# TIGHT LOWER BOUNDS FOR STOCHASTIC NONCONVEX–STRONGLY CONCAVE MINIMAX OPTIMIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We study the stochastic first-order oracle complexity of finding $\epsilon$-stationary points of the primal function in smooth nonconvex-strongly-concave minimax optimization. For sufficiently small $\epsilon$, we establish lower bounds of $\Omega(\kappa L \Delta \sigma^2 \epsilon^{-4})$ under the bounded-variance assumption and $\Omega(\kappa^{3/2} \bar{L} \Delta \sigma \epsilon^{-3})$ under the additional assumption of averaged smoothness. Here, $L$ and $\bar{L}$ denote the smoothness and averaged-smoothness constants, respectively, $\Delta$ is the initial primal gap, $\sigma^2$ bounds the oracle variance, and $\kappa = L/\mu$ or $\bar{L}/\mu$ in the respective settings, where $\mu$ is the strong-concavity parameter. Our bounded-variance lower bound improves the dependence on the condition number from $\kappa^{1/3}$ in previous lower bounds to $\kappa$, while our averaged-smoothness lower bound is the first of its kind. In both settings, the resulting lower bounds match existing upper bounds in their dependence on $\kappa$ and $\epsilon$. Our proofs are based on a unified quadratic lifting construction that transfers a hardness instance for stochastic nonconvex minimization to unconstrained minimax optimization while preserving the required variance and smoothness properties.

## 1 INTRODUCTION

We consider the nonconvex-strongly-concave (NC-SC) minimax problem

$$\min_{x \in \mathbb{R}^{d_x}} \Phi(x), \qquad \Phi(x) := \max_{y \in \mathbb{R}^{d_y}} f(x, y), \tag{1}$$

where $f$ has an $L$-Lipschitz gradient and $f(x, \cdot)$ is $\mu$-strongly concave. Problems of this form arise in a variety of machine learning (Sinha et al., 2017; Liu et al., 2019; Zhu et al., 2023). Strong concavity ensures that the inner maximizer is unique, and the resulting primal function $\Phi$ is differentiable, although generally nonconvex. A central goal is to find a point $\hat{x}$ such that $\|\nabla \Phi(\hat{x})\| \le \epsilon$, using stochastic gradients of $f$. In the deterministic and finite-sum settings, accelerated proximal-point methods achieve the optimal complexity $\mathcal{O}(\sqrt{\kappa}\epsilon^{-2})$, where $\kappa \triangleq L/\mu$ is the condition number, already matching the corresponding lower bounds (Lin et al., 2020b; Zhang et al., 2021; Li et al., 2021; Han et al., 2024).

In the stochastic setting, where $f(x, y) \triangleq \mathbb{E}_\xi[F(x, y; \xi)]$, two common oracle assumptions lead to different complexity guarantees. Under an unbiased stochastic gradient oracle with uniformly bounded variance, existing algorithms achieve complexity $\mathcal{O}(\kappa\epsilon^{-4})$ (Zhang et al., 2024). With the additional assumption of averaged smoothness of the stochastic gradients, the best known upper bounds improve to $\mathcal{O}(\epsilon^{-3})$ (Luo et al., 2020; Zhang et al., 2024; 2026). However, the corresponding lower bounds remain incomplete. Under bounded variance, Li et al. (2021) established an $\Omega(\kappa^{1/3}\epsilon^{-4})$ lower bound for constrained NC–SC problems, leaving a gap in the dependence on $\kappa$ relative to the $\mathcal{O}(\kappa\epsilon^{-4})$ upper bound. Under averaged smoothness, to the best of our knowledge, no lower bound has been established specifically for NC–SC minimax optimization. We therefore ask:

> *Can we establish tight lower bounds for stochastic NC–SC minimax optimization under bounded variance and averaged smoothness?*

**Contributions** In this work, we answer both questions affirmatively. Our contributions are threefold. First, let $\epsilon > 0$ denote the target accuracy and assume $\Phi(0) - \inf_x \Phi(x) \le \Delta$. For an unbiased



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.

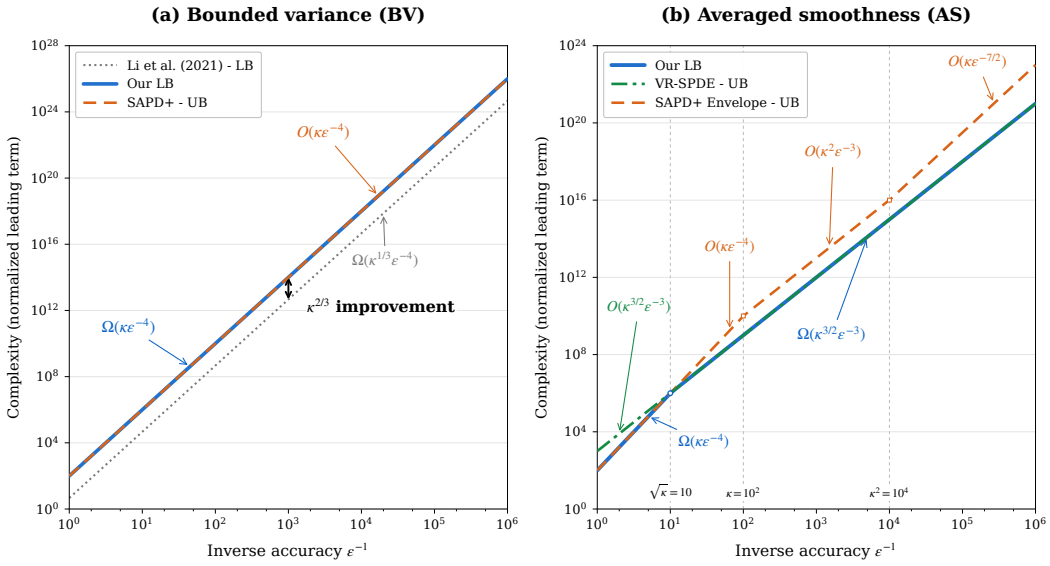



Figure 1: Schematic comparison of previous upper (UB) and lower bounds (LB) with our results under (a) bounded variance and (b) averaged smoothness. Both panels use log–log axes with $r = 1/\varepsilon$, $\kappa = 100$, and $\sigma = 1$, and other constants, parameters and lower-order terms are suppressed. The SAPD+ curve (Zhang et al., 2024) in (b) is the pointwise minimum of its non-VR and VR rate expressions. Overlapping curves indicate matching leading-order dependence, not identical oracle costs.

bounded-variance (BV) oracle with variance bounded by $\sigma^2$, we establish the lower bound, with $\kappa = L/\mu$,

$$\Omega\left(\frac{\kappa L\sigma^2\Delta}{\epsilon^4}\right). \tag{2}$$

Second, for an averaged-smooth (AS) oracle with Lipschitz constant $\bar{L}$, we establish, with $\kappa = \bar{L}/\mu$,

$$\Omega\left(\frac{\bar{L}\Delta}{\epsilon^2}\min\left\{\frac{\kappa\sigma^2}{\epsilon^2}, \frac{\kappa^{3/2}\sigma}{\epsilon}\right\}\right). \tag{3}$$

We refer to Theorems 4.1 and 5.1 for the formal statements, and summarize the comparison with existing results in Table 1 and Figure 1. In the BV setting, our lower bound (2) improves the dependence on $\kappa$ over the previous lower bound of Li et al. (2021) and matches the upper complexity bounds of Zhang et al. (2024) and Zhang et al. (2026) in their dependence on both $\kappa$ and $\epsilon$. To the best of our knowledge, (3) is the first lower bound specifically for the AS setting. Moreover, this lower bound matches the upper complexity bound of Zhang et al. (2026) in its dependence on $\kappa$ and $\epsilon$ when $\epsilon^{-1} \gtrsim \sqrt{\kappa}$, and that of Zhang et al. (2022b) when $\epsilon^{-1} \lesssim \sqrt{\kappa}$.

Finally, we note that the corresponding upper bounds are stated under slightly different problem classes and measures of initial optimality. Specifically, Zhang et al. (2024) and Zhang et al. (2026) use a Moreau-envelope-based stationarity measurement. To facilitate a more direct comparison, we also establish lower bounds using a Moreau envelope stationarity, see Corollary 5.2.

The hard instance is obtained through a quadratic lifting of hard instances for stochastic nonconvex minimization (Carmon et al., 2020; Arjevani et al., 2023). Starting from the hard instance $F_T$ by Carmon et al. (2020), we construct a minimax objective of the form

$$f_T(x, y) = \alpha F_T(\beta y) - \frac{\nu}{2}\|y - \gamma x\|^2,$$

where the parameters $\alpha, \beta, \nu$ and $\gamma$ are chosen so that $f_T(x, \cdot)$ strongly concave. The quadratic coupling is designed so that stationarity of the primal function $\Phi_T(x) = \max_y f_T(x, y)$ implies

Table 1: Summary and Comparison of Lower and Upper Bound Results

| Setting | Method | Metric | Oracle Complexity |
|---|---|---|---|
| BV | SGDA (Lin et al., 2020a) | Primal Stationarity | $\mathcal{O}(\kappa^3\epsilon^{-4})$ |
| | SAPD+ (Zhang et al., 2024) | Moreau Envelope Stationarity | $\mathcal{O}(\kappa\epsilon^{-4})$ |
| | SPDE (Zhang et al., 2026) | Moreau Envelope Stationarity | $\mathcal{O}(\kappa\epsilon^{-4})$ |
| | **Lower Bound** (Li et al., 2021) | Gradient Mapping | $\Omega(\kappa^{1/3}\epsilon^{-4})$ |
| | **Lower Bound (Ours)** Theorem 4.1 | Both Primal and Moreau | $\Omega(\kappa\epsilon^{-4})$ |
| AS | SREDA (Luo et al., 2020) | Primal Stationarity | $\mathcal{O}(\kappa^3\epsilon^{-3})$ |
| | SAPD+ (VR) (Zhang et al., 2024) | Moreau Envelope Stationarity | $O\Big(\kappa\epsilon^{-3}\max\{\kappa,\epsilon^{-1/2}\}\Big)$ |
| | VR-SPDE (Zhang et al., 2026) | Moreau Envelope Stationarity | $\mathcal{O}(\kappa^{3/2}\epsilon^{-3})$ |
| | **Lower Bound (Ours)** Theorem 5.1 | Both Primal and Moreau | $\Omega(\min(\kappa^{3/2}\epsilon^{-3},\kappa\epsilon^{-4}))$ |

**Note:** BV denotes the bounded-variance setting, and AS denotes the averaged-smoothness setting. We use $\kappa = L/\mu$ in the BV setting and $\kappa = \bar{L}/\mu$ in the AS setting. See Theorems 4.1 and 5.1 for the precise assumptions and problem settings. The assumptions in the cited works are not identical; a detailed comparison is provided in Appendix G. AS is a special case of BV, and hence upper bounds established for BV also apply to AS.

approximate stationarity of $F_T$ at the explicitly rescaled point $\beta\gamma x$. In particular, an $\epsilon$-stationary point $\hat{x}$ of $\Phi_T$ yields an appropriately scaled approximate stationary point $\beta\gamma\hat{x}$ of $F_T$. Since this point is obtained directly from $\hat{x}$, the reduction does not require computing the inner maximizer and therefore introduces no additional oracle cost. The construction is compatible with stochastic oracle: we keep the quadratic coupling deterministic and introduce randomness only through the stochastic oracle associated with $F_T$, preventing the oracle variance from growing with $\|x\|$ or $\|y\|$ over the unconstrained domain. The main technical challenge is therefore to establish this stationarity-preserving reduction and tune the scaling parameters so that the lifted instance simultaneously satisfies the required smoothness, strong-concavity and variance conditions.

## 2 Related Literature

**Lower Bounds for Minimax Optimization.** First-order oracle lower bounds for convex–concave and strongly-convex–strongly-concave minimax problems were established by Ouyang & Xu (2021) and Zhang et al. (2022a), while Xie et al. (2020) studied the finite-sum setting. For nonconvex–strongly-concave (NC–SC) problems, Zhang et al. (2021) derived lower bounds for both the deterministic and averaged-smooth finite-sum settings. Han et al. (2024) further studied finite-sum NC–SC problems under an oracle model that allows access to both component gradients and proximal operators. Closest to our setting, Li et al. (2021) established deterministic and stochastic lower bounds for NC–SC problems; under bounded variance, their stochastic lower bound contains the term $\Omega(L\Delta\kappa^{1/3}\sigma^2\epsilon^{-4})$. For a broader overview of oracle-complexity results in minimization and minimax optimization, we refer to Zhang & Hu (2025).

**Nonconvex Minimax Optimization Algorithms.** In the deterministic NC–SC setting, accelerated proximal-point methods achieve near-optimal dependence on the condition number (Lin et al., 2020b; Zhang et al., 2021). In the stochastic setting, two-timescale SGDA attains $\mathcal{O}(\kappa^3\epsilon^{-4})$ stochastic gradient complexity under bounded variance (Lin et al., 2020a; 2025), while SREDA achieves $\mathcal{O}(\kappa^3\epsilon^{-3})$ under averaged smoothness (Luo et al., 2020). Momentum-based and adaptive variants have also been developed (Huang et al., 2022; 2023). SAPD+ combines a proximal-point framework with accelerated saddle-point solvers and also admits a variance-reduced variant (Zhang et al.,

2024). More recently, Zhang et al. (2026) proposed the single-loop methods SPDE and VR-SPDE, with complexities $\mathcal{O}(\kappa\epsilon^{-4})$ under bounded variance and $\mathcal{O}(\kappa^{3/2}\epsilon^{-3})$ under averaged smoothness, respectively. These rates match the $(\kappa, \epsilon)$ dependence of the stochastic terms in our lower bounds. Table 1 summarizes the rates of these methods, with further comparisons deferred to Appendix G. Beyond the NC–SC setting, Li et al. (2026) study smoothing and dual perturbation for nonconvex–concave problems, while Zhang & Xu (2026) develop single-loop methods that attain rates previously achieved by multi-loop schemes. Recent work on NC–SC problems considers generalized smoothness (Gao & Liu, 2026) and regularized Newton methods (Ma et al., 2026).

**Recent Lower Bounds beyond Strong Concavity.** Recent work has extended lower-bound analyses to minimax problems in which strong concavity is relaxed to concavity or the Polyak–Łojasiewicz (PL) condition. For NC–C problems with a compact dual domain, Wu et al. (2026) establish deterministic and stochastic lower bounds of $\Omega(\epsilon^{-3})$ and $\Omega(\epsilon^{-6})$, respectively, for zero-respecting algorithms under Moreau-envelope stationarity. Deterministic lower bounds for projected zero-respecting methods are also developed by Zhang & Xu (2026). Pan et al. (2026) establish the $\Omega(\epsilon^{-3})$ deterministic lower bound for arbitrary first-order algorithms and provide a matching upper bound. For NC–PL problems, Pan & Li (2026) establish an $\Omega(L\Delta\kappa\epsilon^{-2})$ lower bound for arbitrary deterministic first-order methods, matching the known upper bound in the corresponding parameter regime. In particular, their result demonstrates that relaxing strong concavity to the PL condition can lead to a stronger dependence on the condition number $\kappa$ in the deterministic setting. In contrast, our work focuses on stochastic NC–SC problems and characterizes how the bounded-variance and averaged-smoothness oracle models determine the dependence on $\kappa$.

## 3 Problem Setup and Oracle Complexity

We use Euclidean norms, denote $z = (x, y)$, and $\mathrm{supp}(z) \triangleq \{i : z_i \neq 0\}$ denote the support of $z \in \mathbb{R}^d$. For a mapping $T$ between Euclidean spaces, we denote

$$\mathrm{Lip}(T) \triangleq \inf\{C \geq 0 : \|T(u) - T(v)\| \leq C\|u - v\|,\ \forall\, u, v\}$$

as its global Lipschitz constant. Here we study the lower bound under the oracle complexity framework (Nemirovskij & Yudin, 1983), which comes with the following components: a function class, a stochastic oracle, an algorithm class, and a convergence criterion. We will discuss them in details in this section.

**Function class.** We consider nonconvex–strongly-concave (NC–SC) objectives. For $M, \mu, \Delta > 0$, let $\mathcal{F}^{d_x,d_y}(M, \mu, \Delta)$ denote the set of continuously differentiable functions $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ such that, with $z = (x, y)$, $\Phi(x) \triangleq \max_y f(x, y)$, $\Phi_* \triangleq \inf_x \Phi(x) > -\infty$,

$$\|\nabla f(z) - \nabla f(z')\| \leq M\,\|z - z'\| \qquad \forall z, z' \in \mathbb{R}^{d_x+d_y}, \tag{4}$$

$$f(x, y') \leq f(x, y) + \langle \nabla_y f(x, y), y' - y\rangle - \frac{\mu}{2}\,\|y' - y\|^2 \qquad \forall x \in \mathbb{R}^{d_x},\ y, y' \in \mathbb{R}^{d_y}, \tag{5}$$

$$\Phi(0) - \Phi_* \leq \Delta. \tag{6}$$

Both the primal and dual variables are unconstrained, and no convexity in $x$ is imposed. Strong concavity ensures that the inner maximizer $y^*(x)$ exists and is unique for every $x$. Moreover, $\Phi$ is differentiable with $\nabla\Phi(x) = \nabla_x f(x, y^*(x))$, is $M$-weakly convex, and has an $M(1 + M/\mu)$-Lipschitz gradient. These properties hold without any compactness assumption; see Appendix A.

**Oracle class.** We consider *stochastic first-order oracle (SFO)* (Arjevani et al., 2023). An instance is a pair $(f, G)$, including the law of the random seed $\xi$. The oracle is a measurable random vector field and returns a joint gradient estimate $G(z; \xi) = (G_x(z; \xi), G_y(z; \xi))$ such that, for every $z$,

$$\mathbb{E}_\xi G(z; \xi) = \nabla f(z), \qquad \mathbb{E}_\xi\,\|G(z; \xi) - \nabla f(z)\|^2 \leq \sigma^2. \tag{7}$$

The *bounded-variance* (BV) class $\mathcal{I}_{\mathrm{BV}}(L, \mu, \Delta, \sigma)$ requires $f \in \mathcal{F}^{d_x,d_y}(L, \mu, \Delta)$ and Equation (7). The *averaged-smooth* (AS) class $\mathcal{I}_{\mathrm{AS}}(\overline{L}, \mu, \Delta, \sigma)$ requires $f \in \mathcal{F}^{d_x,d_y}(\overline{L}, \mu, \Delta)$, Equation (7), and

$$\mathbb{E}_\xi\,\|G(z; \xi) - G(z'; \xi)\|^2 \leq \overline{L}^2\,\|z - z'\|^2 \qquad \text{for all } z, z'. \tag{8}$$

Here, the same random seed $\xi$ is used to evaluate $G$ at $z$ and $z'$. By unbiasedness and Jensen's inequality, $\|\nabla f(z) - \nabla f(z')\|^2 \le \mathbb{E}_\xi \|G(z;\xi) - G(z';\xi)\|^2$, so Equation (8) implies that $\nabla f$ is $\overline{L}$-Lipschitz. Accordingly, we define the condition number as $\kappa = L/\mu$ in the BV setting and $\kappa = \overline{L}/\mu$ in the AS setting.

### 3.1 Algorithm Class and Lower Bound

**Oracle access.** At round $t$, an algorithm selects $1 \le K_t \le K$ points $z^{t,1}, \dots, z^{t,K_t}$ using its past responses and internal randomness. A fresh independent seed $\xi_t$ is then drawn, and the oracle returns $G(z^{t,1};\xi_t), \dots, G(z^{t,K_t};\xi_t)$. The current points are chosen before any current response is observed; the seed is hidden and cannot be queried again later. We take $K = 1$ in BV and any fixed finite $K$ in AS; $K = 2$ permits paired same-sample differences.

**Algorithm class.** We consider *zero-respecting* stochastic first-order algorithms. Let $g^{t,k} = G(z^{t,k};\xi_t)$ and define the coordinates revealed before round $t$ by

$$\mathcal{S}_t = \bigcup_{s<t} \bigcup_{j=1}^{K_s} \operatorname{supp}(g^{s,j}). \tag{9}$$

An algorithm is zero-respecting if, for every admissible instance, its queries and its output after $R$ rounds satisfy

$$\operatorname{supp}(z^{t,k}) \subseteq \mathcal{S}_t, \qquad \operatorname{supp}((\hat{x},0)) \subseteq \mathcal{S}_{R+1}. \tag{10}$$

Here $(\hat{x},0) \in \mathbb{R}^{d_x+d_y}$ denotes the embedding of the primal output into the joint primal–dual space. The first condition requires every query to use only coordinates revealed by previous oracle gradients, while the second requires every nonzero coordinate of the final output $\hat{x}$ to have been revealed by the end of round $R$. In particular, let $\mathcal{S}_1 = \emptyset$, so first-round queries are at the origin. It covers a lot of algorithms including linear-span algorithms and coordinate-wise updates that preserve zero coordinates.

**Complexity measurement.** Our goal is to find the stationary point $\hat{x}$ in terms of the expectation, i.e.,

$$\mathbb{E}\,\|\nabla\Phi(\hat{x})\|^2 \le \epsilon^2;$$

we also give lower bounds for $\mathbb{E}\,\|\nabla\Phi(\hat{x})\| \le \epsilon$ (refer to Proposition C.2). After $R$ rounds, the number of oracle calls is $N = \sum_{t=1}^{R} K_t$. Each returned joint gradient vector counts once, including all minibatch and refresh samples. Let $\mathcal{A}^{(K)}_{\mathrm{zr},N}$ be the algorithms above using at most $N$ calls on every run. For either instance class $\mathcal{I}$, define

$$\mathcal{C}^{(K)}_{\epsilon,\mathrm{zr}}(\mathcal{I}) = \inf\left\{ N \in \mathbb{N}_0 : \exists A \in \mathcal{A}^{(K)}_{\mathrm{zr},N},\ \sup_{I\in\mathcal{I}} \mathbb{E}_I\,\|\nabla\Phi_I(\hat{x}_A)\|^2 \le \epsilon^2 \right\}. \tag{11}$$

The classes range over all finite dimensions; algorithms may know the dimensions and class parameters. Thus a lower bound rules out a uniform guarantee for every algorithm in this class, not just a particular update rule. Our proof bounds rounds first and uses $R \le N$ to bound calls, without an extra factor $K$. The expected-norm version is defined by replacing the squared norm and $\epsilon^2$ in Equation (11) by the norm and $\epsilon$.

### 3.2 Hard Instance Construction

Let $H : \mathbb{R}^d \to \mathbb{R}$ have an $\ell$-Lipschitz gradient, and let $g_H$ be an unbiased estimator of $\nabla H$. For positive $h, q, \gamma$, set

$$\alpha = q^2/h, \qquad \beta = h/q, \qquad \nu = \mu + \ell h, \tag{12}$$

and define the minimax objective

$$f_H(x,y) = \alpha H(\beta y) - \frac{\nu}{2}\,\|y - \gamma x\|^2. \tag{13}$$

We evaluate the quadratic gradient exactly and randomize only the estimate of $\nabla H$:

$$G_x = \nu\gamma(y - \gamma x), \qquad G_y = q g_H(\beta y;\xi) - \nu(y - \gamma x). \tag{14}$$

If $x_i = y_i = 0$, then $[G_x]_i = 0$ and $[G_y]_i = q[g_H(\beta y; \xi)]_i$. Thus, the quadratic coupling cannot reveal a new coordinate on its own; new coordinates enter the support only through the base oracle $g_H$. The following proposition is the main reduction used in proofs.

**Proposition 3.1** (Lifted Function)**.** *Suppose $H$ is twice continuously differentiable and bounded below, with $H(0) - \inf H \le D$ and $\|\nabla H(0)\| \le g$. Then $f_H(x, \cdot)$ is $\mu$-strongly concave and*

$$\mathrm{Lip}(\nabla f_H) \le \nu(1+\gamma^2) + \ell h, \tag{15}$$

$$\Phi_H(0) - \inf \Phi_H \le \alpha D + \frac{q^2 g^2}{2\mu}, \tag{16}$$

$$\|\nabla H(\beta\gamma x)\| \le \frac{1 + \ell h/\nu}{\gamma q} \|\nabla \Phi_H(x)\| \qquad \forall x. \tag{17}$$

*If $g_H$ has variance at most $v^2$, the lifted oracle has variance at most $q^2 v^2$. If $g_H$ has averaged-smoothness constant $S$, the lifted constant is at most $\nu(1+\gamma^2) + hS$.*

**Remark 3.2.** The above result is important and characterizes several important properties of the lifted function $f_H$. We want to highlight Equation (17) above, with this lower bound result, we transfer the lower bound study on the primal function $\nabla \Phi_H$ to that of $\nabla H$, which is well-studied in nonconvex minimization lower bound literature (Carmon et al., 2020; Arjevani et al., 2023). Moreover, note that

$$G(x, y; \xi) - \nabla f_H(x, y) = \big(0, q[g_H(\beta y; \xi) - \nabla H(\beta y)]\big). \tag{18}$$

Thus the unbounded quadratic gradient contributes no noise, which allows a uniform variance bound on the whole space.

**Choice of $H$ and $g_H$.** Regarding the hard instance function $H$ and oracle $g_H$ for lower bound derivations, we take $H : \mathbb{R}^T \to \mathbb{R}$ to be the hard instance function introduced by Carmon et al. (2020) (i.e., the function $F_T$ therein), which satisfies the zero-chain property, and used in the stochastic nonconvex lower bounds of Arjevani et al. (2023). For an integer $T \ge 2$, it is defined by

$$H(u) = -\Psi(1)\Phi_0(u_1) + \sum_{i=2}^{T} \big[\Psi(-u_{i-1})\Phi_0(-u_i) - \Psi(u_{i-1})\Phi_0(u_i)\big], \tag{19}$$

where

$$\Psi(t) = \begin{cases} 0, & t \le \frac{1}{2}, \\ \exp\big(1 - (2t-1)^{-2}\big), & t > \frac{1}{2}, \end{cases} \qquad \Phi_0(t) = \sqrt{e} \int_{-\infty}^{t} e^{-s^2/2}\, ds.$$

Each summand couples two consecutive coordinates. Starting from the origin, a zero-respecting method using exact gradients can reveal at most one new coordinate per query. Moreover, $u_T = 0$ implies $\|\nabla H(u)\| > 1$, so the method must reach the last coordinate before finding a point with a small gradient. The gradient Lipschitz constant is independent of $T$, and the initial gap is $\mathcal{O}(T)$.

For the estimator $g_H$, we use the stochastic gradient constructions of Arjevani et al. (2023). Additional properties of the resulting stochastic hard instances are collected in Appendix B. In both the BV and AS settings, the lifted constructions preserve the probabilistic zero-chain structure of the underlying hard instance; we describe this property next.

### 3.3 Probabilistic Zero-Chain Mechanism

For $u \in \mathbb{R}^T$ and $a \ge 0$, define

$$\mathrm{prog}_a(u) := \max\big(\{i \in \{1, \dots, T\} : |u_i| > a\} \cup \{0\}\big).$$

In particular, $\mathrm{prog}_0(u)$ is the largest index in $\mathrm{supp}(u)$, with value zero when $u = 0$.

**Definition 3.3** (First-Order Zero-Chain)**.** A differentiable function $H : \mathbb{R}^T \to \mathbb{R}$ is a first-order zero-chain if

$$\mathrm{prog}_0(\nabla H(u)) \le \mathrm{prog}_0(u) + 1 \qquad \text{for every } u \in \mathbb{R}^T.$$

For a zero-respecting method starting from the origin, an exact-gradient query can therefore reveal at most one new coordinate. The stochastic counterpart controls how often a new coordinate can be revealed. We use the following definition from Arjevani et al. (2023).

**Definition 3.4** (Probability-$p$ Zero-Chain)**.** Let $p \in (0, 1]$, and let $g_H(u; \xi)$ be an unbiased gradient estimator of $H$. We call $g_H$ a probability-$p$ zero-chain (PZC) if

$$\begin{aligned} \mathbb{P}_\xi\Big(\exists u \in \mathbb{R}^T : \mathrm{prog}_0(g_H(u;\xi)) = \mathrm{prog}_{1/4}(u) + 1\Big) &\le p, \\ \mathbb{P}_\xi\Big(\exists u \in \mathbb{R}^T : \mathrm{prog}_0(g_H(u;\xi)) > \mathrm{prog}_{1/4}(u) + 1\Big) &= 0. \end{aligned} \tag{20}$$

A single seed is used for all points inside each event. Thus the definition controls an entire same-seed batch, not just each query separately.

**Lemma 3.5** (Progress of the Lifted Oracle)**.** *Suppose $g_H$ is a probability-$p$ zero-chain in the sense of Definition 3.4, and let $G$ be the lifted oracle in Equation (14), with $d_x = d_y = T$. For any zero-respecting algorithm following the oracle protocol of Section 3 and returning $\hat{x}$ after at most $R$ rounds,*

$$\mathbb{P}(\mathrm{supp}(\hat{x}) \subseteq \{1, \dots, T-1\}) \ge 1 - \frac{pR}{T}.$$

*In particular, $R \le T/(4p)$ implies $\mathbb{P}(\hat{x}_T = 0) \ge 3/4$.*

This adapts the PZC progress argument of Arjevani et al. (2023, Lemma 1) to the quadratic lift. Appendix C detailed the discussion and characterizes the PZC property of the two base oracles, proves Lemma 3.5, and combines it with Equation (17) to obtain the gradient lower bound in Proposition C.2.

## 4 The Bounded Variance Case Lower Bound

In this section, we derive the lower complexity bound of stochastic NC-SC minimax optimization problems, under the generic bounded variance setting.

**Theorem 4.1** (Bounded Variance)**.** *There are universal constants $c, c_0 > 0$ such that, for $L, \mu, \Delta > 0$, $\sigma \ge 0$, $\kappa = L/\mu \ge 8$, and $0 < \epsilon^2 \le c_0 L\Delta$,*

$$\mathcal{C}^{(1)}_{\epsilon,\mathrm{zr}}\big(\mathcal{I}_{\mathrm{BV}}(L, \mu, \Delta, \sigma)\big) \ge c\frac{L\Delta}{\epsilon^2} \max\left\{1, \frac{\kappa\sigma^2}{\epsilon^2}\right\}. \tag{21}$$

*The same order holds for the expected-norm target. The hard instances have actual smoothness-to-strong-concavity ratio within universal constant factors of $\kappa$.*

**Proof Sketch.** We use the hard instance function and oracle as discussed in Section 3.2, which are motivated by Carmon et al. (2020); Arjevani et al. (2023). Its gradient stays bounded away from zero until its last coordinate is reached. The oracle reveals the next coordinate with probability at most $p$, with variance $\mathcal{O}(p^{-1})$ independent of $T$. In Equation (12), we choose

$$\gamma^2 \asymp \kappa, \qquad q^2 \asymp \epsilon^2/\kappa, \qquad h \asymp \mu, \qquad p \asymp \min\{1, \epsilon^2/(\kappa\sigma^2)\}, \tag{22}$$

where $p = 1$ when $\sigma = 0$. Proposition 3.1 guarantees that the stationarity transfer fixes $\gamma q$ at a constant multiple of $\epsilon$. A zero-respecting algorithm needs order $T/p$ rounds to reveal the chain. The dependence on $\kappa$ comes from the smaller reveal probability. Appendices B to D present the detailed oracle definition, support argument, and all parameter checks.

**Comparison with Existing Results.** The previous lower bound for the BV setting is $\Omega(\kappa^{1/3}\epsilon^{-4})$ (Li et al., 2021). Our Theorem 4.1 strengthens the dependence on the condition number from $\kappa^{1/3}$ to $\kappa$, while allowing both the primal and dual variables to remain unconstrained. On the upper-bound side, SAPD+ and SPDE achieve a leading stochastic complexity of $\mathcal{O}(\kappa\epsilon^{-4})$ for Moreau-envelope stationarity (Zhang et al., 2024; 2026). In Section 5.2, we extend our lower bound to the same stationarity criterion, while Section 5.3 discusses the remaining differences in assumptions.

## 5 The Averaged-Smooth Lower Bound

Building on the BV analysis, we now impose the additional averaged-smoothness assumption on the stochastic oracle and derive a sharper lower bound.

**Theorem 5.1** (Averaged Smoothness)**.** *There are universal constants $c, c_0, c_1 > 0$ such that the following holds for every fixed finite $K \geq 1$. Let $\overline{L}, \mu, \Delta > 0$, $\sigma \geq 0$, $\kappa = \overline{L}/\mu \geq 8$, and suppose*

$$0 < \epsilon^2 \leq c_0 \overline{L}\Delta, \qquad \epsilon\sigma \leq c_1 \overline{L}\Delta\sqrt{\kappa}. \tag{23}$$

*Then*

$$\mathcal{C}^{(K)}_{\epsilon,\mathrm{zr}}\big(\mathcal{I}_{\mathrm{AS}}(\overline{L}, \mu, \Delta, \sigma)\big) \geq c\frac{\overline{L}\Delta}{\epsilon^2} \max\left\{1, \min\left\{\frac{\kappa\sigma^2}{\epsilon^2}, \frac{\kappa^{3/2}\sigma}{\epsilon}\right\}\right\}. \tag{24}$$

*The same order holds for the expected-norm target.*

### 5.1 Proof Sketch

The stochastic oracle used in the BV analysis does not, in general, satisfy the averaged-smoothness requirement. We therefore adopt the smooth-gating construction of Arjevani et al. (2023), which preserves variance of order $\mathcal{O}(p^{-1})$ while achieving an averaged-smoothness constant of order $\mathcal{O}(p^{-1/2})$. As a consequence, Proposition 3.1 imposes the additional constraint $h \lesssim \min\{\mu, \overline{L}\sqrt{p}\}$. With the same choices $\gamma^2 \asymp \kappa$ and $q^2 \asymp \epsilon^2/\kappa$ as in the BV case, the resulting complexity scales as

$$\frac{T}{p} \asymp \frac{\Delta h}{q^2 p} \asymp \frac{\overline{L}\Delta}{\epsilon^2} \min\left\{p^{-1}, \kappa p^{-1/2}\right\}. \tag{25}$$

Choosing $p \asymp \min\{1, \epsilon^2/(\kappa\sigma^2)\}$ then yields Equation (24). The conditions in Equation (23) ensure that the chain length is admissible and that the initial primal gap remains within $\Delta$. Full calibration details are given in Appendix E.

Returning to Equation (24), the lower bound can be expressed across three regimes:

$$\overline{L}\Delta\epsilon^{-2} \times \begin{cases} 1, & \sigma\epsilon^{-1} \leq \kappa^{-1/2}, \\ \kappa\sigma^2\epsilon^{-2}, & \kappa^{-1/2} < \sigma\epsilon^{-1} < \sqrt{\kappa}, \\ \kappa^{3/2}\sigma\epsilon^{-1}, & \sigma\epsilon^{-1} \geq \sqrt{\kappa}. \end{cases} \tag{26}$$

In particular, in the high-accuracy regime, where $\epsilon$ is sufficiently small, the lower bound reduces to $\Omega(\overline{L}\Delta\kappa^{3/2}\sigma\epsilon^{-3})$. Moreover, when $L = \overline{L}$, the AS lower bound is always no larger than the corresponding BV lower bound.

### 5.2 Lower Bounds for Moreau Envelope Stationarity

Theorems 4.1 and 5.1 measure stationarity using the primal gradient. But recent NC-SC literature of SAPD+ and SPDE (Zhang et al., 2024; 2026) used Moreau envelope stationarity as the measurement. Our primal stationarity lower bound results in Theorem 5.1 does not provide a lower bound for this criterion, so a separate transfer is needed. Set $M = L$ in the BV setting and $M = \overline{L}$ in the AS setting, and define

$$P_f(x) := \Phi_{1/(2M)}(x) = \min_{u \in \mathbb{R}^{d_x}} \{\Phi(u) + M\|u - x\|^2\}. \tag{27}$$

Lemma A.1 ensures that $P_f$ is well-defined and continuously differentiable.

On the hard instances used in our proofs, the two stationarity measures are equivalent up to universal constants. Specifically, Lemma F.2 shows that

$$\frac{123}{128}\|\nabla P_{f_H}(x)\| \leq \|\nabla \Phi_H(x)\| \leq \frac{133}{128}\|\nabla P_{f_H}(x)\| \qquad \text{for every } x. \tag{28}$$

Thus here a small Moreau-envelope gradient would also imply a small primal gradient, with a constant factor independent of $\kappa$. Combining this comparison with the primal-gradient lower bounds in Proposition C.2, we have the following result.

**Corollary 5.2** (Moreau Envelope Lower Bounds)**.** *The lower bounds in Theorems 4.1 and 5.1 also hold, up to universal constant factors, when the required output satisfies either*

$$\mathbb{E}\|\nabla P_f(\hat{x})\|^2 \leq \epsilon^2 \qquad \textit{or} \qquad \mathbb{E}\|\nabla P_f(\hat{x})\| \leq \epsilon.$$

*Here the parameters, oracle protocols, and algorithm class are unchanged.*

Appendix F proves the gradient comparison and the corollary. Both gradients are evaluated at the algorithm's output, so no additional oracle call or proximal computation is needed. The resulting bounds preserve the dependence on $\kappa$ and $\epsilon$ and can be compared directly with upper bounds stated for Moreau-envelope stationarity.

### 5.3 Comparison with Existing Upper Bounds

For primal stationarity, SREDA established $\mathcal{O}(\kappa^3\epsilon^{-3})$ SFO complexity under averaged smoothness (Luo et al., 2020), while it further assumes a compact dual domain and sample-wise concavity. Its expected-norm criterion is also covered by our lower bounds (refer to Proposition C.2). For Moreau envelope stationarity, SAPD+ (Zhang et al., 2024, the revised version) gives

$$\mathcal{O}\left(\frac{L\mathcal{G}_0}{\epsilon^2}\max\{\kappa, \sqrt{\sigma/\epsilon}\}(1+\kappa\sigma/\epsilon)\right) \tag{29}$$

for its variance-reduced variant (Zhang et al., 2024). Consequently, its bound is not uniformly $\mathcal{O}(\kappa^2\epsilon^{-3})$ as accuracy $\epsilon \to 0$.

More recently, Zhang et al. (2026) proposed VR-SPDE and obtained $\mathcal{O}(\overline{L}\kappa^{3/2}\sigma\epsilon^{-3})$ for the AS case. By Corollary 5.2, we can find that the result of Zhang et al. (2026) tightly matches our lower bound in the high-accuracy regime. Also as shown in Table 1 and Figure 1, in the low-accuracy regime, the $\mathcal{O}(\kappa\epsilon^{-4})$ upper bound by SAPD+ (Zhang et al., 2024) also tightly matches our lower bound, so our lower bound reveals the optimal SFO complexity in terms of $\kappa$ and $\epsilon$ for the AS case.

## 6 Conclusion and Future Work

We establish stochastic first-order lower bounds for fully unconstrained NC–SC minimax optimization over zero-respecting algorithms. For sufficiently small $\epsilon$, we prove lower bounds of $\Omega(L\Delta\kappa\sigma^2\epsilon^{-4})$ under bounded variance and $\Omega(\overline{L}\Delta\kappa^{3/2}\sigma\epsilon^{-3})$ under averaged smoothness. Both results follow from a unified construction that couples a nonconvex zero-chain with a deterministic quadratic lifting, and both extend to Moreau envelope stationarity without worsening the dependence on $\kappa$. These lower bounds match recent stochastic upper bounds in terms of their dependence on $\kappa$ and $\epsilon$.

Several questions remain open. First, it would be useful to extend the lower bounds beyond zero-respecting algorithms to more general, possibly randomized, first-order methods. Another direction is to better align the domain and initialization assumptions in upper and lower bounds. Finally, our zero-chain hard instances require a dimension that grows with the target accuracy; whether comparable lower bounds hold in fixed or low dimensions, in the spirit of Chewi et al. (2023), remains open.

## Lean Formalization

The supplementary material also includes a Lean formalization that provides machine-checked verification of key components of the analysis.

## AI USE STATEMENT

In this work, we used generative AI tools, primarily ChatGPT, to assist with exploring candidate hard instance constructions and proof strategies, refining mathematical arguments, surveying relevant literature, and drafting and polishing part of the manuscript. The authors formulated the research questions and problem settings, evaluated and selected among candidate approaches, and independently verified all mathematical claims and proofs included in the final paper. The authors have reviewed the work and take full responsibility for the final content of this paper.

## REFERENCES


Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization: Y. arjevani et al. *Mathematical Programming*, 199(1):165–214, 2023. 1, 3, 3.2, 3.2, 3.2, 3.3, 3.3, 4, 5.1, B.1, B.2, B.3, C.1

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184:71–120, 2020. 1, 3.2, 3.2, 4, B.1, B.1

Sinho Chewi, Sébastien Bubeck, and Adil Salim. On the complexity of finding stationary points of smooth functions in one dimension. In *International Conference on Algorithmic Learning Theory*, pp. 358–374. PMLR, 2023. 6

Yan Gao and Yongchao Liu. An efficient stochastic first-order algorithm for nonconvex-strongly concave minimax optimization beyond lipschitz smoothness. *arXiv preprint arXiv:2603.04940*, 2026. 2

Yuze Han, Guangzeng Xie, and Zhihua Zhang. Lower complexity bounds of finite-sum optimization problems: The results and construction. *Journal of Machine Learning Research*, 25(2):1–86, 2024. 1, 2, G.1

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. 2, G.2

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. 2, G.2

Haochuan Li, Yi Tian, Jingzhao Zhang, and Ali Jadbabaie. Complexity lower bounds for nonconvex-strongly-concave min-max optimization. *Advances in Neural Information Processing Systems*, 34:1792–1804, 2021. 1, 1, 1, 2, 4, G.1

Jiajin Li, Mahesh Nagarajan, Siyu Pan, and Nanxi Zhang. Smoothing meets perturbation: Unified and tight analysis for nonconvex-concave minimax optimization. *arXiv preprint arXiv:2602.14185*, 2026. 2

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, 2020a. 1, 2, A.1, G.2

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on learning theory*, pp. 2738–2779. PMLR, 2020b. 1, 2, A.1, A.1

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. 2, G.2

Mingrui Liu, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. Stochastic auc maximization with deep neural networks. 2019. 1

Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. volume 33, pp. 20566–20577, 2020. 1, 1, 2, 5.3, G.2

Bohao Ma, Nachuan Xiao, and Junyu Zhang. A single-loop regularized newton method for nonconvex-strongly-concave minimax optimization. *arXiv preprint arXiv:2606.07003*, 2026. 2

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983. 3

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems: Y. ouyang, y. xu. *Mathematical Programming*, 185(1): 1–35, 2021. 2

Siyu Pan and Jiajin Li. Lower bounds for nonconvex-p $\{\backslash L\}$ minimax optimization. *arXiv preprint arXiv:2608.26799*, 2026. 2, G.1

Siyu Pan, Taoli Zheng, and Jiajin Li. Optimal deterministic first-order oracle complexity for nonconvex-concave minimax optimization. *arXiv preprint arXiv:2609.14235*, 2026. 2, G.1

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. 2017. 1, A.1, A.1

Qilong Wu, Zhihao Gu, and Junchi Yang. Lower bounds for nonconvex-concave minimax optimization. *arXiv preprint arXiv:2609.14233*, 2026. 2, G.1

Guangzeng Xie, Luo Luo, Yijiang Lian, and Zhihua Zhang. Lower complexity bounds for finite-sum convex-concave minimax optimization problems. In *International Conference on Machine Learning*, pp. 10504–10513. PMLR, 2020. 2

Huiling Zhang, Minhao Zhang, and Zi Xu. Single-loop stochastic projected damped extragradient methods for stochastic nonconvex–(strongly) concave minimax optimization. *arXiv preprint arXiv:2609.21747*, 2026. 1, 1, 1, 2, 4, 5.2, 5.3, G.3

Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, 194(1):901–935, 2022a. 2

Minghao Zhang and Zi Xu. Matching multi-loop complexities with a single loop: Optimal optimization stationarity and best-known game stationarity in nonconvex–concave minimax optimization. *arXiv preprint arXiv:2609.17973*, 2026. 2, 2, G.1

Siqi Zhang and Yifan Hu. Avoid overclaims: Summary of complexity bounds for algorithms in minimization and minimax optimization. 2025. URL https://iclr-blogposts.github.io/2025/blog/opt-summary/. 2

Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in artificial intelligence*, pp. 482–492. PMLR, 2021. 1, 2, 2, G.1

Xuan Zhang, Necdet Serhat Aybat, and Mert Gurbuzbalaban. Sapd+: An accelerated stochastic method for nonconvex-concave minimax problems. volume 35, pp. 21668–21681, 2022b. 1

Xuan Zhang, Necdet Serhat Aybat, and Mert Gürbüzbalaban. SAPD+: An accelerated stochastic method for nonconvex-concave minimax problems. 2024. 1, 1, 1, 1, 2, 4, 5.2, 5.3, 5.3

Dixian Zhu, Bokun Wang, Zhi Chen, Yaxing Wang, Milan Sonka, Xiaodong Wu, and Tianbao Yang. Provable multi-instance deep auc maximization with stochastic pooling. In *International Conference on Machine Learning*, pp. 43205–43227. PMLR, 2023. 1

# A PRELIMINARIES AND PROOF OF THE QUADRATIC LIFT

We collect the standard value-function properties, prove Proposition 3.1, and verify the actual condition numbers of the constructed instances.

## A.1 STANDARD VALUE-FUNCTION PROPERTIES

The following lemma supplies the facts used in Section 3 and in the Moreau-envelope argument of Appendix F. The sensitivity and differentiability statements are standard; see Sinha et al. (2017, Lemma 1) and Lin et al. (2020a, Lemma 4.3). The envelope formula is given in Lin et al. (2020b, Lemma 17).

**Lemma A.1** (Primal function and Moreau envelope)**.** *Let $M, \mu > 0$, and let $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ be continuously differentiable, with an $M$-Lipschitz gradient, and suppose $f(x, \cdot)$ is $\mu$-strongly concave for every $x$. Then $y^*(x) := \arg\max_y f(x, y)$ exists uniquely, the value function $\Phi(x) := \max_y f(x, y)$ is continuously differentiable, and*

$$\mathrm{Lip}(y^*) \le \frac{M}{\mu}, \qquad \nabla\Phi(x) = \nabla_x f(x, y^*(x)), \qquad \mathrm{Lip}(\nabla\Phi) \le M\left(1 + \frac{M}{\mu}\right). \tag{30}$$

*Moreover, $\Phi$ is $M$-weakly convex. If $\inf\Phi > -\infty$, then*

$$u_x := \arg\min_u \{\Phi(u) + M\|u - x\|^2\}$$

*exists uniquely, and the Moreau envelope $P_f := \Phi_{1/(2M)}$ satisfies*

$$\nabla P_f(x) = 2M(x - u_x) = \nabla\Phi(u_x). \tag{31}$$

*Proof.* Strong concavity gives

$$f(x, y) \le f(x, 0) + \langle \nabla_y f(x, 0), y\rangle - \frac{\mu}{2}\|y\|^2 \longrightarrow -\infty \quad (\|y\| \to \infty),$$

so the inner maximum is attained uniquely on the full space. The joint smoothness assumption bounds each blockwise Lipschitz constant by $M$. Applying Sinha et al. (2017, Lemma 1) with these constants gives Equation (30). For each fixed $y$, the function $f(\cdot, y) + M\|\cdot\|^2/2$ is convex; its pointwise supremum is $\Phi + M\|\cdot\|^2/2$, proving weak convexity. When $\inf\Phi > -\infty$, the proximal objective is coercive and $M$-strongly convex, so $u_x$ exists uniquely. Lin et al. (2020b, Lemma 17) gives the first equality in Equation (31); the second is the first-order optimality condition at $u_x$. □

## A.2 PROOF OF PROPOSITION 3.1

The smoothness, gap, and oracle estimates are used in Appendices D and E; the stationarity transfer is used in Proposition C.2.

*Proof of Proposition 3.1. Strong concavity, smoothness, and gap.* Since $\alpha\beta = q$ and $\alpha\beta^2 = h$,

$$\nabla_x f_H = \nu\gamma(y - \gamma x), \qquad \nabla_y f_H = q\nabla H(\beta y) - \nu(y - \gamma x). \tag{32}$$

The bound $\|\nabla^2 H\| \le \ell$ gives $\nabla^2_{yy} f_H = h\nabla^2 H(\beta y) - \nu I \preceq -\mu I$. The Hessian of the quadratic coupling is

$$-\nu\begin{pmatrix}\gamma^2 I & -\gamma I\\ -\gamma I & I\end{pmatrix},$$

whose operator norm is $\nu(1 + \gamma^2)$. The remaining Hessian has norm at most $\ell h$, proving Equation (15). Lemma A.1 now gives a unique inner maximizer and the primal-gradient formula.

At the origin, smoothness of $H$ yields

$$f_H(0, y) \le \alpha H(0) + q\langle\nabla H(0), y\rangle - \frac{\mu}{2}\|y\|^2, \qquad \Phi_H(0) \le \alpha H(0) + \frac{q^2 g^2}{2\mu}.$$

For any $x$, the feasible choice $y = \gamma x$ gives

$$\Phi_H(x) \ge \alpha H(\beta\gamma x) \ge \alpha \inf H > -\infty. \tag{33}$$

Subtracting this lower bound from the bound at the origin proves Equation (16).

*Explicit stationarity transfer.* Set $z^* = \beta y^*(x)$ and $w = \beta\gamma x$. The inner first-order condition and the primal-gradient formula give

$$z^* - w = \frac{h}{\nu}\nabla H(z^*), \qquad \nabla\Phi_H(x) = \gamma q \nabla H(z^*). \tag{34}$$

Consequently,

$$\|\nabla H(w)\| \le \|\nabla H(z^*)\| + \ell\,\|w - z^*\| \le \left(1 + \frac{\ell h}{\nu}\right)\|\nabla H(z^*)\|,$$

which proves Equation (17). For the scaled function $H_{\mathrm{sc}}(u) := \alpha H(\beta u)$, this also gives

$$\|\nabla H_{\mathrm{sc}}(\gamma x)\| \le \frac{1 + \ell h/\nu}{\gamma}\,\|\nabla\Phi_H(x)\|.$$

*Stochastic oracle.* The error identity Equation (18) and unbiasedness of $g_H$ give unbiasedness of $G$ and the variance bound $q^2v^2$. If $g_H$ has root mean-squared Lipschitz constant at most $S$, then Minkowski's inequality gives

$$\begin{aligned}\left(\mathbb{E}\,\|G(z;\xi) - G(z';\xi)\|^2\right)^{1/2} &\le \nu(1+\gamma^2)\,\|z - z'\| + qS\,\|\beta(y - y')\| \\ &\le \left[\nu(1+\gamma^2) + hS\right]\|z - z'\|,\end{aligned}$$

which concludes the proof. □

### A.3 Actual condition numbers

The next lemma is used in the class-membership checks in Appendices D and E. It establishes the actual-condition-number statements in Theorems 4.1 and 5.1.

**Lemma A.2** (Condition numbers of the lift)**.** *Consider the construction of Proposition 3.1. Suppose* $\gamma^2 = M/(4\mu)$, $\ell h \le \mu/4$, *and* $\mathrm{Lip}(\nabla f_H) \le M$. *Let* $\mu_{\mathrm{act}}$ *be the largest uniform inner strong-concavity modulus, and let* $L_{\mathrm{act}} := \mathrm{Lip}(\nabla f_H)$. *Then, for* $\kappa = M/\mu$,

$$\frac{\kappa}{6} \le \frac{L_{\mathrm{act}}}{\mu_{\mathrm{act}}} \le \kappa. \tag{35}$$

*If the lifted oracle also has root mean-squared Lipschitz constant at most* $M$, *its smallest such constant* $\overline{L}_{\mathrm{act}}$ *satisfies the same bounds:* $\kappa/6 \le \overline{L}_{\mathrm{act}}/\mu_{\mathrm{act}} \le \kappa$.

*Proof.* The inner Hessian bounds imply

$$\mu \le \mu_{\mathrm{act}} \le \nu + \ell h \le \frac{3\mu}{2}. \tag{36}$$

Holding $y$ fixed and changing $x$ gives, for every seed,

$$\begin{aligned}\|\nabla f_H(x,y) - \nabla f_H(x',y)\| &= \|G(x,y;\xi) - G(x',y;\xi)\| \\ &= \nu\gamma\sqrt{1+\gamma^2}\,\|x - x'\|.\end{aligned} \tag{37}$$

Since $\nu\gamma\sqrt{1+\gamma^2} \ge \nu\gamma^2 \ge M/4$, we obtain $L_{\mathrm{act}} \ge M/4$ and, when finite, $\overline{L}_{\mathrm{act}} \ge M/4$. Combining these estimates with Equation (36) and the assumed upper bounds proves the claims. □

## B The base chain and its stochastic oracles

### B.1 The nonconvex chain

Recall that we define $H = F_T$ in Equation (19) where $F_T$ is adapted from Carmon et al. (2020): for $T \ge 2$, define

$$\Psi(t) = \begin{cases} 0, & t \le 1/2, \\ \exp\!\left(1 - (2t-1)^{-2}\right), & t > 1/2, \end{cases} \qquad \Phi_0(t) = \sqrt{e}\int_{-\infty}^{t} e^{-s^2/2}\,ds,$$

and

$$F_T(u) = -\Psi(1)\Phi_0(u_1) + \sum_{i=2}^{T}\big[\Psi(-u_{i-1})\Phi_0(-u_i) - \Psi(u_{i-1})\Phi_0(u_i)\big]. \tag{38}$$

For $a \geq 0$, write $\mathrm{prog}_a(u) = \max(\{i : |u_i| > a\} \cup \{0\})$. Both $\Psi$ and $\Phi_0$ are bounded, so $F_T$ is bounded below. We use the following published properties.

**Lemma B.1** (Properties of $F_T$ (Arjevani et al., 2023)). *The function $F_T$ is twice continuously differentiable and satisfies*

$$F_T(0) - \inf F_T \leq 12T, \tag{39}$$

$$\mathrm{Lip}(\nabla F_T) \leq 152, \quad \|\nabla F_T(u)\|_\infty \leq 23, \tag{40}$$

$$\mathrm{prog}_0(\nabla F_T(u)) \leq \mathrm{prog}_{1/2}(u) + 1, \tag{41}$$

$$\mathrm{prog}_1(u) < T \implies \|\nabla F_T(u)\| > 1. \tag{42}$$

*In particular,* $\|\nabla F_T(0)\| \leq 23$.

Based on the chain of Carmon et al. (2020). The bound at zero follows from Equation (41) and the coordinate bound. We take this lemma as the external analytic input; the lift, oracle estimates, progress argument, and parameter choices are proved here.

### B.2 A BOUNDED-VARIANCE ORACLE

Let $Z \sim \mathrm{Bernoulli}(p)$ with $0 < p \leq 1$ and define

$$[g_T^{\mathrm{BV}}(u; Z)]_i = \nabla_i F_T(u)\left[1 + \mathbf{1}\{i > \mathrm{prog}_{1/4}(u)\}(Z/p - 1)\right]. \tag{43}$$

Since $\mathbb{E}(Z/p - 1) = 0$, the estimator is unbiased. Moreover, $\mathrm{prog}_{1/2}(u) \leq \mathrm{prog}_{1/4}(u)$. By Equation (41), its error can have at most one nonzero coordinate, namely $\mathrm{prog}_{1/4}(u) + 1$ when this index is at most $T$. Hence

$$\mathbb{E}\left\|g_T^{\mathrm{BV}}(u; Z) - \nabla F_T(u)\right\|^2 \leq 23^2 \frac{1-p}{p}. \tag{44}$$

The new frontier coordinate can be nonzero only when $Z = 1$. This is the bounded-variance construction in Arjevani et al. (2023, equation (17)); the argument above verifies the properties needed for the lift.

### B.3 AN AVERAGED-SMOOTH ORACLE

We specify a single smooth gate before choosing any problem parameters. Let

$$\Lambda(t) = \begin{cases} \exp\left(-\dfrac{1}{100(t-1/4)(1/2-t)}\right), & 1/4 < t < 1/2, \\ 0, & \text{otherwise}, \end{cases} \qquad \Gamma(t) = \frac{\int_{1/4}^{t} \Lambda(s)\,ds}{\int_{1/4}^{1/2} \Lambda(s)\,ds}.$$

Then $0 \leq \Gamma \leq 1$, $\Gamma = 0$ on $(-\infty, 1/4]$, $\Gamma = 1$ on $[1/2, \infty)$, and $m_\Gamma := \sup_t |\Gamma'(t)| < \infty$ is a universal constant. This is the smooth gate used by Arjevani et al. (2023). For the tail $u_{\geq i} = (u_i, \ldots, u_T)$, set

$$\Theta_i(u) = \Gamma\big(1 - \|\Gamma(|u_{\geq i}|)\|\big), \qquad [g_T^{\mathrm{AS}}(u; Z)]_i = \nabla_i F_T(u)[1 + \Theta_i(u)(Z/p - 1)], \tag{45}$$

where the inner $\Gamma$ is applied coordinatewise. The flat regions of $\Gamma$ give

$$\mathbf{1}\{i > \mathrm{prog}_{1/4}(u)\} \leq \Theta_i(u) \leq \mathbf{1}\{i > \mathrm{prog}_{1/2}(u)\}. \tag{46}$$

Indeed, a tail bounded by $1/4$ gives inner norm zero and outer value one; a tail containing an entry larger than $1/2$ gives inner norm at least one and outer value zero.

**Lemma B.2** (AS oracle moments). *For $g = 23$ and any $\ell \geq 152$, the oracle Equation (45) is unbiased, and for all $u, v$,*

$$\mathbb{E}\left\|g_T^{\mathrm{AS}}(u; Z) - \nabla F_T(u)\right\|^2 \leq g^2 \frac{1-p}{p}, \tag{47}$$

$$\mathbb{E}\left\|g_T^{\mathrm{AS}}(u; Z) - g_T^{\mathrm{AS}}(v; Z)\right\|^2 \leq \frac{4g^2 m_\Gamma^4 + 3\ell^2}{p}\|u - v\|^2. \tag{48}$$

*The constants are independent of $T$.*

*Proof.* Write $a_i(u) = \Theta_i(u)\nabla_i F_T(u)$. By Equation (41) and Equation (46), $a(u)$ has at most one nonzero coordinate, $\mathrm{prog}_{1/2}(u)+1$, with magnitude at most $g$. The centered error is $(Z/p-1)a(u)$, which proves unbiasedness and Equation (47). The map $\Theta_i$ is $m_\Gamma^2$-Lipschitz: absolute value and the norm are 1-Lipschitz, and the inner and outer gates each contribute $m_\Gamma$. The union $J=\mathrm{supp}(a(u))\cup\mathrm{supp}(a(v))$ has at most two indices. On this union,

$$|a_i(u)-a_i(v)|^2 \le 2g^2|\Theta_i(u)-\Theta_i(v)|^2 + 2|\nabla_i F_T(u)-\nabla_i F_T(v)|^2.$$

Summing over $J$ yields

$$\|a(u)-a(v)\|^2 \le (4g^2m_\Gamma^4+2\ell^2)\,\|u-v\|^2\,.$$

The zero mean of $Z/p-1$ eliminates the cross term, so

$$\begin{aligned}\mathbb{E}\left\|g_T^{\mathrm{AS}}(u;Z)-g_T^{\mathrm{AS}}(v;Z)\right\|^2 &= \|\nabla F_T(u)-\nabla F_T(v)\|^2 + \frac{1-p}{p}\,\|a(u)-a(v)\|^2\\ &\le \frac{4g^2m_\Gamma^4+3\ell^2}{p}\,\|u-v\|^2\,.\end{aligned}$$

The deliberately conservative constant avoids relying on a sharper numerical estimate for the gate. □

## C Support bounds and the stationarity lower bound

We prove Lemma 3.5 and the stationarity bound used in Appendices D–F. Write $[m]=\{1,\dots,m\}$ for integers $m\ge 0$, with $[0]=\varnothing$.

**Lemma C.1** (PZC property of the base oracles (Arjevani et al., 2023, Lemmas 3–4))**.** *For every $T\ge 2$ and $p\in(0,1]$, the oracles $g_T^{\mathrm{BV}}$ in Equation (43) and $g_T^{\mathrm{AS}}$ in Equation (45) are probability-$p$ zero-chains in the sense of Definition 3.4. For either oracle, with $Z\sim\mathrm{Bernoulli}(p)$,*

$$\mathrm{prog}_0(g_T(u;Z)) \le \mathrm{prog}_{1/4}(u)+\mathbf{1}\{Z=1\} \qquad \textit{for every } u\in\mathbb{R}^T \textit{ and } Z\in\{0,1\}.$$

*Proof of Lemma 3.5.* Let $\mathcal{E}$ be the first event in Definition 3.4. Outside a null set of seeds, that definition gives, simultaneously for all $u$,

$$\mathrm{prog}_0(g_H(u;\xi)) \le \mathrm{prog}_{1/4}(u)+\mathbf{1}\{\xi\in\mathcal{E}\} \le \mathrm{prog}_0(u)+\mathbf{1}\{\xi\in\mathcal{E}\}.$$

Set $J_0=0$ and $J_t=\sum_{s=1}^t \mathbf{1}\{\xi_s\in\mathcal{E}\}$. Write $G_b^{t,k}=G_b(z^{t,k};\xi_t)$ for $b\in\{x,y\}$. Induction gives, almost surely, for every actual query and response,

$$\begin{aligned}\mathrm{supp}(x^{t,k})\cup\mathrm{supp}(y^{t,k}) &\subseteq [J_{t-1}],\\ \mathrm{supp}(G_x^{t,k})\cup\mathrm{supp}(G_y^{t,k}) &\subseteq [J_t].\end{aligned} \tag{49}$$

The first queries are at the origin; thereafter, zero-respecting gives the first inclusion from the earlier response supports. The coupling terms in Equation (14) are supported within $[J_{t-1}]$. Since $\beta>0$, the PZC inequality gives $\mathrm{supp}(g_H(\beta y^{t,k};\xi_t))\subseteq[J_t]$. This proves the second inclusion for every $k$ with the same seed $\xi_t$.

If the algorithm stops after $\tau\le R$ rounds, the output restriction gives $\mathrm{supp}(\hat{x})\subseteq[J_\tau]\subseteq[J_R]$; unused seeds through round $R$ may be sampled independently for this argument. Since $\mathbb{P}(\xi_t\in\mathcal{E})\le p$, we have $\mathbb{E}J_R\le pR$. Markov's inequality now yields

$$\mathbb{P}(\mathrm{supp}(\hat{x})\not\subseteq[T-1]) \le \mathbb{P}(J_R\ge T) \le \frac{pR}{T}.$$ □

**Proposition C.2** (Stationarity lower bound)**.** *Let $H=F_T$ with $T\ge 2$, let $\epsilon>0$, and use either base oracle from Lemma C.1 in the lift Equation (14). Suppose $\gamma q=8\epsilon$ and $\ell h/\nu\le 1/5$. For every admissible zero-respecting algorithm using at most $R\le T/(4p)$ rounds,*

$$\mathbb{P}\left(\|\nabla\Phi_H(\hat{x})\|>\frac{20}{3}\epsilon\right)\ge\frac{3}{4}. \tag{50}$$

*Consequently,*

$$\mathbb{E}\,\|\nabla\Phi_H(\hat{x})\|\ge 5\epsilon, \qquad \mathbb{E}\,\|\nabla\Phi_H(\hat{x})\|^2\ge\frac{100}{3}\epsilon^2. \tag{51}$$

*Proof.* Lemmas C.1 and 3.5 give $\mathbb{P}(\hat{x}_T = 0) \geq 3/4$. On this event, $(\beta\gamma\hat{x})_T = 0$, so Lemma B.1 gives $\|\nabla F_T(\beta\gamma\hat{x})\| > 1$. The stationarity transfer Equation (17) implies

$$\|\nabla\Phi_H(\hat{x})\| \geq \frac{\gamma q}{1+\ell h/\nu}\,\|\nabla F_T(\beta\gamma\hat{x})\| > \frac{20}{3}\epsilon.$$

This proves Equation (50); taking first and second moments gives Equation (51). □

An algorithm using at most $N$ oracle calls has at most $N$ rounds. Thus both accuracy criteria fail whenever $N \leq T/(4p)$. Appendices D and E apply Proposition C.2 after choosing $T$ and $p$; Appendix F transfers the same event to the envelope gradient.

## D Proof of Theorem 4.1

We keep a conservative constant choice so the parameter checks are shared with AS. Fix universal constants

$$\Delta_0 = 12, \quad \ell = 155, \quad g = 23, \quad s = \max\{336, \sqrt{4g^2 m_\Gamma^4 + 3\ell^2}\}, \quad C = 256g^2. \tag{52}$$

The choices 155 and 336 provide slack; no random-rotation estimates are used. For $M = L$ and $\kappa = M/\mu$, choose

$$\gamma^2 = \kappa/4, \quad q = 8\epsilon/\gamma, \quad q^2 = 256\epsilon^2/\kappa, \quad h = \mu/(4\ell), \quad \nu = 5\mu/4, \tag{53}$$

$$p = \begin{cases} 1, & \sigma = 0, \\ \min\{1, q^2g^2/\sigma^2\}, & \sigma > 0, \end{cases} \qquad B_* := \frac{\Delta h}{4\Delta_0 q^2}, \qquad T = \lfloor B_* \rfloor. \tag{54}$$

Use $H = F_T$, $\alpha = q^2/h$, $\beta = h/q$, and $g_H = g_T^{\mathrm{BV}}$.

**Class membership.** The strong concavity follows from Proposition 3.1. Since $\ell h/\nu = 1/5$ and $\kappa \geq 8$,

$$\nu(1+\gamma^2) + \ell h = \frac{3\mu}{2} + \frac{5M}{16} \leq \frac{M}{2} \leq M.$$

The variance is at most $q^2g^2(1-p)/p \leq \sigma^2$. For $p < 1$, this follows from $q^2g^2/p = \sigma^2$; for $p = 1$, the stochastic error vanishes. The actual condition number satisfies Equation (35).

**Gap and chain length.** The two quantities to control are

$$\frac{q^2g^2}{2\mu} = \frac{128g^2\epsilon^2}{M}, \qquad B_* = \frac{M\Delta}{4096\Delta_0\ell\epsilon^2}. \tag{55}$$

Choose

$$c_0 \leq \min\left\{\frac{1}{256g^2}, \frac{1}{65536\Delta_0\ell}\right\}. \tag{56}$$

Then $\epsilon^2 \leq c_0 M\Delta$ makes the first term in Equation (55) at most $\Delta/2$ and gives $B_* \geq 16$. Thus $T \geq B_*/2 \geq 8$. Also $\alpha\Delta_0 T \leq \Delta/4$, so Equation (16) gives $\Phi_H(0) - \inf\Phi_H \leq 3\Delta/4$. The lower bound Equation (33) verifies that the infimum is finite.

**Query lower bound.** The calibrated parameters satisfy Proposition C.2. Moreover,

$$\begin{aligned}\frac{T}{4p} &\geq \frac{M\Delta}{32768\Delta_0\ell\epsilon^2}\max\left\{1, \frac{\kappa\sigma^2}{C\epsilon^2}\right\} \\ &\geq \frac{1}{32768\Delta_0\ell C}\,\frac{M\Delta}{\epsilon^2}\max\left\{1, \frac{\kappa\sigma^2}{\epsilon^2}\right\}.\end{aligned}$$

Every integer budget not exceeding $T/(4p)$ fails with the uniform margins in Equation (51). This proves Theorem 4.1 for both moment criteria, with $c = (32768\Delta_0\ell C)^{-1}$ or any smaller constant. The hard dimension is $d_x = d_y = T$.

## E Proof of Theorem 5.1

Use the constants Equation (52), let $M = \overline{L}$ and $\kappa = M/\mu$, and keep the definitions of $\gamma, q, p, B_*, T, \alpha, \beta$ in Equation (53)–Equation (54), except that now

$$h = \min\left\{\frac{\mu}{4\ell}, \frac{M\sqrt{p}}{4s}\right\}, \qquad \nu = \mu + \ell h. \tag{57}$$

Use $H = F_T$ and $g_H = g_T^{\mathrm{AS}}$. In particular, $B_*$ is computed using the new $h$.

**Class membership.** We have $\ell h \le \mu/4$, $\nu \le 5\mu/4$, and $\ell h/\nu \le 1/5$. Strong concavity and the BV calculation are unchanged. Lemma B.2 and Proposition 3.1 give the joint averaged-smoothness upper bound

$$\nu(1+\gamma^2) + \frac{hs}{\sqrt{p}} \le \frac{5\mu}{4}\left(1 + \frac{\kappa}{4}\right) + \frac{M}{4} \le \frac{23M}{32} < M. \tag{58}$$

Population smoothness follows by Jensen's inequality. Equation (35) verifies both actual condition numbers.

**Gap and integer chain length.** As before, $q^2 g^2/(2\mu) = 128 g^2 \epsilon^2/M$, while

$$B_* = \frac{M\Delta}{4096\Delta_0\epsilon^2}\min\left\{\frac{1}{\ell}, \frac{\kappa\sqrt{p}}{s}\right\}. \tag{59}$$

A sufficient choice of universal accuracy constants is

$$c_0 \le \min\left\{\frac{1}{256g^2}, \frac{1}{65536\Delta_0\max\{\ell, s\}}\right\}, \qquad c_1 \le \frac{g}{4096\Delta_0 s}. \tag{60}$$

The first condition in Equation (23) bounds the gap correction by $\Delta/2$ and makes the first term in Equation (59), including its prefactor, at least 16. If $p = 1$, it also makes the second term at least 16, since $\kappa \ge 8$. If $p < 1$, then

$$\sqrt{p} = \frac{qg}{\sigma} = \frac{16g\epsilon}{\sqrt{\kappa}\sigma}, \qquad \frac{M\Delta}{4096\Delta_0\epsilon^2}\frac{\kappa\sqrt{p}}{s} = \frac{gM\Delta\sqrt{\kappa}}{256\Delta_0 s\epsilon\sigma} \ge 16$$

by the second condition in Equation (23) and Equation (60). Thus $B_* \ge 16$ in every case, $T \ge B_*/2 \ge 8$, and the initial primal gap is at most $3\Delta/4$ exactly as in Appendix D.

**Rate calculation.** Set $p_* = 1$ when $\sigma = 0$ and $p_* = \min\{1, \epsilon^2/(\kappa\sigma^2)\}$ otherwise. The calibrated probability obeys

$$p_* \le p \le C p_*, \qquad C = 256g^2. \tag{61}$$

Therefore

$$\begin{aligned}\frac{T}{4p} &\ge \frac{M\Delta}{32768\Delta_0\max\{\ell, s\}\epsilon^2}\min\{p^{-1}, \kappa p^{-1/2}\} \\ &\ge \frac{M\Delta}{32768\Delta_0\max\{\ell, s\}C\epsilon^2}\min\{p_*^{-1}, \kappa p_*^{-1/2}\}.\end{aligned} \tag{62}$$

The second inequality uses $C \ge 1$; both terms are bounded below by $C^{-1}$ times their $p_*$ counterparts. For $a = \kappa\sigma^2/\epsilon^2$, elementary case splitting gives

$$\min\{p_*^{-1}, \kappa p_*^{-1/2}\} = \max\{1, \min\{a, \kappa\sqrt{a}\}\}, \qquad \kappa\sqrt{a} = \kappa^{3/2}\sigma/\epsilon.$$

Substitution in Equation (62) and Proposition C.2 prove Theorem 5.1, with $c = (32768\Delta_0\max\{\ell, s\}C)^{-1}$ or smaller. The cases $a \le 1$, $1 < a < \kappa^2$, and $a \ge \kappa^2$ give Equation (26). Neither the construction nor the progress bound depends on the batch-size bound $K$. Thus $d_x = d_y = T$ and all constants are independent of $K$.

## F Proof of the Moreau-envelope lower bounds

We prove Corollary 5.2 in three steps: bound the smoothness of the primal function on the hard family, compare its gradient with the Moreau-envelope gradient, and apply Proposition C.2. Throughout, $M = L$ for BV and $M = \overline{L}$ for AS.

**Proposition F.1** (Primal smoothness)**.** *For the lift in Proposition 3.1,*

$$\mathrm{Lip}(\nabla \Phi_H) \le \frac{\nu\gamma^2 \ell h}{\mu}. \tag{63}$$

*Under either parameter choice in Appendices D–E, this upper bound is at most* $5M/64$*.*

*Proof.* The inner optimality condition is

$$q\nabla H(\beta y^*(x)) = \nu\big(y^*(x) - \gamma x\big).$$

For two points $x, x'$, write $y = y^*(x)$ and $y' = y^*(x')$. Subtracting their optimality conditions gives

$$\nu(y - y') = \nu\gamma(x - x') + q\big(\nabla H(\beta y) - \nabla H(\beta y')\big).$$

Since $H$ has an $\ell$-Lipschitz gradient and $q\beta = h$,

$$\nu \left\|y - y'\right\| \le \nu\gamma \left\|x - x'\right\| + \ell h \left\|y - y'\right\|.$$

Using $\nu - \ell h = \mu$ yields

$$\left\|y^*(x) - y^*(x')\right\| \le \frac{\nu\gamma}{\mu} \left\|x - x'\right\|. \tag{64}$$

The primal-gradient formula in Lemma A.1 and the inner optimality condition imply

$$\nabla \Phi_H(x) = \nu\gamma\big(y^*(x) - \gamma x\big) = \gamma q \nabla H(\beta y^*(x)).$$

Consequently,

$$\begin{aligned}\left\|\nabla \Phi_H(x) - \nabla \Phi_H(x')\right\| &\le \gamma\ell h \left\|y^*(x) - y^*(x')\right\| \\ &\le \frac{\nu\gamma^2 \ell h}{\mu} \left\|x - x'\right\|.\end{aligned}$$

For both parameter choices, $\ell h \le \mu/4$, $\nu \le 5\mu/4$, and $\gamma^2 = M/(4\mu)$. Substituting these inequalities gives

$$\frac{\nu\gamma^2 \ell h}{\mu} \le \frac{1}{\mu}\frac{5\mu}{4}\frac{M}{4\mu}\frac{\mu}{4} = \frac{5M}{64}.$$

□

**Lemma F.2** (Equivalence of stationarity measures on the hard family)**.** *Under either calibration in Appendices D and E, let* $P_{f_H} = (\Phi_H)_{1/(2M)}$*. Then, for every* $x$*,*

$$\frac{123}{128}\|\nabla P_{f_H}(x)\| \le \|\nabla \Phi_H(x)\| \le \frac{133}{128}\|\nabla P_{f_H}(x)\|.$$

*Proof.* Let $u_x$ minimize $\Phi_H(u) + M\|u - x\|^2$. The Moreau-gradient identities give $\nabla P_{f_H}(x) = \nabla \Phi_H(u_x) = 2M(x - u_x)$. By Proposition F.1,

$$\|\nabla \Phi_H(x) - \nabla P_{f_H}(x)\| \le \frac{5M}{64}\|x - u_x\| = \frac{5}{128}\|\nabla P_{f_H}(x)\|.$$

The two bounds follow from the triangle and reverse triangle inequalities. □

*Proof of Corollary 5.2.* Fix either the BV or AS setting and choose exactly the same hard instance and parameters as in its lower-bound proof in Appendix D or E. An algorithm using at most $N \le T/(4p)$ oracle calls also uses at most $T/(4p)$ rounds. Proposition C.2 therefore gives

$$\mathbb{E}\left\|\nabla \Phi_H(\hat{x})\right\| \ge 5\epsilon, \qquad \mathbb{E}\left\|\nabla \Phi_H(\hat{x})\right\|^2 \ge \frac{100}{3}\epsilon^2.$$

Lemma F.2 holds pointwise at the same random output $\hat{x}$. It follows that

$$\mathbb{E}\,\|\nabla P_{f_H}(\hat{x})\| \geq \frac{128}{133}\mathbb{E}\,\|\nabla \Phi_H(\hat{x})\| \geq \frac{640}{133}\epsilon > \epsilon,$$
$$\mathbb{E}\,\|\nabla P_{f_H}(\hat{x})\|^2 \geq \left(\frac{128}{133}\right)^2 \mathbb{E}\,\|\nabla \Phi_H(\hat{x})\|^2$$
$$\geq \frac{100}{3}\left(\frac{128}{133}\right)^2 \epsilon^2 > \epsilon^2.$$

Thus neither Moreau-envelope criterion can be met within the same oracle budget. The bounds on $T/(4p)$ already proved in Appendices D and E give the two claimed complexity lower bounds, with the same parameter ranges. The point $u_{\hat{x}}$ is used only in the gradient comparison; it is neither queried nor returned by the algorithm. □

# G Comparison with existing complexity bounds

## G.1 Earlier lower bounds and the algorithm restriction

Li et al. (2021, PDF p. 5, Theorem 2) give a stochastic term $\Omega(L\Delta\kappa^{1/3}\sigma^2\epsilon^{-4})$ together with a deterministic $\Omega(L\Delta\sqrt{\kappa}\epsilon^{-2})$ term. Their first-order class is defined through projected zero-respecting updates (PDF p. 3, Definition 4), and their stationarity criterion is a projected primal-gradient mapping. Our $\kappa^{2/3}$ improvement concerns the stochastic term; the noise-free part of our construction is weaker. Zhang et al. (2021) study deterministic and averaged-smooth finite-sum oracles with a linear-span class (PDF pp. 5–6). These finite-sum results and those of Han et al. (2024) do not provide a condition-number-dependent lower bound for fresh stochastic AS queries.

The zero-respecting scope is also used in recent minimax lower bounds. For example, Wu et al. (2026, Definition 2.6, PDF p. 7) treat deterministic and stochastic projected zero-respecting algorithms, and Zhang & Xu (2026, Definition 5.1 and Remark 5.1, PDF pp. 12–15) allow randomization subject to that support restriction. Other recent results additionally extend a deterministic lower bound beyond zero-respecting, as in Pan et al. (2026, Theorem 3.2) and Pan & Li (2026, Theorem 3.2). We do not make that extension here. The restriction is part of our lower-bound statement and must also be retained when discussing optimality relative to an upper-bound algorithm.

## G.2 SGDA, SREDA, and SAPD+

**SGDA.** In the ICML 2020 version, Lin et al. (2020a, PDF p. 5, Theorem 4.5) give

$$O\left[\frac{\kappa^2 L\Delta_\Phi + \kappa L^2 D^2}{\epsilon^2}\max\left\{1, \frac{\kappa\sigma^2}{\epsilon^2}\right\}\right] \tag{65}$$

stochastic-gradient evaluations in their bounded-dual setting. Here $D$ is their dual diameter and $\Delta_\Phi$ their initial primal gap. The leading stochastic condition-number exponent is 3 when the other parameters are fixed. The later journal treatment is Lin et al. (2025).

**SREDA.** Luo et al. (2020, NeurIPS PDF p. 6, Theorem 1) prove $\mathbb{E}\,\|\nabla\Phi(\hat{x})\| \leq 1504\epsilon$ with $O(\kappa^3\epsilon^{-3})$ stochastic-gradient evaluations, suppressing the remaining problem data. The subscript 2 printed on their norm is not a square. Their count includes the initial PiSARAH phase, large refresh batches, and inner concave-maximizer calls. Assumptions 1–5 (PDF pp. 2–3) include mean-squared smoothness, a compact dual set, and sample-wise concavity. Rescaling the tolerance by 1504 changes only a universal constant. The expected-norm version of our lower bound is therefore the appropriate moment-level comparison. The separate finite-sum theorem is not used here.

**SAPD+.** Theorems 1 and 3, with Remark 4 (PDF pp. 6–7 of v4), yield the BV bound

$$O\left[\frac{\kappa L\mathcal{G}_0}{\epsilon^2}\max\{1, \sigma^2/\epsilon^2\}\right]. \tag{66}$$

We use a common upper bound for their blockwise smoothness constants and noise levels; this changes only universal factors. The unbounded-domain variant has the additional assumptions stated

in that paper. The quantity

$$\mathcal{G}(x,y) = \sup_{y'} f(x,y') - \inf_{x'} f(x',y) \tag{67}$$

is a global primal–dual gap, and $\mathcal{G}_0$ is its initial value. For AS, Theorem 4 and Remark 8 (PDF p. 9 of v4) give Equation (29); Remark 9 permits mean-squared smoothness in place of the stronger sample-wise condition. With $r = \sigma/\epsilon$ and $\kappa r \gtrsim 1$, its leading expression is

$$\frac{L\mathcal{G}_0}{\epsilon^2} \begin{cases} \kappa^2 r, & r \lesssim \kappa^2, \\ \kappa r^{3/2}, & r \gtrsim \kappa^2. \end{cases}$$

In particular, the second branch eventually applies as $\epsilon \to 0$ with fixed $\kappa$ and positive $\sigma$. We use the revised source rather than carrying over the simplified $\kappa^2\epsilon^{-3}$ entry from an older comparison table.

**Momentum and adaptive methods.** These remain relevant algorithmic alternatives (Huang et al., 2022; 2023), but batch-dependent exponents must be read together with the admissible stepsizes. For example, the discussion following Theorem 12 of Huang et al. (2022) (PDF p. 17) already counts $4bT$ gradient evaluations and imposes $27b\tau/16 \le 1/(6L_f)$, where $\tau$ is their strong-concavity parameter. Thus $b \le 8\kappa/(81L_f^2)$ when $\kappa = L_f/\tau$. Choosing $b$ of order $\kappa^4$ while holding $L_f$ fixed does not satisfy this restriction for arbitrarily large $\kappa$. We therefore do not use that choice to report an unconditional improved exponent in the main table. This observation is about the stated parameter constraints, not a lower bound against those algorithms.

### G.3 SPDE AND VR-SPDE

The relevant result is Zhang et al. (2026), which controls *optimization stationarity* through the $1/(2L)$ Moreau envelope of the original value function. Their game-stationarity theorems concern a different residual.

To avoid confusion with the chain parameter $B_*$, write their initial quantity as $\mathcal{B}$ and its supplied upper bound as $B \ge \mathcal{B}$. With $g_{x,0} = \nabla_x f(x_0,y_0)$ and $g_{y,0} = \nabla_y f(x_0,y_0)$,

$$\mathcal{B} = \Phi(x_0) - \inf\Phi + 2D_Y \|g_{y,0}\| + \frac{\|g_{x,0}\|^2 + 2\|g_{y,0}\|^2}{L} + \left(3 + \frac{1}{64}\right) LD_Y^2, \tag{68}$$

where $D_Y = \max_{y\in Y}\|y\|$ is a radius. Set $b_0 = \beta_{\mathrm{sc}} = \sqrt{2}\,\kappa^{-1/2}/786432$. Equation (154) of their paper gives

$$N_{\mathrm{BV}} \le 3\left(1 + \frac{16LB}{b_0\epsilon^2}\right)\left(1 + \frac{4800\sigma^2}{b_0\epsilon^2}\right). \tag{69}$$

For the variance-reduced method, their equations (155)–(156) set

$$T_0 = \max\{1, \lceil 32LB/(b_0\epsilon^2)\rceil\}, \qquad B_r = \max\{1, \lceil 192C_\star\sigma^2/(b_0^2\epsilon^2)\rceil\},$$
$$A_0 = 16C_\star\chi^2/b_0^2, \qquad q_0 = \max\{1, \lfloor\sqrt{B_r/A_0}\rfloor\}, \qquad b = \lceil A_0 q_0\rceil,$$

where $C_\star$ is their universal constant and $\chi = \ell_{\mathrm{ms}}/L \ge 1$. The bound is

$$N_{\mathrm{AS}} \le B_r + 18T_0 \min\{B_r, \sqrt{A_0 B_r}\}. \tag{70}$$

The initial refresh cost $B_r$ is included. Putting $r = \sigma/\epsilon$, these expressions imply

$$N_{\mathrm{BV}} = O\left[\left(1 + \frac{LB\sqrt{\kappa}}{\epsilon^2}\right)(1 + \sqrt{\kappa}r^2)\right], \tag{71}$$

$$N_{\mathrm{AS}} = O\left[1 + \kappa r^2 + \left(1 + \frac{LB\sqrt{\kappa}}{\epsilon^2}\right)\min\{1+\kappa r^2, \chi\sqrt{\kappa}\sqrt{1+\kappa r^2}\}\right]. \tag{72}$$

The leading VR product is $\chi LB\kappa^{3/2}\sigma\epsilon^{-3}$. The additive $\kappa\sigma^2\epsilon^{-2}$ term matters in joint parameter limits. For the AS comparison, one may take the common valid upper bounds $L = \ell_{\mathrm{ms}} = \overline{L}$, giving $\chi = 1$ and $\kappa = \overline{L}/\mu$. This does not assert that the smallest valid smoothness constants are equal. The published simplification is then $O(\kappa\epsilon^{-2} + \kappa^{3/2}\epsilon^{-3})$ for fixed positive noise and other suppressed parameters.

A paired difference uses points selected before its fresh seed, and costs two gradient evaluations. It is compatible with the AS query protocol here. Independent refresh samples are separate rounds; adaptively chosen requests with independent seeds are also separate rounds. This compatibility is about oracle access, not an assertion that an arbitrary projection in their algorithm is zero-respecting.