrm{dist}^2(\mathbf{y}, \mathcal{Y}^\star(\mathbf{x})) \le C_{\mathrm{qg}} \Delta(\mathbf{x}, \mathbf{y})$, $\forall \mathbf{y} \in \mathcal{Y}$, where $C_{\mathrm{qg}} := \frac{2\lambda_{\mathrm{qg}}}{\left(1 - \sqrt{1 - \lambda_{\mathrm{qg}}\mu}\right)^2}$.*

Lemma 4.6 (Averaged projected-PŁ bound)**.** *Let $\nu_{d_1} := \mathbb{E}_{\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^{d_1})}\|\boldsymbol{u}\|^2 = d_1/(d_1 + 2) \le 1$. For every $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, $0 < \lambda \le \bar{\lambda}$, and $\tau > 0$, $\frac{1}{2}\|\mathcal{G}_{\mathbf{y}, \lambda}(\mathbf{x}, \mathbf{y}, \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))\|^2 \ge \frac{\mu}{1+\tau}\Delta_\delta(\mathbf{x}, \mathbf{y}) - \frac{\nu_{d_1} L_{fy}^2 \delta^2}{2\tau}$. In particular, for $\tau = 1$, we have $\Delta_\delta(\mathbf{x}, \mathbf{y}) \le \frac{1}{\mu}\|\mathcal{G}_{\mathbf{y}, \lambda}(\mathbf{x}, \mathbf{y}, \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))\|^2 + \frac{\nu_{d_1} L_{fy}^2}{\mu}\delta^2$.*

Remark. Lemma 4.5 converts projected-PŁ into distance-to-solution control, which is used to compare the outer gradients at the current inner point and an inner maximizer. Lemma 4.6 bounds the averaged gap Δ_δ by the algorithmically controlled inner projected-gradient residual, up to an $O(\delta^2)$ smoothing error. Together, they enable the outer-gradient tracking and the subsequent value-function stationarity guarantee without requiring $\Phi_\delta(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}} f_\delta(\mathbf{x}, \mathbf{y})$.

4.3 The Zeroth-Order and First-Order SPIDER Gradient Descent-Ascent Method

Equipped with the smoothing surrogate, we now describe our hybrid stochastic algorithm, which combines zeroth-order updates in the nonsmooth outer variable with first-order updates in the smooth inner variable. Specifically, for $\hat{\boldsymbol{u}} \sim \mathrm{Unif}(\mathbb{S}^{d_1 - 1})$, define the single-sample symmetric two-point estimator

$$g_{\mathbf{x}}(\mathbf{x}, \mathbf{y}; \xi, \hat{\boldsymbol{u}}) := \frac{d_1}{2\delta}\big[f(\mathbf{x} + \delta\hat{\boldsymbol{u}}, \mathbf{y}; \xi) - f(\mathbf{x} - \delta\hat{\boldsymbol{u}}, \mathbf{y}; \xi)\big]\hat{\boldsymbol{u}}. \tag{7}$$

For an integer $M \ge 1$, draw i.i.d. oracle samples $\{\xi_i\}_{i=1}^M$ and i.i.d. directions $\{\hat{\boldsymbol{u}}_j\}_{j=1}^M$, where each direction is uniform on $\mathbb{S}^{d_1 - 1}$ and all random variables in the two collections are mutually independent. We define $\bar{\nabla}_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y}; \mathcal{B}) := \frac{1}{M}\sum_{i=1}^M g_{\mathbf{x}}(\mathbf{x}, \mathbf{y}; \xi_i, \hat{\boldsymbol{u}}_i)$.

Our method is presented in Algorithm 1. We refer to Algorithm 1 as the Zeroth-Order and First-Order SPIDER Gradient Descent–Ascent method (ZOFOSGDA). It maintains one recursive estimator for each variable. The estimator w_t targets $\nabla_{\mathbf{x}} f_\delta(\mathbf{x}_t, \mathbf{y}_t)$ through the two-point SZO oracle, whereas v_t targets $\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)$ through the stochastic first-order oracle. Both estimators are refreshed with large batches at epoch boundaries and updated by SPIDER differences between refreshes. The method then combines projected descent in the nonsmooth $\mathbf{x}$-variable with projected ascent in the smooth $\mathbf{y}$-variable.

Algorithm 1 ZOFOSGDA for nonsmooth nonconvex minimax optimization with a smooth inner objective satisfying a projected-PŁ condition

Require: $\mathbf{x}_0 \in \mathcal{X}$, $\mathbf{y}_0 \in \mathcal{Y}$, horizon T, smoothing radius $0 < \delta < \bar{\delta}$, stepsizes γ, λ, epoch length q, refresh batch sizes B_x, B_y, and recursive batch sizes b_x, b_y.

for $t = 0, 1, \ldots, T-1$ **do**
  $\mathbf{z}_t \leftarrow (\mathbf{x}_t, \mathbf{y}_t)$.
  if $t \bmod q = 0$ **then**
    Sample $\mathcal{B}_t^x = \{(\xi_i^x, \hat{\boldsymbol{u}}_i)\}_{i=1}^{B_x}$, with all oracle samples and directions mutually independent and each direction uniform on $\mathbb{S}^{d_1-1}$; sample $\mathcal{B}_t^y = \{\xi_i^y\}_{i=1}^{B_y}$.
    $w_t \leftarrow B_x^{-1} \sum_{i=1}^{B_x} g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t; \xi_i^x, \hat{\boldsymbol{u}}_i)$; $v_t \leftarrow B_y^{-1} \sum_{i=1}^{B_y} \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \xi_i^y)$.
  else
    Sample $\mathcal{S}_t^x = \{(\xi_i^x, \hat{\boldsymbol{u}}_i)\}_{i=1}^{b_x}$, with all oracle samples and directions mutually independent, and sample $\mathcal{S}_t^y = \{\xi_i^y\}_{i=1}^{b_y}$.
    $w_t \leftarrow w_{t-1} + b_x^{-1} \sum_{i=1}^{b_x} [g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t; \xi_i^x, \hat{\boldsymbol{u}}_i) - g_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \xi_i^x, \hat{\boldsymbol{u}}_i)]$.
    $v_t \leftarrow v_{t-1} + b_y^{-1} \sum_{i=1}^{b_y} [\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t; \xi_i^y) - \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \xi_i^y)]$.
  end if
  $\mathbf{x}_{t+1} \leftarrow \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - \gamma w_t)$ and $\mathbf{y}_{t+1} \leftarrow \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_t + \lambda v_t)$
end for
Sample $R \sim \mathrm{Unif}\{0, \ldots, T-1\}$ independently.
return $(\mathbf{x}_R, \mathbf{y}_R)$.

5 Convergence Guarantees

We now establish the convergence and separate SZO/SFO oracle complexities of ZOFOSGDA. The main text states the required assumptions and final guarantees. Technical estimates and main-result proofs are given in Appendices D and E. Here we summarize the additional assumptions and the main results.

Assumption 5.1 (Fresh coupled sampling)**.** Let $\mathcal{F}_t$ contain all randomness revealed before the batch at iteration t is sampled. Conditional on $\mathcal{F}_t$, all newly drawn oracle samples and directions are mutually independent and independent of $\mathcal{F}_t$; the oracle samples are identically distributed according to their respective oracle laws, and the directions are i.i.d. uniform on $\mathbb{S}^{d_1-1}$. In each recursive difference, the two evaluations at $(\mathbf{x}_t, \mathbf{y}_t)$ and $(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$ reuse the same oracle sample and direction within each summand.

Assumption 5.2 (Lower-bounded smoothed value function)**.** We have $\Phi_\delta^\star := \inf_{\mathbf{x} \in \mathcal{X}} \Phi_\delta(\mathbf{x}) > -\infty$.

Under Assumptions 3.1, 3.4, and 5.1, the zeroth-order estimator satisfies the following unbiasedness and variance bounds.

Lemma 5.3 (Unbiasedness and dimension-dependent ZO variance)**.** *Under Assumptions 3.1, 3.4, and 5.1,* $\mathbb{E}[g_{\mathbf{x}}(\mathbf{x}, \mathbf{y}; \xi, \hat{\boldsymbol{u}})] = \nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y})$*. Moreover, it is valid to take the single-pair variance constant* $\bar{\sigma}_x^2 := 2 d_1 L^2$*, so that, for the batch* $\mathcal{B}$ *above, which averages over* M *independent oracle–direction draws,* $\mathbb{E} \left\| \bar{\nabla}_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y}; \mathcal{B}) - \nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y}) \right\|^2 \le \frac{\bar{\sigma}_x^2}{M} = \frac{2 d_1 L^2}{M}$.

Under the assumptions above and the admissible step-size conditions stated below, define $C_r := L_{xy}^2 C_{\mathrm{qg}}$, $\chi := \max\{2, 512 C_r / \mu\}$, $a := \frac{\chi \gamma}{\lambda}$, and the Lyapunov potential

$$\Psi_t := \Phi_\delta(\mathbf{x}_t) + a \Delta_\delta(\mathbf{x}_t, \mathbf{y}_t) = (1 + a) \Phi_\delta(\mathbf{x}_t) - a f_\delta(\mathbf{x}_t, \mathbf{y}_t).$$

Then, we can establish the following convergence properties.

Theorem 5.4 (Simplified convergence guarantee)**.** *Suppose Assumptions 3.1, 3.2, 3.4, 3.3, 5.1, and 5.2 hold. Run Algorithm 1 with parameters satisfying the detailed admissibility conditions in Theorem E.1. If* R *is sampled uniformly from* $\{0, \ldots, T-1\}$*, define the exact smoothed outer mapping by* $\mathcal{G}_{\mathbf{x},\gamma}^\delta(\mathbf{x}, \mathbf{y}) := \mathcal{G}_{\mathbf{x},\gamma}(\mathbf{x}, \mathbf{y}, \nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y}))$*. Then*

$$\mathbb{E}\Big[\|\mathcal{G}_{\mathbf{x},\gamma}^\delta(\mathbf{x}_R, \mathbf{y}_R)\|^2 + \|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}_R, \mathbf{y}_R, \nabla_{\mathbf{y}} f(\mathbf{x}_R, \mathbf{y}_R))\|^2 \Big] \le \mathcal{O}\left(\frac{\Psi_0 - \Phi_\delta^\star}{\gamma T} + \frac{\bar{\sigma}_x^2}{B_x} + \frac{\bar{\sigma}_y^2}{B_y} + \delta^2 \right).$$

The hidden constant depends only on the problem parameters and the fixed ratios in the admissible parameter choice.

Since Definition 4.1 is a pointwise property, we first state the corresponding expected guarantee and then record separately the probability with which the random output satisfies Definition 4.1.

Corollary 5.5 (Expected GGSP guarantee)**.** *If the explicit bound in Theorem E.1 is at most ϵ^2, then $(\mathbf{x}_R, \mathbf{y}_R)$ is an expected $(\gamma, \lambda, \delta, \epsilon)$-GGSP in the sense of Definition 4.2.*

Corollary 5.6 (High-probability GGSP guarantee)**.** *Let $\eta \in (0, 1)$. If the explicit bound in Theorem E.1 is at most $\eta\epsilon^2$, then, with probability at least $1-\eta$, the output $(\mathbf{x}_R, \mathbf{y}_R)$ is a $(\gamma, \lambda, \delta, \epsilon)$-GGSP in the sense of Definition 4.1.*

Remark. Corollaries 5.5 and 5.6 give expected and high-probability versions of the same algorithmic stationarity guarantee. The latter follows from Markov's inequality after tightening the mean-square bound by the factor η.

Corollary 5.7 (Projected value-function Goldstein stationarity)**.** *Let $\mathcal{E}_T$ denote the right-hand side of Eqn. (81), and set $C_\Phi := 1 + C_r/\mu$. Under the conditions of Theorem 5.4,*

$$\mathbb{E}[\mathfrak{r}_\Phi(\mathbf{x}_R)^2] \le C_\Phi \left(\mathcal{E}_T + \nu_{d_1} L_{fy}^2 \delta^2\right). \tag{8}$$

Consequently, if the right-hand side of Eqn. (8) is at most ϵ^2, then $\mathbf{x}_R$ is an expected projected $(\gamma, \delta, \epsilon)$-GGSP of Φ. If it is at most $\eta\epsilon^2$ for some $\eta \in (0, 1)$, then $\mathbf{x}_R$ is a projected $(\gamma, \delta, \epsilon)$-GGSP of Φ with probability at least $1 - \eta$.

Remark. Corollary 5.7 transfers the block-residual bound to the worst-case objective Φ. Its $O(\delta^2)$ smoothing term remains at the target stationarity scale when $\delta = O(\epsilon)$.

Corollary 5.8 (SZO–SFO oracle complexity)**.** *Under the conditions of Theorem 5.4, suppose there exist $\delta_0 \in (0, \bar{\delta})$, $\epsilon_0 > 0$, and $\Delta_{\mathrm{pot}} \ge 1$ such that $\Psi_0 - \Phi_\delta^\star \le \Delta_{\mathrm{pot}}$ for the fixed initialization, $a = 1/2$, and every $\delta \in (0, \delta_0]$. Then, for a sufficiently small problem-dependent $c_\delta > 0$, $0 < \epsilon \le \epsilon_0$, and $0 < \delta \le \min\{\delta_0, c_\delta\epsilon\}$, Algorithm 1 attains both expected stationarity guarantees in Corollaries 5.5 and 5.7, with*

$$Q_{\mathrm{SZO}} = \widetilde{\mathcal{O}}\left(\Delta_{\mathrm{pot}} d_1^{3/2} \delta^{-1} \epsilon^{-3}\right), \quad Q_{\mathrm{SFO}} = \widetilde{\mathcal{O}}\left(\Delta_{\mathrm{pot}} d_1^{1/2} \delta^{-1} \epsilon^{-3}\right). \tag{9}$$

For the canonical choice $\delta = c_\delta\epsilon = \Theta(\epsilon)$, after reducing ϵ_0 if necessary, these bounds become $Q_{\mathrm{SZO}} = \widetilde{\mathcal{O}}\left(\Delta_{\mathrm{pot}} d_1^{3/2} \epsilon^{-4}\right)$ and $Q_{\mathrm{SFO}} = \widetilde{\mathcal{O}}\left(\Delta_{\mathrm{pot}} d_1^{1/2} \epsilon^{-4}\right)$.

Corollary 5.9 (High-confidence oracle complexity (simplified))**.** *Under the conditions of Corollary 5.8, let $\eta \in (0, 1)$, set $\epsilon_\eta = \sqrt{\eta}\,\epsilon$, and choose $0 < \delta \le \min\{\delta_0, c_\delta\epsilon_\eta\}$. Then both the pairwise and value-function GGSP guarantees hold simultaneously with probability at least $1 - \eta$, and*

$$Q_{\mathrm{SZO}} = \widetilde{\mathcal{O}}\left(\Delta_{\mathrm{pot}} d_1^{3/2} \delta^{-1} \eta^{-3/2} \epsilon^{-3}\right), \quad Q_{\mathrm{SFO}} = \widetilde{\mathcal{O}}\left(\Delta_{\mathrm{pot}} d_1^{1/2} \delta^{-1} \eta^{-3/2} \epsilon^{-3}\right). \tag{10}$$

For $\delta = c_\delta\sqrt{\eta}\,\epsilon$, the two bounds reduce to $\widetilde{\mathcal{O}}(\Delta_{\mathrm{pot}} d_1^{3/2} \eta^{-2} \epsilon^{-4})$ and $\widetilde{\mathcal{O}}(\Delta_{\mathrm{pot}} d_1^{1/2} \eta^{-2} \epsilon^{-4})$.

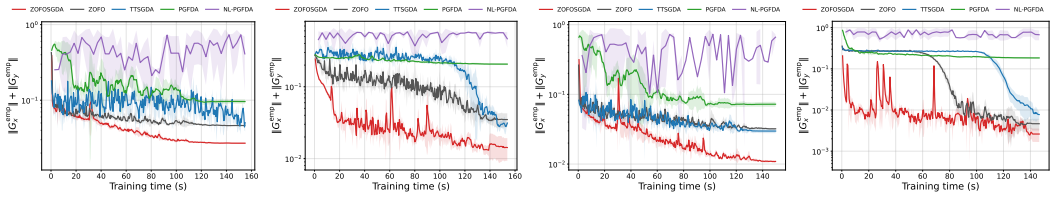
Remark. Here Q_{SZO} and Q_{SFO} count stochastic function-value and stochastic y-gradient queries. The outer query count is measured in the outer batch sizes B_x and b_x; the extra factor d_1 in the resulting rate comes from the two-point outer-gradient estimator. Corollary 5.9 applies the expected result at accuracy $\sqrt{\eta}\,\epsilon$ to obtain both guarantees on the same event.

Table 2: Test accuracy (%) under matched training time (mean ± standard deviation over three random seeds).

Algorithm	a9a	Breast-cancer	Cod-rna	German	Ijcnn1	Splice
TTSGDA	84.15±0.05	97.32±0.42	92.09±0.39	83.17±0.58	91.94±0.16	**85.12±0.07**
PGFDA	76.41±0.12	97.57±1.11	59.82±4.53	74.33±3.62	89.86±0.76	74.87±2.93
NL-PGFDA	68.79±10.46	64.01±7.28	69.56±9.98	67.33±9.73	73.91±7.61	72.87±2.01
ZOFO	83.52±0.17	96.84±0.42	92.14±0.07	80.67±1.89	91.69±0.17	83.80±0.41
ZOFOSGDA	**84.62±0.15**	**97.84±0.42**	**92.22±0.06**	**83.67±0.76**	**92.20±0.04**	**85.12±0.25**

6 Numerical Illustrations

In this section, we provide a numerical illustration of the finite-sample behavior of ZOFOSGDA on a controlled benchmark. Appendix G presents a supplementary AUC-maximization illustration.



(a) a9a (b) Splice (c) Ijcnn1 (d) Breast-cancer

Figure 1: The combined implementation-dependent proxy $\|G_x^{\mathrm{emp}}\| + \|G_y^{\mathrm{emp}}\|$ versus training time for four displayed datasets, where $G_x^{\mathrm{emp}} = \gamma^{-1}[\mathbf{x} - \mathcal{P}_{\mathcal{X}}(\mathbf{x} - \gamma g_x^{\mathrm{sel}})]$ and $G_y^{\mathrm{emp}} = \lambda^{-1}[\mathcal{P}_{\mathcal{Y}}(\mathbf{y} + \lambda \nabla_{\mathbf{y}} f_{\mathrm{syn}}(\mathbf{x}, \mathbf{y})) - \mathbf{y}]$. Here g_x^{sel} is the deterministic automatic-differentiation output for the x-slice (the ordinary gradient when the executed computational branch is smooth and the implementation's fixed convention otherwise), used solely for evaluation.

6.1 Controlled Nonconvex-PŁ Classification Objective

Let $\mathcal{D} = \{(\boldsymbol{a}_i, b_i)\}_{i=1}^N$ be a bounded binary classification dataset. We use the linear classifier $h_{\mathbf{x}}(\boldsymbol{a}) := \mathbf{w}^\top \boldsymbol{a} + \tau$, where $\mathbf{x} = (\mathbf{w}, \tau)$ belongs to a fixed compact convex set $\mathcal{X}$. Define $\ell_i(\mathbf{x}) := [1 - b_i h_{\mathbf{x}}(\boldsymbol{a}_i)]_+$ and $\mathcal{R}_\beta(\mathbf{x}) := \sum_j \min\{|x_j|, \beta\}$. Fix a uniform bound $L_{\max} \geq \sup_{\mathbf{x} \in \mathcal{X}_{\delta}, i} \ell_i(\mathbf{x})$, and set $s_i(\mathbf{x}) := \ell_i(\mathbf{x})/(2L_{\max})$. To obtain a smooth inner problem that is nonconcave but satisfies a uniform projected-PŁ condition, define $T_\alpha(y) := y + \frac{\alpha}{2\pi}\sin(2\pi y)$ and $\psi_{\alpha,s}(y) := \frac{s^2}{2} - \frac{(T_\alpha(y) - s)^2}{2(1-\alpha)^2}$, and fix the experiment-specific warp parameter $\alpha = 0.7$. We study

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}_{\mathrm{syn}}} f_{\mathrm{syn}}(\mathbf{x}, \mathbf{y}) := \frac{1}{N} \sum_{i=1}^N \left[\ell_i(\mathbf{x}) + c\, \psi_{\alpha, s_i(\mathbf{x})}(y_i)\right] + \kappa \mathcal{R}_\beta(\mathbf{x}), \tag{11}$$

where $c > 0$ and $\kappa > 0$ are fixed constants and $\mathcal{Y}_{\mathrm{syn}} = [-1, 1]^N$. Its nonconcavity and the verification of the required assumptions are provided in Appendix F. Full implementation details, parameter settings, and the hyperparameter-search protocol are provided in Appendix F.2.

Illustrative comparison. Table 2 shows that ZOFOSGDA attains the highest reported mean test accuracy on five of the six datasets and matches TTSGDA on Splice. The reported standard deviations for ZOFOSGDA are small across the three seeds. The proxy curves in Figure 1 provide a complementary optimization view. Under the same wall-clock budget, ZOFOSGDA reaches a lower implementation-dependent proxy residual than the competing methods on the four displayed datasets, with particularly clear separation on a9a and Ijcnn1. This comparison is consistent with a benefit from retaining first-order information in the y-block. Appendix F.3 states when the plotted proxy upper-bounds the sum of the two blockwise GGSP residuals.

Ablation against ZOFO. We treat ZOFO as an ablation of ZOFOSGDA: it retains the same two-point zeroth-order estimator for the x block, the same first-order y update, and the same tuning and evaluation pipeline, but removes the SPIDER recursive estimator and its periodic large- batch refresh. Relative to this ablation, ZOFOSGDA has higher mean test accuracy on all six datasets in the reported runs, with an average difference of 1.17 percentage points. Together with the lower proxy values in Figure 1, this comparison is consistent with SPIDER variance reduction benefiting optimization progress. Additional ablation studies are reported in Appendix F.5.

7 Conclusion

We studied stochastic minimax problems with a nonsmooth, nonconvex outer variable and a smooth, possibly nonconcave inner variable satisfying a uniform projected-PŁ condition. ZOFOSGDA combines two-point zeroth-order outer estimation, first-order inner updates, and SPIDER variance reduction. Under the stated assumptions and a bounded initial potential gap, it achieves expected blockwise GGSP and projected value-function GGSP guarantees, with the oracle complexities and high-probability extensions given above. The experiments illustrate finite-sample behavior on the controlled nonconvex-PŁ benchmark and on a supplementary AUC-maximization task outside the verified benchmark.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for manuscript polishing. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, We have reviewed all AI-assisted work. We checked problem setting and the assumption provides by generative AI. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

We provide a detailed description of the experimental setup, including the datasets, hyperparameters, and evaluation metrics used in our experiments. The code and data used in this work will be made publicly available upon publication to facilitate reproducibility. We also include additional details in the appendix to ensure that other researchers can replicate our results accurately.

REFERENCES

Lesi Chen, Jing Xu, and Luo Luo. Faster gradient-free algorithms for nonsmooth nonconvex stochastic optimization. In *International Conference on Machine Learning*, pp. 5219–5233. PMLR, 2023.

Ziyi Chen, Yi Zhou, Tengyu Xu, and Yingbin Liang. Proximal gradient descent-ascent: Variable convergence under KŁ geometry. *arXiv preprint arXiv:2102.04653*, 2021.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning*, pp. 6643–6670. PMLR, 2023.

Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. *Advances in neural information processing systems*, 35:6692–6703, 2022.

Michael Fauß, Abdelhak M Zoubir, and H Vincent Poor. Minimax robust detection: Classic results and recent advances. *IEEE Transactions on signal Processing*, 69:2252–2283, 2021.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Feihu Huang, Xidong Wu, and Heng Huang. Efficient mirror descent ascent methods for nonsmooth minimax problems. *Advances in Neural Information Processing Systems*, 34:10431–10443, 2021.

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *Journal of Machine Learning Research*, 23(36):1–70, 2022.

Feihu Huang, Xidong Wu, and Zhengmian Hu. Adagda: Faster adaptive gradient descent ascent methods for minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2365–2389. PMLR, 2023.

Misha Khodak, Ilya Osadchiy, Keegan Harris, Maria-Florina F Balcan, Kfir Y. Levy, Ron Meir, and Steven Z. Wu. Meta-learning adversarial bandit algorithms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 35441–35471. Curran Associates, Inc., 2023.

Guy Kornowski and Ohad Shamir. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *Journal of Machine Learning Research*, 25(122): 1–14, 2024.

Jiajin Li, Linglingzhi Zhu, and Anthony Man-Cho So. Nonsmooth composite nonconvex-concave minimax optimization. *arXiv preprint arXiv:2209.10825*, 2022.

Zhize Li and Jian Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International conference on machine learning*, pp. 6083–6093. PMLR, 2020.

Tianyi Lin, Zeyu Zheng, and Michael Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 35: 26160–26175, 2022.

Tianyi Lin, Chi Jin, and Michael I Jordan. Two-timescale gradient descent ascent algorithms for nonconvex minimax optimization. *Journal of Machine Learning Research*, 26(11):1–45, 2025.

Zhuanghua Liu, Cheng Chen, Luo Luo, and Bryan Kian Hsiang Low. Zeroth-order methods for constrained nonconvex nonsmooth stochastic optimization. In *Forty-first International Conference on Machine Learning*, 2024a.

Zhuanghua Liu, Luo Luo, and Bryan Kian Hsiang Low. Gradient-free methods for nonconvex nonsmooth stochastic compositional optimization. *Advances in Neural Information Processing Systems*, 37:45438–45461, 2024b.

Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in neural information processing systems*, 33:20566–20577, 2020.

Santiago Mazuelas, Mauricio Romero, and Peter Grunwald. Minimax risk classifiers with 0-1 loss. *Journal of Machine Learning Research*, 24(208):1–48, 2023.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Pranay Sharma, Rohan Panda, Gauri Joshi, and Pramod Varshney. Federated minimax optimization: Improved convergence analyses and algorithms. In *International Conference on Machine Learning*, pp. 19683–19730. PMLR, 2022.

Jinyang Shi and Luo Luo. Nonsmooth nonconvex-concave minimax optimization: Convergence criteria and algorithms. *arXiv preprint arXiv:2604.21371*, 2026.

Zhenyu Sun and Ermin Wei. A communication-efficient algorithm with linear convergence for federated minimax learning. *Advances in Neural Information Processing Systems*, 35:6060–6073, 2022.

Lai Tian, Kaiwen Zhou, and Anthony Man-Cho So. On the finite-time complexity and practical computation of approximate stationarity concepts of lipschitz functions. In *International Conference on Machine Learning*, pp. 21360–21379. PMLR, 2022.

Lianzhe Wang, Shiji Zhou, Shanghang Zhang, Xu Chu, Heng Chang, and Wenwu Zhu. Improving generalization of meta-learning with inverted regularization at inner-level. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7826–7835, 2023a.

Yifei Wang, Liangchen Li, Jiansheng Yang, Zhouchen Lin, and Yisen Wang. Balance, imbalance, and rebalance: Understanding robust overfitting from a minimax game perspective. *Advances in neural information processing systems*, 36:15775–15798, 2023b.

Zhongruo Wang, Krishnakumar Balasubramanian, Shiqian Ma, and Meisam Razaviyayn. Zeroth-order algorithms for nonconvex–strongly-concave minimax problems with improved complexities. *Journal of Global Optimization*, 87(2):709–740, 2023c.

Xidong Wu, Jianhui Sun, Zhengmian Hu, Aidong Zhang, and Heng Huang. Solving a class of non-convex minimax optimization in federated learning. *Advances in Neural Information Processing Systems*, 36:11232–11245, 2023.

Zi Xu, Zi-Qi Wang, Jun-Lin Wang, and Yu-Hong Dai. Zeroth-order alternating gradient descent ascent algorithms for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2211.13668*, 2023.

Zi Xu, Ziqi Wang, Jingjing Shen, and Yuhong Dai. Derivative-free alternating projection algorithms for general nonconvex-concave minimax problems. *SIAM Journal on Optimization*, 34(2):1879–1908, 2024.

Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *Advances in Neural Information Processing Systems*, 33:5789–5800, 2020.

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, 2020a.

Junchi Yang, Siqi Zhang, Negar Kiyavash, and Niao He. A catalyst framework for minimax optimization. *Advances in Neural Information Processing Systems*, 33:5667–5678, 2020b.

Yifan Yang, Zhaofeng Si, Siwei Lyu, and Kaiyi Ji. First-order minimax bilevel optimization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 24990–25035. Curran Associates, Inc., 2024. doi: 10.52202/079017-0788.

Zhenhuan Yang, Shu Hu, Yunwen Lei, Kush R Varshney, Siwei Lyu, and Yiming Ying. Differentially private sgda for minimax problems. In *Uncertainty in Artificial Intelligence*, pp. 2192–2202. PMLR, 2022.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pp. 11173–11182. PMLR, 2020.

Lulu Zhao, Yue Liu, and Yong-Jin Liu. Accelerated stochastic alternating mirror descent ascent algorithm for nonconvex-strongly concave minimax problems. *Journal of Applied Mathematics and Computing*, 71(5):6701–6725, 2025.

A Structural Distinction from Prior Work

Our framework combines nonsmooth outer optimization with a smooth inner objective satisfying a uniform projected-PŁ condition. This condition allows nonconcavity, whereas the closest non-composite nonsmooth minimax guarantees discussed in the main text assume inner concavity or strong concavity. The oracle model is asymmetric: the outer update uses stochastic function values, and the inner update uses stochastic gradients. The analysis controls the projected mapping of $\nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y})$ together with the original inner projected-gradient mapping, yielding a blockwise GGSP. The outer-gradient tracking bound controls $\|\nabla \Phi_\delta(\mathbf{x}) - \nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y})\|^2$ by the averaged inner gap $\Delta_\delta(\mathbf{x}, \mathbf{y})$. It both closes the SPIDER-based Lyapunov argument and transfers these mapping bounds to projected Goldstein stationarity of Φ. This transfer relies on the particular smoothed gradient, which provides a stronger quantitative signal than the minimum residual used in the blockwise GGSP definition.

B Proof Roadmap

The technical proofs are organized as a dependency chain from the inner projected-PŁ geometry to the final oracle-complexity guarantees. The main steps are:

1. **Inner projected-PŁ geometry.** Lemma 4.5 shows that the projected-PŁ condition implies quadratic growth of the inner problem. Lemma 4.6 separately averages the projected-PŁ inequality over outer perturbations to bound the averaged inner gap, with an $\mathcal{O}(\delta^2)$ error.
2. **Smoothing and envelope regularity.** Lemma 3.5 derives a common full-measure differentiability set and cross-variable gradient regularity from Assumption 3.1. Lemma 4.4 establishes the differentiability, Lipschitz properties, and Goldstein-subdifferential inclusion for the randomized smoothing. Lemma D.2 gives smoothness of the smoothed value function and places its gradient in the Goldstein subdifferential, while Lemma D.3 provides the envelope identity at differentiability points.
3. **Tracking the outer gradient.** Combining Lemma D.3 with Lemma 4.5 yields the outer-gradient tracking bound in Lemma D.4. Corollary D.5 converts the inner projected-gradient residual into the corresponding tracking estimate for the smoothed problem. Lemma D.6 then bounds the projected value-function Goldstein residual using the smoothed outer mapping and the original inner mapping.
4. **Estimator and projection estimates.** The SZO estimator properties are established in Lemma 5.3 and Lemma D.1. The joint lower model is given in Corollary D.7, the projection identities in Lemma D.8, and the within-epoch and global SPIDER error bounds in Lemma D.9 and Corollary D.10.
5. **Lyapunov descent.** Lemma D.11 combines the smoothness bounds, projection relations, outer-gradient tracking, and estimator errors to obtain a one-step descent inequality for the Lyapunov potential.
6. **Main guarantees.** Summing the one-step inequality and absorbing the SPIDER errors proves Theorem E.1. The expected and high-probability blockwise GGSP statements follow in Corollary 5.5 and Corollary 5.6, while Corollary 5.7 gives the value-function guarantee. The oracle-complexity bounds and their simultaneous arbitrary-confidence version are proved in Corollary 5.8 and Corollary E.2.

C Proofs for Problem Geometry and Smoothing

C.1 Proof of Lemma 3.5

Proof. Write $U := \mathrm{int}(\mathcal{X}_{\bar{\delta}})$. Since $\bar{\delta} > 0$ and $\mathcal{X}$ is nonempty, closed, and convex, $\mathcal{X}_{\bar{\delta}}$ is a closed convex set with nonempty interior, and its boundary has Lebesgue measure zero. It therefore suffices to construct a common full-measure set in U.

For any $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}$, define $\psi_{\mathbf{y},\mathbf{y}'}(\mathbf{x}) := f(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}')$. Convexity of $\mathcal{Y}$ and continuous differentiability in y give

$$\psi_{\mathbf{y},\mathbf{y}'}(\mathbf{x}) = \int_0^1 \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}' + t(\mathbf{y} - \mathbf{y}')), \mathbf{y} - \mathbf{y}' \right\rangle \mathrm{d}t.$$

By Assumption 3.1(4), for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}_{\bar{\delta}}$,

$$|\psi_{\mathbf{y},\mathbf{y}'}(\mathbf{x}) - \psi_{\mathbf{y},\mathbf{y}'}(\mathbf{x}')| \le L_{fy} \|\mathbf{y} - \mathbf{y}'\| \, \|\mathbf{x} - \mathbf{x}'\|. \tag{12}$$

Choose a countable dense subset $D \subseteq \mathcal{Y}$. For each $\boldsymbol{q} \in D$, Rademacher's theorem gives a measurable full-measure set $U_{\boldsymbol{q}} \subseteq U$ on which $f(\cdot, \boldsymbol{q})$ is differentiable. Their countable intersection $\mathcal{X}_{\mathrm{diff}} := \bigcap_{\boldsymbol{q} \in D} U_{\boldsymbol{q}}$ is measurable and has full measure in U, hence also in $\mathcal{X}_{\bar{\delta}}$. Fix $\mathbf{x} \in \mathcal{X}_{\mathrm{diff}}$. Differentiating Eqn. (12) at this interior point for $\boldsymbol{q}, \boldsymbol{q}' \in D$ yields

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}, \boldsymbol{q}) - \nabla_{\mathbf{x}} f(\mathbf{x}, \boldsymbol{q}')\| \le L_{fy} \|\boldsymbol{q} - \boldsymbol{q}'\|.$$

Thus $\boldsymbol{q} \mapsto \nabla_{\mathbf{x}} f(\mathbf{x}, \boldsymbol{q})$ has a unique L_{fy}-Lipschitz extension $H_{\mathbf{x}} : \mathcal{Y} \to \mathbb{R}^{d_1}$.

To identify this extension with the derivative, fix any $\mathbf{y} \in \mathcal{Y}$ and $\boldsymbol{q} \in D$. For sufficiently small $\boldsymbol{h}$, $\mathbf{x} + \boldsymbol{h} \in U$, and Eqn. (12) gives

$$\begin{aligned} &|f(\mathbf{x} + \boldsymbol{h}, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}) - \langle H_{\mathbf{x}}(\mathbf{y}), \boldsymbol{h} \rangle| \\ &\quad \le |f(\mathbf{x} + \boldsymbol{h}, \boldsymbol{q}) - f(\mathbf{x}, \boldsymbol{q}) - \langle \nabla_{\mathbf{x}} f(\mathbf{x}, \boldsymbol{q}), \boldsymbol{h} \rangle| + 2L_{fy} \|\mathbf{y} - \boldsymbol{q}\| \, \|\boldsymbol{h}\|. \end{aligned}$$

Divide by $\|\boldsymbol{h}\|$ and let $\boldsymbol{h} \to 0$ with $\boldsymbol{q}$ fixed. The limsup is at most $2L_{fy}\|\mathbf{y} - \boldsymbol{q}\|$, which tends to zero as $\boldsymbol{q} \to \mathbf{y}$ in D. Therefore $f(\cdot, \mathbf{y})$ is differentiable at $\mathbf{x}$ and $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) = H_{\mathbf{x}}(\mathbf{y})$. Since $\mathbf{x}$ and $\mathbf{y}$ were arbitrary, the same set works for every $\mathbf{y}$, and the Lipschitz bound for $H_{\mathbf{x}}$ proves the claimed cross-variable bound with $L_{xy} = L_{fy}$. □

Lemma C.1 (Measurability of the Goldstein residuals)**.** *Fix $0 < \delta < \bar{\delta}$. The correspondences*

$$(\mathbf{x}, \mathbf{y}) \mapsto \partial_{\delta,\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \quad \text{on } \mathcal{X} \times \mathcal{Y}, \qquad \mathbf{x} \mapsto \partial_{\delta} \Phi(\mathbf{x}) \quad \text{on } \mathcal{X}$$

have nonempty compact values and are Borel measurable. Consequently, the minimum residuals in Definitions 4.2 and 4.3 are Borel functions of their arguments.

Proof. We first treat the partial Clarke subdifferential. At every $\boldsymbol{s} \in \operatorname{int}(\mathcal{X}_{\bar{\delta}})$, the standard gradient-limit characterization of the Clarke subdifferential may be formed using the common full-measure set $\mathcal{X}_{\text{diff}}$ from Lemma 3.5. The L_{xy} bound on that set therefore gives

$$d_{\mathrm{H}}(\partial_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y}), \partial_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y}')) \le L_{xy}\|\mathbf{y} - \mathbf{y}'\|, \tag{13}$$

where d_{H} is the Hausdorff distance. For fixed $\mathbf{y}$, the Clarke subdifferential has nonempty compact values and a closed graph in the $\boldsymbol{s}$ variable. Combining this fact with Eqn. (13) shows that

$$(\boldsymbol{s}, \mathbf{y}) \mapsto \partial_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y})$$

has a closed graph on $\operatorname{int}(\mathcal{X}_{\bar{\delta}}) \times \mathcal{Y}$. Indeed, for $(\boldsymbol{s}_n, \mathbf{y}_n, \boldsymbol{h}_n) \to (\boldsymbol{s}, \mathbf{y}, \boldsymbol{h})$ with $\boldsymbol{h}_n \in \partial_{\mathbf{x}} f(\boldsymbol{s}_n, \mathbf{y}_n)$, Eqn. (13) provides $\widetilde{\boldsymbol{h}}_n \in \partial_{\mathbf{x}} f(\boldsymbol{s}_n, \mathbf{y})$ with $\|\boldsymbol{h}_n - \widetilde{\boldsymbol{h}}_n\| \to 0$, and closedness at the fixed value $\mathbf{y}$ gives $\boldsymbol{h} \in \partial_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y})$. All these subgradients lie in the closed ball of radius L.

For $\mathbf{x} \in \mathcal{X}$, the radius margin gives $\mathbb{B}_{\delta}(\mathbf{x}) \subset \operatorname{int}(\mathcal{X}_{\bar{\delta}})$. The graph just established, compactness of $\mathbb{B}_{\delta}(\mathbf{x})$, and the uniform L bound imply that

$$\bigcup_{\boldsymbol{s} \in \mathbb{B}_{\delta}(\mathbf{x})} \partial_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y})$$

is compact. Its convex hull is compact in finite dimensions. To see that the resulting Goldstein correspondence has a closed graph, consider a convergent sequence of points in its graph and represent every subgradient by at most $d_1 + 1$ terms using Carathéodory's theorem. Compactness of the sampling balls, the simplex of convex weights, and the closed graph above yield a convergent subsequence whose limit is a valid representation at the limiting $(\mathbf{x}, \mathbf{y})$. Thus $(\mathbf{x}, \mathbf{y}) \mapsto \partial_{\delta,\mathbf{x}} f(\mathbf{x}, \mathbf{y})$ has a closed graph and compact values contained in the radius-L ball. It is consequently upper semicontinuous and Borel measurable.

The value function Φ is L-Lipschitz on $\mathcal{X}_{\bar{\delta}}$. Its Clarke subdifferential therefore has a closed graph, nonempty compact values, and the same uniform bound on $\operatorname{int}(\mathcal{X}_{\bar{\delta}})$. Since every relevant sampling ball lies in this open set, the preceding sampling-ball and Carathéodory argument shows that $\mathbf{x} \mapsto \partial_{\delta} \Phi(\mathbf{x})$ is compact-valued and Borel measurable. The projected residuals are continuous in the base point and candidate subgradient. The measurable minimum theorem applied to the two compact-valued correspondences proves the final claim. □

C.2 Proof of Lemma 4.5

Proof. For brevity, write $\lambda := \lambda_{\text{qg}}$ throughout this proof.

Define the optimality gap

$$\Delta(\mathbf{x}, \mathbf{y}) := \Phi(\mathbf{x}) - f(\mathbf{x}, \mathbf{y}) \ge 0. \tag{14}$$

Starting from an arbitrary $\mathbf{y}_0 \in \mathcal{Y}$, consider the projected gradient-ascent sequence

$$\mathbf{y}_{k+1} = \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_k + \lambda \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_k)), \qquad k \ge 0. \tag{15}$$

Define

$$\mathcal{G}_k := \mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}, \mathbf{y}_k, \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_k)) = \frac{\mathbf{y}_{k+1} - \mathbf{y}_k}{\lambda}. \tag{16}$$

Hence,

$$\mathbf{y}_{k+1} - \mathbf{y}_k = \lambda \mathcal{G}_k. \tag{17}$$

By the characterization of Euclidean projection, if

$$\mathbf{y}_{k+1} = \mathcal{P}_{\mathcal{Y}}(\mathbf{y}_k + \lambda \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_k)), \tag{18}$$

then

$$\langle \mathbf{y}_k + \lambda \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_k) - \mathbf{y}_{k+1}, \mathbf{z} - \mathbf{y}_{k+1} \rangle \leq 0, \qquad \forall \mathbf{z} \in \mathcal{Y}. \tag{19}$$

Setting $\mathbf{z} = \mathbf{y}_k$ gives

$$\langle \mathbf{y}_k + \lambda \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_k) - \mathbf{y}_{k+1}, \mathbf{y}_k - \mathbf{y}_{k+1} \rangle \leq 0. \tag{20}$$

Let

$$\boldsymbol{s}_k := \mathbf{y}_{k+1} - \mathbf{y}_k. \tag{21}$$

Then we have

$$\langle -\boldsymbol{s}_k + \lambda \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_k), -\boldsymbol{s}_k \rangle \leq 0, \tag{22}$$

or equivalently,

$$\langle \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_k), \boldsymbol{s}_k \rangle \geq \frac{1}{\lambda} \|\boldsymbol{s}_k\|^2. \tag{23}$$

Since $\nabla_{\mathbf{y}} f(\mathbf{x}, \cdot)$ is L_{fy}-Lipschitz continuous and $\mathcal{Y}$ is convex, the standard smoothness inequality yields

$$f(\mathbf{x}, \mathbf{y}_{k+1}) \geq f(\mathbf{x}, \mathbf{y}_k) + \langle \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}_k), \boldsymbol{s}_k \rangle - \frac{L_{fy}}{2} \|\boldsymbol{s}_k\|^2. \tag{24}$$

Then we have

$$f(\mathbf{x}, \mathbf{y}_{k+1}) - f(\mathbf{x}, \mathbf{y}_k) \geq \left(\frac{1}{\lambda} - \frac{L_{fy}}{2} \right) \|\boldsymbol{s}_k\|^2. \tag{25}$$

Because $\lambda \leq 1/L_{fy}$,

$$\frac{1}{\lambda} - \frac{L_{fy}}{2} \geq \frac{1}{2\lambda}.$$

Therefore,

$$f(\mathbf{x}, \mathbf{y}_{k+1}) - f(\mathbf{x}, \mathbf{y}_k) \geq \frac{1}{2\lambda} \|\boldsymbol{s}_k\|^2. \tag{26}$$

Using $\boldsymbol{s}_k = \lambda \mathcal{G}_k$, we obtain

$$f(\mathbf{x}, \mathbf{y}_{k+1}) - f(\mathbf{x}, \mathbf{y}_k) \geq \frac{\lambda}{2} \|\mathcal{G}_k\|^2. \tag{27}$$

Equivalently, in terms of the optimality gap,

$$\Delta(\mathbf{x}, \mathbf{y}_{k+1}) \leq \Delta(\mathbf{x}, \mathbf{y}_k) - \frac{\lambda}{2} \|\mathcal{G}_k\|^2. \tag{28}$$

Applying Assumption 3.2 at $\mathbf{y}_k$ gives

$$\frac{1}{2}\|\mathcal{G}_k\|^2 \geq \mu\Delta(\mathbf{x}, \mathbf{y}_k). \tag{29}$$

Combining the two inequalities gives

$$\Delta(\mathbf{x}, \mathbf{y}_{k+1}) \leq \Delta(\mathbf{x}, \mathbf{y}_k) - \lambda\mu\Delta(\mathbf{x}, \mathbf{y}_k) = (1 - \lambda\mu)\Delta(\mathbf{x}, \mathbf{y}_k). \tag{30}$$

Let $\vartheta := 1 - \lambda\mu$. We have $\frac{1}{2} \leq \vartheta < 1$. Thus

$$\Delta(\mathbf{x}, \mathbf{y}_k) \leq \vartheta^k \Delta(\mathbf{x}, \mathbf{y}_0), \qquad k \geq 0. \tag{31}$$

Therefore, if $k \longrightarrow \infty$, we have $\Delta(\mathbf{x}, \mathbf{y}_k) \longrightarrow 0$.

From Eqn. (28) and the nonnegativity of $\Delta(\mathbf{x}, \mathbf{y}_{k+1})$,

$$\frac{\lambda}{2}\|\mathcal{G}_k\|^2 \leq \Delta(\mathbf{x}, \mathbf{y}_k). \tag{32}$$

Therefore,

$$\|\mathbf{y}_{k+1} - \mathbf{y}_k\| = \lambda\|\mathcal{G}_k\| \leq \sqrt{2\lambda\,\Delta(\mathbf{x}, \mathbf{y}_k)}. \tag{33}$$

Using Eqn. (31), we have

$$\|\mathbf{y}_{k+1} - \mathbf{y}_k\| \leq \sqrt{2\lambda\,\Delta(\mathbf{x}, \mathbf{y}_0)}\,\vartheta^{k/2}. \tag{34}$$

Consequently, we have

$$\sum_{k=0}^{\infty} \|\mathbf{y}_{k+1} - \mathbf{y}_k\| \leq \sqrt{2\lambda\,\Delta(\mathbf{x}, \mathbf{y}_0)} \sum_{k=0}^{\infty} \vartheta^{k/2} = \frac{\sqrt{2\lambda\,\Delta(\mathbf{x}, \mathbf{y}_0)}}{1 - \sqrt{\vartheta}} < \infty. \tag{35}$$

Hence $\{\mathbf{y}_k\}$ is a Cauchy sequence. Since $\mathcal{Y}$ is closed and complete, there exists some $\bar{\mathbf{y}} \in \mathcal{Y}$ such that $\mathbf{y}_k \to \bar{\mathbf{y}}$.

Since $f(\mathbf{x}, \cdot)$ is continuously differentiable, by $\Delta(\mathbf{x}, \mathbf{y}_k) \longrightarrow 0.$, we have

$$\Phi(\mathbf{x}) - f(\mathbf{x}, \bar{\mathbf{y}}) = \lim_{k\to\infty} \big[\Phi(\mathbf{x}) - f(\mathbf{x}, \mathbf{y}_k)\big] = 0. \tag{36}$$

Therefore, we have $f(\mathbf{x}, \bar{\mathbf{y}}) = \Phi(\mathbf{x})$, and hence

$$\bar{\mathbf{y}} \in \mathcal{Y}^\star(\mathbf{x}). \tag{37}$$

Since $\bar{\mathbf{y}} \in \mathcal{Y}^\star(\mathbf{x})$, we have

$$\mathrm{dist}(\mathbf{y}_0, \mathcal{Y}^\star(\mathbf{x})) \leq \|\mathbf{y}_0 - \bar{\mathbf{y}}\|. \tag{38}$$

Using the triangle inequality we can obtain,

$$\|\mathbf{y}_0 - \bar{\mathbf{y}}\| \leq \sum_{k=0}^{\infty} \|\mathbf{y}_{k+1} - \mathbf{y}_k\| \leq \frac{\sqrt{2\lambda\,\Delta(\mathbf{x}, \mathbf{y}_0)}}{1 - \sqrt{\vartheta}}. \tag{39}$$

Recalling that $\vartheta = 1 - \lambda\mu$, we obtain

$$\mathrm{dist}(\mathbf{y}_0, \mathcal{Y}^\star(\mathbf{x})) \leq \frac{\sqrt{2\lambda}}{1 - \sqrt{1 - \lambda\mu}} \sqrt{\Delta(\mathbf{x}, \mathbf{y}_0)}. \tag{40}$$

Squaring both sides gives

$$\mathrm{dist}^2(\mathbf{y}_0, \mathcal{Y}^\star(\mathbf{x})) \leq \frac{2\lambda}{\big(1 - \sqrt{1 - \lambda\mu}\big)^2} \big[\Delta(\mathbf{x}, \mathbf{y}_0)\big]. \tag{41}$$

Since $\mathbf{y}_0 \in \mathcal{Y}$ was arbitrary, the result holds for every $\mathbf{y} \in \mathcal{Y}$, with

$$C_{\mathrm{qg}} = \frac{2\lambda}{\big(1 - \sqrt{1 - \lambda\mu}\big)^2}.$$

This completes the proof. □

C.3 Proof of Lemma 4.4

Proof. Throughout the proof, all displayed outer points $\mathbf{x}, \mathbf{x}'$ lie in $\mathcal{X}$ and all inner points $\mathbf{y}, \mathbf{y}'$ lie in $\mathcal{Y}$. In particular, every perturbed point used below belongs to $\mathbb{B}_\delta(\mathbf{x}) \subseteq \mathcal{X}_\delta \subset \mathrm{int}(\mathcal{X}_{\bar{\delta}})$ (and likewise for $\mathbf{x}'$), where the last inclusion uses $\delta < \bar{\delta}$.

For the first statement, since $f(\mathbf{x}, \mathbf{y})$ is L-Lipschitz, we have

$$|f_\delta(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}, \mathbf{y})| = \left|\mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^{d_1})}[f(\mathbf{x} + \delta\boldsymbol{u}, \mathbf{y})] - f(\mathbf{x}, \mathbf{y})\right| \le \delta L \mathbb{E}[\|\boldsymbol{u}\|] \le \delta L \tag{42}$$

For the second statement, joint L-Lipschitz continuity gives

$$\begin{aligned} |f_\delta(\mathbf{x}, \mathbf{y}) - f_\delta(\mathbf{x}', \mathbf{y}')| &\le \mathbb{E}_{\boldsymbol{u}}|f(\mathbf{x} + \delta\boldsymbol{u}, \mathbf{y}) - f(\mathbf{x}' + \delta\boldsymbol{u}, \mathbf{y}')| && (43)\\ &\le L\|(\mathbf{x}, \mathbf{y}) - (\mathbf{x}', \mathbf{y}')\|. && (44)\end{aligned}$$

For fixed $\mathbf{y}$, the restriction of $f(\cdot, \mathbf{y})$ to $\mathcal{X}_{\bar{\delta}}$ is L-Lipschitz by Assumption 3.1(3), and Assumption 3.1(1) gives local Lipschitzness on an open neighborhood needed for the convolution theorem. Since every sampling ball $\mathbb{B}_\delta(\mathbf{x})$ lies in $\mathcal{X}_{\bar{\delta}}$, the convolution and its boundary representation use only the stated values of f. Rademacher's theorem and weak differentiation of convolution then give

$$\nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\boldsymbol{u}}[\nabla_{\mathbf{x}} f(\mathbf{x} + \delta\boldsymbol{u}, \mathbf{y})], \tag{45}$$

where the integrand is the almost-everywhere classical gradient. For fixed $\mathbf{y}$, this gradient is measurable as the a.e. derivative of the locally Lipschitz function $f(\cdot, \mathbf{y})$; joint measurability in $(\mathbf{x}, \mathbf{y})$ is not needed here. The same convolution also satisfies the boundary identity

$$\nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y}) = \frac{d_1}{\delta}\mathbb{E}_{\boldsymbol{v}\sim\mathrm{Unif}(\mathbb{S}^{d_1-1})}[f(\mathbf{x} + \delta\boldsymbol{v}, \mathbf{y})\boldsymbol{v}].$$

For $\mathbf{x}_1, \mathbf{x}_2$ and a unit vector $\boldsymbol{q}$,

$$\begin{aligned} &|\langle\nabla_{\mathbf{x}} f_\delta(\mathbf{x}_1, \mathbf{y}) - \nabla_{\mathbf{x}} f_\delta(\mathbf{x}_2, \mathbf{y}), \boldsymbol{q}\rangle| \\ &\le \frac{d_1}{\delta}\mathbb{E}_{\boldsymbol{v}}[|f(\mathbf{x}_1 + \delta\boldsymbol{v}, \mathbf{y}) - f(\mathbf{x}_2 + \delta\boldsymbol{v}, \mathbf{y})||\langle\boldsymbol{v}, \boldsymbol{q}\rangle|] \\ &\le \frac{d_1 L}{\delta}\|\mathbf{x}_1 - \mathbf{x}_2\|\mathbb{E}|v_1|. \end{aligned}$$

Taking the supremum over unit $\boldsymbol{q}$ gives the stated $L_{x,\delta}$ bound. Also $\mathbb{E}|v_1| \le \sqrt{\mathbb{E}v_1^2} = 1/\sqrt{d_1}$.

For the third result, use Eqn. (45). Since $\mathcal{X}_{\mathrm{diff}}$ has full measure, the integrand is evaluated in $\mathcal{X}_{\mathrm{diff}}$ almost surely. Hence, for arbitrary $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$,

$$\begin{aligned} &\|\nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y}_1) - \nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y}_2)\| \\ &\le \mathbb{E}_{\boldsymbol{u}}\|\nabla_{\mathbf{x}} f(\mathbf{x} + \delta\boldsymbol{u}, \mathbf{y}_1) - \nabla_{\mathbf{x}} f(\mathbf{x} + \delta\boldsymbol{u}, \mathbf{y}_2)\| \\ &\le L_{xy}\|\mathbf{y}_1 - \mathbf{y}_2\|, \end{aligned}$$

where the last inequality follows from Lemma 3.5.

For the fourth statement, dominated differentiation gives $\nabla_{\mathbf{y}} f_\delta(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\boldsymbol{u}}[\nabla_{\mathbf{y}} f(\mathbf{x} + \delta\boldsymbol{u}, \mathbf{y})]$. Jensen's inequality and Assumption 3.1(4) then give

$$\begin{aligned} &\|\nabla_{\mathbf{y}} f_\delta(\mathbf{x}_1, \mathbf{y}_1) - \nabla_{\mathbf{y}} f_\delta(\mathbf{x}_2, \mathbf{y}_2)\| && (46)\\ &\quad\le \mathbb{E}_{\boldsymbol{u}}\|\nabla_{\mathbf{y}} f(\mathbf{x}_1 + \delta\boldsymbol{u}, \mathbf{y}_1) - \nabla_{\mathbf{y}} f(\mathbf{x}_2 + \delta\boldsymbol{u}, \mathbf{y}_2)\| && (47)\\ &\quad\le L_{fy}\|(\mathbf{x}_1, \mathbf{y}_1) - (\mathbf{x}_2, \mathbf{y}_2)\|. && (48)\end{aligned}$$

Moreover,

$$\begin{aligned} \|\nabla_{\mathbf{y}} f_\delta(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\| &\le \mathbb{E}_{\boldsymbol{u}}\|\nabla_{\mathbf{y}} f(\mathbf{x} + \delta\boldsymbol{u}, \mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\| \\ &\le L_{fy}\delta\mathbb{E}\|\boldsymbol{u}\| \le L_{fy}\delta. \end{aligned}$$

Finally, each almost-everywhere gradient in Eqn. (45) lies in the Clarke subdifferential at its sampling point (the sampled boundary has measure zero). By Assumption 3.1(1), the Clarke subdifferential is

well-defined for the locally Lipschitz outer slice. Local boundedness and outer semicontinuity of the Clarke subdifferential on the compact ball imply that

$$\operatorname{conv}\left(\bigcup_{\boldsymbol{s}\in\mathbb{B}_\delta(\mathbf{x})}\partial_{\mathbf{x}} f(\boldsymbol{s},\mathbf{y})\right)$$

is compact and hence closed. The expectation in Eqn. (45) belongs to this convex set, proving $\nabla_{\mathbf{x}} f_\delta(\mathbf{x},\mathbf{y}) \in \partial_{\delta,\mathbf{x}} f(\mathbf{x},\mathbf{y})$.

□

D Technical Lemmas for the Convergence Analysis

For the bounds below, set $L_{x,\delta} = L_{\Phi,\delta} := L\sqrt{d_1}/\delta$ and $M_{x,\delta} := d_1 L/\delta$; the single oracle–direction variance constant $\bar{\sigma}_x^2 = 2d_1L^2$ is given in Lemma 5.3. For the technical estimates, let $e_t^x := w_t - \nabla_{\mathbf{x}} f_\delta(\mathbf{x}_t,\mathbf{y}_t)$ and $e_t^y := v_t - \nabla_{\mathbf{y}} f(\mathbf{x}_t,\mathbf{y}_t)$ denote the estimator errors, and let $\widetilde{\mathcal{G}}_{\mathbf{x},t} := (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma$ and $\widetilde{\mathcal{G}}_{\mathbf{y},t} := (\mathbf{y}_{t+1} - \mathbf{y}_t)/\lambda$ denote the computed projected mappings. These quantities are used only in the proofs below. We also write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$ for conditional expectation under Assumption 5.1.

For the Lyapunov estimates, set

$$C_r := L_{xy}^2 C_{\mathrm{qg}}, \qquad \chi := \max\{2, 512C_r/\mu\}, \qquad a := \frac{\chi\gamma}{\lambda},$$

$$C_e := 4 + 4\chi + 32C_r/\mu, \qquad C_\delta := L_{fy}^2(4\chi + 16C_r/\mu), \qquad \Psi_t := \Phi_\delta(\mathbf{x}_t) + a\Delta_\delta(\mathbf{x}_t,\mathbf{y}_t).$$

D.1 Proof of Lemma 5.3: SZO estimator variance

We prove Lemma 5.3.

Proof. By symmetry of the uniform sphere distribution,

$$\begin{aligned}\mathbb{E}_{\xi,\hat{\boldsymbol{u}}}[g_{\mathbf{x}}(\mathbf{z};\xi,\hat{\boldsymbol{u}})] &= \mathbb{E}_{\xi,\hat{\boldsymbol{u}}}\left[\frac{d_1}{2\delta}(f(\mathbf{x}+\delta\hat{\boldsymbol{u}},\mathbf{y};\xi) - f(\mathbf{x}-\delta\hat{\boldsymbol{u}},\mathbf{y};\xi))\hat{\boldsymbol{u}}\right] \\ &= \mathbb{E}_{\xi,\hat{\boldsymbol{u}}}\left[\frac{d_1}{\delta} f(\mathbf{x}+\delta\hat{\boldsymbol{u}},\mathbf{y};\xi)\hat{\boldsymbol{u}}\right] \\ &= \frac{d_1}{\delta}\mathbb{E}_{\hat{\boldsymbol{u}}}[f(\mathbf{x}+\delta\hat{\boldsymbol{u}},\mathbf{y})\hat{\boldsymbol{u}}] = \nabla_{\mathbf{x}} f_\delta(\mathbf{x},\mathbf{y}),\end{aligned}$$

which proves the unbiasedness claim.

We next recall a standard spherical Poincaré inequality: if $h : \mathbb{S}^{d_1-1} \to \mathbb{R}$ is L_h-Lipschitz and $d_1 \geq 2$, then

$$\operatorname{Var}(h(\hat{\boldsymbol{u}})) \leq \frac{L_h^2}{d_1 - 1}.$$

For fixed ξ, let $h_\xi(\hat{\boldsymbol{u}}) := f(\mathbf{x}+\delta\hat{\boldsymbol{u}},\mathbf{y};\xi)$. By Assumption 3.4, h_ξ is $\delta L(\xi)$-Lipschitz. Since $\hat{\boldsymbol{u}}$ and $-\hat{\boldsymbol{u}}$ have the same distribution,

$$\mathbb{E}_{\hat{\boldsymbol{u}}}|h_\xi(\hat{\boldsymbol{u}}) - h_\xi(-\hat{\boldsymbol{u}})|^2 \leq 4\operatorname{Var}_{\hat{\boldsymbol{u}}}(h_\xi(\hat{\boldsymbol{u}})) \leq \frac{4\delta^2 L(\xi)^2}{d_1 - 1}.$$

Consequently, for $d_1 \geq 2$,

$$\mathbb{E}_{\xi,\hat{\boldsymbol{u}}}\|g_{\mathbf{x}}(\mathbf{z};\xi,\hat{\boldsymbol{u}})\|^2 = \frac{d_1^2}{4\delta^2}\mathbb{E}_{\xi,\hat{\boldsymbol{u}}}|h_\xi(\hat{\boldsymbol{u}}) - h_\xi(-\hat{\boldsymbol{u}})|^2 \leq \frac{d_1^2}{d_1 - 1}\mathbb{E}[L(\xi)^2] \leq 2d_1L^2.$$

For $d_1 = 1$, the two-point difference is bounded directly by $L(\xi)$, so the same order holds. In particular,

$$\mathbb{E}\|g_{\mathbf{x}}(\mathbf{z};\xi,\hat{\boldsymbol{u}})\|^2 \leq 2d_1L^2 = \bar{\sigma}_x^2, \tag{49}$$

which is the dimension-dependent constant used in the manuscript.

For the M independent oracle–direction evaluations $g_i = g_{\mathbf{x}}(\mathbf{z}; \xi_i, \hat{\boldsymbol{u}}_i)$,

$$\begin{aligned}
&\mathbb{E}\left\| \frac{1}{M}\sum_{i=1}^{M} g_i - \nabla_{\mathbf{x}} f_\delta(\mathbf{z}) \right\|^2 \\
&= \frac{1}{M^2}\sum_{i=1}^{M} \mathbb{E}\|g_i - \nabla_{\mathbf{x}} f_\delta(\mathbf{z})\|^2 \\
&\le \frac{1}{M}\mathbb{E}\|g_{\mathbf{x}}(\mathbf{z}; \xi, \hat{\boldsymbol{u}})\|^2 \le \frac{\bar{\sigma}_x^2}{M},
\end{aligned}$$

because the centered cross terms vanish. This establishes the batch variance bound. □

D.2 Mean-square continuity of the SZO estimator

Lemma D.1 (Mean-square continuity of the ZO estimator)**.** *Under Assumption 3.4, if the same pair* $(\xi, \hat{\boldsymbol{u}})$ *is used at* $\mathbf{z}$ *and* $\mathbf{z}'$*, then*

$$\mathbb{E}_{\xi,\hat{\boldsymbol{u}}}\|g_{\mathbf{x}}(\mathbf{z}; \xi, \hat{\boldsymbol{u}}) - g_{\mathbf{x}}(\mathbf{z}'; \xi, \hat{\boldsymbol{u}})\|^2 \le M_{x,\delta}^2 \|\mathbf{z} - \mathbf{z}'\|^2, \qquad M_{x,\delta} := \frac{d_1 L}{\delta}. \tag{50}$$

Proof. Let $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ and $\mathbf{z}' = (\mathbf{x}', \mathbf{y}')$. For fixed $(\xi, \hat{\boldsymbol{u}})$, define

$$\begin{aligned}
D_+(\xi, \hat{\boldsymbol{u}}) &:= f(\mathbf{x} + \delta\hat{\boldsymbol{u}}, \mathbf{y}; \xi) - f(\mathbf{x}' + \delta\hat{\boldsymbol{u}}, \mathbf{y}'; \xi), \\
D_-(\xi, \hat{\boldsymbol{u}}) &:= f(\mathbf{x} - \delta\hat{\boldsymbol{u}}, \mathbf{y}; \xi) - f(\mathbf{x}' - \delta\hat{\boldsymbol{u}}, \mathbf{y}'; \xi).
\end{aligned}$$

Then

$$g_{\mathbf{x}}(\mathbf{z}; \xi, \hat{\boldsymbol{u}}) - g_{\mathbf{x}}(\mathbf{z}'; \xi, \hat{\boldsymbol{u}}) = \frac{d_1}{2\delta}[D_+(\xi, \hat{\boldsymbol{u}}) - D_-(\xi, \hat{\boldsymbol{u}})]\hat{\boldsymbol{u}}.$$

By Assumption 3.4,

$$|D_+(\xi, \hat{\boldsymbol{u}})| \le L(\xi)\|\mathbf{z} - \mathbf{z}'\|, \qquad |D_-(\xi, \hat{\boldsymbol{u}})| \le L(\xi)\|\mathbf{z} - \mathbf{z}'\|.$$

Since $\|\hat{\boldsymbol{u}}\| = 1$,

$$\begin{aligned}
\|g_{\mathbf{x}}(\mathbf{z}; \xi, \hat{\boldsymbol{u}}) - g_{\mathbf{x}}(\mathbf{z}'; \xi, \hat{\boldsymbol{u}})\| &\le \frac{d_1}{2\delta}(|D_+| + |D_-|) \\
&\le \frac{d_1 L(\xi)}{\delta}\|\mathbf{z} - \mathbf{z}'\|.
\end{aligned}$$

Squaring and taking expectation gives

$$\mathbb{E}\|g_{\mathbf{x}}(\mathbf{z}; \xi, \hat{\boldsymbol{u}}) - g_{\mathbf{x}}(\mathbf{z}'; \xi, \hat{\boldsymbol{u}})\|^2 \le \frac{d_1^2 L^2}{\delta^2}\|\mathbf{z} - \mathbf{z}'\|^2,$$

which is Eqn. (50). □

D.3 Regularity of the smoothed value function

Lemma D.2 (Value-function smoothing and Goldstein inclusion)**.** *Under Assumption 3.1,* Φ *is* L*-Lipschitz on* $\mathcal{X}_{\bar{\delta}}$*,* $\nabla\Phi_\delta(\mathbf{x}) \in \partial_\delta\Phi(\mathbf{x})$ *for every* $\mathbf{x} \in \mathcal{X}$*, and, for every* $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$*,*

$$\|\nabla\Phi_\delta(\mathbf{x}_1) - \nabla\Phi_\delta(\mathbf{x}_2)\| \le \frac{L\sqrt{d_1}}{\delta}\|\mathbf{x}_1 - \mathbf{x}_2\|. \tag{51}$$

Proof. For $\boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{X}_{\bar{\delta}}$, choose $\mathbf{y}_1^\star \in \mathcal{Y}^\star(\boldsymbol{s}_1)$. Then

$$\Phi(\boldsymbol{s}_1) - \Phi(\boldsymbol{s}_2) \le f(\boldsymbol{s}_1, \mathbf{y}_1^\star) - f(\boldsymbol{s}_2, \mathbf{y}_1^\star) \le L\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|. \tag{52}$$

Interchanging $\boldsymbol{s}_1, \boldsymbol{s}_2$ shows that Φ is L-Lipschitz on $\mathcal{X}_{\bar{\delta}}$. Choose any $r \in (0, \bar{\delta}-\delta)$. If $\operatorname{dist}(\mathbf{x}', \mathcal{X}) < r$, then $\mathbb{B}_\delta(\mathbf{x}') \subset \operatorname{int}(\mathcal{X}_{\bar{\delta}})$. Hence the local convolution Φ_δ is well-defined on an open neighborhood of $\mathcal{X}$, and the same boundary identity as in the proof of Lemma 4.4 gives, for $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$,

$$\|\nabla\Phi_\delta(\mathbf{x}_1) - \nabla\Phi_\delta(\mathbf{x}_2)\| \le \frac{\sqrt{d_1}L}{\delta}\|\mathbf{x}_1 - \mathbf{x}_2\|. \tag{53}$$

Rademacher's theorem and weak differentiation of convolution also give

$$\nabla\Phi_\delta(\mathbf{x}) = \mathbb{E}_{\boldsymbol{u}}[\nabla\Phi(\mathbf{x} + \delta\boldsymbol{u})],$$

where the integrand is the classical gradient at almost every sampled point. Each such gradient belongs to the Clarke subdifferential at its sampling point. On the open sampling neighborhood, the Clarke subdifferential has a closed graph and all its values lie in the closed ball of radius L. Its graph restricted to the compact sampling ball is therefore compact, as is the union of its values; in finite dimensions, the convex hull of this union is also compact. The displayed expectation therefore belongs to $\partial_\delta\Phi(\mathbf{x})$.

□

D.4 Envelope identity at differentiability points

Lemma D.3 (Envelope identity at differentiability points)**.** *Under Assumption 3.1, let $\mathcal{X}_{\mathrm{diff}}$ be the common set from Lemma 3.5. For every $\boldsymbol{s} \in \mathcal{X}_{\mathrm{diff}}$ at which Φ is differentiable,*

$$\nabla\Phi(\boldsymbol{s}) = \nabla_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y}^\star), \qquad \forall \mathbf{y}^\star \in \mathcal{Y}^\star(\boldsymbol{s}). \tag{54}$$

Proof. Fix any $\mathbf{y}^\star \in \mathcal{Y}^\star(\boldsymbol{s})$. For every direction $\boldsymbol{d}$ and all sufficiently small $t > 0$, interiority of $\boldsymbol{s}$ ensures that $\boldsymbol{s} + t\boldsymbol{d} \in \mathcal{X}_{\bar{\delta}}$, and

$$\Phi(\boldsymbol{s} + t\boldsymbol{d}) - \Phi(\boldsymbol{s}) \ge f(\boldsymbol{s} + t\boldsymbol{d}, \mathbf{y}^\star) - f(\boldsymbol{s}, \mathbf{y}^\star).$$

Dividing by t and letting $t \downarrow 0$, differentiability of Φ and $f(\cdot, \mathbf{y}^\star)$ at $\boldsymbol{s}$ gives

$$\langle \nabla\Phi(\boldsymbol{s}), \boldsymbol{d}\rangle \ge \langle \nabla_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y}^\star), \boldsymbol{d}\rangle, \qquad \forall \boldsymbol{d}.$$

Set $\boldsymbol{d} = \nabla_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y}^\star) - \nabla\Phi(\boldsymbol{s})$. The preceding inequality becomes

$$-\|\nabla\Phi(\boldsymbol{s}) - \nabla_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y}^\star)\|^2 \ge 0,$$

so the two gradients are equal. Since $\mathbf{y}^\star$ was arbitrary, the identity holds for every active maximizer. □

D.5 Outer-gradient tracking through the inner gap

Lemma D.4 (Outer-gradient tracking)**.** *Under Assumptions 3.1 and 3.2, for every $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$,*

$$\|\nabla\Phi_\delta(\mathbf{x}) - \nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y})\|^2 \le C_r \Delta_\delta(\mathbf{x}, \mathbf{y}). \tag{55}$$

Proof. Let $\boldsymbol{s} = \mathbf{x} + \delta\boldsymbol{u}$, where $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^{d_1})$. Since $\|\boldsymbol{u}\| < 1$ almost surely,

$$\mathbb{B}_{\delta(1-\|\boldsymbol{u}\|)}(\boldsymbol{s}) \subseteq \mathcal{X}_\delta,$$

and hence $\boldsymbol{s} \in \operatorname{int}(\mathcal{X}_\delta)$ almost surely. Moreover, $\mathcal{X}_{\mathrm{diff}}$ has full measure, and the locally Lipschitz function Φ is differentiable almost everywhere on $\operatorname{int}(\mathcal{X}_\delta)$. Therefore, $\boldsymbol{s} \in \mathcal{X}_{\mathrm{diff}} \cap \operatorname{int}(\mathcal{X}_\delta)$ and Φ is differentiable at $\boldsymbol{s}$ for almost every $\boldsymbol{u}$. At every such point, Lemmas D.3 and 3.5 give, for every $\mathbf{y}^\star \in \mathcal{Y}^\star(\boldsymbol{s})$,

$$\|\nabla\Phi(\boldsymbol{s}) - \nabla_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y})\| = \|\nabla_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y}^\star) - \nabla_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y})\| \le L_{xy}\|\mathbf{y}^\star - \mathbf{y}\|.$$

Taking the infimum over $\mathbf{y}^\star$, squaring, and applying Lemma 4.5 yields

$$\|\nabla\Phi(\boldsymbol{s}) - \nabla_{\mathbf{x}} f(\boldsymbol{s}, \mathbf{y})\|^2 \le L_{xy}^2 \operatorname{dist}^2(\mathbf{y}, \mathcal{Y}^\star(\boldsymbol{s})) \le C_r \Delta(\boldsymbol{s}, \mathbf{y}). \tag{56}$$

Convolution commutes with weak differentiation. Hence, by Jensen,

$$\begin{aligned}
\|\nabla\Phi_\delta(\mathbf{x}) - \nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y})\|^2 &= \|\mathbb{E}_{\boldsymbol{u}}[\nabla\Phi(\mathbf{x} + \delta\boldsymbol{u}) - \nabla_{\mathbf{x}} f(\mathbf{x} + \delta\boldsymbol{u}, \mathbf{y})]\|^2 && (57)\\
&\le \mathbb{E}_{\boldsymbol{u}}\|\nabla\Phi(\mathbf{x} + \delta\boldsymbol{u}) - \nabla_{\mathbf{x}} f(\mathbf{x} + \delta\boldsymbol{u}, \mathbf{y})\|^2 && (58)\\
&\le C_r \mathbb{E}_{\boldsymbol{u}}[\Delta(\mathbf{x} + \delta\boldsymbol{u}, \mathbf{y})] && (59)\\
&= C_r \Delta_\delta(\mathbf{x}, \mathbf{y}). && (60)
\end{aligned}$$

□

D.6 Proof of Lemma 4.6

Proof. For $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{B}^{d_1})$, let $\mathcal{G}_{\boldsymbol{u}} := \mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}+\delta\boldsymbol{u}, \mathbf{y}, \nabla_{\mathbf{y}} f(\mathbf{x}+\delta\boldsymbol{u}, \mathbf{y}))$. Projected PŁ at $\mathbf{x}+\delta\boldsymbol{u}$ gives

$$\frac{1}{2}\mathbb{E}\|\mathcal{G}_{\boldsymbol{u}}\|^2 \geq \mu\Delta_\delta(\mathbf{x},\mathbf{y}).$$

Projection nonexpansiveness and Assumption 3.1(4) imply

$$\|\mathcal{G}_{\boldsymbol{u}} - \mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x},\mathbf{y},\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}))\| \leq \|\nabla_{\mathbf{y}} f(\mathbf{x}+\delta\boldsymbol{u},\mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\| \leq L_{fy}\delta\|\boldsymbol{u}\|.$$

For any $\tau > 0$,

$$\|a+b\|^2 \leq (1+\tau)\|a\|^2 + (1+\tau^{-1})\|b\|^2.$$

Therefore

$$\mathbb{E}\|\mathcal{G}_{\boldsymbol{u}}\|^2 \leq (1+\tau)\|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x},\mathbf{y},\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}))\|^2 + (1+\tau^{-1})L_{fy}^2\delta^2\nu_{d_1}.$$

Combining the last two displays and dividing by $1+\tau$ proves the general inequality stated in Lemma 4.6. Setting $\tau = 1$ yields the particular bound given there. □

D.7 Outer-Gradient Tracking via the Projected-PŁ Residual

Let

$$\nu_{d_1} := \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(\mathbb{B}^{d_1})}\|\boldsymbol{u}\|^2 = \frac{d_1}{d_1+2}.$$

Corollary D.5 (Outer-gradient tracking)**.** *Under Assumptions 3.1 and 3.2, for every* $(\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$ *and* $0 < \lambda \leq \bar{\lambda}$*,*

$$\|\nabla\Phi_\delta(\mathbf{x}) - \nabla_{\mathbf{x}} f_\delta(\mathbf{x},\mathbf{y})\|^2 \leq \frac{C_r}{\mu}\|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x},\mathbf{y},\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}))\|^2 + \frac{C_r\nu_{d_1}L_{fy}^2}{\mu}\delta^2. \tag{61}$$

Remark. Thus the inner projected-gradient residual controls the mismatch between the smoothed outer gradient at the current inner point and the gradient of the smoothed value function, up to the $O(\delta^2)$ smoothing error.

Proof. Substitute the particular bound in Lemma 4.6 into Lemma D.4. □

Lemma D.6 (Transfer to value-function Goldstein stationarity)**.** *Let* $\rho := C_r/\mu$*. For every* $(\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{Y}$ *and* $0 < \lambda \leq \bar{\lambda}$*,*

$$\mathfrak{r}_\Phi(\mathbf{x})^2 \leq (1+\rho)\Big(\|\mathcal{G}^\delta_{\mathbf{x},\gamma}(\mathbf{x},\mathbf{y})\|^2 + \|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x},\mathbf{y},\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}))\|^2 + \nu_{d_1}L_{fy}^2\delta^2\Big). \tag{62}$$

Proof. Lemma D.2 gives $\nabla\Phi_\delta(\mathbf{x}) \in \partial_\delta\Phi(\mathbf{x})$. Therefore,

$$\mathfrak{r}_\Phi(\mathbf{x}) \leq \|\mathcal{G}_{\mathbf{x},\gamma}(\mathbf{x},\mathbf{y},\nabla\Phi_\delta(\mathbf{x}))\|.$$

Nonexpansiveness of Euclidean projection further gives

$$\begin{aligned}&\|\mathcal{G}_{\mathbf{x},\gamma}(\mathbf{x},\mathbf{y},\nabla\Phi_\delta(\mathbf{x})) - \mathcal{G}^\delta_{\mathbf{x},\gamma}(\mathbf{x},\mathbf{y})\| \\ &\quad\leq \|\nabla\Phi_\delta(\mathbf{x}) - \nabla_{\mathbf{x}} f_\delta(\mathbf{x},\mathbf{y})\|.\end{aligned}$$

Let $A := \|\mathcal{G}^\delta_{\mathbf{x},\gamma}(\mathbf{x},\mathbf{y})\|$ and $D := \|\nabla\Phi_\delta(\mathbf{x}) - \nabla_{\mathbf{x}} f_\delta(\mathbf{x},\mathbf{y})\|$. Then $\mathfrak{r}_\Phi(\mathbf{x}) \leq A + D$. By Corollary D.5,

$$D^2 \leq \rho\|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x},\mathbf{y},\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}))\|^2 + \rho\nu_{d_1}L_{fy}^2\delta^2.$$

Applying $(A+D)^2 \leq (1+\rho)A^2 + (1+\rho^{-1})D^2$ proves Eqn. (62). □

D.8 JOINT LOWER MODEL FOR THE TWO VARIABLES

Corollary D.7 (Joint lower model for the two variables)**.** *Under Assumption 3.1, for* $(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}') \in \mathcal{X} \times \mathcal{Y}$*, let* $\boldsymbol{d}_x := \mathbf{x}' - \mathbf{x}$ *and* $\boldsymbol{d}_y := \mathbf{y}' - \mathbf{y}$*. Then*

$$f_\delta(\mathbf{x}', \mathbf{y}') \geq f_\delta(\mathbf{x}, \mathbf{y}) + \langle \nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y}), \boldsymbol{d}_x \rangle + \langle \nabla_{\mathbf{y}} f_\delta(\mathbf{x}, \mathbf{y}), \boldsymbol{d}_y \rangle \tag{63}$$

$$- \frac{L_{x,\delta}}{2} \|\boldsymbol{d}_x\|^2 - \frac{L_{fy}}{2} \|\boldsymbol{d}_y\|^2 - L_{fy} \|\boldsymbol{d}_x\| \|\boldsymbol{d}_y\|. \tag{64}$$

Remark. The mixed term accounts for the change in the y-gradient caused by moving the outer variable; this coupling is used in the one-step Lyapunov estimate.

Proof. First change $\mathbf{x}$ at fixed $\mathbf{y}$:

$$f_\delta(\mathbf{x}', \mathbf{y}) \geq f_\delta(\mathbf{x}, \mathbf{y}) + \langle \nabla_{\mathbf{x}} f_\delta(\mathbf{x}, \mathbf{y}), \boldsymbol{d}_x \rangle - \frac{L_{x,\delta}}{2} \|\boldsymbol{d}_x\|^2.$$

Then change $\mathbf{y}$ at fixed $\mathbf{x}'$:

$$f_\delta(\mathbf{x}', \mathbf{y}') \geq f_\delta(\mathbf{x}', \mathbf{y}) + \langle \nabla_{\mathbf{y}} f_\delta(\mathbf{x}', \mathbf{y}), \boldsymbol{d}_y \rangle - \frac{L_{fy}}{2} \|\boldsymbol{d}_y\|^2.$$

Since $\nabla_{\mathbf{y}} f_\delta$ is L_{fy}-Lipschitz in $\mathbf{x}$,

$$\langle \nabla_{\mathbf{y}} f_\delta(\mathbf{x}', \mathbf{y}), \boldsymbol{d}_y \rangle \geq \langle \nabla_{\mathbf{y}} f_\delta(\mathbf{x}, \mathbf{y}), \boldsymbol{d}_y \rangle - L_{fy} \|\boldsymbol{d}_x\| \|\boldsymbol{d}_y\|.$$

Combining proves Eqn. (64). □

D.9 PROJECTION RELATIONS

Lemma D.8 (Projection relations)**.** *For every iteration* t*,*

$$\langle w_t, \mathbf{x}_{t+1} - \mathbf{x}_t \rangle \leq -\frac{1}{\gamma} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2, \tag{65}$$

$$\langle v_t, \mathbf{y}_{t+1} - \mathbf{y}_t \rangle \geq \frac{1}{\lambda} \|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2, \tag{66}$$

and

$$\|\mathcal{G}^\delta_{\mathbf{x},\gamma}(\mathbf{x}_t, \mathbf{y}_t) - \widetilde{\mathcal{G}}_{\mathbf{x},t}\| \leq \|e_t^x\|, \qquad \|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}_t, \mathbf{y}_t, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)) - \widetilde{\mathcal{G}}_{\mathbf{y},t}\| \leq \|e_t^y\|. \tag{67}$$

Proof. The Euclidean projection identity $\boldsymbol{z}^+ = \mathcal{P}_C(\boldsymbol{z})$ implies

$$\langle \boldsymbol{z} - \boldsymbol{z}^+, \boldsymbol{w} - \boldsymbol{z}^+ \rangle \leq 0, \qquad \forall \boldsymbol{w} \in C.$$

For the x-update, use $\boldsymbol{z} = \mathbf{x}_t - \gamma w_t$, $\boldsymbol{z}^+ = \mathbf{x}_{t+1}$, and $\boldsymbol{w} = \mathbf{x}_t$ to obtain Eqn. (65). For the y-update, use $\boldsymbol{z} = \mathbf{y}_t + \lambda v_t$, $\boldsymbol{z}^+ = \mathbf{y}_{t+1}$, and $\boldsymbol{w} = \mathbf{y}_t$ to obtain Eqn. (66).

Projection nonexpansiveness gives

$$\begin{aligned} &\|\mathcal{G}^\delta_{\mathbf{x},\gamma}(\mathbf{x}_t, \mathbf{y}_t) - \widetilde{\mathcal{G}}_{\mathbf{x},t}\| \\ &= \frac{1}{\gamma} \|\mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - \gamma \nabla_{\mathbf{x}} f_\delta(\mathbf{x}_t, \mathbf{y}_t)) - \mathcal{P}_{\mathcal{X}}(\mathbf{x}_t - \gamma w_t)\| \\ &\leq \|\nabla_{\mathbf{x}} f_\delta(\mathbf{x}_t, \mathbf{y}_t) - w_t\| = \|e_t^x\|. \end{aligned}$$

The y inequality follows identically. □

D.10 SPIDER ESTIMATOR TRACKING

Lemma D.9 (SPIDER tracking within an epoch)**.** *Let* s *be the first iteration of a SPIDER epoch. For* $t \in \{s, \ldots, \min\{s + q - 1, T - 1\}\}$*,*

$$\mathbb{E}\|e_t^x\|^2 \leq \frac{\bar{\sigma}_x^2}{B_x} + \frac{M_{x,\delta}^2}{b_x} \sum_{j=s+1}^{t} \mathbb{E}\|\mathbf{z}_j - \mathbf{z}_{j-1}\|^2, \tag{68}$$

$$\mathbb{E}\|e_t^y\|^2 \leq \frac{\bar{\sigma}_y^2}{B_y} + \frac{L_{y,\mathrm{ms}}^2}{b_y} \sum_{j=s+1}^{t} \mathbb{E}\|\mathbf{z}_j - \mathbf{z}_{j-1}\|^2. \tag{69}$$

Corollary D.10 (Global SPIDER error)**.** *Define*

$$\Sigma_B := \frac{2d_1L^2}{B_x} + \frac{\bar{\sigma}_y^2}{B_y}, \tag{70}$$

$$A_q := q\left(\frac{M_{x,\delta}^2}{b_x} + \frac{L_{y,\mathrm{ms}}^2}{b_y}\right). \tag{71}$$

Then

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}[\|e_t^x\|^2 + \|e_t^y\|^2] &\le T\Sigma_B + A_q\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{z}_{t+1}-\mathbf{z}_t\|^2 \\ &= T\Sigma_B + A_q\gamma^2\sum_{t=0}^{T-1}\mathbb{E}\|\widetilde{\mathcal{G}}_{\mathbf{x},t}\|^2 + A_q\lambda^2\sum_{t=0}^{T-1}\mathbb{E}\|\widetilde{\mathcal{G}}_{\mathbf{y},t}\|^2.\end{aligned} \tag{72}$$

Remark. The first term accumulates the refresh-batch variance, whereas the second captures the estimator drift caused by movement of the iterates within each epoch.

Proof. We first prove the x bound. At an epoch start s, Lemma 5.3 gives

$$\mathbb{E}\|e_s^x\|^2 \le \frac{\bar{\sigma}_x^2}{B_x}.$$

At a non-refresh iteration t, conditioning on the past,

$$\begin{aligned}e_t^x = e_{t-1}^x + \frac{1}{b_x}\sum_{i=1}^{b_x}\Big(&g_{\mathbf{x}}(\mathbf{z}_t;\xi_i,\hat{\boldsymbol{u}}_i) - g_{\mathbf{x}}(\mathbf{z}_{t-1};\xi_i,\hat{\boldsymbol{u}}_i) \\ &- [\nabla_{\mathbf{x}}f_\delta(\mathbf{z}_t) - \nabla_{\mathbf{x}}f_\delta(\mathbf{z}_{t-1})]\Big).\end{aligned}$$

The mini-batch summands have conditional mean zero because the same $(\xi_i, \hat{\boldsymbol{u}}_i)$ is used at $\mathbf{z}_t$ and $\mathbf{z}_{t-1}$. Therefore the cross term with e_{t-1}^x vanishes and independence within the mini-batch yields

$$\begin{aligned}\mathbb{E}_t\|e_t^x\|^2 &\le \|e_{t-1}^x\|^2 + \frac{1}{b_x}\mathbb{E}_t\|g_{\mathbf{x}}(\mathbf{z}_t;\xi,\hat{\boldsymbol{u}}) - g_{\mathbf{x}}(\mathbf{z}_{t-1};\xi,\hat{\boldsymbol{u}})\|^2 \\ &\le \|e_{t-1}^x\|^2 + \frac{M_{x,\delta}^2}{b_x}\|\mathbf{z}_t-\mathbf{z}_{t-1}\|^2.\end{aligned}$$

Iterating from s to t proves Eqn. (68).

For y, Assumption 3.3 gives the refresh bound $\mathbb{E}\|e_s^y\|^2 \le \bar{\sigma}_y^2/B_y$. The same conditional-mean-zero argument and the mean-square Lipschitz condition in Assumption 3.3 yield

$$\mathbb{E}_t\|e_t^y\|^2 \le \|e_{t-1}^y\|^2 + \frac{L_{y,\mathrm{ms}}^2}{b_y}\|\mathbf{z}_t-\mathbf{z}_{t-1}\|^2,$$

which proves Eqn. (69).

To obtain the global bound, let $I_s = \{s,\dots,m_s\}$ be an epoch and let $n_s := m_s - s + 1 \le q$ be its actual length. Every movement $\|\mathbf{z}_j - \mathbf{z}_{j-1}\|^2$ appears in at most n_s later error bounds. Therefore

$$\begin{aligned}\sum_{t\in I_s}\mathbb{E}[\|e_t^x\|^2 + \|e_t^y\|^2] \le\ & n_s\left(\frac{\bar{\sigma}_x^2}{B_x} + \frac{\bar{\sigma}_y^2}{B_y}\right) \\ &+ n_s\left(\frac{M_{x,\delta}^2}{b_x} + \frac{L_{y,\mathrm{ms}}^2}{b_y}\right)\sum_{j=s+1}^{m_s}\mathbb{E}\|\mathbf{z}_j-\mathbf{z}_{j-1}\|^2.\end{aligned}$$

Now use $n_s \le q$ in the movement term and $\sum_s n_s = T$ in the refresh term. Summing all epochs proves the first line of Eqn. (72). The second follows from

$$\|\mathbf{z}_{t+1}-\mathbf{z}_t\|^2 = \gamma^2\|\widetilde{\mathcal{G}}_{\mathbf{x},t}\|^2 + \lambda^2\|\widetilde{\mathcal{G}}_{\mathbf{y},t}\|^2.$$

□

D.11 One-Step Lyapunov Descent

Lemma D.11 (One-step Lyapunov descent)**.** *Suppose*

$$0 < \lambda \le \min\left\{\bar{\lambda}, \frac{1}{L_{fy}}\right\}, \qquad 0 < \gamma \le \frac{\lambda}{\chi}, \tag{73}$$

and

$$\gamma \le \frac{1}{2[(1+a)L_{\Phi,\delta} + aL_{x,\delta}]}, \qquad \gamma \le \frac{1}{\sqrt{32\chi}\, L_{fy}}. \tag{74}$$

Then $a \le 1$ *and*

$$\begin{aligned}\mathbb{E}_t[\Psi_{t+1}] \le \Psi_t - \frac{\gamma}{2}\mathbb{E}_t\left[\|\widetilde{\mathcal{G}}_{\mathbf{x},t}\|^2 + \|\widetilde{\mathcal{G}}_{\mathbf{y},t}\|^2\right] \\ + C_e\gamma\mathbb{E}_t[\|e_t^x\|^2 + \|e_t^y\|^2] + C_\delta\gamma\delta^2.\end{aligned} \tag{75}$$

Proof. Let

$$\boldsymbol{d}_t^x := \mathbf{x}_{t+1} - \mathbf{x}_t, \qquad \boldsymbol{d}_t^y := \mathbf{y}_{t+1} - \mathbf{y}_t,$$
$$\boldsymbol{r}_t := \nabla\Phi_\delta(\mathbf{x}_t) - \nabla_{\mathbf{x}} f_\delta(\mathbf{x}_t, \mathbf{y}_t), \qquad \boldsymbol{b}_t^\delta := \nabla_{\mathbf{y}} f_\delta(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t).$$

By Lemma 4.4(4),

$$\|\boldsymbol{b}_t^\delta\| \le L_{fy}\delta. \tag{76}$$

Since $\Psi_t = (1+a)\Phi_\delta(\mathbf{x}_t) - af_\delta(\mathbf{x}_t, \mathbf{y}_t)$, the smoothness of Φ_δ and Corollary D.7 imply

$$\begin{aligned}\Psi_{t+1} - \Psi_t \le \langle \nabla_{\mathbf{x}} f_\delta(\mathbf{x}_t, \mathbf{y}_t) + (1+a)\boldsymbol{r}_t, \boldsymbol{d}_t^x\rangle - a\langle \nabla_{\mathbf{y}} f_\delta(\mathbf{x}_t, \mathbf{y}_t), \boldsymbol{d}_t^y\rangle \\ + \frac{(1+a)L_{\Phi,\delta} + aL_{x,\delta}}{2}\|\boldsymbol{d}_t^x\|^2 + \frac{aL_{fy}}{2}\|\boldsymbol{d}_t^y\|^2 + aL_{fy}\|\boldsymbol{d}_t^x\|\|\boldsymbol{d}_t^y\|.\end{aligned} \tag{77}$$

Using $\nabla_{\mathbf{x}} f_\delta = w_t - e_t^x$ and $\nabla_{\mathbf{y}} f_\delta = v_t - e_t^y + \boldsymbol{b}_t^\delta$ together with Lemma D.8 gives

$$\begin{aligned}\Psi_{t+1} - \Psi_t \le -\left[\frac{1}{\gamma} - \frac{(1+a)L_{\Phi,\delta} + aL_{x,\delta}}{2}\right]\|\boldsymbol{d}_t^x\|^2 \\ - a\left[\frac{1}{\lambda} - \frac{L_{fy}}{2}\right]\|\boldsymbol{d}_t^y\|^2 + aL_{fy}\|\boldsymbol{d}_t^x\|\|\boldsymbol{d}_t^y\| \\ + \langle (1+a)\boldsymbol{r}_t - e_t^x, \boldsymbol{d}_t^x\rangle + a\langle e_t^y - \boldsymbol{b}_t^\delta, \boldsymbol{d}_t^y\rangle.\end{aligned} \tag{78}$$

By the first smoothness-related stepsize restriction,

$$\frac{(1+a)L_{\Phi,\delta} + aL_{x,\delta}}{2} \le \frac{1}{4\gamma},$$

and by $\lambda \le 1/L_{fy}$,

$$\frac{1}{\lambda} - \frac{L_{fy}}{2} \ge \frac{1}{2\lambda}.$$

Young's inequality yields

$$\begin{aligned}aL_{fy}\|\boldsymbol{d}_t^x\|\|\boldsymbol{d}_t^y\| &\le \frac{1}{8\gamma}\|\boldsymbol{d}_t^x\|^2 + 2a^2\gamma L_{fy}^2\|\boldsymbol{d}_t^y\|^2, \\ \langle (1+a)\boldsymbol{r}_t - e_t^x, \boldsymbol{d}_t^x\rangle &\le \frac{1}{8\gamma}\|\boldsymbol{d}_t^x\|^2 + 4\gamma(1+a)^2\|\boldsymbol{r}_t\|^2 + 4\gamma\|e_t^x\|^2, \\ a\langle e_t^y - \boldsymbol{b}_t^\delta, \boldsymbol{d}_t^y\rangle &\le \frac{a}{8\lambda}\|\boldsymbol{d}_t^y\|^2 + 4a\lambda\|e_t^y\|^2 + 4a\lambda\|\boldsymbol{b}_t^\delta\|^2.\end{aligned}$$

Substituting gives

$$\begin{aligned}\Psi_{t+1} - \Psi_t \le -\frac{1}{2\gamma}\|\boldsymbol{d}_t^x\|^2 - \frac{3a}{8\lambda}\|\boldsymbol{d}_t^y\|^2 + 2a^2\gamma L_{fy}^2\|\boldsymbol{d}_t^y\|^2 \\ + 4\gamma(1+a)^2\|\boldsymbol{r}_t\|^2 + 4\gamma\|e_t^x\|^2 + 4a\lambda\|e_t^y\|^2 + 4a\lambda L_{fy}^2\delta^2.\end{aligned} \tag{79}$$

By Corollary D.5 and $\nu_{d_1} \le 1$,

$$\|\boldsymbol{r}_t\|^2 \le \frac{C_r}{\mu}\|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}_t,\mathbf{y}_t,\nabla_{\mathbf{y}} f(\mathbf{x}_t,\mathbf{y}_t))\|^2 + \frac{C_r L_{fy}^2}{\mu}\delta^2.$$

Lemma D.8 gives

$$\|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}_t,\mathbf{y}_t,\nabla_{\mathbf{y}} f(\mathbf{x}_t,\mathbf{y}_t))\|^2 \le 2\|\widetilde{\mathcal{G}}_{\mathbf{y},t}\|^2 + 2\|e_t^y\|^2 = \frac{2}{\lambda^2}\|\boldsymbol{d}_t^y\|^2 + 2\|e_t^y\|^2.$$

Therefore

$$4\gamma(1+a)^2\|\boldsymbol{r}_t\|^2 \le \frac{8\gamma(1+a)^2 C_r}{\mu\lambda^2}\|\boldsymbol{d}_t^y\|^2 + \frac{8\gamma(1+a)^2 C_r}{\mu}\|e_t^y\|^2 + \frac{4\gamma(1+a)^2 C_r L_{fy}^2}{\mu}\delta^2. \tag{80}$$

Now $a = \chi\gamma/\lambda$ and $\gamma \le \lambda/\chi$ imply $a \le 1$. Since $\chi \ge 512 C_r/\mu$ and $(1+a)^2 \le 4$,

$$\frac{8\gamma(1+a)^2 C_r}{\mu\lambda^2} \le \frac{a}{16\lambda}.$$

The second smoothness-related stepsize restriction gives

$$2\chi\gamma^2 L_{fy}^2 \le \frac{1}{16},$$

which is equivalent to

$$2a^2\gamma L_{fy}^2 \le \frac{a}{16\lambda}.$$

Thus the coefficient of $\|\boldsymbol{d}_t^y\|^2$ in Eqn. (79) is at most

$$-\frac{3a}{8\lambda} + \frac{a}{16\lambda} + \frac{a}{16\lambda} = -\frac{a}{4\lambda}.$$

Moreover,

$$\frac{a}{4\lambda}\|\boldsymbol{d}_t^y\|^2 = \frac{\chi\gamma}{4}\|\widetilde{\mathcal{G}}_{\mathbf{y},t}\|^2 \ge \frac{\gamma}{2}\|\widetilde{\mathcal{G}}_{\mathbf{y},t}\|^2,$$

because $\chi \ge 2$, while $\|\boldsymbol{d}_t^x\|^2/(2\gamma) = \gamma\|\widetilde{\mathcal{G}}_{\mathbf{x},t}\|^2/2$. Finally, $a\lambda = \chi\gamma$, $(1+a)^2 \le 4$, and Eqn. (76) bound all remaining terms by

$$C_e\gamma(\|e_t^x\|^2 + \|e_t^y\|^2) + C_\delta\gamma\delta^2.$$

Taking conditional expectation proves Eqn. (75). □

E Proofs of the Main Results

Theorem E.1 (Detailed convergence bound)**.** *Suppose Assumptions 3.1, 3.2, 3.4, 3.3, 5.1, and 5.2 hold. With $\chi := \max\{2, 512C_r/\mu\}$, $a := \chi\gamma/\lambda$, $\Sigma_B := \bar{\sigma}_x^2/B_x + \bar{\sigma}_y^2/B_y$ and $A_q := q(M_{x,\delta}^2/b_x + L_{y,\mathrm{ms}}^2/b_y)$, set $C_e := 4 + 4\chi + 32C_r/\mu$ and $C_\delta := L_{fy}^2(4\chi + 16C_r/\mu)$. If Algorithm 1 uses parameters satisfying*

$$0 < \lambda \le \min\{\bar{\lambda}, 1/L_{fy}\}, \qquad 0 < \gamma \le \lambda/\chi,$$
$$\gamma \le \frac{1}{2[(1+a)L_{\Phi,\delta} + aL_{x,\delta}]}, \qquad \gamma \le \frac{1}{\sqrt{32\chi}\,L_{fy}},$$

and

$$C_e A_q \gamma^2 \le \frac{1}{4}, \qquad C_e A_q \lambda^2 \le \frac{1}{4},$$

then

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\mathcal{G}_{\mathbf{x},\gamma}^{\delta}(\mathbf{x}_t,\mathbf{y}_t)\|^2 + \|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}_t,\mathbf{y}_t,\nabla_{\mathbf{y}} f(\mathbf{x}_t,\mathbf{y}_t))\|^2\Big] \le \frac{12(\Psi_0 - \Phi_\delta^\star)}{\gamma T} + (12C_e + 2)\Sigma_B + 12C_\delta\delta^2. \tag{81}$$

For a uniformly sampled $R \in \{0,\dots,T-1\}$, the same right-hand side also bounds the expected sum of the two squared exact mapping norms at $(\mathbf{x}_R,\mathbf{y}_R)$; the subsequent corollaries convert these mappings to the stated Goldstein residuals.

E.1 Proof of Theorem E.1

Proof. Summing Eqn. (75) from $t = 0$ to $T - 1$ and taking expectation gives

$$\frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\widetilde{\mathcal{G}}_{\mathbf{x},t}\|^2 + \|\widetilde{\mathcal{G}}_{\mathbf{y},t}\|^2] \le \Psi_0 - \mathbb{E}\Psi_T + C_e\gamma\sum_{t=0}^{T-1}\mathbb{E}[\|e_t^x\|^2 + \|e_t^y\|^2] + C_\delta\gamma T\delta^2. \tag{82}$$

Since $\Delta_\delta \ge 0$,

$$\Psi_T = \Phi_\delta(\mathbf{x}_T) + a\Delta_\delta(\mathbf{x}_T, \mathbf{y}_T) \ge \Phi_\delta(\mathbf{x}_T) \ge \Phi_\delta^\star.$$

Using Corollary D.10,

$$\sum_t \mathbb{E}[\|e_t^x\|^2 + \|e_t^y\|^2] \le T\Sigma_B + A_q\gamma^2\sum_t \mathbb{E}\|\widetilde{\mathcal{G}}_{\mathbf{x},t}\|^2 + A_q\lambda^2\sum_t \mathbb{E}\|\widetilde{\mathcal{G}}_{\mathbf{y},t}\|^2.$$

The absorption conditions in Theorem E.1 imply

$$\frac{\gamma}{4}\sum_t \mathbb{E}[\|\widetilde{\mathcal{G}}_{\mathbf{x},t}\|^2 + \|\widetilde{\mathcal{G}}_{\mathbf{y},t}\|^2] \le \Psi_0 - \Phi_\delta^\star + C_e\gamma T\Sigma_B + C_\delta\gamma T\delta^2. \tag{83}$$

Hence

$$\frac{1}{T}\sum_t \mathbb{E}[\|\widetilde{\mathcal{G}}_{\mathbf{x},t}\|^2 + \|\widetilde{\mathcal{G}}_{\mathbf{y},t}\|^2] \le \frac{4(\Psi_0 - \Phi_\delta^\star)}{\gamma T} + 4C_e\Sigma_B + 4C_\delta\delta^2. \tag{84}$$

The global SPIDER bound also gives

$$\begin{aligned}\frac{1}{T}\sum_t \mathbb{E}[\|e_t^x\|^2 + \|e_t^y\|^2] &\le \Sigma_B + A_q\gamma^2\frac{1}{T}\sum_t \mathbb{E}\|\widetilde{\mathcal{G}}_{\mathbf{x},t}\|^2 + A_q\lambda^2\frac{1}{T}\sum_t \mathbb{E}\|\widetilde{\mathcal{G}}_{\mathbf{y},t}\|^2 \\ &\le \Sigma_B + \frac{1}{4C_e}\frac{1}{T}\sum_t \mathbb{E}[\|\widetilde{\mathcal{G}}_{\mathbf{x},t}\|^2 + \|\widetilde{\mathcal{G}}_{\mathbf{y},t}\|^2].\end{aligned}$$

By Eqn. (67) and $(a+b)^2 \le 2a^2 + 2b^2$,

$$\begin{aligned}&\|\mathcal{G}_{\mathbf{x},\gamma}^\delta(\mathbf{x}_t, \mathbf{y}_t)\|^2 + \|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}_t, \mathbf{y}_t, \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t))\|^2 \\ &\le 2[\|\widetilde{\mathcal{G}}_{\mathbf{x},t}\|^2 + \|\widetilde{\mathcal{G}}_{\mathbf{y},t}\|^2] + 2[\|e_t^x\|^2 + \|e_t^y\|^2].\end{aligned}$$

Since $C_e \ge 1$, the averaged sum of squared exact-mapping norms is at most three times the corresponding sum of squared computed-mapping norms plus $2\Sigma_B$. Substituting Eqn. (84) proves Eqn. (81). Uniform random sampling of R gives the final statement of Theorem E.1. □

E.2 Proof of Corollary 5.5

Proof. Lemma C.1 ensures that the parameterized Goldstein set and its minimum projected residual are Borel measurable. Hence

$$(\mathbf{x}_R, \mathbf{y}_R) \mapsto \min_{\boldsymbol{h}\in\partial_{\delta,\mathbf{x}} f(\mathbf{x}_R,\mathbf{y}_R)} \|\mathcal{G}_{\mathbf{x},\gamma}(\mathbf{x}_R, \mathbf{y}_R, \boldsymbol{h})\|$$

is measurable. Let

$$\mathfrak{r}_x := \min_{\boldsymbol{h}\in\partial_{\delta,\mathbf{x}} f(\mathbf{x}_R,\mathbf{y}_R)} \|\mathcal{G}_{\mathbf{x},\gamma}(\mathbf{x}_R, \mathbf{y}_R, \boldsymbol{h})\|, \qquad \mathfrak{r}_y := \|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}_R, \mathbf{y}_R, \nabla_{\mathbf{y}} f(\mathbf{x}_R, \mathbf{y}_R))\|.$$

Lemma 4.4(5) gives $\mathfrak{r}_x^2 \le \|\mathcal{G}_{\mathbf{x},\gamma}^\delta(\mathbf{x}_R, \mathbf{y}_R)\|^2$, while $\mathfrak{r}_y = \|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}_R, \mathbf{y}_R, \nabla_{\mathbf{y}} f(\mathbf{x}_R, \mathbf{y}_R))\|$. Hence Theorem E.1 and Jensen's inequality imply

$$\mathbb{E}[\mathfrak{r}_x] \le \sqrt{\mathbb{E}[\mathfrak{r}_x^2]} \le \epsilon, \qquad \mathbb{E}[\mathfrak{r}_y] \le \sqrt{\mathbb{E}[\mathfrak{r}_y^2]} \le \epsilon.$$

These are precisely the two separate conditions in Definition 4.2. □

E.3 Proof of Corollary 5.6

Proof. Define the measurable nonnegative residual

$$\widetilde{Z}_R := \|\mathcal{G}^{\delta}_{\mathbf{x},\gamma}(\mathbf{x}_R, \mathbf{y}_R)\|^2 + \|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}_R, \mathbf{y}_R, \nabla_{\mathbf{y}} f(\mathbf{x}_R, \mathbf{y}_R))\|^2.$$

Theorem E.1 gives $\mathbb{E}[\widetilde{Z}_R] \le \eta\epsilon^2$. Markov's inequality therefore yields

$$\Pr(\widetilde{Z}_R > \epsilon^2) \le \frac{\mathbb{E}[\widetilde{Z}_R]}{\epsilon^2} \le \eta.$$

On the event $\widetilde{Z}_R \le \epsilon^2$, both squared residuals are at most ϵ^2. Thus the y-condition in Definition 4.1 holds, and Lemma 4.4(5) gives

$$\min_{\boldsymbol{h} \in \partial_{\delta,\mathbf{x}} f(\mathbf{x}_R, \mathbf{y}_R)} \|\mathcal{G}_{\mathbf{x},\gamma}(\mathbf{x}_R, \mathbf{y}_R, \boldsymbol{h})\| \le \|\mathcal{G}^{\delta}_{\mathbf{x},\gamma}(\mathbf{x}_R, \mathbf{y}_R)\| \le \epsilon.$$

Hence the output satisfies Definition 4.1 with probability at least $1 - \eta$. □

E.4 Proof of Corollary 5.7

Proof. Lemma C.1 shows that $\mathfrak{r}_\Phi(\mathbf{x}_R)$ is measurable. Taking expectations in Lemma D.6 and applying Theorem E.1 gives

$$\mathbb{E}[\mathfrak{r}_\Phi(\mathbf{x}_R)^2] \le \left(1 + \frac{C_r}{\mu}\right)\left(\mathcal{E}_T + \nu_{d_1} L_{fy}^2 \delta^2\right),$$

which is Eqn. (8). If the right-hand side is at most ϵ^2, Jensen's inequality yields $\mathbb{E}[\mathfrak{r}_\Phi(\mathbf{x}_R)] \le \epsilon$. If it is at most $\eta\epsilon^2$, Markov's inequality gives

$$\Pr(\mathfrak{r}_\Phi(\mathbf{x}_R) > \epsilon) \le \frac{\mathbb{E}[\mathfrak{r}_\Phi(\mathbf{x}_R)^2]}{\epsilon^2} \le \eta.$$

These are the expected and pointwise statements in the corollary. □

E.5 Full statement and proof of Corollary 5.8

Detailed statement. The assumptions on the fixed ambient domain make $L, L_{fy}, L_{xy}, \mu, \bar{\lambda}, \bar{\sigma}_y$, and $L_{y,\mathrm{ms}}$ independent of the smoothing radius (and hence make the derived constants $C_{\mathrm{qg}}, C_r, \chi, C_e, C_\delta$ uniform as well). Suppose there are $\delta_0 \in (0, \bar{\delta})$, $\epsilon_0 > 0$, and $\Delta_{\mathrm{pot}} \ge 1$, independent of δ and ϵ, such that for the fixed initialization and $a = 1/2$,

$$\Psi_0 - \Phi^\star_\delta \le \Delta_{\mathrm{pot}}, \qquad 0 < \delta \le \delta_0. \tag{85}$$

This is a uniform-in-δ complexity convention rather than an oracle assumption. It prevents the hidden initial-gap factor from depending on the target accuracy. For a fixed initialization it is mild whenever the original value function has a finite initial gap and the standard smoothing bias is uniformly $\mathcal{O}(L\delta)$ (Lin et al., 2022; Shi & Luo, 2026).

For sufficiently small δ, choose

$$\gamma = \Theta\left(\frac{\delta}{L\sqrt{d_1}}\right), \qquad \lambda = 2\chi\gamma, \qquad a = \frac{1}{2}, \tag{86}$$

$$q = \Theta(\epsilon^{-1}), \qquad B_x = \Theta(d_1 L^2 \epsilon^{-2}), \qquad B_y = \Theta(\epsilon^{-2}), \tag{87}$$

Here the hidden constant in B_y may depend on the fixed variance bound $\bar{\sigma}_y^2$; equivalently, an explicit choice is $B_y = \Theta(\bar{\sigma}_y^2 \epsilon^{-2})$.

$$b_x = \Theta(d_1 \epsilon^{-1}), \qquad b_y = \Theta(\epsilon^{-1}), \qquad T = \Theta\left(\frac{\Delta_{\mathrm{pot}} L \sqrt{d_1}}{\delta \epsilon^2}\right). \tag{88}$$

The outer batch sizes B_x and b_x, together with q, B_y, b_y, T, are rounded up to integers and lower-bounded by one. There exists a sufficiently small problem-dependent constant $c_\delta > 0$ such that, for

$0 < \epsilon \le \epsilon_0$ and $0 < \delta \le \min\{\delta_0, c_\delta \epsilon\}$, appropriately chosen hidden constants make $(\mathbf{x}_R, \mathbf{y}_R)$ an expected $(\gamma, \lambda, \delta, \epsilon)$-GGSP in the sense of Definition 4.2, and $\mathbf{x}_R$ is an expected projected $(\gamma, \delta, \epsilon)$-GGSP of Φ in the sense of Definition 4.3. The expected sum of squared block residuals is $\mathcal{O}(\epsilon^2)$. The exact query counts satisfy

$$Q_{\mathrm{SZO}} \le 2\left\lceil \frac{T}{q} \right\rceil B_x + 4Tb_x, \qquad Q_{\mathrm{SFO}} \le \left\lceil \frac{T}{q} \right\rceil B_y + 2Tb_y, \tag{89}$$

and consequently obey the query-complexity rates stated in the main text.

Proof. First, the refresh term satisfies

$$\Sigma_B = \frac{2d_1 L^2}{B_x} + \frac{\bar{\sigma}_y^2}{B_y} = \mathcal{O}(\epsilon^2)$$

under Eqn. (87).

Next, $L_{x,\delta}, L_{\Phi,\delta} = \mathcal{O}(L\sqrt{d_1}/\delta)$. Hence a sufficiently small constant in $\gamma = \Theta(\delta/(L\sqrt{d_1}))$ satisfies the smoothing-related stepsize restriction in Theorem E.1. Choosing $\lambda = 2\chi\gamma$ gives $a = 1/2$ and automatically satisfies $\gamma \le \lambda/\chi$. For sufficiently small δ, $\lambda \le \min\{\bar{\lambda}, 1/L_{fy}\}$ and the final stepsize restriction in Theorem E.1 also holds.

The ZO increment constant is

$$M_{x,\delta}^2 = \frac{d_1^2 L^2}{\delta^2}.$$

Since $\lambda = \Theta(\gamma)$,

$$M_{x,\delta}^2 \lambda^2 = \Theta(d_1)$$

up to problem-dependent constants. Therefore, with $q = \Theta(\epsilon^{-1})$ and $b_x = \Theta(d_1 q)$,

$$q\frac{M_{x,\delta}^2}{b_x}\lambda^2 = \mathcal{O}(1).$$

Increasing the hidden constant in the batch size b_x makes this term as small as required by the absorption conditions in Theorem E.1. Likewise, $b_y = \Theta(q)$ makes $qL_{y,\mathrm{ms}}^2\lambda^2/b_y$ bounded (indeed vanishing as $\delta \to 0$). Thus the absorption conditions hold.

By Eqn. (85), it suffices to choose T so that

$$\frac{\Delta_{\mathrm{pot}}}{\gamma T} = \mathcal{O}(\epsilon^2),$$

which gives

$$T = \Theta\left(\frac{\Delta_{\mathrm{pot}} L\sqrt{d_1}}{\delta\epsilon^2}\right).$$

Finally, $0 < \delta \le c_\delta \epsilon$ makes both smoothing floors of order $\mathcal{O}(\epsilon^2)$. Set $C_\Phi := 1 + C_r/\mu$. By taking c_δ sufficiently small and increasing the fixed constants in T, B_x, and B_y, without changing their orders, we can ensure simultaneously that

$$\mathcal{E}_T \le \frac{\epsilon^2}{2C_\Phi}, \qquad C_\Phi \nu_{d_1} L_{fy}^2 \delta^2 \le \frac{\epsilon^2}{2}.$$

Because $C_\Phi \ge 1$, Corollary 5.5 gives the expected blockwise GGSP guarantee. Corollary 5.7 and the two displayed bounds give $\mathbb{E}[\mathfrak{r}_\Phi(\mathbf{x}_R)^2] \le \epsilon^2$, hence the expected projected GGSP guarantee for Φ.

There are at most $\lceil T/q \rceil$ refresh iterations. Each outer batch sample uses two function evaluations per stochastic pair in a refresh and four in a recursive difference (at both $\mathbf{z}_t$ and $\mathbf{z}_{t-1}$). Hence

$$Q_{\mathrm{SZO}} \le 2\left\lceil \frac{T}{q} \right\rceil B_x + 4Tb_x.$$

Using $B_x = \Theta(d_1 L^2 \epsilon^{-2})$, $q = \Theta(\epsilon^{-1})$, and $b_x = \Theta(d_1\epsilon^{-1})$ gives an average SZO cost $\mathcal{O}(d_1/\epsilon)$ per outer iteration up to problem constants. Multiplying by $T = \mathcal{O}(\Delta_{\mathrm{pot}}\sqrt{d_1}/(\delta\epsilon^2))$ yields

$$Q_{\mathrm{SZO}} = \widetilde{\mathcal{O}}(\Delta_{\mathrm{pot}} d_1^{3/2} \delta^{-1} \epsilon^{-3}).$$

For the first-order oracle, each refresh sample uses one stochastic y-gradient and each recursive sample uses two stochastic y-gradients, one at $\mathbf{z}_t$ and one at $\mathbf{z}_{t-1}$. Therefore,

$$Q_{\mathrm{SFO}} \le \left\lceil \frac{T}{q} \right\rceil B_y + 2Tb_y.$$

Using $B_y = \Theta(\epsilon^{-2})$, $q = \Theta(\epsilon^{-1})$, $b_y = \Theta(\epsilon^{-1})$, and treating $\bar{\sigma}_y$ as a problem-dependent constant gives an average SFO cost $\mathcal{O}(\epsilon^{-1})$ per outer iteration. Hence

$$Q_{\mathrm{SFO}} = \widetilde{\mathcal{O}}(\Delta_{\mathrm{pot}} d_1^{1/2} \delta^{-1} \epsilon^{-3}).$$

With the canonical choice $\delta = c_\delta \epsilon = \Theta(\epsilon)$, the two query bounds become

$$Q_{\mathrm{SZO}} = \widetilde{\mathcal{O}}(\Delta_{\mathrm{pot}} d_1^{3/2} \epsilon^{-4}), \qquad Q_{\mathrm{SFO}} = \widetilde{\mathcal{O}}(\Delta_{\mathrm{pot}} d_1^{1/2} \epsilon^{-4}).$$

The factor $d_1^{1/2}$ in Q_{SFO} enters only through the shared iteration horizon T imposed by the smoothed x-update; a single SFO query itself has no d_1-dependent query cost. □

Corollary E.2 (Arbitrary-confidence oracle complexity)**.** *Under the conditions of Corollary 5.8, let $\eta \in (0,1)$ be an arbitrary failure probability, suppose $0 < \epsilon \le \epsilon_0$, and set $\epsilon_\eta := \sqrt{\eta}\,\epsilon$. Apply the construction of Corollary 5.8 with target accuracy ϵ_η and choose $0 < \delta \le \min\{\delta_0, c_\delta \epsilon_\eta\}$. Then, with probability at least $1-\eta$, $(\mathbf{x}_R, \mathbf{y}_R)$ is a $(\gamma, \lambda, \delta, \epsilon)$-GGSP in the sense of Definition 4.1 and $\mathbf{x}_R$ is a projected $(\gamma, \delta, \epsilon)$-GGSP of Φ in the sense of Definition 4.3, simultaneously. The corresponding query complexities are*

$$Q_{\mathrm{SZO}} = \widetilde{\mathcal{O}}\left(\Delta_{\mathrm{pot}} d_1^{3/2} \delta^{-1} \eta^{-3/2} \epsilon^{-3}\right), \tag{90}$$

$$Q_{\mathrm{SFO}} = \widetilde{\mathcal{O}}\left(\Delta_{\mathrm{pot}} d_1^{1/2} \delta^{-1} \eta^{-3/2} \epsilon^{-3}\right). \tag{91}$$

With the canonical choice $\delta = c_\delta \sqrt{\eta}\,\epsilon$, these become

$$Q_{\mathrm{SZO}} = \widetilde{\mathcal{O}}\left(\Delta_{\mathrm{pot}} d_1^{3/2} \eta^{-2} \epsilon^{-4}\right), \qquad Q_{\mathrm{SFO}} = \widetilde{\mathcal{O}}\left(\Delta_{\mathrm{pot}} d_1^{1/2} \eta^{-2} \epsilon^{-4}\right). \tag{92}$$

Remark. This result makes the confidence dependence explicit. Both stationarity guarantees are obtained on one event by controlling their combined residual, so no separate union bound is needed.

E.6 Proof of Corollary E.2

Proof. Set $\epsilon_\eta = \sqrt{\eta}\,\epsilon$ and $C_\Phi := 1 + C_r/\mu$, and write

$$Z_R := \|\mathcal{G}^{\delta}_{\mathbf{x},\gamma}(\mathbf{x}_R, \mathbf{y}_R)\|^2 + \|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x}_R, \mathbf{y}_R, \nabla_{\mathbf{y}} f(\mathbf{x}_R, \mathbf{y}_R))\|^2.$$

Repeating the proof of Corollary 5.8 with target accuracy ϵ_η, choose the hidden constants so that

$$\mathbb{E}[Z_R] \le \frac{\epsilon_\eta^2}{2(1+C_\Phi)}, \qquad C_\Phi \nu_{d_1} L_{fy}^2 \delta^2 \le \frac{\epsilon_\eta^2}{2}.$$

These choices only change problem-dependent constants. By Lemma D.6,

$$\mathbb{E}[Z_R + \mathfrak{r}_\Phi(\mathbf{x}_R)^2] \le (1+C_\Phi)\mathbb{E}[Z_R] + C_\Phi \nu_{d_1} L_{fy}^2 \delta^2 \le \epsilon_\eta^2 = \eta \epsilon^2.$$

Markov's inequality therefore shows that $Z_R + \mathfrak{r}_\Phi(\mathbf{x}_R)^2 \le \epsilon^2$ with probability at least $1-\eta$. On this event, both block residuals and the value-function residual are at most ϵ; Lemma 4.4(5) converts the outer block residual to the pointwise condition in Definition 4.1. This proves the two claimed guarantees simultaneously. Substituting $\epsilon_\eta = \sqrt{\eta}\,\epsilon$ into the query-complexity rates stated in the main text gives

$$\epsilon_\eta^{-3} = \eta^{-3/2} \epsilon^{-3},$$

which proves Eqn. (90) and Eqn. (91). Taking $\delta = c_\delta \sqrt{\eta}\,\epsilon$ in these two bounds gives Eqn. (92). □

F Controlled Nonconvex-PŁ Classification Numerical Illustration

This section collects the analytical verification and implementation details for the controlled objective used in Section 6.1.

F.1 BENCHMARK VERIFICATION

We verify the claims made for the problem in Section 6.1. Let $L_{\max} > 0$ be a deterministic uniform upper bound on the hinge losses over $\mathcal{X}_{\tilde{\delta}}$, and let $s_i(\mathbf{x}) = \ell_i(\mathbf{x})/(2L_{\max})$. Then $s_i(\mathbf{x}) \in [0, 1/2]$. Write

$$T_\alpha(y) = y + \frac{\alpha}{2\pi}\sin(2\pi y), \qquad m_\alpha := 1 - \alpha, \qquad M_\alpha := 1 + \alpha,$$

where $\alpha = 0.7$. We have

$$T'_\alpha(y) = 1 + \alpha\cos(2\pi y) \in [m_\alpha, M_\alpha]$$

and $T_\alpha(-1) = -1$, $T_\alpha(1) = 1$. Thus T_α is a bijection from $[-1, 1]$ onto itself. For fixed $\mathbf{x}$, maximizing the inner objective is equivalent to solving

$$\min_{\mathbf{y}\in\mathcal{Y}_{\text{syn}}} \varphi_{\mathbf{x}}(\mathbf{y}) := \frac{c}{2Nm_\alpha^2}\sum_{i=1}^{N}\big(T_\alpha(y_i) - s_i(\mathbf{x})\big)^2. \tag{93}$$

Inner solution and nonconvexity. The strict monotonicity of T_α implies that the unique inner maximizer is

$$y_i^\star(\mathbf{x}) = T_\alpha^{-1}(s_i(\mathbf{x})), \qquad i = 1, \dots, N.$$

Since $s_i(\mathbf{x}) \in [0, 1/2]$, this maximizer lies in the interior of the inner feasible interval. Substitution gives the value function

$$\Phi_{\text{syn}}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\Big[\ell_i(\mathbf{x}) + \frac{c}{2}s_i(\mathbf{x})^2\Big] + \kappa\mathcal{R}_\beta(\mathbf{x}).$$

To expose the nonconcave inner geometry, note that $T''_\alpha(y) = -2\pi\alpha\sin(2\pi y)$. At $y_i = -1/4$, direct differentiation gives

$$\frac{\partial^2 f_{\text{syn}}}{\partial y_i^2}(\mathbf{x}, \mathbf{y}) = \frac{c}{Nm_\alpha^2}\Big[-1 + \alpha^2 + \frac{\pi\alpha}{2} + 2\pi\alpha s_i(\mathbf{x})\Big]. \tag{94}$$

For $\alpha = 0.7$, the bracket is bounded below by $-1 + 0.7^2 + 0.35\pi \approx 0.5896 > 0$. Hence $f_{\text{syn}}(\mathbf{x}, \cdot)$ has a positive-curvature direction at $\mathbf{y} = (-1/4)\mathbf{1}$ for every $\mathbf{x}$. At the inner maximizer, in contrast, the same coordinatewise second derivative equals $-c[T'_\alpha(y_i^\star(\mathbf{x}))]^2/(Nm_\alpha^2) < 0$. The inner objective is therefore nonconcave, despite having a unique global maximizer.

Outer-slice nonconvexity. The warped coupling also gives nonconvexity in an outer slice. For the benchmark verification, we impose the mild nondegeneracy condition that $\mathcal{X}$ contains a relative open line segment on which the active-hinge set is nonempty and fixed, and that its direction $\boldsymbol{v}$ has a nonzero slope for at least one active hinge. We choose the segment away from the capped-penalty breakpoints. Fix the admissible inner point $\mathbf{y} = 0$ and consider such a segment. Writing $h = \ell_i(\mathbf{x})$, we have $s_i(\mathbf{x}) = h/(2L_{\max})$ and h is affine in $\mathbf{x}$ on this segment. Therefore, for the corresponding summand,

$$\frac{\mathrm{d}^2}{\mathrm{d}h^2}\big[h + c\,\psi_{\alpha,h/(2L_{\max})}(0)\big] = \frac{c}{4L_{\max}^2}\left(1 - \frac{1}{m_\alpha^2}\right) < 0.$$

Consequently, on the region above and for the chosen direction $\boldsymbol{v}$,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} f_{\text{syn}}(\mathbf{x} + t\boldsymbol{v}, 0) = \frac{c}{4NL_{\max}^2}\left(1 - \frac{1}{m_\alpha^2}\right)\sum_{i\in\mathcal{A}} d_i^2 < 0, \qquad d_i := -b_i\widetilde{\boldsymbol{a}}_i^\top\boldsymbol{v},$$

where $\widetilde{\boldsymbol{a}}_i := (\boldsymbol{a}_i, 1)$ is the feature vector augmented for the bias coordinate and $\mathcal{A}$ is the active-hinge set. Thus the outer slice $f_{\text{syn}}(\cdot, 0)$ contains a strictly concave segment and is not convex. The capped-ℓ_1 term is piecewise affine away from $x_j \in \{0, -\beta, \beta\}$ and supplies the nonsmoothness (and is itself nonconvex as a capped penalty), but it is not needed for the above outer-slice nonconvexity argument.

The fixed-slice calculation establishes the required outer nonconvexity of the minimax objective. It does not by itself determine the convexity of the value function displayed above, and we make no separate convexity claim about that function.

Projected-PŁ condition. The exact inner gap and the ith coordinate of the population inner gradient are

$$\Delta_{\mathrm{syn}}(\mathbf{x},\mathbf{y}) := \Phi_{\mathrm{syn}}(\mathbf{x}) - f_{\mathrm{syn}}(\mathbf{x},\mathbf{y}) = \frac{c}{2Nm_\alpha^2}\sum_{i=1}^{N}\big(T_\alpha(y_i) - s_i(\mathbf{x})\big)^2, \tag{95}$$

$$[\nabla_{\mathbf{y}} f_{\mathrm{syn}}(\mathbf{x},\mathbf{y})]_i = -\frac{c}{Nm_\alpha^2}\big(T_\alpha(y_i) - s_i(\mathbf{x})\big)T'_\alpha(y_i). \tag{96}$$

It follows from $T'_\alpha(y_i) \geq m_\alpha$ that

$$\frac{1}{2}\|\nabla_{\mathbf{y}} f_{\mathrm{syn}}(\mathbf{x},\mathbf{y})\|^2 \geq \frac{c}{N}\Delta_{\mathrm{syn}}(\mathbf{x},\mathbf{y}).$$

It remains to account for the projection. If $y_i \neq y_i^\star(\mathbf{x})$, the mean-value theorem gives a point ξ_i between these two values such that

$$T_\alpha(y_i) - s_i(\mathbf{x}) = T'_\alpha(\xi_i)(y_i - y_i^\star(\mathbf{x})).$$

Using Eqn. (96), the unprojected ascent point is

$$y_i + \lambda[\nabla_{\mathbf{y}} f_{\mathrm{syn}}(\mathbf{x},\mathbf{y})]_i = (1-\theta_i)y_i + \theta_i y_i^\star(\mathbf{x}), \qquad \theta_i := \frac{\lambda c}{Nm_\alpha^2}T'_\alpha(y_i)T'_\alpha(\xi_i).$$

For

$$0 < \lambda \leq \bar{\lambda} := \frac{Nm_\alpha^2}{cM_\alpha^2}, \tag{97}$$

we have $\theta_i \in [0,1]$. The same conclusion is immediate when $y_i = y_i^\star(\mathbf{x})$. Thus every updated coordinate lies on the line segment between y_i and $y_i^\star(\mathbf{x})$, the projection is inactive, and the projected-gradient mapping equals the population gradient. Therefore,

$$\frac{1}{2}\|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x},\mathbf{y},\nabla_{\mathbf{y}} f_{\mathrm{syn}}(\mathbf{x},\mathbf{y}))\|^2 \geq \frac{c}{N}\left[\Phi_{\mathrm{syn}}(\mathbf{x}) - f_{\mathrm{syn}}(\mathbf{x},\mathbf{y})\right].$$

This verifies Assumption 3.2 uniformly in $\mathbf{x}$ with

$$\mu = \frac{c}{N}, \qquad \bar{\lambda} = \frac{N(1-\alpha)^2}{c(1+\alpha)^2}.$$

For $\alpha = 0.7$, one admissible step range is $0 < \lambda \leq 9N/(289c) \approx 0.03114N/c$.

Regularity and stochastic oracles. Because the dataset is finite, $\mathcal{X}_{\bar{\delta}}$ is compact, and the score $h_{\mathbf{x}}$ is linear in $\mathbf{x}$, the scores and hinge losses have finite uniform Lipschitz constants on $\mathcal{X}_{\bar{\delta}}$. The capped-$\ell_1$ term is also Lipschitz. These finitely many locally Lipschitz functions are differentiable on a common full-measure subset of $\mathcal{X}_{\bar{\delta}}$, and the maps s_i are uniformly Lipschitz and differentiable on that same set. Since T_α and its first two derivatives are bounded on $[-1,1]$, $\nabla_{\mathbf{y}} f_{\mathrm{syn}}$ is jointly Lipschitz in $(\mathbf{x},\mathbf{y})$. At common differentiability points, the dependence of the outer gradient on $\mathbf{y}$ is through

$$\frac{c}{Nm_\alpha^2}\sum_{i=1}^{N} T_\alpha(y_i)\nabla s_i(\mathbf{x}),$$

whose difference between two inner points is bounded by a finite constant times $\|\mathbf{y} - \mathbf{y}'\|$. This also gives a problem-specific instance of the constant in Lemma 3.5.

Finally, sample $i \sim \mathrm{Unif}\{1,\dots,N\}$ and define

$$F_{\mathrm{syn}}(\mathbf{x},\mathbf{y};i) := \ell_i(\mathbf{x}) + c\,\psi_{\alpha,s_i(\mathbf{x})}(y_i) + \kappa\mathcal{R}_\beta(\mathbf{x}). \tag{98}$$

Then $\mathbb{E}_i[F_{\mathrm{syn}}(\mathbf{x},\mathbf{y};i)] = f_{\mathrm{syn}}(\mathbf{x},\mathbf{y})$, and

$$\nabla_{\mathbf{y}} F_{\mathrm{syn}}(\mathbf{x},\mathbf{y};i) = -\frac{c}{m_\alpha^2}\big(T_\alpha(y_i) - s_i(\mathbf{x})\big)T'_\alpha(y_i)\boldsymbol{e}_i$$

is an unbiased stochastic y-gradient. Boundedness of the domain and of the displayed derivatives gives finite variance and mean-square Lipschitz constants. The same bounds give a deterministic Lipschitz modulus for each sample function, verifying Assumptions 3.3 and 3.4. The displayed value function is nonnegative, so Assumption 5.2 also holds.

F.2 Experimental Settings and Hyperparameter Search

Data, model, and problem parameters. We use six binary-classification datasets from the LIBSVM binary-data repository. Their dimensions and the exact train–test files used in our experiments are summarized in Table 3. We use the official split whenever separate train and test files are available; otherwise, we construct a deterministic stratified 80/20 split with seed 42. There is no additional validation split because hyperparameters are selected using the primal objective and inner-gap criterion, never test accuracy. Feature standardization is fitted on the training partition only and then applied to the test partition, after which each row is normalized to have norm at most one.

Table 3: Statistics of the binary-classification datasets. "Official" denotes the train–test files distributed by LIBSVM; the other datasets use a stratified 80/20 split of the corresponding LIBSVM file.

Dataset	Train	Test	Features	Split
a9a	32,561	16,281	123	Official
Breast-cancer	546	137	10	Stratified 80/20
Cod-rna	59,535	271,617	8	Official
German	800	200	24	Stratified 80/20
Ijcnn1	49,990	91,701	22	Official
Splice	1,000	2,175	60	Official

All experiments use a linear classifier with hinge loss and the warped nonconvex-PŁ objective described in Section 6.1. The common problem parameters are $\alpha = 0.7$, $c = 1$, initial $y_i = -1/4$, $\kappa = 10^{-4}$, $\beta = 1$, and smoothing radius $\delta = 10^{-4}$. The weight and bias blocks are projected onto Euclidean balls of radius 5, and $\mathcal{Y}_{\mathrm{syn}} = [-1, 1]^N$. The deterministic loss bound is computed with the fixed ambient radius $\bar{\delta} = 0.2$, which contains the smoothing neighborhood. Parameters are initialized with He initialization, and all runs use single-precision arithmetic. Metrics are recorded every 100 iterations and reported over seeds $\{42, 43, 44\}$. The hyperparameters is presented in Table 4. The step-size search and confirmation procedure is described in the next paragraph.

Table 4: Algorithm-specific settings used in the search, confirmation, and matched-time experiments. K_{zo} is the number of zeroth-order directions, while B_{samp} and b_{samp} are stochastic sample counts for refresh and recursive batches.

Algorithm	K_{zo}	q	B_{samp}	b_{samp}	Inner steps	Refresh probability	Additional settings
ZOFOSGDA	5	30	512	1	–	–	deterministic SPIDER refresh/recursion
ZOFO	5	–	–	1	–	–	no SPIDER estimator
TTSGDA	–	–	–	1	–	–	first-order baseline
PGFDA	5	30	512	1	–	$1/30$	100 warm-start and 100 refinement steps
NL–PGFDA	5	–	–	1	10	–	nested loop; inner radius 10^{-4}

The table separates the number of stochastic samples from the number of zeroth-order directions. For ZOFOSGDA, ZOFO, PGFDA, and NL–PGFDA, the refresh and recursive batch sizes are B_{samp} and b_{samp} are the numbers of stochastic sample draws in refresh and recursive batches, respectively, whereas $K_{\mathrm{zo}} = 5$ is the number of independent spherical directions used for each sampled outer point. Consequently, the implementation evaluates $K_{\mathrm{zo}}B_{\mathrm{samp}}$ and $K_{\mathrm{zo}}b_{\mathrm{samp}}$ two-point estimators in the two batch types. Directions attached to the same sampled outer point share its stochastic sample, so these estimator counts are computational workload counts rather than effective theoretical batch sizes. In particular, we do not invoke a $1/(K_{\mathrm{zo}}B_{\mathrm{samp}})$ or $1/(K_{\mathrm{zo}}b_{\mathrm{samp}})$ variance bound. The theory instead uses B_x and b_x for mutually independent oracle–direction pairs and does not introduce a separate direction multiplier.

Search and confirmation. For both block step sizes, we search the natural-exponential grid

$$\mathcal{H} = \{\exp(-k) : k = 0, 1, \ldots, 10\}.$$

For ZOFOSGDA, ZOFO, TTSGDA, and PGFDA, we impose $\eta_x < \eta_y$, leaving 55 pairs per method. NL–PGFDA retains the full $11 \times 11 = 121$ grid as the unconstrained reference baseline.

Only PGFDA uses a stochastic refresh event, with probability $1/q$; NL–PGFDA instead uses its standard nested-loop update with 10 inner steps and no refresh probability. The smoothing radius is fixed at 10^{-4} rather than tuned. Each candidate is first run for 1,000 iterations with seed 42 and constant step sizes. Candidates with inner gap in Eqn. (95) at most 0.1 are ranked by increasing primal objective; if none satisfies the threshold, all successful candidates are ranked by the primal objective. The top three candidates are then confirmed for 5,000 iterations with each of the three seeds, again using constant step sizes. We rerank these candidates using the three-seed mean primal objective under the same inner-gap rule. Test accuracy is never used for hyperparameter selection.

Matched-time evaluation. For each dataset, the mean runtime of the selected NL–PGFDA1 confirmation runs defines the wall-clock budget. The other four methods are rerun independently for the same budget on all three seeds using their confirmed initial step sizes. We use wall-clock time here because SZO and SFO queries have different computational structures and are therefore not naturally comparable after an unweighted sum. All methods are run in the same software and hardware environment; the resulting time comparisons are empirical efficiency measurements rather than replacements for the separate theoretical oracle counts. Table 2 reports the final logged test accuracy, whereas Figures 1 and 2 plot the three-seed mean of $\|G_x^{\mathrm{emp}}\|+\|G_y^{\mathrm{emp}}\|$ against training time with one-sample-standard-deviation bands. The same convention-selected first-order x-derivative and smooth y-gradient described in Section 6.1 are used only to evaluate this proxy.

F.3 Relation between the empirical proxy and GGSP stationarity

The plotted x-quantity is an implementation-dependent diagnostic rather than an exact Goldstein residual. For evaluation only, let $g_x^{\mathrm{sel}}(\mathbf{x},\mathbf{y})$ denote the value returned by automatic differentiation under the implementation's fixed conventions for nonsmooth primitives. We compute

$$G_x^{\mathrm{emp}}(\mathbf{x},\mathbf{y}) := \mathcal{G}_{\mathbf{x},\gamma}(\mathbf{x},\mathbf{y},g_x^{\mathrm{sel}}(\mathbf{x},\mathbf{y})), \qquad G_y^{\mathrm{emp}}(\mathbf{x},\mathbf{y}) := \mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x},\mathbf{y},\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})).$$

If every nonsmooth primitive in the executed computational graph is evaluated away from its kink, the ordinary chain rule applies, so $g_x^{\mathrm{sel}}(\mathbf{x},\mathbf{y}) = \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})$ and hence $g_x^{\mathrm{sel}}(\mathbf{x},\mathbf{y}) \in \partial_{\delta,\mathbf{x}} f(\mathbf{x},\mathbf{y})$. Outside this smooth-branch case, neither piecewise smoothness nor differentiability of the composite slice alone guarantees this inclusion. For example,

$$r(t) := \mathrm{ReLU}(t) - \mathrm{ReLU}(-t) = t$$

is differentiable at $t=0$ with $\partial r(0) = \{1\}$, whereas an implementation using $\mathrm{ReLU}'(0) = 0$ returns the automatic- differentiation value $0 \notin \partial r(0)$. At any remaining evaluation point, the required condition is therefore the separately verified inclusion

$$g_x^{\mathrm{sel}}(\mathbf{x},\mathbf{y}) \in \partial_{\delta,\mathbf{x}} f(\mathbf{x},\mathbf{y}). \tag{99}$$

Membership in the Clarke subdifferential at the current point is one sufficient way to establish Eqn. (99), but it does not follow merely from a piecewise-smooth or ReLU representation.

Define the two blockwise residuals from Definition 4.1 by

$$r_x^{\mathrm{GGSP}}(\mathbf{x},\mathbf{y}) := \min_{\boldsymbol{h}\in\partial_{\delta,\mathbf{x}} f(\mathbf{x},\mathbf{y})} \|\mathcal{G}_{\mathbf{x},\gamma}(\mathbf{x},\mathbf{y},\boldsymbol{h})\|,$$

$$r_y^{\mathrm{GGSP}}(\mathbf{x},\mathbf{y}) := \|\mathcal{G}_{\mathbf{y},\lambda}(\mathbf{x},\mathbf{y},\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}))\|.$$

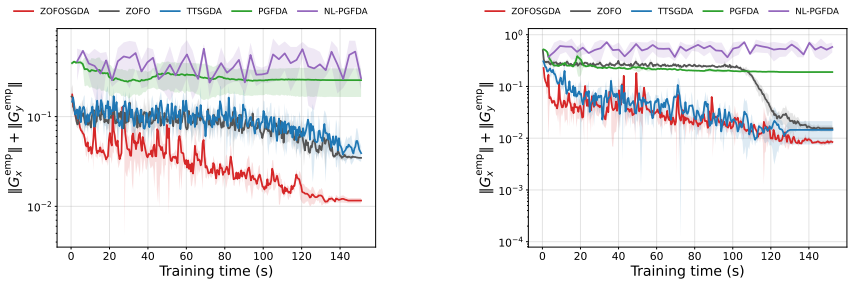
Whenever Eqn. (99) has been established,

$$r_x^{\mathrm{GGSP}}(\mathbf{x},\mathbf{y}) + r_y^{\mathrm{GGSP}}(\mathbf{x},\mathbf{y}) \le \|G_x^{\mathrm{emp}}\| + \|G_y^{\mathrm{emp}}\|.$$

Thus the plotted sum upper-bounds the sum of the two blockwise GGSP residuals at such verified points, rather than directly computing the minimum over the Goldstein subdifferential. At every unverified point it has no certified GGSP upper-bound interpretation and is only an implementation-dependent diagnostic. In neither case does it directly measure the value-function residual in Definition 4.3. The x-updates of ZOFOSGDA still use only zeroth-order queries; the selected derivative is used solely for evaluation.

F.4 Additional Empirical-Proxy Plots

For completeness, Figure 2 reports the same empirical proxy residual for the two remaining benchmark datasets under the matched-time protocol. These datasets are not included in Figure 1 and are shown side by side.



(a) Cod-rna (b) German

Figure 2: The combined empirical proxy residual $\|G_x^{\mathrm{emp}}\| + \|G_y^{\mathrm{emp}}\|$, defined in Figure 1, versus training time on the two remaining benchmark datasets.

F.5 Hyperparameter Ablations

We study the refresh period $q \in \{10, 20, 30, 50\}$, smoothing radius $\delta \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$, and large refresh batch $B \in \{128, 256, 512, 1024\}$. When varying one parameter, we keep the other parameters at their defaults $q = 30$, $\delta = 10^{-4}$, and $B = 512$, and $b = 1$. For every nondefault configuration, the two block step sizes are selected using the same search and three-seed confirmation protocol described in Appendix F.2. We then run seeds $\{42, 43, 44\}$ and compare the residual trajectories over the training-time interval common to all four settings. In the updated plots, the default $q = 30$, $\delta = 10^{-4}$, and $B = 512$ trajectories are taken directly from the five-method matched-time experiment used in Figure 1; the other trajectories are taken from their corresponding ablation runs. The default entries marked by † in Tables 5– 7 reproduce the values reported in Table 2; all other entries are obtained from the corresponding ablation runs.

Table 5: Test accuracy (%) for the refresh-period ablation (mean ± sample standard deviation over three seeds). Bold and underlined entries are the best and second-best displayed means in each column, respectively.

q	a9a	Breast-cancer	Cod-rna	German	Ijcnn1	Splice
10	83.95±0.06	96.84±0.42	92.25±0.02	83.00±0.50	**92.21±0.05**	85.07±0.16
20	83.87±0.06	97.08±0.00	**92.28±0.11**	81.67±0.76	91.70±0.05	84.98±0.25
30†	**84.62±0.15**	**97.84±0.42**	92.22±0.06	**83.67±0.76**	92.20±0.04	**85.12±0.25**
50	83.92±0.11	97.32±0.42	92.19±0.19	81.67±0.76	90.57±0.11	84.23±0.60

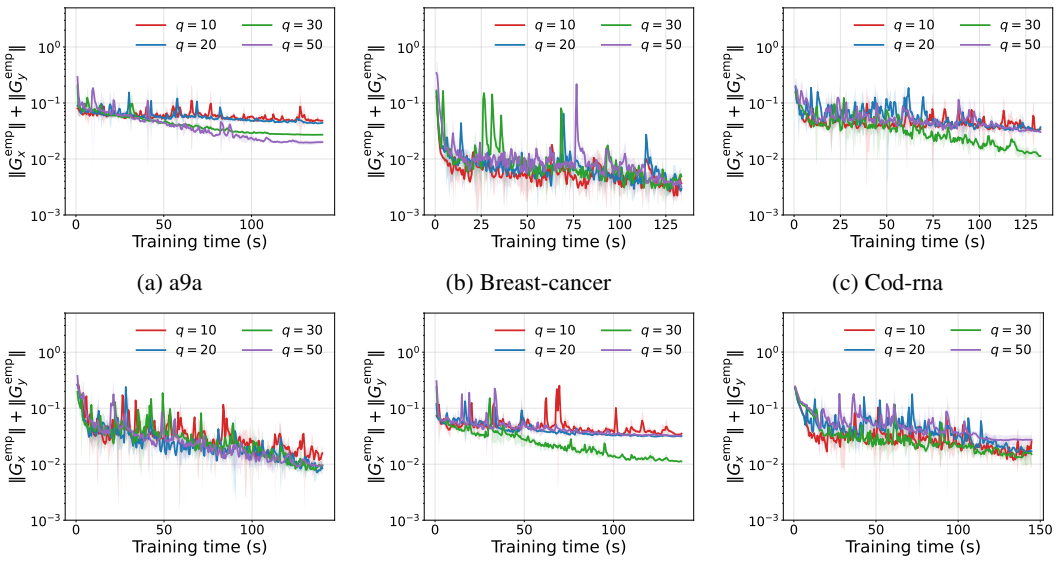
Table 6: Test accuracy (%) for the smoothing-radius ablation. Reporting and highlighting follow Table 5.

δ	a9a	Breast-cancer	Cod-rna	German	Ijcnn1	Splice
10^{-2}	84.04±0.08	97.08±0.00	92.28±0.02	82.83±0.76	91.81±0.20	84.78±0.12
10^{-3}	83.62±0.39	97.08±0.00	**92.29±0.05**	82.83±1.04	92.08±0.20	85.07±0.31
10^{-4}†	**84.62±0.15**	**97.84±0.42**	92.22±0.06	**83.67±0.76**	**92.20±0.04**	**85.12±0.25**
10^{-5}	76.67±0.35	96.59±0.42	92.19±0.10	82.50±0.50	90.50±0.00	84.60±1.08

Table 7: Test accuracy (%) for the large-refresh-batch ablation. Reporting and highlighting follow Table 5.

B	a9a	Breast-cancer	Cod-rna	German	Ijcnn1	Splice
128	83.41±0.16	96.59±0.42	**92.35±0.15**	80.67±1.89	90.51±0.02	84.48±0.45
256	83.47±0.17	97.08±0.00	92.26±0.07	82.83±1.26	90.51±0.02	84.66±0.33
$512^{\dagger}$	**84.62±0.15**	**97.84±0.42**	92.22±0.06	**83.67±0.76**	**92.20±0.04**	**85.12±0.25**
1024	83.14±0.47	97.08±0.73	92.25±0.10	82.33±0.58	90.50±0.00	84.44±0.10

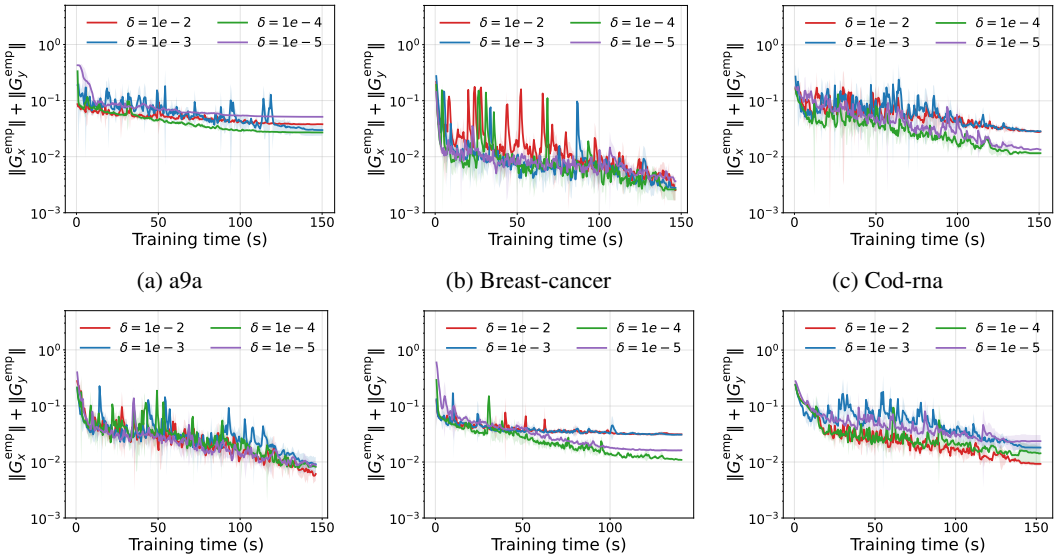
The accuracy results favor the default configuration overall: $q = 30$, $\delta = 10^{-4}$, and $B = 512$ attain the highest displayed accuracy on most datasets. The effect is not uniform across tasks—for example, $q = 10$ is slightly better on Ijcnn1, while $\delta = 10^{-3}$ and $B = 128$ are best on Cod-rna. The very small radius $\delta = 10^{-5}$ causes a pronounced degradation on a9a. Figures 3– 5 complement the terminal accuracy by showing the optimization trajectories. They reveal dataset-dependent tradeoffs between variance reduction and per-iteration cost, so no nondefault setting uniformly dominates the default in wall-clock convergence. All panels use the fixed vertical range $[10^{-3}, 5]$; standard-deviation shading is omitted wherever its lower endpoint would fall below 10^{-3} rather than being artificially clipped to the plotting boundary.



(a) a9a (b) Breast-cancer (c) Cod-rna

(d) German (e) Ijcnn1 (f) Splice

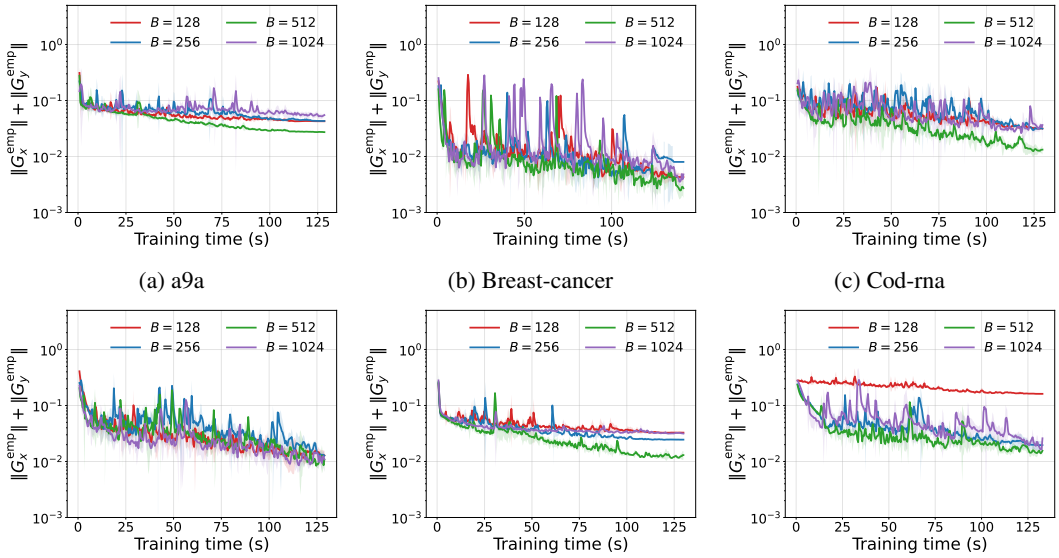
Figure 3: Effect of the refresh period q on the combined empirical proxy residual versus training time. Curves and bands are the three-seed mean and one sample standard deviation, respectively.



(a) a9a (b) Breast-cancer (c) Cod-rna

(d) German (e) Ijcnn1 (f) Splice

Figure 4: Effect of the smoothing radius δ on the combined empirical proxy residual versus training time.



(a) a9a (b) Breast-cancer (c) Cod-rna

(d) German (e) Ijcnn1 (f) Splice

Figure 5: Effect of the large refresh batch B on the combined empirical proxy residual versus training time.

G AUC-MAXIMIZATION NUMERICAL ILLUSTRATION

As a supplementary numerical illustration, we evaluate ZOFOSGDA on AUC maximization with the least-squares surrogate (Yang et al., 2022). The problem admits the minimax formulation

$$\min_{\mathbf{x}=(\theta,\mu_1,\mu_2)} \max_{\mathbf{y}\in\mathcal{Y}} \mathbb{E}_{(\boldsymbol{a},b)}[f(\theta,\mu_1,\mu_2,\mathbf{y};\boldsymbol{a},b)], \qquad \mathcal{Y}=[-C_{\mathrm{AUC}},C_{\mathrm{AUC}}], \tag{100}$$

where $C_{\mathrm{AUC}} > 0$ is fixed before optimization and

$$\begin{aligned} f = {} & (1-p)\big(h(\theta;\boldsymbol{a})-\mu_1\big)^2\mathbb{I}[b=1] + p\big(h(\theta;\boldsymbol{a})-\mu_2\big)^2\mathbb{I}[b=-1] \\ & + 2(1+\mathbf{y})\Big(p\,h(\theta;\boldsymbol{a})\mathbb{I}[b=-1] - (1-p)h(\theta;\boldsymbol{a})\mathbb{I}[b=1]\Big) - p(1-p)\mathbf{y}^2, \end{aligned}$$

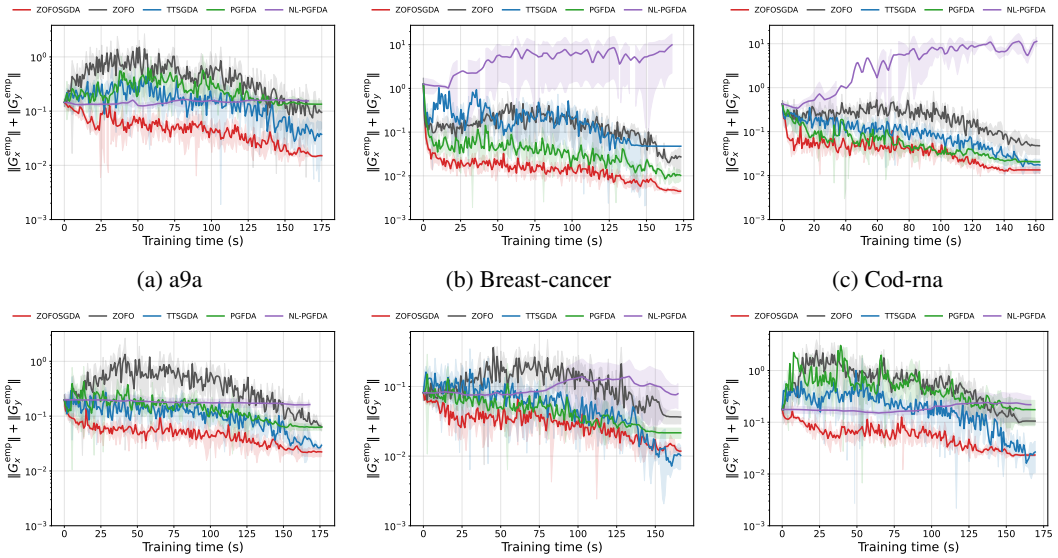
with $p = \mathbb{P}(b=1)$. In the implementation, p is the empirical positive-class proportion in the training split, $C_{\mathrm{AUC}} = 5$, and $(\mu_1,\mu_2) \in [-5,5]^2$. The score function $h(\theta;\boldsymbol{a})$ is a one-hidden-layer multilayer perceptron with 32 ReLU units; each layer's weight block and bias block is projected onto a Euclidean ball of radius 5. We compare ZOFOSGDA with ZOFO, TTSGDA, PGFDA, and NL–PGFDA on the same six binary-classification datasets listed in Table 3: a9a, Breast-cancer, Cod-rna, German, Ijcnn1, and Splice. We use the same train–test splits, preprocessing, algorithm-specific batch sizes, and natural-exponential step-size grid as in Appendix F.2. In particular, the search stage uses 1,000 iterations with constant step sizes, and the top three feasible candidates under the inner-gap criterion are confirmed for 5,000 iterations over seeds $\{42,43,44\}$. Test AUC is not used for hyperparameter selection. The selected methods are evaluated under the matched-time protocol.

Table 8: Matched-time test AUC (%, mean $\pm$ sample standard deviation over seeds $\{42,43,44\}$). The highest mean in each column is bold.

Algorithm	a9a	Breast-cancer	Cod-rna	German	Ijcnn1	Splice
ZOFOSGDA	$\mathbf{89.87 \pm 0.06}$	$\mathbf{99.64 \pm 0.14}$	$\mathbf{99.10 \pm 0.05}$	$\mathbf{83.90 \pm 0.58}$	97.83 ± 0.64	$\mathbf{97.94 \pm 0.08}$
ZOFO	87.27 ± 0.41	99.63 ± 0.19	98.72 ± 0.09	82.58 ± 0.57	89.84 ± 1.86	90.99 ± 0.48
TTSGDA	89.82 ± 0.05	99.36 ± 0.79	99.08 ± 0.02	81.12 ± 1.11	$\mathbf{97.93 \pm 0.12}$	97.07 ± 0.19
PGFDA	84.95 ± 1.38	98.73 ± 0.70	98.66 ± 0.14	82.64 ± 1.43	90.59 ± 1.02	86.48 ± 3.44
NL-PGFDA	51.50 ± 11.12	57.63 ± 34.21	50.97 ± 13.57	53.48 ± 11.26	52.69 ± 8.44	45.23 ± 4.21

Table 8 reports the matched-time test AUC. ZOFOSGDA attains the highest mean on five of the six datasets and ranks among the top two methods on all six. TTSGDA attains the highest mean on Ijcnn1. This comparison provides a second illustration of finite-sample behavior on an AUC-maximization objective with a nonsmooth, nonconvex ReLU score model.

Figure 6 reports the corresponding empirical proxy residual, using the same convention-selected x-derivative and smooth y-gradient described above. Each curve is the three-seed mean of $\|G_x^{\mathrm{emp}}\| + \|G_y^{\mathrm{emp}}\|$, and each shaded region denotes one sample standard deviation. For readability on the logarithmic scale, plot elements below 10^{-3} are not displayed. The conditional upper-bound interpretation in Appendix F.3 applies at smooth-branch points and at other points where membership in Eqn. (99) is separately verified. Otherwise, including unverified ReLU kinks or cancellation points, the plotted quantity is only an implementation-dependent diagnostic.



(a) a9a (b) Breast-cancer (c) Cod-rna

(d) German (e) Ijcnn1 (f) Splice

Figure 6: Combined empirical proxy residual $\|G_x^{\mathrm{emp}}\| + \|G_y^{\mathrm{emp}}\|$ versus training time in the AUC-maximization experiments. Curves and shaded bands show the mean and one sample standard deviation over seeds $\{42, 43, 44\}$, respectively.