Optimal Local Convergence Rates of Stochastic First-Order Methods Under Local Alpha-PŁ

Saeed Masiha[1] **Saber Salehkaleybar**[2] **Niao He**[3]

Negar Kiyavash[1] **Patrick Thiran**[4]

[1]EPFL School of Management of Technology
[2]Leiden University Computer Science institute (LIACS)
[3]ETH Department of Computer Science
[4]EPFL Department of Computer and Communications Sciences

Abstract

We study the local oracle complexity of stochastic first-order methods under a local α–Polyak–Łojasiewicz (α–PŁ) condition in a neighborhood of a target connected component $\mathcal{M}'$ of the local minimizer set. The parameter $\alpha \in [1, 2]$ is the exponent of the gradient norm in the α–PŁ inequality: $\alpha = 2$ recovers the classical PŁ case, $\alpha = 1$ corresponds to Hölder-type error bounds, and intermediate values interpolate between these regimes. Our performance criterion is the number of oracle queries required to output $\hat{x}$ with $F(\hat{x}) - l \leq \varepsilon$, where $l := F(y)$ for any $y \in \mathcal{M}'$. We work in a local regime where the algorithm is initialized near $\mathcal{M}'$ and, with high probability, its iterates remain in that neighborhood. We establish a lower bound $\Omega(\varepsilon^{-2/\alpha})$ for all stochastic first-order methods in this regime, and we obtain a matching upper bound $\mathcal{O}(\varepsilon^{-2/\alpha})$ for $1 \leq \alpha < 2$ via a SARAH-type variance-reduced method with time-varying batch sizes and step sizes. Thus, for $1 \leq \alpha < 2$, the optimal dependence on ε is $\Theta(\varepsilon^{-2/\alpha})$. In the convex setting, assuming a local α–PŁ condition on the ε-sublevel set, we further show a complexity lower bound $\widetilde{\Omega}(\varepsilon^{-2/\alpha})$ for reaching an ε-global optimum, matching the ε-dependence of known accelerated stochastic subgradient methods.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

Stochastic first-order methods are the workhorses of modern large-scale learning [Bottou et al., 2018]. Classical analyses for such methods typically assume global convexity or impose global growth conditions [Karimi et al., 2016]. However, such assumptions do not reflect the training landscape of deep and structured models, where the objectives are highly nonconvex, yet once the iterates enter a favorable local region, simple gradient-based schemes often converge rapidly. A more realistic assumption is perhaps *local* growth conditions, such as the Polyak–Łojasiewicz (PŁ) and Kurdyka–Łojasiewicz (KŁ) inequalities, or their Hölder-type generalizations [Attouch et al., 2010, Drusvyatskiy and Lewis, 2018, Li and Pong, 2018, Jiang and Li, 2022]. Unlike strong convexity, these conditions do not imply the uniqueness of the minimizer. This distinction is crucial in deep learning, where permutation and rescaling symmetries, as well as over-parameterization, typically results in connected sets of minimizers rather than a single point [Freeman and Bruna, 2016, Şimşekli et al., 2021, Neyshabur et al., 2015, Garipov et al., 2018, Draxler et al., 2018]. We formalize this behavior by working with a *local α-PŁ* condition: there exist $\tau > 0$, an exponent α with $1 \leq \alpha \leq 2$, and a neighborhood $\mathcal{U}$ of the target component $\mathcal{M}$ of the local minimizers of F such that, with $l := F(y)$ for any $y \in \mathcal{M}$, $F(x) - l \leq \tau \|\nabla F(x)\|^\alpha$ for all $x \in \mathcal{U}$.

In this work, we consider the following setting: Fix a connected component $\mathcal{M}$ of local minimizers and a neighborhood $\mathcal{U} \supset \mathcal{M}$ where F satisfies the local α-PŁ inequality and ask the question:

> *How does the exponent α quantitatively influence the optimal number of oracle calls needed to output $\hat{x} \in \mathcal{U}$ with $\mathbb{E}[F(\hat{x})] - l \leq \varepsilon$?*

Setting. We work with a *batch-smooth* stochastic first-order oracle (SFO), denoted by $\mathsf{O}_\sigma^{\tilde{L}}$, which (i) returns unbiased gradient estimates with variance bounded by σ^2 (Eq. (4), Sec. 3.1), (ii) is $\tilde{L}$-mean-square smooth across query points (Eq. (5), Sec. 3.1), and (iii) upon K simultaneous queries returns gradients that share the same randomness (Eq. (6), Sec. 3.1).

Let A be any stochastic first-order method interacting with $O \in \mathsf{O}_\sigma^{\tilde{L}}$. Assume A is initialized sufficiently close to a target connected component $\mathcal{M}$ of local minimizers and satisfies a *stay-in-neighborhood* property: with high probability, all iterates remain in a fixed neighborhood $\mathcal{U}$ of $\mathcal{M}$. This assumption is standard in analyses of local growth/error–bound conditions and can be justified for gradient-type methods under local smoothness and bounded-variance noise [Weissmann et al., 2025, Mertikopoulos et al., 2020].

The *oracle complexity* of A is the number of oracle calls required to obtain $\hat{x}$ with $\mathbb{E}[F(\hat{x})] - l \le \varepsilon$, where $l := F(y)$ for every $y \in \mathcal{M}$. We assume that, in a neighborhood of $\mathcal{M}$, the objective F is locally Lipschitz and locally smooth, and that it satisfies a local α–PŁ condition.

Throughout, we focus on $1 \le \alpha \le 2$; the case $\alpha = 2$ recovers the classical PŁ inequality [Karimi et al., 2016], while $\alpha = 1$ corresponds to gradient-dominance with exponent 1 [Masiha et al., 2022, Liu et al., 2020].

Motivation for studying local α-PŁ setting. The local α-PŁ assumption holds for a wide class of objectives. By the Kurdyka–Łojasiewicz (KŁ) theory, *every real-analytic function* satisfies a local KŁ inequality around each critical point; in particular, for any critical point $x_\star$ there exist $\tau > 0$, $\alpha \in (1, 2]$[1], and a neighborhood $\mathcal{N}$ of $x_\star$ such that $F(x) - F(x_\star) \le \tau \|\nabla F(x)\|^\alpha$ for all $x \in \mathcal{N}$. This follows from the classical Łojasiewicz gradient inequality and modern KŁ extensions [Bolte et al., 2007a, Attouch et al., 2010, Kurdyka, 1998]. In particular, a nondegenerate local minimum typically has $\alpha = 2$, recovering the classical PŁ regime [Attouch et al., 2010]. When the Hessian is rank-deficient and the first nonzero term in the Taylor expansion is of order $p > 2$, the local PŁ exponent satisfies $\alpha = \frac{p}{p-1} \in (1, 2)$; see Appendix C for more details and closed-form examples (such as isolated and manifold minimizers, and a deep-linear product loss).

The KŁ property extends to broad *tame* classes—semialgebraic/*definable*[2] objectives—that encompass many standard ML losses [Bolte et al., 2007b, Attouch et al., 2010]. They include least-squares/ridge, logistic or cross-entropy, ReLU networks (often with ℓ_1/ℓ_2 regularization), and polynomial matrix factorization (see Appendix C.3 for more details). In all these cases the empirical risk is definable, hence satisfies a local KŁ inequality and therefore exhibits local α-PŁ growth near critical points (see related references in Appendix C.3). Moreover, in Section 5, we empirically validate the local α–PŁ property on two training losses for binary classification and dictionary learning.

Contributions. We show that the exponent α in the local α-PŁ condition determines the ε–dependence of the oracle/sample complexity. In particular, to output an ε-local optimum $\hat{x}$ (i.e., $F(\hat{x}) - l \le \varepsilon$ where l is the target value of function[3]) on average, the optimal number of oracle calls is $\Theta\big(\varepsilon^{-2/\alpha}\big)$ for $1 \le \alpha \le 2$. This tight rate holds both a *nonconvex* setting with a local convergence in a neighborhood of a connected component of minimizers, and in a *convex* setting with a global convergence to an ε-optimal solution.

Nonconvex setting (local convergence): *(i) Lower bound:* Under local Lipschitzness, local smoothness, and local α-PŁ near any connected component $\mathcal{M}$ of local minimizers (see Assumption 2, Sec. 2), any stochastic first-order algorithm with a batch-smooth SFO requires $\Omega\big(\varepsilon^{-2/\alpha}\big)$ oracle calls for $1 \le \alpha \le 2$ in order to reach an ε-local optimum on average, i.e., finding $\hat{x}$ such that $\mathbb{E}[F(\hat{x})] - l \le \varepsilon$. We derive the lower bound by reducing the search for an ε-local optimum to a sequential hypothesis testing problem with noisy observations. We subsequently establish a connection between the probability of error in the hypothesis testing problem and the complexity of finding an ε-local optimum point. *(ii) Upper bound:*[4] We show that the lower bound is tight by proving that a variance-reduced method (a SARAH-style scheme [Nguyen et al., 2017]) with varying step-sizes and batch-sizes in Algorithm 1 achieves $\mathcal{O}\big(\varepsilon^{-2/\alpha}\big)$ for $1 \le \alpha < 2$. The analysis establishes a high-probability "stay-in-neighborhood" property for the updates of the algorithm and then proves convergence of $F(x_t)$ to the level l.

Technically, localization for SARAH requires controlling non-martingale cross-terms induced by a biased gradient estimator via a coupled choice of time-varying step sizes and batch sizes. **Convex setting (global convergence):** For convex functions, under local

[1] Note that the exponent α is local and may vary from point to point.

[2] see Appendix C.3 for the definition.

[3] In the nonconvex setting, let $l := F(y)$ for any $y \in \mathcal{M}$, where $\mathcal{M}$ is a connected component of local minimizers. In the convex setting, set $l := \min_{y \in \mathbb{R}^d} F(y)$.

[4] When $\alpha = 2$, SGD with a decaying step-size attains an ε-local optimum *in expectation* using $\mathcal{O}(\varepsilon^{-1})$ oracle calls [Weissmann et al., 2025]. Thus, for $\alpha = 2$, an $\mathcal{O}(\varepsilon^{-1})$ complexity upper bound is available.

$(\alpha, \tau, \varepsilon)$-PŁ property (see Assumption 6), we provide a lower bound $\tilde{\Omega}(\tau^{2/\alpha}\varepsilon^{-2/\alpha})$[5] $(1 \le \alpha \le 2)$ for first-order algorithms with a stochastic first-order oracle and bounded stochastic gradients[6] in order to reach an ε-global-optimum point. We establish this bound by a reduction to the noisy binary search (NBS) problem [Karp and Kleinberg, 2007]. When $1 \le \alpha \le 2$, this lower bound matches the oracle complexity of accelerated stochastic subgradient methods [Xu et al., 2017] in terms of dependence on ε and τ.

To the best of our knowledge, this is the first work to study the *optimal* local convergence rates of stochastic first-order algorithms under a *local* α-PŁ condition. Related results on upper and lower bounds under (global) PŁ assumptions are summarized in Appendix A.

Organization. The rest of the paper is organized as follows: Section 2 provides preliminaries and definitions regarding local α-PŁ. Section 3 gives lower and upper bounds on the oracle complexity of stochastic first-order methods under local α-PŁ for $1 \le \alpha < 2$. The lower bound for the stochastic first-order methods under convexity and local $(\alpha, \tau, \varepsilon)$-PŁ property is given in Section 4. Section 5 provides the empirical evaluation of local α-PŁ. Full proofs of our main results, together with additional discussion, appear in the appendices.

2 PRELIMINARIES

For a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$, define

$$\mathcal{M}_F \;:=\; \big\{\, x \in \mathbb{R}^d :\; x \text{ is a local minimizer of } F \,\big\}.$$

Assumption 1 (Compactness)**.** *We assume that the collection of all local minima $\mathcal{M}_F$ is contained within a compact set.*

We now formalize the local geometry assumption used throughout this work.

Assumption 2 (Local α-PŁ)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function. Fix α with $1 \le \alpha \le 2$. We say that F satisfies the local α-PŁ condition if there exists a constant $\tau > 0$[7] such that for every isolated connected component $\mathcal{M} \subseteq \mathcal{M}_F$, there exists an open neighborhood $\mathcal{N}(\mathcal{M})$ for which*

$$F(x) - l(\mathcal{M}) \;\le\; \tau \,\|\nabla F(x)\|^{\alpha} \quad \forall x \in \mathcal{N}(\mathcal{M}), \qquad (1)$$

where $l(\mathcal{M}) := F(y)$ for any $y \in \mathcal{M}$[8]*.*

Relevant ML examples of local α-PŁ functions. The local α–PŁ inequality (1) quantifies how "flat" the objective is near a target connected component $\mathcal{M}$ of local minimizers. By the Kurdyka–Łojasiewicz (KŁ) inequality, many standard ML losses admit such a *local* bound; see Appendix C.3 for a brief summary. As a concrete example, for a one-dimensional d-layer linear neural network and a single data point (a, b), the squared loss $F(x) = \big((\prod_{i=1}^{d} x_i)\, a - b\big)^2$ satisfies a local α–PŁ inequality with exponent $\alpha = \frac{2d}{2d-1} \in (1, 2)$ (see further examples in Appendix C). We further empirically validate the local α–PŁ inequality in Section 5 for two training losses used for binary classification and dictionary learning on synthetic data.

Remark 1. *In Assumption 2, we impose the α–PŁ condition only in neighborhoods of* isolated *connected components of the set of local minimizers. Concretely, if $\mathcal{M}$ is a connected component of $\mathcal{M}_F$ that is isolated, then there exists an open neighborhood $\mathcal{N}(\mathcal{M})$ containing no other minimizer. In particular, isolation induces a positive function–value margin ($F(x) - l > 0$ for $x \in \mathcal{N}(\mathcal{M}) \setminus \mathcal{M}$), which is key to the "stay-in-neighborhood" arguments for SARAH-type updates (Appendix G.2). Moreover, if $F \in \mathcal{C}^2$ and every critical point outside a neighborhood of $\mathcal{M}$ is a strict maximum (i.e., $\nabla^2 F(x) \prec 0$), then the set of local minimizers is connected and coincides with the set of global minimizers [Masiha et al., 2025, Proposition 2.1] and therefore the connected set of minimizers is isolated. Empirically, connected minimizer sets are frequently observed in over-parameterized models [Draxler et al., 2018, Nguyen, 2019, Kuditipudi et al., 2019]; in such regimes, the isolation requirement is essentially redundant.*

For a set $\mathcal{S} \subset \mathbb{R}^d$ and $R > 0$, define a *tube* with radius R around set $\mathcal{S}$ as $\mathbb{B}_2(\mathcal{S}; R) := \big\{x \in \mathbb{R}^d :\, \mathrm{dist}(x, \mathcal{S}) \le R\big\}$, where $\mathrm{dist}(x, \mathcal{S}) := \inf_{y \in \mathcal{S}} \|x - y\|$. In the following lemma, using the compactness and isolation of the set of local minimizers of F, we show that there exists a tube $\mathbb{B}_2(\mathcal{M}; R) \subseteq \mathcal{N}(\mathcal{M})$ for $\mathcal{N}(\mathcal{M})$ defined in Assumption 2.

Lemma 1. *Suppose that for an isolated connected component $\mathcal{M} \subseteq \mathcal{M}_F$, there exists an open neighborhood $\mathcal{N}(\mathcal{M})$ on which F satisfies the local α-PŁ in-*

[5] In this paper we use $\tilde{\mathcal{O}}$ and $\tilde{\Omega}$ to hide poly-logarithmic factors.

[6] This is a standard assumption in stochastic convex non-smooth optimization [Xu et al., 2017, Yang and Lin, 2018].

[7] We assume a single constant $\tau > 0$ that works uniformly across all connected components of local minimizers. In Appendix B, we show that such a universal τ exists provided each component $\mathcal{M} \subset \mathcal{M}_F$ admits its own constant $\tau_{\mathcal{M}} > 0$.

[8] W.L.O.G, we assume that $F(y) > l(\mathcal{M})$ for all $y \in \mathcal{N}(\mathcal{M}) \setminus \mathcal{M}$, as $\mathcal{M}$ is an isolated compact connected set of local minima (otherwise we choose $\mathcal{N}(\mathcal{M})$ small enough).

equality (1)*, and that* $\mathcal{M}$ *is compact. Then there exists* $R > 0$ *with* $\mathbb{B}_2(\mathcal{M}; R) \subseteq \mathcal{N}(\mathcal{M})$.

The proof is given in Appendix E.1.

Remark 2. *Our analysis concerns the* post-transient *phase (after the iterates enter a basin of attraction), where rates are determined by the geometry near minimizers. Empirically, dynamics in modern overparameterized models are typically observed not to remain at saddles or maxima [Dauphin et al., 2014]. Theory supports this picture: (i) gradient descent with random initialization almost surely avoids strict saddles [Lee et al., 2016]; (ii) small perturbations enable polynomial-time escape from saddle neighborhoods [Jin et al., 2017]; and (iii) the stochasticity of SGD behaves as a Langevin-type diffusion that helps leave saddle regions and reach basins [Mandt et al., 2017, Raginsky et al., 2017]. It is therefore natural to impose the* α*–PŁ condition* locally *around minimizers rather than at arbitrary critical points.*

We assume the following local Lipschitzness and smoothness properties for the objective function F.

Assumption 3 (Local Lipschitzness)**.** *A function* $F : \mathbb{R}^d \to \mathbb{R}$ *is* locally Lipschitz *if for every compact set* $\mathcal{K} \subset \mathbb{R}^d$*, there exists a constant* $L_{\mathcal{K}} \geq 0$ *such that* $|F(x) - F(y)| \leq L_{\mathcal{K}} \|x - y\|$ *for all* $x, y \in \mathcal{K}$.

Assumption 4 (Local Smoothness)**.** *A differentiable function* $F : \mathbb{R}^d \to \mathbb{R}$ *is* locally smooth *(has a* locally Lipschitz *gradient) if for every compact set* $\mathcal{K} \subset \mathbb{R}^d$ *there exists* $\hat{L}_{\mathcal{K}} \geq 0$ *such that* $\|\nabla F(x) - \nabla F(y)\| \leq \hat{L}_{\mathcal{K}} \|x - y\|$ *for all* $x, y \in \mathcal{K}$.

2.1 Geometric Notation and Constants

Fix an isolated connected component $\mathcal{M} \subseteq \mathcal{M}_F$ of local minimizers, and let $l := F(y)$ for any $y \in \mathcal{M}$. By Lemma 1, choose $R > 0$ such that $\mathbb{B}_2(\mathcal{M}; R) \subseteq \mathcal{N}(\mathcal{M})$ (so the local regularity assumptions hold on $\mathbb{B}_2(\mathcal{M}; R)$). We use the nested sets

$$\begin{aligned} \mathcal{U} &:= \mathbb{B}_2(\mathcal{M}; R/2), \\ \mathcal{U}_0 &:= \{x \in \mathcal{U} : \ F(x) - l \leq s/2\}, \\ \mathcal{R} &:= \mathbb{B}_2(\mathcal{M}; 3R/4) \setminus \mathbb{B}_2(\mathcal{M}; R/2). \end{aligned}$$

Since $\mathcal{M}$ is compact and isolated, the closure of $\mathcal{R}$ is compact and disjoint from $\mathcal{M}$, hence continuity of F implies the positive barrier $s_0 := \inf_{x \in \mathcal{R}} (F(x) - l) > 0$. We fix any $s \in (0, s_0)$; as long as the iterates satisfy $F(x_t) - l < s$, no iterate can enter $\mathcal{R}$, and therefore they remain in $\mathcal{U}$.

3 UPPER AND LOWER BOUNDS ON LOCAL CONVERGENCE RATE UNDER LOCAL PŁ

Let $\mathcal{U}$ be a sufficiently small neighborhood of the target isolated connected component $\mathcal{M}$ of local minimizers of F (i.e., $\mathcal{U} \subseteq \mathcal{N}(\mathcal{M})$) so that the local α-PŁ holds, and let $l := F(y)$ for any $y \in \mathcal{M}$. For $\varepsilon > 0$, a (possibly random) output $\hat{x}$ is an *ε-local optimum relative to* $\mathcal{M}$ if

$$\mathbb{E}[\, F(\hat{x}) - l \mid \{\hat{x} \in \mathcal{U}\}] \leq \varepsilon. \tag{2}$$

We call $F(\hat{x}) - l$ the *local optimality gap* relative to $\mathcal{M}$.

Let $\mathcal{E}$ be the event that (i) the algorithm is initialized in $\mathcal{U}$ within a neighborhood of $\mathcal{M}$, and (ii) all subsequent iterates remain in $\mathcal{U}$. Our performance metric is the *expected number of oracle queries* required to produce an output $\hat{x}$, evaluated *conditional on* $\mathcal{E}$. We later show that for Algorithm 1, if the initialization is sufficiently close to $\mathcal{M}$, then $\mathcal{E}$ holds with high probability and the returned $\hat{x}$ is an ε-local optimum relative to $\mathcal{M}$.

3.1 Problem setting

Function class. For $1 \leq \alpha \leq 2$, define

$$\mathcal{F}_\alpha = \left\{ F : \mathbb{R}^d \to \mathbb{R} \;\middle|\; \begin{array}{c} F \text{ is locally Lipschitz,} \\ F \text{ is locally smooth,} \\ F \text{ is local } \alpha\text{-PŁ,} \\ \mathcal{M}_F \text{ is compact} \end{array} \right\}. \tag{3}$$

Equivalently, $F \in \mathcal{F}_\alpha$ if and only if Assumptions 1 to 4 hold.

Batch-smooth stochastic first-order oracle. We access a stochastic oracle specified by a measurable map $\mathbf{g} : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^d$ and a distribution P_Z on $\mathcal{Z}$ such that, for all $x, y \in \mathbb{R}^d$, $\mathbb{E}_{Z \sim P_Z}\big[\mathbf{g}(x, Z)\big] = \nabla F(x)$,

$$\mathbb{E}_{Z \sim P_Z}\big[\|\mathbf{g}(x, Z) - \nabla F(x)\|^2\big] \leq \sigma^2, \tag{4}$$

$$\mathbb{E}_{Z \sim P_Z}\big[\|\mathbf{g}(x, Z) - \mathbf{g}(y, Z)\|^2\big] \leq \tilde{L}^2 \|x - y\|^2. \tag{5}$$

The $\tilde{L}$-average smoothness (5) is a standard assumption in variance-reduction analyses [Fang et al., 2018, Lei et al., 2017, Cutkosky and Orabona, 2019]. The family of such oracles is denoted by $\mathsf{O}_\sigma^{\tilde{L}}$.
A *batch* query at points $x^{(1)}, \ldots, x^{(K)}$ draws a single $Z \sim P_Z$ and returns shared-noise gradients; for every $O \in \mathsf{O}_\sigma^{\tilde{L}}$,

$$O\big(x^{(1)}, \ldots, x^{(K)}\big) = \big(\mathbf{g}(x^{(1)}, Z), \ldots, \mathbf{g}(x^{(K)}, Z)\big), \tag{6}$$

where $\mathbf{g}$ satisfies (4)–(5).

First-order optimization algorithm. A stochastic first-order algorithm A initialized at x_0 proceeds in rounds $t = 1, 2, \ldots$. At round t, it (possibly adaptively) chooses a batch size K_t and a set of query points $x_t := \big(x_t^{(1)}, \ldots, x_t^{(K_t)}\big) \in (\mathbb{R}^d)^{K_t}$, then receives noisy gradients from the batch oracle O in (6). Formally, for $t \geq 1$,

$$x_t \;=\; \mathsf{A}_t\Big(O\big(x_1^{(1)}, \ldots, x_1^{(K_1)}\big),\; \ldots,\; O\big(x_{t-1}^{(1)}, \ldots, x_{t-1}^{(K_{t-1})}\big)\Big),$$

where A_t is a measurable policy that may depend on $\{x_0, x_1, \ldots, x_{t-1}\}$ and the algorithm's internal randomness. We write $\mathcal{A}_T$ for the class of algorithms that perform at most T oracle rounds.

Assumption 5 ($(1-\delta)$-stay-in-neighborhood)**.** *Let $F \in \mathcal{F}_\alpha$ and let $\mathcal{M}_F$ be the set of its local minimizers. Fix an isolated connected component $\mathcal{M} \subset \mathcal{M}_F$, a horizon $T \in \mathbb{N}$, and a $\delta \in (0, 1)$. An algorithm $\mathsf{A} \in \mathcal{A}_T$ satisfies "$(1-\delta)$-stay-in-neighborhood" assumption on $\mathcal{M}$ if there exist neighborhoods[9] $\mathcal{U}_0$ and $\mathcal{U}$ where $\mathcal{U}_0 \subseteq \mathcal{U}$ and $\mathcal{M} \subset \mathcal{U}_0 \subseteq \mathcal{U}$ such that, for any initialization $x_0 \in \mathcal{U}_0$, the algorithm's iterates satisfy*

$$\mathbb{P}(x_t \in \mathcal{U} \text{ for all } t = 1, \ldots, T) \;\geq\; 1 - \delta,$$

where the probability is taken over the internal randomness of the algorithm and the oracle.

Assumption 5 is natural: empirically, training dynamics in modern ML often enter a basin and thereafter behave *locally*; see, e.g., [Li et al., 2018, Garipov et al., 2018, Draxler et al., 2018]. In our lower-bound analysis, we restrict our attention to algorithms that satisfy Assumption 5 on $(\mathcal{U}_0, \mathcal{U})$; equivalently, for fixed $(\mathcal{M}, T, \delta)$, we define

$$\mathcal{A}_T(\mathcal{U}_0, \mathcal{U}, \delta) := \Big\{\mathsf{A} \in \mathcal{A}_T : \forall x_0 \in \mathcal{U}_0, \mathbb{P}\big(x_t \in \mathcal{U}, \forall t \leq T\big) \geq 1 - \delta\Big\}. \quad (7)$$

In the upper-bound analysis, we show that a SARAH-type method (see Algorithm 1) indeed satisfies this assumption.

3.2 Complexity lower bound

The main result of this section is stated in the following theorem.

Theorem 3.1. *Fix $\alpha \in (1, 2]$ and $\delta \in (0, 1/2)$. There exist (i) an instance (F, O) with $F \in \mathcal{F}_\alpha$ and a batch-smooth SFO $O \in \mathsf{O}_\sigma^{\tilde{L}}$, (ii) a connected component $\mathcal{M} \subseteq \mathcal{M}_F$ of local minimizers with level $l := F(y)$ for $y \in \mathcal{M}$, and (iii) constants $\tau > 0$, $R > 0$, together with an open neighborhood $\mathcal{N}(\mathcal{M}) \supseteq \mathbb{B}_2(\mathcal{M}; R)$ on which the local α–PŁ condition holds with parameter τ, such that the following holds.*
For any first–order algorithm A that (i) is initialized at $x_0 \in \mathcal{N}(\mathcal{M})$ and (ii) satisfies the stay–in–neighborhood property up to the iteration horizon T with $\mathcal{U}_0 \subseteq \mathcal{U} \subseteq \mathcal{N}(\mathcal{M})$ (i.e., $\mathsf{A} \in \mathcal{A}_T(\mathcal{U}_0, \mathcal{U}, \delta)$), the number of oracle queries m required to output $\hat{x}$ with $\mathbb{E}[F(\hat{x}) - l \mid \mathcal{E}_T(\mathcal{U})] \leq \varepsilon$, must satisfy $m = \Omega(\tau^{2/\alpha}\sigma^2\epsilon^{-2/\alpha})$ for $\varepsilon \leq R^\alpha \tau^{1/(\alpha-1)}(\frac{\alpha-1}{\alpha})^{\frac{\alpha}{\alpha-1}}$ where $\mathcal{E}_T(\mathcal{U}) := \{x_k \in \mathcal{U}$ for all iterates $k \leq T\}$.

*Sketch of proof.*We follow a standard information–theoretic reduction for minimax lower bounds [Raginsky and Rakhlin, 2009]. The idea is to cast optimization as a noisy binary hypothesis test [Tsybakov, 2008, Ch. 2] between two objective functions that are statistically hard to distinguish.
We construct two one–dimensional $\mathcal{C}^1$ functions that (i) satisfy local Lipschitzness, local smoothness, and a local α–PŁ condition; (ii) have unique minimizers separated by a distance ρ; and (iii) coincide outside a small neighborhood. Under a noisy gradient oracle with variance σ^2, the two functions induce statistically close distributions: choosing $\rho \asymp \sigma^{\alpha-1}/m^{(\alpha-1)/2}$ ensures the Kullback–Leibler divergence between the distributions of m oracle responses is $\mathcal{O}(1)$. Therefore, by Fano's inequality, any algorithm has a constant probability of outputting a point at least $\rho/2$ distance away from the true minimizer. Using the local error bound implied by the local α–PŁ condition, this distance lower bound translates into a function–value suboptimality (the local optimality gap defined in (2)) of order $\frac{\tau\sigma^\alpha}{m^{\alpha/2}}$. Since the stay-in-neighborhood event has probability at least $1 - \delta$ with $\delta < 1/2$, this unconditional testing error also yields a constant conditional error probability given $\mathcal{E}_T(\mathcal{U})$. Hence the expected local optimality gap is $\Omega\big(\tau\sigma^\alpha m^{-\alpha/2}\big)$, yielding the claimed complexity lower bound. □

Remark 3. *The case $\alpha = 1$ is not included in Theorem 3.1, as it follows from known results. In particular, Foster et al. (2019) proved that, under convexity and smoothness, any stochastic first-order method needs $\Omega(\varepsilon^{-2})$ oracle calls to find an ε-stationary point in the sense that $\mathbb{E}[\|\nabla F(x)\|] \leq \varepsilon$. In Section F.1, we showed that their hard instance function belongs to our class $\mathcal{F}_{\alpha=1}$: its set of stationary points coincides with the set of global minimizers, and it satisfies the local 1-PŁ, Lipschitz, and smoothness properties. Moreover, the stochastic gradients in their construction can be realized by an oracle $O \in \mathsf{O}_\sigma^{\tilde{L}}$. Therefore, when $\alpha = 1$, their lower bound of $\Omega(\varepsilon^{-2})$ holds in the setting considered in this section.*

[9] A convenient special case is the "tubular" choice $\mathcal{U}_0 = \mathbb{B}_2(\mathcal{M}; r_0)$ and $\mathcal{U} = \mathbb{B}_2(\mathcal{M}; r_1)$ with $0 < r_0 \leq r_1$.

3.3 Complexity upper bound

We introduce Algorithm 1, a SARAH–type variance–reduced method with *time-varying* batch-sizes and step-sizes. Define the mini-batch estimator $\mathbf{g}_{\mathcal{J}}(x) := \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} \mathbf{g}(x,\xi_j)$, where $\{\xi_j\}_{j\in\mathcal{J}}$ are i.i.d. samples and $\mathbb{E}[\mathbf{g}(x,\xi)] = \nabla F(x)$. The SARAH gradient update is given in line 4 of Algorithm 1. When Algorithm 1 is initialized in a small neighborhood of a fixed connected component $\mathcal{M}$ of local minimizers, the method attains the optimal sample complexity $\Theta(\varepsilon^{-2/\alpha})$ for obtaining an ε-local minimum point relative to $\mathcal{M}$, matching the lower bound in Theorem 3.1.

A key technical step is to show that, under a local α-PŁ condition and a batch-smooth oracle, the iterates of Algorithm 1 remain with high probability within a sufficiently small neighborhood of the target component $\mathcal{M}$. Even in the case that $\mathcal{M} = \{x_{\text{loc}}\}$ is a singleton local minimizer, this is a challenging task to keep updates of Algorithm 1 $\{x_t\}_{t\geq 1}$ close to x_{loc} *without* convexity. Our time-varying batch-sizes and step-sizes ensure small per-iteration steps and control the distance between two consecutive updates; together with the local α-PŁ geometry, this prevents escape from the neighborhood.

Conditioned on the event that the entire trajectory remains in a neighborhood of $\mathcal{M}$, we establish *function-value* convergence to the level of $\mathcal{M}$: $\lim_{t\to\infty}\mathbb{E}[F(x_t)] = l$, together with non-asymptotic bounds on $\mathbb{E}[F(x_t)] - l$ along the run. Since a local α-PŁ condition does not enforce convexity and $\mathcal{M}$ may be a manifold of minimizers, one should not expect x_t to converge to a single point in $\mathcal{M}$. In this case, function-value convergence is a natural notion of convergence.[10] Our main theorem, stated next, shows that Algorithm 1 indeed achieves the tight upper bound under local α-PŁ.

Algorithm 1 SARAH

Input: Maximum number of iterations T, batch sizes $\{n_g^t\}_{t=1}^T$, and the period length S, initial point x_0.

1: $t \leftarrow 0$
2: **while** $t \leq T-1$ **do**
3: Sample index set $\mathcal{J}_t$ with $|\mathcal{J}_t| = n_g^t$.
4: $$\mathbf{v}_t \leftarrow \begin{cases} \mathbf{g}_{\mathcal{J}_t}(x_t), & \text{mod}(t,S)=0 \\ \mathbf{g}_{\mathcal{J}_t}(x_t) - \mathbf{g}_{\mathcal{J}_t}(x_{t-1}) + \mathbf{v}_{t-1}, & \text{else} \end{cases}$$
5: $x_{t+1} \leftarrow x_t - \eta_t \mathbf{v}_t$
6: $t \leftarrow t+1$
7: **end while**
8: **return** x_T

Theorem 3.2. *Fix $\alpha \in [1,2)$ and $\delta \in (0,1)$, and let $\mathcal{M}$ be an isolated connected component of local minima of F with level $l = F(y)$ for all $y \in \mathcal{M}$. Assume $F \in \mathcal{F}_\alpha$ and choose $R > 0$ so that $\mathbb{B}_2(\mathcal{M};R) \subseteq \mathcal{N}(\mathcal{M})$ (cf. Assumption 2). Let the oracle be batch-smooth, $O \in \mathsf{O}_\sigma^{\tilde{L}}$, and run Algorithm 1 from x_0 with step sizes $\eta_t = \eta_0(t+1)^{-\frac{\alpha}{2}-(2-\alpha)\mathsf{x}}$ and batch-sizes $n_g^t = (t+1)^{1-2\mathsf{x}}$, for an arbitrary small value $\mathsf{x} > 0$. Denote the iterates by $\{x_t\}_{t\geq 1}$ and set $\hat{x} := x_T$. The following statements hold.*

- *There exists $s > 0$ such that, with $\mathcal{U} := \{x : \text{dist}(x,\mathcal{M}) \leq R/2\}$, $\mathcal{U}_0 := \{x : \text{dist}(x,\mathcal{M}) < R/2,\ F(x) - l \leq s/2\}$, and $\eta_0 = \mathcal{O}(\min\{\sqrt{s}, s^{-1}\delta, R\sqrt{\delta}\})$, if $x_0 \in \mathcal{U}_0$, the event $\mathcal{E}_T(\mathcal{U}) := \{\ x_t \in \mathcal{U}$ for all $t = 1,\dots,T\ \}$ occurs with probability at least $1-\delta$.*
- *Let $N := \mathbb{E}\Big[\sum_{t=1}^T n_g^t \mid \mathcal{E}_T(\mathcal{U})\Big]$ denote the expected total number of oracle queries used up to iteration T given $\mathcal{E}_T(\mathcal{U})$ defined in the first statement. Then for $\varepsilon \leq s/2$, $\mathbb{E}[F(\hat{x}) - l \mid \mathcal{E}_T(\mathcal{U})] \leq \varepsilon$, with $N = \mathcal{O}\big(\varepsilon^{-2/\alpha}\big)$.*

Remark 4 (On tuning and batch sizes)**.** *The α-dependent schedules in Theorem 3.2 are used to obtain the rate-optimal ε-dependence; if α is unknown, one may use a generic polynomially decaying step size (yielding a suboptimal worst-case rate under the same local assumptions). The increasing mini-batch sizes are likewise used to match the optimal oracle complexity; designing a localized method with comparable guarantees using single-sample (recursive-momentum) gradients is an interesting direction for future work.*

*Proof sketch.*We show that (i) the iterates *stay* near the target set of minimizers $\mathcal{M}$ with high probability, and (ii) conditioning on this event, the function value gap $D_t := F(x_t) - l$ decays at the optimal rate under the local α-PŁ.

(i) High–probability stay-in-neighborhood of $\mathcal{M}$. Let $R > 0$, the neighborhoods $\mathcal{U}_0 \subset \mathcal{U}$, the forbidden annulus $\mathcal{R}$, and the barrier levels $0 < s < s_0$ be as in Section 2.1. Starting from $x_0 \in \mathcal{U}_0$, we ensure (a) that updates are small so the iterates cannot jump across $\mathcal{R}$, and (b) that the function gaps stay below s so the barrier prevents entrance into $\mathcal{R}$. To ensure the event $\mathcal{E}_T(\mathcal{U}) = \{x_t \in \mathcal{U}$ for all $t = 1,...,T\}$ happens with high probability, we control the growth of the following two events by appropriate choice of our time-varying step/batch sizes.

(1) Small–error event $\mathcal{E}_{\text{err}}$. Following Appendix G.2,

[10] The function-value convergence is a symmetry-invariant target and does not depend on choosing a representative minimizer. See [Liu and Zhou, 2023] for more details.

we collect all stochastic terms in the descent recursion into an accumulator R_n (a "noise budget"). Under our decaying step sizes and growing batch sizes, its expected increments are summable, which yields $\mathbb{P}(\mathcal{E}_{\text{err}}) \geq 1 - \delta/2$ for $\mathcal{E}_{\text{err}} := \{\mathsf{R}_k < s,\ \forall k \leq T\}$. In turn, $\mathcal{E}_{\text{err}}$ implies $D_t < s$ for all $t \leq T$.

(2) Small–steps event $\mathcal{E}_{\text{step}}$. Because SARAH estimator (line 4 in Algorithm 1) is bounded on the neighborhood and step-sizes are tuned, each update is small: $\|x_t - x_{t-1}\| \leq R/4$ for $t \geq 1$. Thus no iterate can escape from $\mathbb{B}_2(\mathcal{M}; 3R/4)$ in a single step. We ensure this holds with probability at least $1-\delta/2$. By a union bound, both events occur simultaneously with probability at least $1-\delta$. On $\mathcal{E}_{\text{err}} \cap \mathcal{E}_{\text{step}}$, the preceding argument prevents the iterates from entering $\mathcal{R}$, hence $\mathcal{E}_T(\mathcal{U})$ holds with probability at least $1-\delta$.

(ii) Optimal complexity inside the neighborhood. Conditioned on staying in $\mathcal{U}$ (which holds with probability $\geq 1-\delta$), the same local smoothness and α-PŁ geometry imply the following recursion inequality for $\delta_t := \mathbb{E}[F(x_t) \mid \mathcal{E}_T(\mathcal{U})] - l$:

$$\delta_{t+1} \leq \delta_t - \frac{\eta_t}{2\tau^{\frac{2}{\alpha}}}\delta_t^{\frac{2}{\alpha}} + \frac{\eta_t}{2}\mathbb{E}[\|\mathbf{v}_t - \nabla F(x_t)\|^2 \mid \mathcal{E}_T(\mathcal{U})]. \quad (8)$$

The first term is a geometric drift induced by local α-PŁ, and the last term is the (decaying) noise level of the SARAH estimator; solving this nonlinear recursion amounts to balancing these two effects (see Lemma 18 in Appendix G.3). Using the time–varying batch sizes in Theorem 3.2 and the SARAH-update of gradients, we ensure that $\mathbb{E}[\|\mathbf{v}_t - \nabla F(x_t)\|^2 \mid \mathcal{E}_T(\mathcal{U})] = \mathcal{O}\big(\frac{1}{t}\big)$. Together with the step sizes from Theorem 3.2, the recursion in (8) yields $\delta_t = \mathcal{O}\big(t^{-\alpha/2+\alpha\times}\big)$. Therefore, to achieve $\delta_t = \mathcal{O}(\varepsilon)$, the total number of stochastic gradient calls (i.e., the sum of all mini–batch sizes) satisfies, on average, $\sum_{t=0}^{n-1} n_g^t = \mathcal{O}\big(\varepsilon^{-2/\alpha}\big)$. Consequently, the method attains the *optimal* oracle complexity $\Theta\big(\varepsilon^{-2/\alpha}\big)$, matching the lower bound in Theorem 3.1. □

Remark 5. *Our high–probability stability argument follows the localization scheme used for SGD in [Weissmann et al., 2025, Mertikopoulos et al., 2020] which proves that, once the iterate enters a small basin around a target component $\mathcal{M}$, it stays there with high probability. However, the proof for Algorithm 1 differs in two essential ways: (i) Different descent dynamics. For SGD with an* unbiased *gradient estimator $\mathbf{g}_t$, the descent inequality features a variance term scaled by η_t^2, of the form $\eta_t^2\ \mathrm{E}\big[\|\mathbf{g}_t - \nabla F(x_t)\|^2\big]$. For SARAH, the recursion involves the variance of the* biased *estimator $\mathbf{v}_t$ with a* linear *step-size factor, namely $\eta_t\ \mathrm{E}\big[\|\mathbf{v}_t - \nabla F(x_t)\|^2\big]$. The linear (rather than quadratic) scaling makes the stochastic error more difficult to control; noise decays more slowly and requires tighter coupling between step-size and batch-size (see Lemmas 17 and 18 in Appendix G.3). (ii) Non–martingale cross term. In the SGD analysis, the inner-product "cross term" $\langle \nabla F(x_t), \mathbf{g}_t - \nabla F(x_t)\rangle$ forms a martingale difference with zero conditional mean; its weighted sum is controlled directly from this zero-mean property. In contrast, SARAH's estimator $\mathbf{v}_t$ is* biased*, hence $\langle \nabla F(x_t), \mathbf{v}_t - \nabla F(x_t)\rangle$ is not a martingale difference. Our analysis instead controls the* weighted accumulation *of cross terms $\sum_{t=0}^{T-1} \eta_t \left\langle \nabla F(x_t), \mathbf{v}_t - \nabla F(x_t)\right\rangle$, by using the SARAH variance–reduction recursion with carefully chosen, time–varying step sizes $\{\eta_t\}$ and batch sizes $\{n_g^t\}$. This lets us bound simultaneously the bias–variance aggregate $\sum_{t=0}^{T-1} \eta_t \|\mathbf{v}_t - \nabla F(x_t)\|^2$ and $\sum_{t=0}^{T-1} \eta_t \left\langle \nabla F(x_t), \mathbf{v}_t - \nabla F(x_t)\right\rangle$, so that both remain uniformly small with high probability on the "small–error" event. This replaces the martingale step used in SGD and is the key technical difference in our localization analysis; see Lemma 13 in Appendix G.2 for the precise high–probability bound.*

4 LOWER BOUND ON GLOBAL CONVERGENCE RATE UNDER LOCAL PŁ AND CONVEXITY

In this section, we consider the problem of finding an ε-global-optimum point when the objective function $F : \mathcal{X} \to \mathbb{R}$ is convex and satisfies the local $(\alpha, \tau, \varepsilon)$-PŁ property (refer to Assumption 6). Our goal is to find a point $\hat{x} \in \mathcal{X}$ such that $F(\hat{x}) - \min_{x\in\mathcal{X}} F(x) \leq \varepsilon$ with probability at least $1-\delta$, with access to F through a stochastic first-order oracle with bounded stochastic gradients.

We first summarize the setting we use to establish the complexity lower bound.

Function class.

Assumption 6 (Local $(\alpha, \tau, \varepsilon)$-PŁ)**.** *Function $F : \mathcal{X} \to \mathbb{R}$ (where $\mathcal{X} \subseteq \mathbb{R}^d$) satisfies the local $(\alpha, \tau, \varepsilon)$-PŁ property when for all $x \in \mathcal{X} \cap \mathcal{S}_\varepsilon$, we have*

$$F(x) - \min_{x\in\mathcal{X}} F(x) \leq \tau \|\nabla F(x)\|^\alpha,$$

where $\mathcal{S}_\varepsilon := \{x :\ F(x) - \min_{x\in\mathcal{X}} F(x) \leq \varepsilon\}$, $\tau > 0$, and $1 \leq \alpha \leq 2$.

$\mathcal{F}^{\mathcal{X}}_{\alpha,\tau,\varepsilon}$ includes all convex functions that satisfy Assumptions 6, i.e.,

$$\mathcal{F}^{\mathcal{X}}_{\alpha,\tau,\varepsilon} = \left\{ \begin{array}{l} F : \mathcal{X} \to \mathbb{R} \\ \mathcal{X} \subset \mathbb{R}^d \end{array} \middle| \begin{array}{c} F \text{ is convex,} \\ F \text{ satisfies } (\alpha,\tau,\varepsilon)\text{-PŁ} \end{array} \right\}. \quad (9)$$

Domain class. Denote by $\mathbb{S}_R$, the class of convex, closed, and bounded sets in $\mathbb{R}^d$ whose diameter $\mathrm{diam}(\mathcal{X}) \leq R$ for every $\mathcal{X} \in \mathbb{S}_R$.

Stochastic first-order oracle with bounded stochastic gradients. We denote a family of stochastic first-order oracles satisfying the following properties by O^G: (i) satisfying the conditions in (4), and (ii) bounded stochastic gradients, i.e., $\|\mathbf{g}(x,z)\| \leq G$ for every $x \in \mathcal{X}$ and $z \in \mathcal{Z}$ where $G > 0$ is some constant.

We now provide a tight lower bound for the probability-based minimax oracle complexity of the function class $\mathcal{F}^{\mathcal{X}}_{\alpha,\tau,\varepsilon}$ and the family of oracles O^G for stochastic projected first-order methods.

Theorem 4.1. *For the family of domain sets $\mathbb{S}_R$, there exists a function F in the function class $\mathcal{F}^{\mathcal{X}}_{\alpha,\tau,\varepsilon}$, an oracle O in the family of oracles O^G, and $\alpha \in (1,2]$ and $\varepsilon \leq \min\{((\alpha-1)/\alpha)^\alpha \tau, 1\}$, such that the number of oracle queries for any stochastic first-order algorithm $\mathcal{A} \in \mathcal{A}_m$ required to output an estimate $\hat{x}$ satisfying $F(\hat{x}) - F^* \leq \varepsilon$ with probability at least $1-\delta$ is*

$$\Omega\left(G^2 \tau^{\frac{2}{\alpha}} \varepsilon^{-\frac{2}{\alpha}} \cdot \log\left(\frac{(\alpha-1)R}{2\alpha\varepsilon^{\frac{\alpha-1}{\alpha}}\tau^{\frac{1}{\alpha}}}\right)\right). \tag{10}$$

*Sketch of proof.*We reduce from noisy binary search problem (NBS): given a sorted set of N keys and a comparison oracle that is correct with probability $1/2 + p$, any (adaptive) strategy needs $\Omega(p^{-2}\log N)$ queries [Feige et al., 1994, Karp and Kleinberg, 2007]. Let $X = [0, R]$ and partition it into N equal intervals $[a_j, a_{j+1})$. The unknown target interval index j^* plays the role of the hidden position in NBS. From the NBS comparison outcomes we define a stochastic gradient oracle $g(x, Z)$ that is (i) unbiased for the gradient of a convex function F, and (ii) uniformly bounded $\|g(x,Z)\| \leq G$. The construction ensures that on the ε-sublevel set S_ε, $F(x) - F^* \leq \tau\,\|\nabla F(x)\|^\alpha$ and that any ε-minimizer must lie in the unique target interval $[a_{j^*}, a_{j^*+1})$, so solving the optimization problem to accuracy ε identifies j^*. We set $p \asymp \frac{\varepsilon^{1/\alpha}}{G\,\tau^{1/\alpha}}$ and $N \asymp \frac{\alpha-1}{2\alpha} \cdot \frac{R}{\varepsilon^{(\alpha-1)/\alpha}\,\tau^{1/\alpha}}$, so the gradient signal on S_ε matches the comparison reliability p, and ε-accuracy is equivalent to interval identification. If a stochastic first-order algorithm uses T oracle calls and outputs $\hat{x}$ with $F(\hat{x}) - F^* \leq \varepsilon$, then it can solve the NBS instance with at most a constant-factor more queries and the same success probability. Hence substituting our choices of p and N into the lower bound $\Omega(p^{-2}\log N)$ gives the result (for $\alpha \in (1,2]$ and suitably small ε so the construction lies in the stated class). □

Remark 6. *Note that every convex function satisfies the local $(\alpha = 1, \tau, \varepsilon)$-PŁ property. It is well-known that for bounded domain convex functions, stochastic first-order methods achieve a tight lower bound of $\Omega(\varepsilon^{-2})$ with access to O^G oracle [Nemirovskij and Yudin, 1983, Agarwal et al., 2009]. Therefore, similarly to Remark 3, we did not include the case $\alpha = 1$ in the statement of Theorem 4.1.*

Remark 7 (Complexity upper bound)**.** *In Xu et al. [2017, Theorem 1], the authors showed that for function $F \in \mathcal{F}^{\mathcal{X}}_{\alpha,\tau,\varepsilon}$ and oracle class O^G, a constrained version of the Accelerated Stochastic Subgradient Method (see Algorithm 1 in [Xu et al., 2017]) guarantees that $F(x_T) - \min_{x\in\mathcal{X}} F(x) \leq \varepsilon$ with probability $1-\delta$, for some $\delta > 0$, and $T = \mathcal{O}\left(G^2\tau^{2/\alpha} \cdot \log(1/\delta) \cdot \log\left(\varepsilon^{-(\alpha-1)/\alpha}\tau^{-1/\alpha}\right)/\varepsilon^{2/\alpha}\right)$ which matches with our lower bound in* (10) *in terms of dependency on ε, τ, and G.*

5 EMPIRICAL EVALUATION OF LOCAL α–PŁ

In this section, we numerically probe local α–PŁ-type behavior near terminal iterates in two learning tasks: (i) classification with a ReLU network, and (ii) dictionary learning. We also consider real-data versions of these experiments in Appendix J and additional learning tasks (e.g., polynomial and deep linear matrix factorization) in Appendix K; in some setups, the estimated envelope at $\alpha = 2$ explodes under tighter localization, suggesting that $\tau(\alpha)$ may be unbounded at $\alpha = 2$. Let $\hat{x}$ be the last iterate returned by full-batch gradient descent. In our experiments, we consider the runs for which $\|\nabla F(\hat{x})\|$ is small, so $\hat{x}$ is an *almost critical point*. We then consider a small ball $\mathbb{B}_2(\hat{x}; R)$ and compute, for $1 \leq \alpha \leq 2$, $\mathsf{R}_\alpha(x) := (F(x) - F(\hat{x}))\|\nabla F(x)\|^{-\alpha}$. Let $\tau(\alpha) := \sup_{x\in\mathbb{B}_2(\hat{x};R)} \mathsf{R}_\alpha(x)$. In both tasks (classification and dictionary learning), we observe that $\tau(\alpha)$ is finite on $\mathbb{B}_2(\hat{x}; R)$ for all $1 \leq \alpha \leq 2$ (see Figure 1). Since $\tau(\alpha)$ is evaluated on a fixed-radius neighborhood around an almost critical point, rather than on a certified neighborhood of an isolated connected component of local minimizers, these plots should be interpreted as empirical evidence for local α-PŁ-type behavior and not as a verification of Assumption 2. This also does not by itself prove that $\tau(\alpha) \to \infty$ as $\alpha \to 2$; however, we do observe a sharp growth of $\tau(\alpha)$ as α approaches 2 in both experiments. Accordingly, the experiments indicate that a finite local envelope of the form $F(x) - F(\hat{x}) \leq \tau(\alpha)\,\|\nabla F(x)\|^\alpha$ can describe the sampled neighborhood around $\hat{x}$. The complexity upper bound scales polynomially with $\tau(\alpha)$ and depends on ε as $\varepsilon^{-2/\alpha}$. Therefore, as α increases, the term depending on $\tau(\alpha)$ increases while the $\varepsilon^{-2/\alpha}$ term decreases. Since the estimated envelope is finite over the probed neighborhood for all $1 \leq \alpha \leq 2$, for sufficiently small ε, the optimal pair $(\tau(\alpha), \alpha)$ for the tightest upper bound lies near $\alpha = 2$. If ε is not sufficiently small, however, the optimal α moves towards $\alpha = 1$. These

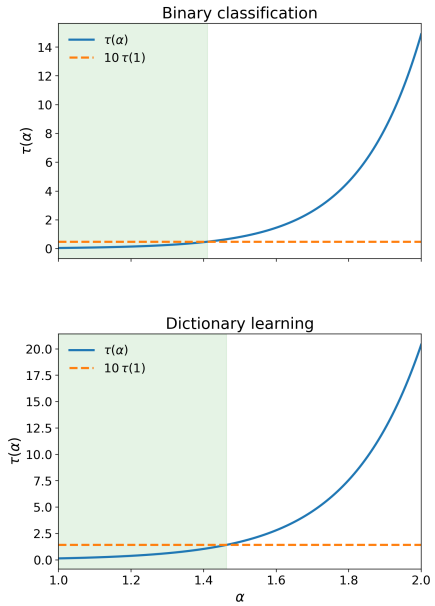

Figure 1: $\tau(\alpha)$ versus $1 \leq \alpha \leq 2$ on the ball $\mathbb{B}_2(\hat{x}; R = 0.1)$. The green region highlights $\{\alpha : \tau(\alpha) \leq 10\tau(1)\}$, illustrating a near-flat regime around $\alpha = 1$ and the sharp growth as $\alpha \uparrow 2$.

curves therefore provide heuristic guidance on how, for a given ε, one might tune α in the learning rate of Algorithm 1 to target a smaller predicted number of oracle queries. Further experimental details for these two tasks are provided in Appendix J, and additional empirical results are reported in Appendix K.

Limitations and future work. Our guarantees are local around an isolated connected component of minimizers and assume initialization in a neighborhood where the local α–PŁ condition holds. The constants (e.g., the neighborhood radius and τ) are problem-dependent, and achieving the optimal rate relies on schedules that depend on (or estimate) α and may use increasing mini-batches. Extending these results to more global guarantees and to batch-free variance-reduced updates are natural directions for future work.

References

A. Agarwal, M. J. Wainwright, P. Bartlett, and P. Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

A. Aich, A. B. Aich, and B. Wade. From sublinear to linear: Fast convergence in deep networks via locally polyak–łojasiewicz regions. *arXiv preprint arXiv:2507.21429*, 2025. URL `https://arxiv.org/abs/2507.21429`.

Z. Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex SGD. *Advances in Neural Information Processing Systems*, 31, 2018.

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *ICML*, pages 242–252. PMLR, 2019. URL `https://proceedings.mlr.press/v97/allen-zhu19a.html`.

H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

J. Bolte, A. Daniilidis, and A. Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4): 1205–1223, 2007a.

J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007b.

J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1–2):459–494, 2014.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

X. Chen, N. He, Y. Hu, and Z. Ye. Efficient algorithms for minimizing compositions of convex functions and random functions and its applications in network revenue management. *arXiv preprint arXiv:2205.01774*, 2022. URL `https://arxiv.org/abs/2205.01774`.

A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex SGD. *Advances in Neural Information Processing Systems*, 32, 2019.

Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. *arXiv preprint arXiv:1703.04933*, 2017. URL `https://arxiv.org/abs/1703.04933`.

F. Draxler, K. Veschgini, M. Salmhofer, and F. Hamprecht. Essentially no barriers in neural network energy landscape. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on*

Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, pages 1309–1318. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/draxler18a.html.

D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018. doi: 10.1287/moor.2017.0889.

S. S. Du, X. Zhai, B. Póczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1eK3i09YQ. ICLR.

C. Fang, C. J. Li, Z. Lin, and T. Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

I. Fatkhullin, J. Etesami, N. He, and N. Kiyavash. Sharp analysis of stochastic optimization under global kurdyka–łojasiewicz inequality. *arXiv preprint arXiv:2210.01748*, 2022. URL https://arxiv.org/abs/2210.01748.

U. Feige, P. Raghavan, D. Peleg, and E. Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, 1994.

X. Fontaine, V. De Bortoli, and A. Durmus. Convergence rates and approximation results for SGD and its continuous-time counterpart. In *Conference on Learning Theory*, pages 1965–2058. PMLR, 2021.

D. J. Foster, A. Sekhari, and K. Sridharan. Uniform convergence of gradients for non-convex learning and optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

D. J. Foster, A. Sekhari, O. Shamir, N. Srebro, K. Sridharan, and B. Woodworth. The complexity of making the gradient small in stochastic convex optimization. In *Conference on Learning Theory*, pages 1319–1345. PMLR, 2019.

C. D. Freeman and J. Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016. URL https://arxiv.org/abs/1611.01540.

T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, and A. G. Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

B. Ghorbani, S. Krishnan, and Y. Xiao. An investigation into neural net hessians. In *International Conference on Machine Learning (ICML)*, 2019.

B. Hanin and D. Rolnick. Complexity of linear regions in deep networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

M. Hardt and T. Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016. URL https://arxiv.org/abs/1611.04231.

R. Islamov, N. Ajroldi, A. Orvieto, and A. Lucchi. Loss landscape characterization of neural networks without over-parametrization. *arXiv preprint arXiv:2410.12455*, 2024. URL https://arxiv.org/abs/2410.12455.

R. Jiang and X. Li. Hölderian error bounds and kurdyka–łojasiewicz inequality for the trust region subproblem. *Mathematics of Operations Research*, 47(4):3140–3167, 2022. doi: 10.1287/moor.2021.1243.

C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1724–1732, 2017.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

R. M. Karp and R. Kleinberg. Noisy binary search and its applications. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 881–890. SIAM, 2007.

A. Khaled and P. Richtárik. Better theory for SGD in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020. URL https://arxiv.org/abs/2002.03329.

R. Kuditipudi, X. Wang, H. Lee, Y. Zhang, Z. Li, W. Hu, R. Ge, and S. Arora. Explaining landscape connectivity of low-cost solutions for multilayer nets. *Advances in neural information processing systems*, 32, 2019.

K. Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'Institut Fourier*, 48(3):769–783, 1998.

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1246–1257. PMLR, 2016. URL https://proceedings.mlr.press/v49/lee16.html.

L. Lei, C. Ju, J. Chen, and M. I. Jordan. Non-convex finite-sum optimization via SCSG methods. *Advances in Neural Information Processing Systems*, 30, 2017.

G. Li and T. K. Pong. Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of Computational Mathematics*, 18(5):1199–1232, 2018. doi: 10.1007/s10208-017-9370-0.

H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Y. Li and Y. Yuan. Convergence analysis of two-layer neural networks with ReLU activation. *Advances in Neural Information Processing Systems*, 30, 2017.

J. Liu, S. Wright, C. Ré, V. Bittorf, and S. Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. In *International Conference on Machine Learning*, pages 469–477. PMLR, 2014.

Y. Liu, K. Zhang, T. Basar, and W. Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020.

Z. Liu and Z. Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. *arXiv preprint arXiv:2312.08531*, 2023. URL `https://arxiv.org/abs/2312.08531`.

S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate bayesian inference. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

S. Masiha, S. Salehkaleybar, N. He, N. Kiyavash, and P. Thiran. Stochastic second-order methods improve best-known sample complexity of SGD for gradient-dominated functions. *Advances in Neural Information Processing Systems*, 35:10862–10875, 2022.

S. Masiha, Z. Shen, N. Kiyavash, and N. He. Superquantile-gibbs relaxation for minima-selection in bi-level optimization. *arXiv preprint arXiv:2505.05991*, 2025. URL `https://arxiv.org/abs/2505.05991`.

P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. *Advances in Neural Information Processing Systems*, 33:1117–1128, 2020.

G. F. Montúfar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 2924–2932, 2014.

I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1): 69–107, 2019.

A. S. Nemirovskij and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

B. Neyshabur, R. Salakhutdinov, and N. Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2613–2621. PMLR, 2017. URL `https://proceedings.mlr.press/v70/nguyen17b.html`.

P. H. Nguyen, L. Nguyen, and M. van Dijk. Tight dimension independent lower bound on the expected convergence rate for diminishing step sizes in SGD. *Advances in Neural Information Processing Systems*, 32, 2019.

Q. Nguyen. On connected sublevel sets in deep learning. In *International conference on machine learning*, pages 4790–4799. PMLR, 2019.

B. T. Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

M. Raginsky and A. Rakhlin. Information complexity of black-box convex optimization: A new look via feedback information theory. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 803–810. IEEE, 2009.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Nonconvex learning via stochastic gradient langevin dynamics: A nonasymptotic analysis. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1674–1683, 2017.

Q. Rebjock and N. Boumal. Fast convergence to non-isolated minima: four equivalent conditions for C^2 functions. *Mathematical Programming*, pages 1–49, 2024.

L. Sagun, U. Evci, V. U. Guney, Y. Dauphin, and L. Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017. URL `https://arxiv.org/abs/1706.04454`.

A. Shapiro and A. Nemirovski. On complexity of stochastic programming problems. *Continuous optimization: Current trends and modern applications*, pages 111–146, 2005.

C. Şimşekli, L. Sagun, and C. Hongler. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pages 9722–9732, 2021. URL `https://proceedings.mlr.press/v139/simsek21a/simsek21a.pdf`.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519.

L. van den Dries. *Tame Topology and O-minimal Structures*. Cambridge University Press, 1998.

L. van den Dries, A. Macintyre, and D. Marker. The elementary theory of restricted analytic functions with exponentiation. *Annals of Mathematics*, 140 (1):183–205, 1994.

S. Weissmann, S. Klein, W. Azizian, and L. Döring. Almost sure convergence of stochastic gradient methods under gradient domination. *Transactions on Machine Learning Research*, 2025.

A. J. Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function. *Journal of the American Mathematical Society*, 9(4):1051–1094, 1996.

Y. Xu, Q. Lin, and T. Yang. Stochastic convex optimization: Faster local growth implies faster global convergence. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2017.

T. Yang and Q. Lin. Rsg: Beating subgradient method without smoothness and strong convexity. *The Journal of Machine Learning Research*, 19(1):236–268, 2018.

Q. Yu, Y. Wang, B. Huang, Q. Lei, and J. D. Lee. Optimal sample complexity bounds for non-convex optimization under kurdyka-lojasiewicz condition. In *International Conference on Artificial Intelligence and Statistics*, pages 6806–6821. PMLR, 2023.

P. Yue, C. Fang, and Z. Lin. On the lower bound of minimizing polyak–łojasiewicz functions. *arXiv preprint arXiv:2212.13551*, 2022. URL `https://arxiv.org/abs/2212.13551`.

H. Zhang, J. Jiang, and Z.-Q. Luo. On the linear convergence of a proximal gradient method for a class of nonsmooth convex minimization problems. *Journal of the Operations Research Society of China*, 1 (2):163–186, 2013.

D. Zhou and Q. Gu. Stochastic recursive variance-reduced cubic regularization methods. In *International Conference on Artificial Intelligence and Statistics*, pages 3980–3990. PMLR, 2020.

D. Zou, Y. Cao, D. Zhou, and Q. Gu. Stochastic gradient descent optimizes over-parameterized deep ReLU networks. *arXiv preprint arXiv:1811.08888*, 2018. URL `https://arxiv.org/abs/1811.08888`.

CHECKLIST

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**/No/Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes/No/**Not Applicable**]

   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]

   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

Optimal Local Convergence Rates of Stochastic First-Order Methods Under Local α-PŁ: Supplementary Materials

Notations

We adopt the following notation in the sequel. Calligraphic letters (e.g., $\mathcal{S}$) denote sets. Lowercase bold letters (e.g., x) denote vectors. $\|\cdot\|$ denotes the ℓ_2-norm of a vector. We use $KL(\mu\|\nu) := \int \log\left(\frac{d\mu}{d\nu}(x)\right)\mu(dx)$ to denote the Kullback–Leibler (KL) divergence between two probability measures μ and ν. The diameter of the subset $\mathcal{X}$ of $\mathbb{R}^d$ is defined by $\operatorname{diam}(\mathcal{X}) := \sup_{x,y\in\mathcal{X}} \|x-y\|$. $\mathcal{C}$ denotes the class of continuous real functions. That is, f is in differentiability class $\mathcal{C}^k$ if and only if there exists a k-th derivative of f which is continuous. Given functions $f, g : \mathcal{A} \to [0,\infty)$ where $\mathcal{A}$ could be any set, we use non-asymptotic big-O notation: $f = \mathcal{O}(g)$ if there exists a constant $c < \infty$ such that $f(a) \le c \cdot g(a)$ for all $a \in \mathcal{A}$ and $f = \Omega(g)$ if there is a constant $c > 0$ such that $f(a) \ge c \cdot g(a)$. We write $f = \tilde{\mathcal{O}}(g)$ as a shorthand for $f = \mathcal{O}(g \cdot \max\{1, (\log(g))^k\})$ for some integer $k > 0$ and $\tilde{\Omega}$ is similarly defined. The d-dimensional ball with radius R around the center x_0 with respect to ℓ_2-norm is denoted by $\mathbb{B}_2(x_0; R) := \left\{x \in \mathbb{R}^d : \|x - x_0\| \le R\right\}$. For a subset $\mathcal{A} \subseteq \mathbb{R}^d$, its closure is denoted $\overline{\mathcal{A}}$ (and sometimes $\operatorname{cl}(\mathcal{A})$) and is the smallest closed set containing $\mathcal{A}$. We use $\mathbf{1}(\cdot)$ for the indicator function: $\mathbf{1}(E) = 1$ if event E occurs and 0 otherwise. We use $\sigma(\mathcal{S})$ to denote the sigma-field generated by a collection $\mathcal{S}$, i.e., the smallest sigma-field containing $\mathcal{C}$. For notational convenience, throughout all proofs in appendices, we write L_1 and L_2 for the local regularity constants of F on the compact neighborhood $\mathcal{N}(\mathcal{M})$ around the connected component $\mathcal{M}$ of local minimizers of F: for all $x, y \in \mathcal{N}(\mathcal{M})$,

$$
\begin{aligned}
&|F(x) - F(y)| \le L_1 \|x - y\| \\
&\|\nabla F(x) - \nabla F(y)\| \le L_2 \|x - y\|.
\end{aligned}
$$

Organization of the Appendix. Appendix A reviews PŁ-type conditions and oracle-complexity upper/lower bounds in both convex and nonconvex settings. Appendix B motivates using a single (uniform) constant τ in Assumption 2. Appendix C explains why $\alpha \in (1, 2)$ captures flatness and why many ML losses satisfy the KŁ inequality, hence admit a local α–PŁ property; Appendix D further shows that at nondegenerate minima one recovers the classical local ($\alpha = 2$)–PŁ. Proofs of the main results from Section 2 are deferred to Appendix E. The proof of Theorem 3.1 appears in Appendix F.2, and the proof of Theorem 3.2 appears in Appendix G. The results in Section 4 are proved in Appendix H. Appendix I compares Theorem 4.1 with the lower bound of Foster et al. [2019]. Appendix J provides details for Section 5, and Appendix K reports additional empirical evaluations.

A RELATED WORK

PŁ property and its applications: The ($\alpha = 2$)-PŁ property (commonly called PŁ condition) was initially introduced by [Polyak, 1963]. Karimi et al. [2016] showed that the ($\alpha = 2$)-PŁ condition is less restrictive than several known global optimality conditions in the literature of machine learning [Liu et al., 2014, Necoara et al., 2019, Zhang et al., 2013]. The ($\alpha = 2$)-PŁ property is satisfied (sometimes locally rather than globally, and also under distributional assumptions) for the population risk in some learning models including neural networks with one hidden layer [Li and Yuan, 2017], ResNets with linear activation [Hardt and Ma, 2016], generalized linear models and robust regression [Foster et al., 2018].
The global PŁ inequality is often too restrictive for modern neural networks: even when optimization behaves well in practice, verifying a uniform global PL constant typically relies on substantial—and often impractical—over-parameterization, as global convergence guarantees are proved in very wide-network regimes [Allen-Zhu et al., 2019, Du et al., 2019, Zou et al., 2018]. Islamov et al. [2024] make this point explicit and construct settings where classical PL-type (and related "aiming") conditions fail, motivating weaker, geometry-aware

alternatives; they propose landscape conditions that avoid heavy over-parameterization and tolerate saddle regions. This evidence justifies focusing on *local* α-PL behavior near basins of attraction—where training actually settles—rather than assuming a global PL landscape that rarely holds in realistic deep models.
Aich et al. [2025] show that neighborhoods around initialization that are *locally quasi-convex* with a *stable neural tangent kernel (NTK)* in fact satisfy a *local PL* inequality; they term such neighborhoods *Locally Polyak–Łojasiewicz Regions (LPLRs)* and prove that gradient descent attains *linear* convergence once iterates enter an LPLR. Their analysis provides a concrete mechanism (NTK stability) by which local PL—the special case $\alpha = 2$ of our local α-PL framework—emerges at finite number of hidden neurons per layer (width). Our results are complementary: while Aich et al. [2025] establish deterministic linear rates for GD under local PL, we derive *optimal local convergence rates for stochastic first-order methods* under the more general local α-PL condition ($\alpha \in [1, 2]$), thereby covering noisy/mini-batch regimes.

Complexity lower bounds: In the convex setting, several complexity lower bounds have been derived by establishing a connection between stochastic optimization and hypothesis testing. For instance, Shapiro and Nemirovski [2005] reduced a class of one-dimensional linear optimization problems to a binary hypothesis testing problem. Later on, this approach was used to derive the minimax oracle complexity of stochastic convex optimization in several works [Agarwal et al., 2009, Raginsky and Rakhlin, 2009]. As an example, Agarwal et al. [2009] obtained a lower bound of $\Omega(\varepsilon^{-2})$ for the minimax oracle complexity of stochastic first-order methods in order to achieve an ε-global-optimum point of a bounded-domain Lipschitz convex function. This bound is derived through a reduction to a Bernoulli vector parameter estimation problem. For the same function class in [Agarwal et al., 2009], Raginsky and Rakhlin [2009] derived a complexity lower bound of $\Omega(\varepsilon^{-2})$ by a reduction to hypothesis testing with feedback, where the oracle provides noisy gradients by adding Gaussian noise to the true gradients[11]. If the function is smooth (instead of Lipschitz) and convex, and the initial optimality gap is bounded (instead of the domain being bounded), a lower bound of $\Omega(\varepsilon^{-2})$ exists for the oracle complexity of stochastic first-order methods, according to Foster et al.'s complexity analysis [Foster et al., 2019]. This bound is derived through a reduction to a noisy binary search problem.

In the non-convex setting, under ($\alpha = 2$)-PŁ and L-smoothness, Yue et al. [2022] established a lower bound of $\Omega(L\tau \log(\varepsilon^{-1}))$ on the deterministic first-order methods to achieve an ε-global-optimum point[12]. The main idea is based on a "zero-chain" function[13] proposed as a hard instance, which is composed of the worst convex function designed by [Nesterov, 2003] and a coordinate-wise function that makes the function non-convex. More recently, Yu et al. [2023] obtained lower bounds on the oracle complexity of zeroth-order methods for non-convex smooth and α-PŁ functions with an additive noise oracle. This lower bound is tight in terms of the dependence on ε for dimensions less than six.

For our lower bound in the non-convex setting (Theorem 3.1), akin to Raginsky and Rakhlin [2009] we use a reduction to hypothesis testing with an additive Gaussian noise oracle. We benefit from a set of mutual information bounds to establish a tight lower bound on the complexity of stochastic first-order optimization algorithms for smooth and gradient-dominated functions. What distinguishes Theorem 3.1 from [Raginsky and Rakhlin, 2009, Theorem 2] is the construction of hard instances that satisfy local α-PŁ. These instances allow us to derive the optimal dependence on the precision $\varepsilon > 0$ in the complexity lower bound.

In the convex setting, under local $(\alpha, \tau, \varepsilon)$-PŁ property, we use a reduction to the noisy binary search problem in order to obtain a tight lower bound for first-order algorithms. In Appendix I, we discuss in more detail how our approach for deriving the lower bound in Theorem 4.1 compares to [Foster et al., 2019].

Complexity upper bounds: In the non-convex unconstrained optimization setting, Khaled and Richtárik [2020] showed that under ($\alpha = 2$)-PŁ condition, stochastic gradient descent (SGD) with time-varying step-size reaches an ε-global-optimum point with an oracle complexity of $\mathcal{O}(1/\varepsilon)$. Furthermore, it was shown that this dependency of the oracle complexity on ε is optimal for SGD [Nguyen et al., 2019]. Recently, Fontaine et al. [2021] obtained an oracle complexity $\mathcal{O}(\varepsilon^{-4/\alpha+1})$ for SGD under smoothness and α-PŁ property for $1 \leq \alpha \leq 2$. Fatkhullin et al. [2022] establish an oracle complexity of $\mathcal{O}(\varepsilon^{-2/\alpha})$ for the variance-reduced method PAGER under a batch-smooth stochastic first-order oracle. Their analysis assumes that the entire trajectories of SGD

[11] Note that Agarwal et al. [2009] considered noisy first-order oracles which do not allow additive noise due to a coin-tossing construction.

[12] In this lower bound, the dependencies on L, τ, and ε are the same as the ones in gradient descent's iteration complexity.

[13] For a zero-chain function having a sufficiently high dimension, $d - T$ entries of update vector will never reach their optimal values after the execution of any first-order algorithm for a given T number of iterations.

and PAGER remain within the region where the objective satisfies the local-α PŁ condition. In contrast, we *show* that the trajectory of Algorithm 1 stays inside the local neighborhood on which the function satisfies local α-PŁ. For convex functions, when α-PŁ holds on an ε-sub-level set of a global minimizer (see Assumption 6), stochastic first-order algorithms achieve an ε-global-optimum point with $\tilde{\mathcal{O}}(\varepsilon^{-2/\alpha})$ samples of stochastic gradients [Xu et al., 2017, Yang and Lin, 2018]. In Theorem 4.1, we show that the dependency of number of queries $\tilde{\mathcal{O}}(\varepsilon^{-2/\alpha})$ on ε is tight.

B UNIVERSALITY OF THE CONSTANT τ IN ASSUMPTION 2

In this part, we first show that α-KŁ property implies α-PŁ one as follows:

Lemma 2 (KŁ $\Rightarrow$ local α–PŁ near a compact isolated component)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}$ be $\mathcal{C}^1$. Let $\mathcal{M}$ be a* compact, isolated, connected *component of local minimizers of F, and set $l := F(y)$ for any $y \in \mathcal{M}$. Assume that F satisfies a local KŁ inequality at each $x_* \in \mathcal{M}$: there exist $r_{x_*} > 0$, $\alpha_{x_*} \in [1, 2]$, and $\tau_{x_*} > 0$ such that*

$$F(x) - l \;\le\; \tau_{x_*} \|\nabla F(x)\|^{\alpha_{x_*}} \qquad \textit{for all } x \in \mathbb{B}_2(x_*; r_{x_*}) \textit{ with } F(x) > l.$$

Then there exist $r > 0$, $\alpha \in [1, 2]$, and $\tau > 0$ such that, for every

$$x \in \mathbb{B}_r(\mathcal{M}) \;:=\; \bigcup_{x^* \in \mathcal{M}} \mathbb{B}_2(x^*; r),$$

we have $F(x) > l$ for $x \notin \mathcal{M}$ and the uniform local α–PŁ bound

$$F(x) - l \;\le\; \tau \|\nabla F(x)\|^{\alpha}.$$

Proof. By the assumption, for each $x^* \in \mathcal{M}$ there exist $r_{x^*} > 0$, $\alpha_{x^*} \in [1, 2]$, and $\tau_{x^*} > 0$ with

$$F(x) - l \;\le\; \tau_{x^*} \|\nabla F(x)\|^{\alpha_{x^*}} \qquad \forall x \in \mathbb{B}_2(x^*; r_{x^*}) \text{ with } F(x) > l. \tag{11}$$

Since $\mathcal{M}$ is an isolated compact set of local minima, by shrinking r_{x^*} if needed, we may assume that

$$F(x) > l \quad \text{for all } x \in \mathbb{B}_2(x^*; r_{x^*}) \setminus \mathcal{M},$$

and also, by continuity of F at $\mathcal{M}$, that

$$F(x) - l \;\le\; 1 \quad \text{for all } x \in \mathbb{B}_2(x^*; r_{x^*}).$$

The compactness of $\mathcal{M}$ yields that there is a *finite* subcover $\mathcal{A} = \{y_1^*, \ldots, y_m^*\} \subset \mathcal{M}$ such that

$$\mathcal{U} \;:=\; \bigcup_{y^* \in \mathcal{A}} \mathbb{B}_2(y^*; r_{y^*}) \;\supset\; \mathcal{M}.$$

Define

$$\alpha \;:=\; \min_{y^* \in \mathcal{A}} \alpha_{y^*} \in [1, 2], \qquad \tau \;:=\; \max_{y^* \in \mathcal{A}} \tau_{y^*}^{\alpha/\alpha_{y^*}} \;>\; 0.$$

Fix any $x \in \mathcal{U}$ and choose $y^* \in \mathcal{A}$ with $x \in \mathbb{B}_2(y^*; r_{y^*})$. From (11) and $F(x) - l \le 1$ we get

$$\|\nabla F(x)\| \;\ge\; \tau_{y^*}^{-1/\alpha_{y^*}} \big(F(x) - l\big)^{1/\alpha_{y^*}} \;\ge\; \tau^{-1/\alpha} \big(F(x) - l\big)^{1/\alpha},$$

where we used (i) $\alpha \le \alpha_{y^*}$ so $(F(x) - l)^{1/\alpha_{y^*}} \ge (F(x) - l)^{1/\alpha}$ (since $F(x) - l \le 1$), and (ii) $\tau \ge \tau_{y^*}^{\alpha/\alpha_{y^*}}$ so $\tau^{-1/\alpha} \le \tau_{y^*}^{-1/\alpha_{y^*}}$. Rearranging the terms gives the desired uniform bound

$$F(x) - l \;\le\; \tau \|\nabla F(x)\|^{\alpha}, \qquad \forall x \in \mathcal{U}.$$

Finally, since $\mathcal{U}$ is open and contains $\mathcal{M}$, choose $r > 0$ with $\mathbb{B}_r(\mathcal{M}) \subset \mathcal{U}$. As $\mathcal{M}$ is an isolated component of local minimizers of F, $F(x) > l$ for $x \in \mathbb{B}_r(\mathcal{M}) \setminus \mathcal{M}$. $\square$

Lemma 3 (Uniformization of the local α–PŁ constant across components)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}$ be $\mathcal{C}^1$ and fix $\alpha \in [1, 2]$. Let $\mathcal{M}_F$ denote the set of local minimizers of F, and suppose $\mathcal{M}_F$ decomposes into* isolated *connected components. For each connected component $\mathcal{M} \subset \mathcal{M}_F$, set $l_{\mathcal{M}} := F(y)$ for all $y \in \mathcal{M}$, and assume that there exist an open neighborhood $\mathcal{U}_{\mathcal{M}}$ of $\mathcal{M}$ and a constant $\tau_{\mathcal{M}} > 0$ such that*

$$F(x) - l_{\mathcal{M}} \;\le\; \tau_{\mathcal{M}} \|\nabla F(x)\|^{\alpha} \qquad \text{and} \qquad F(x) > l_{\mathcal{M}} \text{ for all } x \in \mathcal{U}_{\mathcal{M}} \setminus \mathcal{M}. \tag{12}$$

Then the following hold.

(a) If $\mathcal{M}_F$ has finitely many connected components $\{\mathcal{M}_1, \dots, \mathcal{M}_m\}$, then let

$$\tau \;:=\; \max_{1 \le j \le m} \tau_{\mathcal{M}_j}, \qquad \mathcal{U} \;:=\; \bigcup_{j=1}^{m} \mathcal{U}_{\mathcal{M}_j},$$

and for all $x \in \mathcal{U}$ let $\mathcal{M}$ be the (unique) component with $x \in \mathcal{U}_{\mathcal{M}}$. Then we have that

$$F(x) - l_{\mathcal{M}} \;\le\; \tau \|\nabla F(x)\|^{\alpha} \quad \text{and} \quad F(x) > l_{\mathcal{M}} \text{ if } x \notin \mathcal{M}.$$

(b) More generally, let $K \subset \mathbb{R}^d$ be compact. Then only finitely many components have neighborhoods intersecting K; denote this finite index set by

$$\mathsf{I}(K) \;:=\; \{\mathcal{M} \text{ component}\colon\; \mathcal{U}_{\mathcal{M}} \cap K \neq \emptyset\}.$$

Setting

$$\tau(K) \;:=\; \max_{\mathcal{M} \in \mathsf{I}(K)} \tau_{\mathcal{M}}, \qquad \mathcal{U}(K) \;:=\; \bigcup_{\mathcal{M} \in \mathsf{I}(K)} \mathcal{U}_{\mathcal{M}},$$

we have that, for all $x \in \mathcal{U}(K)$ let $\mathcal{M}$ be the (unique) component with $x \in \mathcal{U}_{\mathcal{M}}$,

$$F(x) - l_{\mathcal{M}} \;\le\; \tau(K) \|\nabla F(x)\|^{\alpha} \quad \text{and} \quad F(x) > l_{\mathcal{M}} \text{ if } x \notin \mathcal{M}.$$

Proof. Since each component $\mathcal{M}$ is isolated, by shrinking the neighborhoods $\{\mathcal{U}_{\mathcal{M}}\}$ if necessary we may assume they are pairwise disjoint and hence that (12) holds on each $\mathcal{U}_{\mathcal{M}}$.

(a) Finite-component case. Let $\tau := \max_j \tau_{\mathcal{M}_j}$. Fix $x \in \mathcal{U}$ and let $\mathcal{M}$ be the unique component with $x \in \mathcal{U}_{\mathcal{M}}$. Then (12) gives $F(x) - l_{\mathcal{M}} \le \tau_{\mathcal{M}} \|\nabla F(x)\|^{\alpha} \le \tau \|\nabla F(x)\|^{\alpha}$, and the strict inequality of $\mathcal{M}$ also follows from (12).

(b) Compact-region case. Because the neighborhoods $\{\mathcal{U}_{\mathcal{M}}\}$ are pairwise disjoint and each contains its (closed) component $\mathcal{M}$, only finitely many of them can intersect a compact set K (otherwise we would obtain an infinite family of pairwise disjoint open sets with points in K, contradicting compactness via a standard finite-subcover argument). Thus $\mathsf{I}(K)$ is finite, and the same max argument as in part (a) with $\tau(K) := \max_{\mathcal{M} \in \mathsf{I}(K)} \tau_{\mathcal{M}}$ yields the claim on $\mathcal{U}(K) = \bigcup_{\mathcal{M} \in \mathsf{I}(K)} \mathcal{U}_{\mathcal{M}}$. □

C α-PŁ FUNCTIONS

In this section, we start in Section C.1 by motivating the local α-PŁ for the regime $\alpha \in (1, 2)$. Then we introduce three closed-form examples that satisfy the α-PŁ property for $1 < \alpha < 2$. Next we provide in Section C.2 an example of ($\alpha = 1$)-PŁ function in the network revenue management problem. Finally, we discuss the implications of KŁ theory in machine learning in Section C.3.

C.1 Local alpha PŁ regime with $\alpha \in (1, 2)$

In this section, we explain why exponents $\alpha \in (1, 2)$ capture local flatness: if the first nonzero variation of F at a minimizer occurs at order $p > 2$ (i.e., all derivatives up to order $p - 1$ vanish), then F exhibits local α–PŁ growth with $\alpha = \frac{p}{p-1} \in (1, 2)$. We also present explicit examples of functions that satisfy a local α–PŁ inequality for $1 < \alpha < 2$.

Definition (order-p contact at a minimizer). Let $F : \mathbb{R}^d \to \mathbb{R}$ be $\mathcal{C}^p$ near a local minimizer $x^\star$ with value $l := F(x^\star)$. We say that F has *order-p contact* at $x^\star$ (with $p > 2$) if all derivatives up to order $p-1$ vanish at $x^\star$,

$$\nabla F(x^\star) = 0, \quad \nabla^2 F(x^\star) = 0, \ \ldots, \ D^{p-1}F(x^\star) = 0,$$

and the first nonzero term in the Taylor expansion of F is of the order p with a positive coefficient in the sense that, for h small,

$$F(x^\star + h) = l + c_p\|h\|^p + o(\|h\|^p) \quad \text{for some } c_p > 0.$$

(For manifolds of minimizers $\mathcal{M}$, the same definition applies *in the normal directions* to $\mathcal{M}$; tangential directions remain flat.)
Consequences. For sufficiently small $\|h\|$,

$$c_1\|h\|^p \le F(x^\star + h) - l \le c_2\|h\|^p,$$
$$c_3\|h\|^{p-1} \le \|\nabla F(x^\star + h)\| \le c_4\|h\|^{p-1}.$$

Eliminating $\|h\|$ yields the local error bound

$$F(x) - l \ \le \ \tau\,\|\nabla F(x)\|^\alpha \qquad \text{with} \quad \alpha = \frac{p}{p-1} \in (1,2), \quad \tau = \frac{c_2}{c_3^\alpha}.$$

Thus $\alpha \in (1,2)$ corresponds to *degenerate curvature* of F at local minimizers (the Hessian vanishes or is rank-deficient). For a minimizer manifold $\mathcal{M}$, writing $x = (u, v)$ with u normal to $\mathcal{M}$ and v tangential to $\mathcal{M}$, if $F(u, v) = \phi(u)$ with $\phi(u) = \Theta(\|u\|^p)$, the same relation i.e. $\alpha = p/(p-1)$ holds in a small tube around $\mathcal{M}$.

Explicit models (with $p > 2$ and $\alpha = p/(p-1) \in (1,2)$).

Example 1 (Isolated minimizer)**.** *Consider $f(x) = c \cdot \|x\|^q$, where $q > 2$ and $c > 0$. $f(x)$ is α-PŁ with $\alpha = q/(q-1)$ and $\tau = q^{q/(1-q)} \cdot C^{1/(1-q)}$.*

Example 2 (Manifold of minimizers)**.** *$F(u, v) = \|u\|^p$ on $\mathbb{R}^r \times \mathbb{R}^{d-r}$ has minimizers $\mathcal{M} = \{(0, v) \mid v \in \mathbb{R}^{d-r}\}$; the same local α–PŁ bound holds in a tubular neighborhood of $\mathcal{M}$ (normal directions behave as the isolated case).*

Example 3 (Product loss)**.** *In this example, $F : \mathbb{R}^d \to \mathbb{R}$ is defined as $F(x) = (x_1x_2\cdots x_d)^2$. It appears for instance as the squared loss of a one-dimensional d-layer linear neural network model on a 2-dimensional data point $(z, y) = (1, 0)$ where $z = 1$ is the data sample (feature) and $y = 0$ the label so that the squared loss on $(x_1x_2\cdots x_dz - y)^2$. $F(x)$ satisfies the global ($\alpha = 2d/(2d-1)$)-PŁ property because:*

$$F(\mathbf{x}) - F^* \le \frac{1}{(4d)^{\frac{d}{2d-1}}}\|\nabla F(\mathbf{x})\|^{\frac{2d}{2d-1}}. \tag{13}$$

Indeed,

$$\begin{aligned}\|\nabla F(\mathbf{x})\|^2 &= 4(x_1x_2\cdots x_d)^4 \cdot \left[\frac{1}{x_1^2} + \ldots + \frac{1}{x_d^2}\right] \\ &\overset{(a)}{\ge} 4(x_1x_2\cdots x_d)^4 \cdot \frac{d}{(x_1\ldots x_d)^{2/d}} \\ &= 4d(F(x) - F^*)^{2-1/d},\end{aligned} \tag{14}$$

where (a) follows because the harmonic mean is always upper bounded by the geometric mean (HM-GM inequality).

Why does $\alpha \in (1,2)$ PL appear in over-parametrized models. Empirically and structurally, The loss function landscape of trained network exhibits many *flat* directions in the neighborhood of solutions. Flatness means that the Hessian is rank-deficient and that the first nonzero variation normal to the minimizer set can be of order $p > 2$. As explained above, this yields the local growth law $F(x) - l \lesssim \|\nabla F(x)\|^\alpha$ with $\alpha = p/(p-1) \in (1,2)$. The following observations support this picture.
(i) Symmetries create flat directions. Common architectures admit parameter symmetries (e.g., permutations of hidden units; positive-homogeneous layer rescaling for ReLU), producing connected families of equivalent

solutions and rank-deficient Hessians at minima [Dinh et al., 2017, Neyshabur et al., 2015]. *(ii) Empirical curvature.* Hessian spectra measured at trained solutions show many near-zero eigenvalues with a few outliers, indicating extended flat directions [Sagun et al., 2017, Ghorbani et al., 2019]. *(iii) connectivity of solution set.* Independently trained solutions are connected by simple low-loss curves, revealing wide, flat basins [Garipov et al., 2018, Draxler et al., 2018].

C.2 $(\alpha = 1)$-PŁ property in network revenue management problem

Another class of examples where an $(\alpha = 1)$-PŁ property appears is found in supply chain and revenue management. As shown in Chen et al. [2022], problems in these domains can often be formulated as

$$\min_{x \in \mathcal{X}} F(x) := \mathbb{E}[\phi(x \wedge \xi)],$$

where $\mathcal{X}$ is a convex and compact subset of $\mathbb{R}^d$, $\xi \in \mathbb{R}^d$ is a non-negative random vector, the symbol $\wedge$ denotes component-wise minimum, and $\phi(\cdot)$ is a convex function. As a result, $F(\cdot)$ becomes non-convex. However, such a problem often admits a convex reformulation

$$\min_{y \in \mathcal{Y}} G(y) := F(g^{-1}(y)),$$

where $g(x) = x \wedge \xi$ and $g^{-1}(\mathbf{u}) := (g_1^{-1}(u_1), \ldots, g_d^{-1}(u_d))$ with $g_i^{-1}(u_i) := \min_{x \in \mathcal{X}}\{x_i : g_i(x_i) \geq u_i\}$. Therefore, the function $G(\cdot)$ is convex. Suppose $g : \mathcal{X} \to \mathcal{Y}$ is a bijective differentiable map with $\nabla g(x) \succeq \lambda I$, $\lambda \geq 0$ for all $x \in \mathcal{X}$. Then function $F(\cdot)$ satisfies $(\alpha = 1)$-PŁ property. Indeed, for any $x \in \mathcal{X}$ with $g(x) = y$,

$$\begin{aligned} F(x) - F^* &= G(y) - G^* \\ &\leq \langle \nabla G(y), y - y^* \rangle \\ &\leq \|\nabla G(y)\| \cdot \|y - y^*\| \\ &= \|\nabla g^{-1}(y) \nabla F(x)\| \cdot \|y - y^*\| \\ &\leq \frac{D_{\mathcal{Y}}}{\lambda} \|\nabla F(x)\|, \end{aligned}$$

where $D_{\mathcal{Y}}$ is the diameter of the set $\mathcal{Y}$. Therefore, $F(\cdot)$ is a $(\alpha = 1)$-PŁ function.

C.3 Implications of KŁ/α-PŁ

Kurdyka–Łojasiewicz (KŁ) inequality. Let $F : \mathbb{R}^d \to (-\infty, +\infty]$ be proper and lower semicontinuous, and assume F is $\mathcal{C}^1$ on a neighborhood of a compact set $\mathcal{C} \subset \mathbb{R}^d$. Suppose $\mathcal{C} \subset \operatorname{crit} F$ where $\operatorname{crit} F := \{x \in \mathbb{R}^d : \nabla F(x) = 0\}$ and that F is constant on $\mathcal{C}$; denote this constant by $F_{\mathcal{C}}$. We say that F *satisfies the KŁ inequality on* $\mathcal{C}$ if there exist $\eta > 0$, a concave function $\phi : [0, \eta) \to \mathbb{R}_+$ with $\phi(0) = 0$, $\phi \in \mathcal{C}^1((0, \eta))$, and $\phi'(s) > 0$ for $s \in (0, \eta)$, and a neighborhood $\mathcal{N}$ of $\mathcal{C}$ such that, for all $x \in \mathcal{N}$ with $F_{\mathcal{C}} < F(x) < F_{\mathcal{C}} + \eta$,

$$\phi'\big(F(x) - F_{\mathcal{C}}\big)\, \|\nabla F(x)\| \;\geq\; 1.$$

A common parametrization uses the *KŁ exponent* $\theta \in [0, 1)$: there exist $c > 0$ and $\eta > 0$ such that $\phi(s) = c\, s^{1-\theta}$ on $[0, \eta)$, in which case for all $x \in \mathcal{N}$ with $F_{\mathcal{C}} < F(x) < F_{\mathcal{C}} + \eta$,

$$\|\nabla F(x)\| \;\geq\; \kappa \left(F(x) - F_{\mathcal{C}}\right)^{\theta} \qquad \text{for some } \kappa > 0. \tag{15}$$

(For nonsmooth F, replace $\|\nabla F(x)\|$ by $\operatorname{dist}(0, \partial F(x))$.) Real-analytic functions and, more broadly, definable/tame functions satisfy the KŁ inequality around each compact set of critical points [Kurdyka, 1998, Bolte et al., 2007a, Attouch et al., 2010].

Local α-PŁ condition. Given a target set of minimizers $\mathcal{M}$, the function F obeys a *local α-PŁ* inequality on a neighborhood $\mathcal{U}$ of $\mathcal{M}$ if there exist $\tau > 0$ and $\alpha \in [1, 2]$ such that

$$\begin{gathered} F(x) - F^* \;\leq\; \tau \|\nabla F(x)\|^{\alpha} \qquad \forall x \in \mathcal{U}, \\ \text{where } F^* := F(y) \text{ for any } y \in \mathcal{M}. \end{gathered} \tag{16}$$

When $\alpha = 2$ we recover the classical PŁ inequality. Matching equations (15) and (16) with each other require to take $\alpha = 1/\theta$: if F has KŁ exponent $\theta \in (0, 1]$ near $\mathcal{M}$, then (16) holds with $\alpha = 1/\theta$ (up to rescaling of constants); in particular, nondegenerate minima (a nondegenerate minimum is a local minimizer x such that $\nabla^2 F(x) \succ 0$) have $\theta = 1/2$ and hence $\alpha = 2$ (see more details in Appendix D).

o-minimal structures, definable sets, and definable functions. An *o-minimal structure* $\mathcal{S}$ on the ordered field $(\mathbb{R}, +, \cdot, \leq)$ is a sequence $\{\mathcal{S}_n\}_{n\in\mathbb{N}}$ where each $\mathcal{S}_n$ is a collection of subsets of $\mathbb{R}^n$ (called the *definable sets of arity* n) satisfying:

- (Closure) Each $\mathcal{S}_n$ is closed under finite unions, finite intersections, and complements; if $A \in \mathcal{S}_n$ and $B \in \mathcal{S}_m$, then $A \times B \in \mathcal{S}_{n+m}$; and if $A \in \mathcal{S}_{n+m}$, then any linear image or k-dimensional coordinate projection of A belongs to the appropriate $\mathcal{S}_k$.
- (Basic sets) All algebraic sets are definable: for every polynomial $p : \mathbb{R}^n \to \mathbb{R}$, the zero set $\{x \in \mathbb{R}^n : p(x) = 0\}$ lies in $\mathcal{S}_n$.
- (o-minimality) Every set in $\mathcal{S}_1$ is a finite union of points and open intervals. In particular, there are no "wild" (e.g., fractal or infinitely oscillatory) definable subsets of $\mathbb{R}$.

A set $A \subset \mathbb{R}^n$ is *definable (in $\mathcal{S}$)* if $A \in \mathcal{S}_n$.

A function $f : \mathbb{R}^d \to \mathbb{R}^m$ is *definable (in $\mathcal{S}$)* if its *graph*

$$\operatorname{gph} f \;:=\; \{(x, y) \in \mathbb{R}^d \times \mathbb{R}^m : y = f(x)\}$$

is a definable set; equivalently, $\operatorname{gph} f \in \mathcal{S}_{d+m}$. Thus, when we say "$f$ is definable", we are not placing f itself inside a set collection; rather, we certify definability via its graph being one of the sets admitted by the structure.

Semialgebraic sets and functions (the structure $\mathbb{R}_{\text{alg}}$). A set $S \subset \mathbb{R}^n$ is *semialgebraic* if it can be constructed from finitely many polynomial equalities and inequalities using finitely many unions and intersections. Typical examples include intervals, halfspaces, Euclidean balls/ellipsoids, polynomial sublevel sets, and finite unions of these. The family of all semialgebraic sets $\{\mathcal{S}_n\}_{n\geq 1}$ forms an o-minimal structure, denoted $\mathbb{R}_{\text{alg}}$. A function $f : \mathbb{R}^d \to \mathbb{R}^m$ is *semialgebraic* precisely when $\operatorname{gph} f$ is a semialgebraic subset of $\mathbb{R}^{d+m}$.

Two canonical o-minimal structures are

1. $\mathbb{R}_{\text{alg}}$, whose definable sets/functions are exactly the semialgebraic ones,
2. $\mathbb{R}_{\text{an,exp}}$, an o-minimal expansion of the real field obtained by adjoining (i) all *restricted real-analytic* functions (i.e., real-analytic on a box and extended by 0 outside it) and (ii) the exponential function. In particular, any set/function definable using these primitives (together with the field operations and order) is definable in $\mathbb{R}_{\text{an,exp}}$.

Many ML loss functions used in practice are definable: polynomial losses/constraints; piecewise-linear maps such as ReLU and max-pooling (built by finitely many linear pieces); compositions of these with affine layers; and standard regularizers like ℓ_1/ℓ_2. Definability is stable under the operations used to assemble models (addition, composition, taking products, projections), which is why these classes are called *tame* [van den Dries, 1998, van den Dries et al., 1994, Wilkie, 1996]. In the rest of this section, we detail why some of the most popular ML loss functions are tame.

- **Least squares / ridge / polynomial models.** Polynomial maps and norms are semialgebraic, and the semialgebraic class is closed under finite unions/intersections, products, addition/multiplication, and composition. Hence, empirical risks built from squares and polynomial penalties are semialgebraic (thus definable and tame) and, by the KŁ theory for definable/subanalytic functions, satisfy a local KŁ inequality [Bolte et al., 2007a, Attouch et al., 2010, van den Dries, 1998].

- **Logistic / cross-entropy.** We consider two standard cases.

  (a) Binary logistic regression. Given data $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$, the empirical logistic risk

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} \log\big(1 + \exp(-y_i\, x_i^\top w)\big) + \lambda\, R(w)$$

  is *real-analytic* on $\mathbb{R}^d$ when R is analytic (e.g., $R(w) = \frac{1}{2}\|w\|_2^2$), because $t \mapsto \log(1 + \exp t)$ is analytic on $\mathbb{R}$, and compositions/sums of analytic functions are analytic. By the classical Łojasiewicz/KŁ gradient inequality, every real-analytic function satisfies a local KŁ inequality around each critical point [Bolte et al., 2007a, Kurdyka, 1998]. Hence L enjoys a local KŁ inequality, which is equivalent to a local α–PŁ inequality for some $\alpha \in (1, 2]$ near its critical points.

  (b) Multiclass softmax cross-entropy. With logits $z_i = Wx_i \in \mathbb{R}^K$ and labels $y_i \in \{1, \dots, K\}$, the loss

$$L(W) = \frac{1}{n}\sum_{i=1}^{n}\left(\log\Big(\sum_{k=1}^{K} e^{z_{ik}}\Big) - z_{i,y_i}\right) + \lambda\, R(W)$$

  is real-analytic. Here z_{ik} is the k-th entry of vector z_i. The map $z \mapsto \log\sum_k e^{z_k}$ is analytic on $\mathbb{R}^K$, and $z_i = Wx_i$ is linear in W. Therefore, L satisfies a local KŁ (hence local α–PŁ) inequality around its critical points [Bolte et al., 2007a, Kurdyka, 1998].

  (c) Nonsmooth regularizers. If R is nonsmooth but *definable* (e.g., $R(w) = \|w\|_1$, which is semialgebraic), then L remains definable in the o-minimal structure $\mathrm{R}_{\mathrm{an,exp}}$ (which contains restricted analytic functions and exponential functions). Definable functions satisfy the KŁ property in the (sub)differential sense [Bolte et al., 2007a, Kurdyka, 1998], so the same conclusion holds with $\|\nabla L\|$ replaced by $\mathrm{dist}(0, \partial L)$.

- **Deep nets with ReLU with ℓ_1/ℓ_2 regularization.** Absolute value and finite maxima of polynomials are semialgebraic, hence so are ReLU and max.
  Semialgebraic functions/sets are closed under finite sums, products, and *composition* with affine maps. Therefore each layer map

$$x \mapsto \mathrm{ReLU}(Ax + b),$$

  is semialgebraic, and so is their finite composition. This also implies that the network is piecewise linear (polyhedral), see [Montúfar et al., 2014, Hanin and Rolnick, 2019].
  The penalties $\|w\|_1$ (piecewise linear) and $\frac{1}{2}\|w\|_2^2$ (polynomial) are semialgebraic. For squared loss (and other polynomial losses), the empirical risk

$$\frac{1}{n}\sum_i \tfrac{1}{2}\|f_\theta(x_i) - y_i\|^2 \;+\; \lambda_1\|w\|_1 + \lambda_2\tfrac{1}{2}\|w\|_2^2$$

  is semialgebraic because it is a finite sum of semialgebraic functions. More generally, if the data loss uses exp/log (e.g., cross-entropy), the whole objective is *definable* in the o-minimal structure $\mathrm{R}_{\mathrm{an,exp}}$ since semialgebraic functions (the network) and restricted analytic/exp/log are definable and closed under composition. Hence these objectives admit a local KŁ (therefore local α–PŁ) inequality.

- **Matrix factorization.** Objectives of the form $\frac{1}{2}\|UV^\top - M\|_F^2$ (with polynomial regularizers) are polynomials in the entries of U and V, and hence are semialgebraic. Semialgebraic functions are definable and satisfy a local KŁ inequality; therefore these problems admit a local α–PŁ bound near their critical points. This is precisely the setting used in KŁ-based convergence analyses (e.g., PALM) [Bolte et al., 2014, Attouch et al., 2010].

D WHY IS α EQUAL TO 2 AT NONDEGENERATE MINIMA

Let $x_\star$ be a nondegenerate local minimum of F, i.e., such that $\nabla F(x_\star) = 0$ and $\nabla^2 F(x_\star) \succ 0$. By continuity of the Hessian, there exist $\mu > 0$ and a neighborhood U of $x_\star$ such that $\nabla^2 F(x) \succeq \mu I$ for all $x \in U$; therefore F is (locally) μ-strongly convex on U. Local strong convexity implies the PŁ inequality with exponent 2:

$$F(x) - F(x_\star) \;\le\; \frac{1}{2\mu}\,\|\nabla F(x)\|^2 \qquad \text{for all } x \in U, \tag{17}$$

which is exactly the α-PŁ condition with $\alpha = 2$ [see, e.g., Karimi et al., 2016].

E PROOFS OF SECTION 2

For notational convenience, throughout all proofs in appendices, we write L_1 and L_2 for the local regularity constants of F on the compact neighborhood $\mathcal{N}(\mathcal{M})$: for all $x, y \in \mathcal{N}(\mathcal{M})$,

$$|F(x) - F(y)| \leq L_1 \|x - y\| \tag{18}$$

$$\|\nabla F(x) - \nabla F(y)\| \leq L_2 \|x - y\|. \tag{19}$$

E.1 Existence of tube $\mathbb{B}_2(\mathcal{M}; R)$ inside $N(\mathcal{M})$.

Let $\mathcal{M} \subset \mathbb{R}^d$ be a (nonempty) compact set and let $\mathcal{N}(\mathcal{M})$ be any open set containing $\mathcal{M}$. In Lemma 4, we show that there exists $R > 0$ such that

$$\mathbb{B}_2(\mathcal{M}; R) \ \subset \ \mathcal{N}(\mathcal{M}).$$

In particular, if F satisfies the local α-PŁ inequality on the open neighborhood $\mathcal{N}(\mathcal{M})$, then it also holds on the metric tube $\mathbb{B}_2(\mathcal{M}; R)$ for some $R > 0$.

Lemma 4. *Let $\mathcal{S} \subset \mathbb{R}^d$ be compact and $\mathcal{U} \subset \mathbb{R}^d$ be an open set containing $\mathcal{S}$. Define the distance to the complement*

$$d^* \ := \ \inf_{x \in \mathcal{S}} \ \mathrm{dist}\big(x, \mathbb{R}^d \setminus \mathcal{U}\big) \ = \ \inf_{x \in \mathcal{S}} \ \inf_{y \notin \mathcal{U}} \ \|x - y\|.$$

Then $d^* > 0$ and, for any $0 < R < d^*$, $\mathbb{B}_2(S; R) \subset \mathcal{U}$.

*Proof.*The function $x \mapsto \mathrm{dist}\big(x, \mathbb{R}^d \setminus \mathcal{U}\big)$ is continuous and strictly positive on $\mathcal{S}$, because $\mathcal{S} \subset \mathcal{U}$ and $\mathcal{U}$ is open. By the extreme value theorem, the compact set $\mathcal{S}$ attains a positive minimum $d^* > 0$. If $0 < R < d^*$ and $z \in \mathbb{B}_2(\mathcal{S}; R)$, then $\mathrm{dist}\big(z, \mathbb{R}^d \setminus \mathcal{U}\big) \geq d^* - R > 0$, and therefore $z \in U$. □

Remark 8. *Suppose that for a connected component $\mathcal{M} \subseteq \mathcal{M}$ there exists an open neighborhood $\mathcal{N}(\mathcal{M})$ on which F satisfies the local α-PŁ inequality* (1)*, and that $\mathcal{M}$ is compact. Then there exists $R > 0$ with $\mathbb{B}_2(\mathcal{M}; R) \subseteq \mathcal{N}(\mathcal{M})$, so Assumption 2 holds with this R.*

F PROOFS OF SECTION 3

F.1 Proof of Remark 3

In this part, we show that the hard instance of function in Foster et al. [2019, Theorem 4] satisfies Assumption 2 for $x \in \mathbb{B}_2^d(0; R)$. Moreover, the set of stationary points of this function coincides with its set of global minimizers. In addition, the stochastic gradients of their worst-case function can be produced by an oracle $O \in \mathsf{O}_\sigma^{\tilde{L}}$. Let m be the number of iterations of a given stochastic first-order algorithm. Foster et al. [2019, Theorem 4] use the following hard instance of function:

$$\tilde{F}(x) = \frac{\sigma}{m} \sum_{i=1}^{m} \langle x, z_i \rangle + \frac{b}{2} \|x\|^2, \tag{20}$$

where $\{z_1, \ldots, z_m\}$ are orthonormal vectors in $\mathbb{R}^d$ $(d \geq m)$ and $b = 2\sigma/(R\sqrt{m})$. $\tilde{F}$ attains its minimum at $x^* = -\sigma/(bm) \sum_{i=1}^m z_i$ which has norm $\|x^*\| = \sigma/(b\sqrt{m}) = R/2 < R$. $\tilde{F}$ satisfies $L_1(R)$-Lipschitzness (18) and $L_2(R)$-smoothness (19) with

$$L_1(R) \leq \sigma + bR, \quad L_2(R) \leq b.$$

The stochastic gradient is as follows:

$$\tilde{\mathbf{g}}(x, z) = \sigma z + bx,$$

where z is a random variable with the uniform distribution over $\{z_1, \ldots, z_m\}$. Note that $\mathbb{E}[\tilde{\mathbf{g}}(x, z)] = \nabla \tilde{F}(x)$,

$$\begin{aligned} \mathbb{E}[\|\tilde{\mathbf{g}}(x, z) - \nabla \tilde{F}(x)\|^2] &= \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[\|\sigma z_i - \frac{\sigma}{m} \sum_{j=1}^{m} z_j\|^2] \\ &= \sigma^2 \left(1 - \frac{1}{m}\right) \leq \sigma^2, \end{aligned} \tag{21}$$

and

$$\mathbb{E}[\|\tilde{\mathbf{g}}(x,z)-\tilde{\mathbf{g}}(y,z)\|^2]=b^2\|x-y\|^2.$$

Therefore, the stochastic gradient $\tilde{\mathbf{g}}$ satisfies the properties of $\mathsf{O}_\sigma^{\tilde{L}}$. Since $\tilde{F}$ is convex, we have

$$\begin{aligned}\tilde{F}(x)-\tilde{F}^* &\le \langle \nabla\tilde{F}(x), x^*-x\rangle \\ &\le \sup_{y\in\mathbb{B}_2^d(0;R)} \|x^*-y\|\cdot\|\nabla\tilde{F}(x)\| \le 2R\|\nabla\tilde{F}(x)\|. \end{aligned} \tag{22}$$

Thus $\tilde{F}$ satisfies Assumption 2 with $\tau\ge 2R$ and then $\tilde{F}\in\mathcal{F}_{L_1,L_2}^\alpha$. Note that

$$\begin{aligned}\|\nabla\tilde{F}(x)\|^2 &= b^2\|x\|^2+\frac{\sigma^2}{m}+\frac{2b\sigma}{m}\sum_{i=1}^m\langle x,z_i\rangle \\ &= 2b(\tilde{F}(x)-\tilde{F}^*)\end{aligned} \tag{23}$$

where $\tilde{F}^*=-\sigma^2/(2bm)$. Foster et al. [2019, Theorem 4] prove that $\mathbb{E}[\|\nabla\tilde{F}(\hat{x})\|^2]\ge\sigma^2/(8m)$ where $\hat{x}$ is the output of any randomized algorithm whose input is $S=\{z_1,\dots,z_{m/2-1}\}$. Then

$$\mathbb{E}[\tilde{F}(\hat{x})]-\tilde{F}^*=\frac{1}{2b}\mathbb{E}[\|\nabla\tilde{F}(\hat{x})\|^2]\ge\frac{R\sqrt{m}}{2\sigma}\cdot\frac{\sigma^2}{8m}=\frac{\sigma^2R}{16\sqrt{m}}.$$

Therefore, when $\alpha=1$, their lower bound of $\Omega(\varepsilon^{-2})$ holds in the same setting as Theorem 3.1.

F.2 Proof of Theorem 3.1

Proof of Theorem 3.1. Let $\mathcal{F}_\alpha^{\text{uni}}$ be a subset of $\mathcal{F}_\alpha$ such that every $f\in\mathcal{F}_\alpha^{\text{uni}}$ has a unique local minimizer. For two functions f_0 and f_1 in $\mathcal{F}_\alpha^{\text{uni}}$, let us define $\delta(f_0,f_1):=\|x_{f_1}^*-x_{f_0}^*\|$ where $x_{f_i}^*=\arg\min_{x\in\mathbb{R}^d}f_i(x)$ for $i\in\{0,1\}$. For a fixed algorithm $\mathsf{A}\in\mathcal{A}_T$, let x_T be the output of the T-th iteration of A and $\hat{F}_T$ be a function in $\mathcal{F}_\alpha^{\text{uni}}$ whose minimizer is x_T.

[Rebjock and Boumal, 2024, Proposition 2.2] implies then that for $F\in\mathcal{F}_\alpha^{\text{uni}}$, there exists a neighborhood of x_F^*, $\tilde{\mathcal{N}}(x_F^*)\subseteq\mathcal{N}(x_F^*)$ such that $\lambda\cdot\|x-x_F^*\|^{\alpha/(\alpha-1)}\le F(x)-F(x_F^*)$ for all $x\in\tilde{\mathcal{N}}(x_F^*)$, where $\lambda=((\alpha-1)/\alpha)^{\alpha/(\alpha-1)}\tau^{-1/(\alpha-1)}$ and $x_F^*=\arg\min_{x\in\mathbb{R}^d}F(x)$.
We assume that for every algorithm $\mathsf{A}\in\mathcal{A}_T$, there are two subsets $\mathcal{U}_0^F$ and $\mathcal{U}^F$ of $\tilde{\mathcal{N}}(x_F^*)$ such that if $x_0\in\mathcal{U}_0^F$ we have $\{x_t\}_{t=1}^T$ lie in $\mathcal{U}^F$ with high probability $1-\delta$, i.e., $\mathsf{A}\in\mathcal{A}_T(\mathcal{U}_0^F,\mathcal{U}^F,\delta)$. We define the following event:

$$\mathcal{E}_m(\mathcal{U}^F):=\left\{\{x_t\}_{t=1}^m\in\mathcal{U}^F\right\}.$$

Conditioned on $\mathcal{E}_m(\mathcal{U}^F)$, for $0<\rho<1/2$, we obtain

$$\begin{aligned}&\sup_{F\in\mathcal{F}_\alpha^{\text{uni}}}\inf_{\mathsf{A}\in\mathcal{A}_m(\mathcal{U}_0^F,\mathcal{U}^F,\delta)}\mathbb{E}[F(x_m)\mid\mathcal{E}_m(\mathcal{U}^F)]-F(x_F^*) \\ &\ge\lambda\cdot\sup_{F\in\mathcal{F}_\alpha^{\text{uni}}}\inf_{\mathsf{A}\in\mathcal{A}_m(\mathcal{U}_0^F,\mathcal{U}^F,\delta)}\mathbb{E}\left[\|x_m-x_F^*\|^{\frac{\alpha}{\alpha-1}}\mid\mathcal{E}_m(\mathcal{U}^F)\right] \\ &\overset{(a)}{\ge}\lambda\cdot\left(\sup_{F\in\mathcal{F}_\alpha^{\text{uni}}}\inf_{\mathsf{A}\in\mathcal{A}_m(\mathcal{U}_0^F,\mathcal{U}^F,\delta)}\mathbb{E}\left[\|x_m-x_F^*\|\mid\mathcal{E}_m(\mathcal{U}^F)\right]\right)^{\frac{\alpha}{\alpha-1}} \\ &\overset{(b)}{\ge}\lambda\cdot\left(\frac{\rho}{2}\cdot\sup_{F\in\mathcal{F}_\alpha^{\text{uni}}}\inf_{\mathsf{A}\in\mathcal{A}_m(\mathcal{U}_0^F,\mathcal{U}^F,\delta)}\mathbb{P}\left[\delta(\hat{F}_m,F)>\frac{\rho}{2}\mid\mathcal{E}_m(\mathcal{U}^F)\right]\right)^{\frac{\alpha}{\alpha-1}},\end{aligned} \tag{24}$$

where (a) comes from Jensen's inequality, and (b) from Markov's inequality and $\delta(\hat{F}_m,F)=\|x_m-x_F^*\|$ as $\hat{F}_m$ is a function in $\mathcal{F}_\alpha^{\text{uni}}$ whose minimizer is x_m.

In order to give a lower bound on (24), we use Fano's inequality given in the following lemma.

Lemma 5 (Fano-type inequality)**.** *Let $\mathcal{F}$ be a non-parametric class of functions, $\delta(\cdot,\cdot) : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ be a semi-distance*[14]*, and $\{P_f : f \in \mathcal{F}\}$ be a family of probability distribution indexed by $f \in \mathcal{F}$. Assume that there are $f_0, f_1 \in \mathcal{F}$ such that $\delta(f_0, f_1) \geq \rho > 0$ and $KL(P_{f_0} \| P_{f_1}) \leq \gamma$ for some $\gamma > 0$. Then,*

$$\sup_{f\in\mathcal{F}} \inf_{\hat{f}} P_f\left(\left\{\delta(\hat{f}, f) > \frac{\rho}{2}\right\}\right) \geq \max\left\{\frac{e^{-\gamma}}{4}, \frac{1-\sqrt{\gamma/2}}{2}\right\}, \tag{25}$$

where $\hat{f}$ is an estimator of f from samples generated by P_f.

Lemma 5 follows from Tsybakov [2008, Theorem 2.5]. In order to apply Lemma 5, we need to specify $f_0, f_1 \in \mathcal{F}_\alpha^{\text{uni}}$ and corresponding P_{f_0}, P_{f_1} such that $\delta(f_0, f_1) \geq \rho$ and $KL(P_{f_0} \| P_{f_1}) \leq \gamma$.

Construction of f_0, f_1: We construct two continuously differentiable 1-dimensional functions $f_0, f_1 : \mathbb{R} \to \mathbb{R}$ as follows:

$$f_0(x) = \begin{cases} C|x|^{\frac{\alpha}{\alpha-1}} & -R \leq x \leq R \\ C\frac{\alpha}{\alpha-1}R^{\frac{1}{\alpha-1}}x + D & R < x \\ -C\frac{\alpha}{\alpha-1}R^{\frac{1}{\alpha-1}}x + D & x < -R \end{cases}, \tag{26}$$

$$f_1(x) = \begin{cases} 2^{\frac{1}{\alpha-1}}C(|x-\rho|^{\frac{\alpha}{\alpha-1}} + |\rho|^{\frac{\alpha}{\alpha-1}}) & 0 \leq x \leq 2\rho \\ f_0(x) & 2\rho \leq x \\ -\frac{\alpha}{\alpha-1}2^{\frac{1}{\alpha-1}}C\rho^{\frac{1}{\alpha-1}}x + 2^{\frac{\alpha}{\alpha-1}}C\rho^{\frac{\alpha}{\alpha-1}} & x \leq 0, \end{cases} \tag{27}$$

where $C > 0$ is a constant and $D = -(\alpha-1)^{-1}CR^{\alpha/(\alpha-1)}$. We choose C as follows:

$$C = \tau^{-\frac{1}{\alpha-1}}\left(\frac{\alpha-1}{\alpha}\right)^{\frac{\alpha}{\alpha-1}}. \tag{28}$$

Based on Lemma 1, for every function $F \in \mathcal{F}_\alpha^{\text{uni}}$, $\mathcal{U}_0^F, \mathcal{U}^F \subseteq \mathbb{B}_2(x_F^*; R)$. Let the neighborhoods of the minimizers of f_0 and f_1 be defined as $\tilde{\mathcal{N}}_{f_0}(0) = [-R, R]$ and $\tilde{\mathcal{N}}_{f_1}(\rho) = [0, R]$, respectively. Then $\mathcal{U}_0^{f_i}, \mathcal{U}^{f_i}$ attributed to f_i, are subset of $\tilde{\mathcal{N}}_{f_i}(x_{f_i}^*)$ for $i \in \{0, 1\}$. We also pick ρ such that $2\rho < R$.
In Lemma 7 (refer to Appendix F.3), we prove that $f_0, f_1 \in \mathcal{F}_\alpha^{\text{uni}}$ with the following constants for smoothness and Lipschitzness of both functions in their own neighborhoods:

$$\begin{aligned} L_2(R) &\geq \frac{(\alpha-1)^{\frac{2-\alpha}{\alpha-1}}}{(\alpha\tau)^{\frac{1}{\alpha-1}}}R^{\frac{2-\alpha}{\alpha-1}}, \\ L_1(R) &\geq \left(\frac{\alpha-1}{\alpha}\right)^{\frac{1}{\alpha-1}}\left(\frac{R}{\tau}\right)^{\frac{1}{\alpha-1}}. \end{aligned} \tag{29}$$

Specification of the oracle: We first specify the oracle O^*, needed to define P_{f_0} and P_{f_1}, and which simply adds a standard normal noise to the gradient values. Let $f \in \mathcal{F}_\alpha^{\text{uni}}$. Then

$$O^*(x) = (f'(x) + Z), \tag{30}$$

where Z is a zero-mean normal noise with variance σ^2. Therefore, $O^* \in \mathsf{O}_\sigma^{\tilde{L}}$ as $f'(x, Z) := f'(x) + Z$ is unbiased, $\mathbb{E}[|f'(x, Z) - \mathbb{E}[f'(x, Z)]|^2] = \sigma^2$, and this oracle is $\tilde{L}$-average smooth with $\tilde{L} = L_2(R)$,

$$\begin{aligned} \mathbb{E}[|f'(x, Z) - f'(y, Z)|^2] &= |f'(x) - f'(y)|^2 \\ &\leq L_2(R)^2|x-y|^2. \end{aligned}$$

Specification of P_{f_0} and P_{f_1}: For $i \in \{0, 1\}$, $P_{f_i}^m$ denotes the distribution of $\{X_t, f_i'(X_t, Z_t)\}_{t=1}^m$ where X_t denotes the output of stochastic first-order algorithm A at iteration t.

[14] $\delta(\cdot,\cdot)$ is a semi-distance if it satisfies the symmetry property and the triangle inequality but not the separation property (i.e., for every $f, g \in \mathcal{F}$, $\delta(f, g) = 0 \Leftrightarrow f = g$).

Lemma 6. *Let $P_{f_i}^m$ be the distribution of $\{X_t, f_i'(X_t, Z_t)\}_{t=1}^m$ for $i = \{0, 1\}$ and f_0, f_1 are defined in* (26) *and* (27)*, respectively. Then for $0 < \rho \le 1/2$, we have*

$$KL(P_{f_0}^m \| P_{f_1}^m) = \mathcal{O}\left(\frac{C^2 m}{\sigma^2}\left(\frac{\alpha}{\alpha-1}\right)^2 \rho^{\frac{2}{\alpha-1}}\right).$$

The proof of Lemma 6 is given in Appendix F.3. Since $\delta < 1/2$, choose any constant

$$\gamma_\delta \in \big(0,\ 2(1-2\delta)^2\big).$$

Lemma 6 then shows that by taking

$$\rho = \Theta\left(\gamma_\delta^{\frac{\alpha-1}{2}} m^{-\frac{\alpha-1}{2}} \left(\frac{\sigma}{C}\right)^{\alpha-1} \left(\frac{\alpha-1}{\alpha}\right)^{\alpha-1}\right),$$

we ensure $KL(P_{f_0}^m \| P_{f_1}^m) \le \gamma_\delta$. Hence, given $\delta(f_0, f_1) \ge \rho$, Lemma 5 implies that

$$\sup_{F \in \mathcal{F}_\alpha^{\text{uni}}} \inf_{\mathsf{A} \in \mathcal{A}_m(\mathcal{U}_0^F, \mathcal{U}^F, \delta)} \mathbb{P}\left[\delta(\hat{F}_m, F) > \frac{\rho}{2}\right] \ge c_\delta, \tag{31}$$

where

$$c_\delta := \max\left\{\frac{e^{-\gamma_\delta}}{4}, \frac{1-\sqrt{\gamma_\delta/2}}{2}\right\} > \delta.$$

For every admissible pair (F, A), the stay-in-neighborhood property gives

$$\mathbb{P}\big(\mathcal{E}_m(\mathcal{U}^F)\big) \ge 1 - \delta.$$

Therefore,

$$\begin{aligned}
\mathbb{P}\Big[\delta(\hat{F}_m, F) > \frac{\rho}{2} \,\Big|\, \mathcal{E}_m(\mathcal{U}^F)\Big] &= \frac{\mathbb{P}\Big(\Big\{\delta(\hat{F}_m, F) > \frac{\rho}{2}\Big\} \cap \mathcal{E}_m(\mathcal{U}^F)\Big)}{\mathbb{P}(\mathcal{E}_m(\mathcal{U}^F))} \\
&\ge \frac{\mathbb{P}\Big[\delta(\hat{F}_m, F) > \frac{\rho}{2}\Big] - \mathbb{P}\big[\mathcal{E}_m(\mathcal{U}^F)^c\big]}{\mathbb{P}(\mathcal{E}_m(\mathcal{U}^F))} \\
&\ge \frac{\mathbb{P}\Big[\delta(\hat{F}_m, F) > \frac{\rho}{2}\Big] - \delta}{1-\delta}.
\end{aligned} \tag{32}$$

Taking first $\inf_{\mathsf{A}}$, then $\sup_F$, and using (31), we obtain

$$\sup_{F \in \mathcal{F}_\alpha^{\text{uni}}} \inf_{\mathsf{A} \in \mathcal{A}_m(\mathcal{U}_0^F, \mathcal{U}^F, \delta)} \mathbb{P}\left[\delta(\hat{F}_m, F) > \frac{\rho}{2} \,\middle|\, \mathcal{E}_m(\mathcal{U}^F)\right] \ge \kappa_\delta, \tag{33}$$

where

$$\kappa_\delta := \frac{c_\delta - \delta}{1-\delta} > 0.$$

We return to (24), and finish the proof by plugging (33) in (24) to get

$$\begin{aligned}
&\sup_{F \in \mathcal{F}_\alpha^{\text{uni}}} \inf_{\mathsf{A} \in \mathcal{A}_m(\mathcal{U}_0^F, \mathcal{U}^F, \delta)} \mathbb{E}[F(x_m) \mid \mathcal{E}_m(\mathcal{U}^F)] - F(x_F^*) \\
&\quad \ge \lambda\left(\frac{\rho}{2} \cdot \sup_{F \in \mathcal{F}_\alpha^{\text{uni}}} \inf_{\mathsf{A} \in \mathcal{A}_m(\mathcal{U}_0^F, \mathcal{U}^F, \delta)} \mathbb{P}\left[\delta(\hat{F}_m, F) > \frac{\rho}{2} \,\middle|\, \mathcal{E}_m(\mathcal{U}^F)\right]\right)^{\frac{\alpha}{\alpha-1}} \\
&\quad \overset{(c)}{\ge} \lambda\left[\Omega\left(\frac{1}{m^{\frac{\alpha-1}{2}}}\left(\frac{\sigma}{C}\right)^{\alpha-1}\left(\frac{\alpha-1}{\alpha}\right)^{\alpha-1}\right)\right]^{\frac{\alpha}{\alpha-1}}
\end{aligned}$$

$$= \Omega\left(\frac{\lambda\sigma^\alpha\left(\frac{\alpha-1}{\alpha}\right)^\alpha}{C^\alpha m^{\frac{\alpha}{2}}}\right) \overset{(d)}{=} \Omega\left(\frac{\tau\sigma^\alpha}{m^{\frac{\alpha}{2}}}\right) \tag{34}$$

where (c) follows from (33) and the above choice of ρ, and (d) results from the choices of $\lambda = ((\alpha-1)/\alpha)^{\alpha/(\alpha-1)}\tau^{-1/(\alpha-1)}$ and C in Equation (28). From (34), $\mathsf{m}_\varepsilon(\mathcal{F}^{\mathcal{X}}_{\alpha,\tau,L}, O^*) = \Omega\left(\tau^{2/\alpha}\sigma^2/\varepsilon^{2/\alpha}\right)$, which concludes the proof. □

F.3 Proof of Lemma 6

Let X_i, Y_i denote the updating point and the gradient sample observed at iteration i of the stochastic first-order algorithm A, respectively. Note that

$$P^m_{f_i}(Y_t|X_t = x) = \mathbb{P}(f_i'(X_t, Z_t)|X_t = x) = \mathcal{N}(f_i'(x), \sigma^2). \tag{35}$$

Let us define $X^m := \{X_i\}_{i=1}^m, Y^m := \{Y_i\}_{i=1}^m$. Then, we have:

$$\begin{aligned}
KL(P^m_{f_1}\|P^m_{f_0}) &= \mathbb{E}_{P^m_{f_1}}\left[\log\frac{P^m_{f_1}(X^m, Y^m)}{P^m_{f_0}(X^m, Y^m)}\right] \\
&= \mathbb{E}_{P^m_{f_1}}\left[\log\frac{\prod_{t=1}^m P^m_{f_1}(Y_t|X_t)\cdot P(X_t|X^{t-1}, Y^{t-1})}{\prod_{t=1}^m P^m_{f_0}(Y_t|X_t)\cdot P(X_t|X^{t-1}, Y^{t-1})}\right] && (36)\\
&= \mathbb{E}_{P^m_{f_1}}\left[\log\frac{\prod_{t=1}^m P^m_{f_1}(Y_t|X_t)}{\prod_{t=1}^m P^m_{f_0}(Y_t|X_t)}\right] \\
&= \sum_{t=1}^m \mathbb{E}_{X_t\sim P^m_{f_1}}\left[\mathbb{E}_{P^m_{f_1}}\left[\log\frac{P^m_{f_1}(Y_t|X_t)}{P^m_{f_0}(Y_t|X_t)}\middle| X_t\right]\right] \\
&\le m\cdot\max_{x\in\mathcal{N}_{f_1}(\rho)}\mathbb{E}_{P^m_{f_1}}\left[\log\frac{P^m_{f_1}(Y_t|X_t)}{P^m_{f_0}(Y_t|X_t)}\middle| X_t = x\right] && (37)\\
&= \frac{m}{2\sigma^2}\left(\max_{x\in\mathcal{N}_{f_1}(\rho)}|f_0'(x) - f_1'(x)|^2\right) && (38)\\
&= \frac{C^2 m}{2\sigma^2}\left(\frac{\alpha}{\alpha-1}\right)^2\cdot\left[\max_{x\in[0,2\rho]}\left(2^{\frac{1}{\alpha-1}}|x-\rho|^{\frac{1}{\alpha-1}}\mathrm{sgn}(x-\rho) - x^{\frac{1}{\alpha-1}}\right)^2\right] && (39)\\
&= \mathcal{O}\left(\frac{C^2 m}{\sigma^2}\left(\frac{\alpha}{\alpha-1}\right)^2\rho^{\frac{2}{\alpha-1}}\right), && (40)
\end{aligned}$$

where (36) comes from the fact that given (X^{t-1}, Y^{t-1}), stochastic first-order algorithm's updated point X_t is independent of the choice of the objective function. Equation (37) follows by $\mathrm{supp}(P_{X_t}) = \mathcal{U}_0^{f_1} \subseteq \mathcal{N}_{f_1}(\rho)$ when $X^m \sim P^m_{f_1}$, Equation (38) follows from (35), and (39) from the construction of f_0 (refer to (26)) and of f_1 (refer to (27)). In (40), we use the fact that $x = 0$ achieves the maximum value in (39).

Lemma 7. *Functions f_0 and f_1, defined in (26) and (27), are elements of $\mathcal{F}^{\alpha,uni}_{L_1,L_2}$ with $L_2 \ge C\alpha(\alpha-1)^{-2}R^{(2-\alpha)/(\alpha-1)}$ and $\tau \ge C^{1-\alpha}((\alpha-1)/\alpha)^\alpha$.*

Proof of Lemma 7. Recall

$$f_0(x) = C|x|^{\frac{\alpha}{\alpha-1}}, \tag{41}$$

$$f_1(x) = \begin{cases} 2^{\frac{1}{\alpha-1}}C(|x-\rho|^{\frac{\alpha}{\alpha-1}} + |\rho|^{\frac{\alpha}{\alpha-1}}) & 0 \le x \le 2\rho \\ f_0(x) & 2\rho \le x \\ -\frac{\alpha}{\alpha-1}2^{\frac{1}{\alpha-1}}C\rho^{\frac{1}{\alpha-1}}x + 2^{\frac{\alpha}{\alpha-1}}C\rho^{\frac{\alpha}{\alpha-1}} & x \le 0 \end{cases}. \tag{42}$$

Note that each of f_0 and f_1 has a unique minimizer. Specifically, $x^*_{f_0} =_x f_0(x) = 0$ and $x^*_{f_1} =_x f_1(x) = \rho$.

L_1-local smoothness of f_0 and f_1: we have

$$|f_0'(x)| = C\frac{\alpha}{\alpha-1}|x|^{\frac{1}{\alpha-1}}, \tag{43}$$

and

$$|f_1''(x)| = \begin{cases} 2^{\frac{1}{\alpha-1}}C\frac{\alpha}{\alpha-1}|x-\rho|^{\frac{1}{\alpha-1}} & 0 < x < 2\rho \\ |f_0'(x)| & 2\rho \le x \\ \frac{\alpha}{\alpha-1}2^{\frac{1}{\alpha-1}}C\rho^{\frac{1}{\alpha-1}} & x \le 0 \end{cases}. \tag{44}$$

Then $|f_0'(x)| \le \frac{\alpha C}{\alpha-1}R^{\frac{1}{\alpha-1}}$ for all $x \in \mathcal{N}_{f_0}(0) = [-R, R]$ and $|f_1'(x)| \le |f_0'(x)| \le L_1(R) = \frac{\alpha C}{\alpha-1}R^{\frac{1}{\alpha-1}}$ for all $x \in \mathcal{N}_{f_1}(\rho) = [0, R]$.

L_2-local smoothness of f_0 and f_1:

$$|f_0''(x)| = C\frac{\alpha}{(\alpha-1)^2}|x|^{\frac{2-\alpha}{\alpha-1}}. \tag{45}$$

Let $L_2(R) = CR^{(2-\alpha)/(\alpha-1)}\alpha/(\alpha-1)^2$. Hence f_0 is L_2-locally smooth in $[-R, R]$.

$$|f_1''(x)| = \begin{cases} 2^{\frac{1}{\alpha-1}}C\frac{\alpha}{(\alpha-1)^2}|x-\rho|^{\frac{2-\alpha}{\alpha-1}} & 0 < x < 2\rho \\ |f_0''(x)| & 2\rho < x \\ 0 & x \le 0 \end{cases}. \tag{46}$$

Hence f_1 is smooth with constant $L_2(R)$. Then both f_0 and f_1 are $L_2(R)$-locally smooth in $[0, R]$.

Local α-PŁ: For f_0 with the neighborhood $[-R, R]$ around $x^*_{f_0} = 0$, we have

$$\begin{aligned} f_0(x) - \min_{y\in[-R,R]} f_0(y) &= C|x|^{\frac{\alpha}{\alpha-1}} \\ &\le \frac{(\alpha-1)^\alpha}{\alpha^\alpha C^\alpha} \cdot \left| C\frac{\alpha}{\alpha-1}|x|^{\frac{1}{\alpha-1}} \right|^\alpha \\ &= \frac{(\alpha-1)^\alpha}{\alpha^\alpha C^\alpha} \cdot |f_0'(x)|^\alpha. \end{aligned} \tag{47}$$

For f_1 with the neighborhood $[0, R]$ around $x^*_{f_1} = \rho$, we have

$$\begin{aligned} f_1(x) - \min_{y\in[0,R]} f_1(y) &= 2^{\frac{1}{\alpha-1}}C|x-\rho|^{\frac{\alpha}{\alpha-1}} \\ &\le \frac{(\alpha-1)^\alpha}{C^\alpha\alpha^\alpha 2^{\frac{\alpha}{\alpha-1}}} \cdot \left| 2^{\frac{1}{\alpha-1}}\frac{\alpha}{\alpha-1} \cdot C|x-\rho|^{\frac{1}{\alpha-1}} \right|^\alpha \\ &= \frac{(\alpha-1)^\alpha}{C^\alpha\alpha^\alpha 2^{\frac{\alpha}{\alpha-1}}} \cdot |f_1'(x)|^\alpha. \end{aligned} \tag{48}$$

□

G PROOFS OF THEOREM 3.2

Throughout the proof of Theorem 3.2, we use the following parameterization of the time–varying step size and batch size as

$$\eta_k = \frac{\eta_0}{(k+1)^\eta}, \qquad n_g^k = (k+1)^\beta, \quad k \ge 0, \tag{49}$$

where $\eta > 0$ and $\beta > 0$ are some exponent constants and $\eta_0 > 0$ is the initial stepsize.

G.1 Variance of the gradient estimator of SARAH

Let us define $\mathbf{g}_{\mathcal{J}_t}(x_t) := \frac{1}{|\mathcal{J}_t|}\sum_{j\in\mathcal{J}_t}\mathbf{g}(x_t,\xi_j)$. By adapting the variance reduction method in [Zhou and Gu, 2020] (see Algorithm 1), we have the following bound on the errors of the estimated gradient $\mathbf{v}_t$:

Lemma 8. *Let n_g^t denote the number of stochastic gradient samples used at iteration t, and fix $\varepsilon > 0$. Suppose that at every* checkpoint *$t = kS$ with $k \in \mathbb{N}$,*

$$n_g^t \;\geq\; \frac{2\sigma^2}{\varepsilon},$$

and at every non-checkpoint *iteration,*

$$n_g^t \;\geq\; \frac{4\tilde{L}^2 S\,\|x_t - x_{t-1}\|^2}{\varepsilon}. \tag{50}$$

Then, under the $\tilde{L}$-average smoothness assumption (5)*, the variance-reduced gradient estimator $\mathbf{v}_t$ satisfies*

$$\mathbb{E}\big[\|\nabla F(x_t) - \mathbf{v}_t\|^2\big] \;\leq\; \varepsilon \tag{51}$$

for all iterations t.

Proof. First the gradient estimator $\mathbf{v}_t$ is such that $\mathbf{v}_t - \nabla F(x_t) = \sum_{k=\lfloor t/S\rfloor S}^{t}\mathbf{u}_k$ where

$$\begin{aligned}\mathbf{u}_k &= \mathbf{1}(k = \lfloor t/S\rfloor S)\cdot[\mathbf{g}_{\mathcal{J}_k}(x_k) - \nabla F(x_k)]\\ &+ \mathbf{1}(k > \lfloor t/S\rfloor S)\cdot[\mathbf{g}_{\mathcal{J}_k}(x_k) - \mathbf{g}_{\mathcal{J}_k}(x_{k-1}) - \nabla F(x_k) + \nabla F(x_{k-1})].\end{aligned}$$

We know that $\mathbb{E}[\mathbf{u}_k|\mathcal{F}_k] = 0$ where $\mathcal{F}_k = \sigma(x_0,\ldots,x_k,\mathbf{u}_0,\cdots,\mathbf{u}_{k-1})$. Conditioned on $\mathcal{F}_k$ and for $k > \lfloor t/S\rfloor S$, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{u}_k\|^2|\mathcal{F}_k] &= \mathbb{E}\Big\|\frac{1}{n_g^k}\sum_{i=1}^{n_g^k}\mathbf{g}(x_k,\xi_i) - \mathbf{g}(x_{k-1},\xi_i) - \nabla F(x_k) + \nabla F(x_{k-1})\Big\|^2\\ &\overset{(a)}{=} \frac{1}{n_g^k}\mathbb{E}\Big\|\mathbf{g}(x_k,\xi_1) - \mathbf{g}(x_{k-1},\xi_1) - \nabla F(x_k) + \nabla F(x_{k-1})\Big\|^2\\ &\overset{(b)}{\leq} \frac{2}{n_g^k}\mathbb{E}\|\mathbf{g}(x_k,\xi_1) - \mathbf{g}(x_{k-1},\xi_1)\|^2 + \frac{2}{n_g^k}\mathbb{E}\|\nabla F(x_k) - \nabla F(x_{k-1})\|^2\\ &\overset{(c)}{\leq} \frac{4\tilde{L}^2}{n_g^k}\|x_k - x_{k-1}\|^2\end{aligned} \tag{52}$$

where (a) follows because $[\mathbf{g}(x_k,\xi_i) - \mathbf{g}(x_{k-1},\xi_i) - \nabla F(x_k) + \nabla F(x_{k-1})]$ are i.i.d conditioned on $\mathcal{F}_k$ for $1 \leq i \leq n_g^k$, (b) from $(a+b)^2 \leq 2a^2 + 2b^2$ and (c) from the $\tilde{L}$-average smooothness Equation (5). For $k = \lfloor t/S\rfloor S$, $\mathbb{E}[\|\mathbf{u}_k\|^2] \leq \frac{\sigma^2}{n_g^k}$.

$$\begin{aligned}\mathbb{E}[\|\nabla F(x_t) - \mathbf{v}_t\|^2] &= \mathbb{E}\left\|\sum_{k=\lfloor t/S\rfloor S}^{t}\mathbf{u}_k\right\|^2 \leq \sum_{k=\lfloor t/S\rfloor S}^{t}\mathbb{E}[\mathbb{E}[\|\mathbf{u}_k\|^2|\mathcal{F}_k]]\\ &\leq \frac{\sigma^2}{n_g^{\lfloor t/S\rfloor S}} + \sum_{k=\lfloor t/S\rfloor S+1}^{t}4\tilde{L}^2\mathbb{E}\left[\frac{\|x_k - x_{k-1}\|^2}{n_g^k}\right]\end{aligned} \tag{53}$$

where the first inequality comes from the fact that $\mathbb{E}[\mathbf{u}_k^T\mathbf{u}_{k'}] = \mathbb{E}[\mathbf{u}_{k'}^T\mathbb{E}[\mathbf{u}_k|\mathcal{F}_k]] = 0$ for $k > k'$.
If we take $n_g^k \geq \frac{2\sigma^2}{\varepsilon}$ samples at checkpoints ($k = \lfloor t/S\rfloor S$) and and at the other iterations, we take $n_g^k \geq \frac{4\tilde{L}^2 S\|x_k - x_{k-1}\|^2}{\varepsilon}$ samples, we have

$$\mathbb{E}[\|\nabla F(x_t) - \mathbf{v}_t\|^2] \leq \varepsilon.$$

□

Now by using (53) in the proof of the previous lemma and the batch-size in (49), we have

$$\mathbb{E}[\|\mathbf{v}_t - \nabla F(x_t)\|^2] \leq \frac{\sigma^2}{(\lfloor t/S \rfloor S+1)^\beta} + \frac{4\tilde{L}^2 \cdot [t - \lfloor t/S \rfloor S]}{(\lfloor t/S \rfloor S+1)^\beta} \cdot \max_{\lfloor t/S \rfloor S+1 \leq k \leq t} \mathbb{E}[\|x_k - x_{k-1}\|^2]. \tag{54}$$

We later use this bound to show to prove high probability stability of updates of SARAH near a connected component of local minima.

G.2 Proof of stability of SARAH

Fix any $\delta \in (0,1)$ and let $\mathcal{M}$ be an isolated, compact, connected set of local minimizers with common value $l := F(x^\star)$ for all $x^\star \in \mathcal{M}$. Define the neighborhoods

$$\begin{aligned} \mathcal{U}_0 &:= \Big\{x : \ \operatorname{dist}(x, \mathcal{M}) < \tfrac{R}{2}, \ \ F(x) - l \leq \tfrac{s}{2}\Big\}, \\ \mathcal{U} &:= \Big\{x : \ \operatorname{dist}(x, \mathcal{M}) < \tfrac{R}{2}\Big\}, \end{aligned} \tag{55}$$

and the event $\mathcal{E}_n(\mathcal{U}) := \{x_k \in \mathcal{U} \text{ for all } k \leq n\}$ and abbreviate it as $\mathcal{E}_n$. We claim that $\mathbb{P}(\mathcal{E}_n) \geq 1 - \delta$ whenever $x_0 \in \mathcal{U}_0$.
Assume that $x_k \in \mathcal{U}$ and choose the step-size (and batching policy) so that the one–step move is small with high probability; i.e. $\mathbb{P}(C_n) \geq 1 - \delta/2$ where

$$C_n := \Big\{ \ \|x_{k+1} - x_k\| \leq \tfrac{R}{4} \text{ for all } k \leq n \ \Big\}.$$

By the triangle inequality, on C_n we have $x_{k+1} \in \mathbb{B}_2(\mathcal{M};\, 3R/4)$. We define the "forbidden region"

$$\mathcal{R} := \mathbb{B}_2(\mathcal{M};\, 3R/4) \setminus \mathbb{B}_2(\mathcal{M};\, R/2),$$

and then as $\mathcal{M}$ is a compact isolated set, the next lemma shows that $f(x') - l > 0$ for all $x \in \mathcal{R}$.

Lemma 9. *Let*

$$s_0 := \inf\{\, F(x) - l : \ x \in \mathcal{R} \,\} > 0.$$

Then we have that $s_0 > 0$.

*Proof.*Suppose, by contradiction, that $s_0 = 0$. Then there exists a sequence $\{x^k\} \subset \mathcal{R}$ with $F(x^k) \to l$. Because $\mathcal{R} \subset \mathbb{B}_2(\mathcal{M};\, 3R/4)$ and $\mathcal{M}$ is compact, the sequence $\{x^k\}$ is bounded:

$$\|x^k\| \leq \tfrac{3R}{4} + \sup_{y \in \mathcal{M}} \|y\| < \infty.$$

Hence there is a subsequence $x^{k_i} \to x$. By continuity, $F(x) = l$. Moreover, $\operatorname{dist}(x^k, \mathcal{M}) \geq R/2$ for all k, so $\operatorname{dist}(x, \mathcal{M}) \geq R/2$. On the other hand, $x \in \overline{\mathbb{B}_2(\mathcal{M};\, 3R/4)} \subset \mathbb{B}_2(\mathcal{M};\, R)$, and as $\mathcal{M}$ is an isolated set, we have $F(y) > l$ for all $y \in \mathbb{B}_2(\mathcal{M};\, R) \setminus \mathcal{M}$. Thus $F(x) = l$ forces $x \in \mathcal{M}$, which contradicts $\operatorname{dist}(x, \mathcal{M}) \geq R/2$. Therefore $s_0 > 0$. □

Next, we invoke the following lemma and verify its assumptions later to guarantee that the SARAH iterates remain in $\mathcal{U}$.

Lemma 10. *Suppose* $x, y \in \mathbb{R}^d$ *such that (i)* $\operatorname{dist}(x, \mathcal{M}) < \frac{R}{2}$, *(ii)* $\|x - y\| \leq \frac{R}{4}$, *and (iii)* $F(y) - l < s_0$. *Then* $\operatorname{dist}(y, \mathcal{M}) < \frac{R}{2}$.

*Proof.*Pick $x^* \in \mathcal{M}$ with $\|x - x^*\| < \frac{R}{2}$. By the triangle inequality,

$$\|y - x^*\| \leq \|y - x\| + \|x - x^*\| \leq \tfrac{R}{4} + \tfrac{R}{2} = \tfrac{3R}{4},$$

so $\operatorname{dist}(y, \mathcal{M}) \leq \frac{3R}{4}$. Suppose by a contradiction that $\operatorname{dist}(y, \mathcal{M}) \geq \frac{R}{2}$, then

$$y \in \mathcal{R} = \mathbb{B}_2(\mathcal{M};\, 3R/4) \setminus \mathbb{B}_2(\mathcal{M};\, R/2).$$

By the definition of s_0, this implies $f(y) - l \geq s_0$, contradicting assumption (ii). Hence $\operatorname{dist}(y, \mathcal{M}) < \frac{R}{2}$, as claimed. □

The plan is to exclude $x_{k+1} \in \mathcal{R}$ by showing that, with high probability, $F(x_{k+1}) - l < s_0$ once $x_0 \in \mathcal{U}_0$. Let $D_n := F(x_n) - l$ and define

$$\mathsf{A}_k := -\langle \nabla F(x_k),\, \mathbf{v}_k - \nabla F(x_k)\rangle, \qquad \mathsf{B}_k := \|\mathbf{v}_k\|^2,$$

where $\mathbf{v}_k$ is the SARAH gradient estimator in Algorithm 1.

Lemma 11. *Under L_2-local smoothness and local α-PŁ and for $\eta_0 \le \tau/2$, on $\mathcal{E}_n$, we have*

$$\textit{(i)}\ \alpha = 2\textit{:}\quad D_{n+1}\mathbf{1}_{\mathcal{E}_n} \;\le\; D_0 \cdot \prod_{k=0}^{n}(1-\tau^{-1}\eta_k) + \sum_{k=0}^{n}\Big(\prod_{j=k+1}^{n}(1-\tau^{-1}\eta_j)\Big)\,\eta_k\,\mathsf{A}_k\,\mathbf{1}_{\mathcal{E}_k} + \frac{L_2}{2}\sum_{k=0}^{n}\eta_k^2\,\mathsf{B}_k\,\mathbf{1}_{\mathcal{E}_k},$$

$$\textit{(ii)}\ \alpha \in [1,2)\textit{:}\quad D_{n+1}\mathbf{1}_{\mathcal{E}_n} \;\le\; D_0 \cdot \prod_{k=0}^{n}(1-\tau^{-2/\alpha}\eta_k^q) + \tilde{c}\sum_{k=0}^{n}\eta_k^{\frac{2q-2}{2-\alpha}} + \sum_{k=0}^{n}\Big(\prod_{j=k+1}^{n}(1-\tau^{-2/\alpha}\eta_j^q)\Big)\,\eta_k\,\mathsf{A}_k\,\mathbf{1}_{\mathcal{E}_k} + \frac{L_2}{2}\sum_{k=0}^{n}\eta_k^2\,\mathsf{B}_k\,\mathbf{1}_{\mathcal{E}_k},$$

for any $1 \le q$.

*Proof.*Since F is L_2-smooth and the update rule is $x_{n+1} = x_n - \eta_n \mathbf{v}_n$, we have

$$\begin{aligned} D_{n+1} &\le D_n - \eta_n\langle \nabla F(x_n), \mathbf{v}_n\rangle + \frac{L_2\eta_n^2}{2}\|\mathbf{v}_n\|^2 \\ &= D_n - \eta_n\|\nabla F(x_n)\|^2 - \eta_n\langle \nabla F(x_n), \mathbf{v}_n - \nabla F(x_n)\rangle + \frac{L_2\eta_n^2}{2}\|\mathbf{v}_n\|^2 \\ &= D_n - \eta_n\|\nabla F(x_n)\|^2 + \eta_n\mathsf{A}_n + \frac{L_2\eta_n^2}{2}\mathsf{B}_k. \end{aligned} \tag{56}$$

Multiplying both hand-sides of (56) by $\mathbf{1}_{\mathcal{E}_n}$ and noting that $\mathbf{1}_{\mathcal{E}_{n+1}} \le \mathbf{1}_{\mathcal{E}_n}$, we obtain the recursive bounds below. We distinguish the two cases.

Case $\alpha = 2$. On $\mathcal{E}_n$ the unified gradient domination gives $\|\nabla F(x_n)\|^2 \ge \tau^{-1}\,(F(x_n) - l) = \tau^{-1}D_n$, hence

$$D_{n+1}\mathbf{1}_{\mathcal{E}_n} \;\le\; (1-\eta_n\tau^{-1})\,D_n\mathbf{1}_{\mathcal{E}_n} \;+\; \eta_n\mathsf{A}_n\mathbf{1}_{\mathcal{E}_n} \;+\; \frac{L\eta_n^2}{2}\,\mathsf{B}_n\mathbf{1}_{\mathcal{E}_n}. \tag{57}$$

Iterating this affine recursion and using $\mathbf{1}_{\mathcal{E}_n} \le \mathbf{1}_{\mathcal{E}_k}$ for $k \le n$ gives the unfolded product–sum representation stated in the lemma.

Case $\alpha \in [1,2)$. On $\mathcal{E}_n$, the local PŁ property yields that $\|\nabla F(x_n)\| \ge \tau^{-\frac{1}{\alpha}}\,D_n^{\frac{1}{\alpha}}$ and therefore (56) becomes

$$\begin{aligned} D_{n+1} &\le D_n - \eta_n\|\nabla F(x_n)\|^2 + \eta_n\mathsf{A}_n + \frac{L\eta_n^2}{2}\mathsf{B}_k \\ &\le D_n - \eta_n\tau^{-\frac{2}{\alpha}}D_n^{\frac{2}{\alpha}} + \eta_n\mathsf{A}_n + \frac{L\eta_n^2}{2}\mathsf{B}_k \\ &= (1-\eta_n^q\tau^{-\frac{2}{\alpha}})D_n + \eta_n\tau^{-\frac{2}{\alpha}}(\eta_n^{q-1}D_n - D_n^{\frac{2}{\alpha}}) + \eta_n\mathsf{A}_n + \frac{L\eta_n^2}{2}\mathsf{B}_k \\ &\overset{(a)}{\le} (1-\eta_n^q\tau^{-\frac{2}{\alpha}})D_n + \tilde{c}\eta_n^{\frac{2q-2}{2-\alpha}} + \eta_n\mathsf{A}_n + \frac{L\eta_n^2}{2}\mathsf{B}_k. \end{aligned} \tag{58}$$

In step (a), to linearize $D_n^{2/\alpha}$ and bound the term $\eta_n^q\tau^{-2/\alpha}D_n - \eta_n\tau^{-2/\alpha}D_n^{2/\alpha}$, we use

$$\max_{x\in\mathrm{R}}\{\,c\,x - x^{2/\alpha}\,\} = \left(\tfrac{\alpha}{2}\right)^{\frac{\alpha}{2-\alpha}}\,\frac{2-\alpha}{\alpha}\,c^{\frac{2}{2-\alpha}},$$

applied to $x = D_n$ and $c = \eta_n^{q-1}$. Hence, for

$$\tilde{c} \;:=\; \left(\tfrac{\alpha}{2}\right)^{\frac{\alpha}{2-\alpha}}\left(1-\tfrac{\alpha}{2}\right)\tau^{-2/\alpha},$$

we obtain

$$\eta_n^q \tau^{-2/\alpha} D_n - \eta_n \tau^{-2/\alpha} D_n^{2/\alpha} \;\le\; \tilde{c}\, \eta_n^{\frac{2q-2}{2-\alpha}}.$$

Multiplying (58) by $\mathbf{1}_{\mathcal{E}_n}$ and iterating this affine recursion, and next using $\mathbf{1}_{\mathcal{E}_n} \le \mathbf{1}_{\mathcal{E}_k}$ for $k \le n$, we obtain the unfolded product–sum representation stated in the lemma. □

We denote the aggregated noises with the natural weights:

$$\tilde{\mathsf{A}}_k := \sum_{j=0}^{k} \Big(\prod_{i=j+1}^{k} \theta_i \Big) \eta_j\, \mathsf{A}_j\, \mathbf{1}_{\mathcal{E}_j},\; \tilde{\mathsf{B}}_k := \frac{L}{2} \sum_{j=0}^{k} \eta_j^2\, \mathsf{B}_j\, \mathbf{1}_{\mathcal{E}_j}, \tag{59}$$

where $\theta_i = 1 - \tau^{-1}\eta_i$ for $\alpha = 2$ and $\theta_i = 1 - \tau^{-2/\alpha}\eta_i^q$ for $\alpha \in [1,2)$. Let us define for $n \ge 1$, $\mathsf{R}_n := \tilde{\mathsf{A}}_n^2 + \tilde{\mathsf{B}}_n$ and the "small-error" event

$$E_n := \big\{\; \mathsf{R}_k < s \text{ for all } k \le n \;\big\}.$$

Lemma 12. *Let us define*

$$\widehat{E}_n := E_n \cap C_n.$$

Assume $x_0 \in \mathcal{U}_0$ almost surely. Pick s such that $2s + \sqrt{s} < s_0$. If $\alpha = 2$, let $\eta_0 > 0$ be an arbitrary constant; if $\alpha \in [1,2)$ choose $2 - \alpha/2 \le q$ and η_0 small enough so that

$$\sum_{k=1}^{\infty} \eta_k^{\frac{2q-2}{2-\alpha}} = \eta_0^{\frac{2q-2}{2-\alpha}} \sum_{n=1}^{\infty} \frac{1}{(n+1)^{\frac{2q-2}{2-\alpha}}} < \frac{s}{2\tilde{c}}, \tag{60}$$

with $\tilde{c} = (2/\alpha)^{-\frac{\alpha}{2-\alpha}}\big(1 - \frac{\alpha}{2}\big)\tau^{-2/\alpha}$. Then the following holds for all $n \in \mathbb{N}$:

$$\widehat{E}_n \subset \mathcal{E}_{n+1}.$$

*Proof.*By definition, $\{\mathsf{R}_k < s\}_{k \le n+1}$ implies $\{\mathsf{R}_k < s\}_{k \le n}$, so $E_{n+1} \subset E_n$. Similarly, $\{x_k \in \mathcal{U}\}_{k \le n+1}$ implies $\{x_k \in \mathcal{U}\}_{k \le n}$, hence $\mathcal{E}_{n+1} \subset \mathcal{E}_n$. Since $C_{n+1} \subset C_n$, we also have $\widehat{E}_{n+1} = E_{n+1} \cap C_{n+1} \subset E_n \cap C_n = \widehat{E}_n$.
We prove $\widehat{E}_n \subset \mathcal{E}_{n+1}$ by induction. Because $x_0 \in \mathcal{U}_0 \subset \mathcal{U}$ a.s., the claim holds for $n = 0$. Assume $\widehat{E}_{n-1} \subset \mathcal{E}_n$ and pick $\omega \in \widehat{E}_n$. Then $\omega \in \mathcal{E}_n$ (induction hypothesis) and $\omega \in C_n$ (by definition). Applying Lemma 10 with $x = x_n(\omega)$ and $y = x_{n+1}(\omega)$: on $\mathcal{E}_n(\omega)$,

$$\inf_{x^* \in \mathcal{M}} \|x_n(\omega) - x^*\| < \tfrac{R}{2} \quad \text{and} \quad \|x_{n+1}(\omega) - x_n(\omega)\| \le \tfrac{R}{4},$$

so conditions (i) and (iii) of Lemma 10 are satisfied. It remains to ensure that $f(x_{n+1}(\omega)) - l < s$, where $s > 0$ is as in Lemma 9. We distinguish two cases $\alpha = 2$ and $\alpha \in [1,2)$ in the rest of the proof:
Case $\alpha = 2$. From the recursion of Lemma 11 (valid on $\mathcal{E}_n$),

$$D_{n+1}\mathbf{1}_{\mathcal{E}_n} \;\le\; D_0 \prod_{k=0}^{n}(1 - \tau^{-1}\eta_k) + \sum_{k=0}^{n}\Big(\prod_{j=k+1}^{n} (1 - \tau^{-1}\eta_j)\Big)\, \eta_k\, \mathsf{A}_k\, \mathbf{1}_{\mathcal{E}_k} + \frac{L}{2}\sum_{k=0}^{n} \eta_k^2\, \mathsf{B}_k\, \mathbf{1}_{\mathcal{E}_k}. \tag{61}$$

Since $D_0 \le s/2$ (because $x_0 \in \mathcal{U}_0$) and all products in (61) are less than or equal to 1,

$$D_{n+1}(\omega) \;\le\; \frac{s}{2} + |\tilde{\mathsf{A}}_n(\omega)| + \tilde{\mathsf{B}}_n(\omega) \;\le\; \frac{s}{2} + \sqrt{\mathsf{R}_n(\omega)} + \mathsf{R}_n(\omega).$$

As $\omega \in E_n$ gives $\mathsf{R}_n(\omega) < s$, we get $D_{n+1}(\omega) \le 3s/2 + \sqrt{s}$. Choosing $s > 0$ so that $2s + \sqrt{s} < s_0$ yields $f(x_{n+1}(\omega)) - l < s_0$.
Case $\alpha \in [1,2)$. Lemma 11 yields that

$$\begin{aligned} D_{n+1}\mathbf{1}_{\mathcal{E}_n} \;\le\; & D_0 \prod_{k=0}^{n}(1 - \tau^{-2/\alpha}\eta_k^q) \;+\; \tilde{c}\sum_{k=0}^{n} \eta_k^{\frac{2q-\alpha}{2-\alpha}} \\ & + \sum_{k=0}^{n}\Big(\prod_{j=k+1}^{n} (1 - \tau^{-2/\alpha}\eta_j^q)\Big)\, \eta_k\, \mathsf{A}_k\, \mathbf{1}_{\mathcal{E}_k} + \frac{L}{2}\sum_{k=0}^{n} \eta_k^2\, \mathsf{B}_k\, \mathbf{1}_{\mathcal{E}_k}. \end{aligned} \tag{62}$$

Using $D_0 \le s/2$, the step–size condition $\sum_{k=0}^{n} \eta_k^{\frac{2q-2}{2-\alpha}} \le s/(2\tilde{c})$, and $|\tilde{\mathsf{A}}_n| \le \sqrt{\mathsf{R}_n}$, $\tilde{\mathsf{B}}_n \le \mathsf{R}_n$, we obtain $D_{n+1}(\omega) \le 2s + \sqrt{s} < s_0$. Thus Lemma 11 applies and $x_{n+1}(\omega) \in \mathcal{U}$, i.e., $\omega \in \mathcal{E}_{n+1}$.

□

Lemma 13. *Under the same conditions in Lemma 12, and the following condition on η_0*

$$\eta_0 \le \delta(2s\mathsf{C}_1 \cdot L_2)^{-1} \times \Big[L_1^2 + \sigma^2 + \tilde{L}^2(R^2 + (\mathrm{diam}(\mathcal{M}))^2)\Big]^{-1} \times \left(\sum_{k=1}^{\infty} k^{-3\eta/2}(\lfloor k/S \rfloor S + 1)^{-\beta/2}\right)^{-1}, \tag{63}$$

where C_1 is a positive constant independent of parameters of the problem, we have

$$\mathbb{P}(E_n) \ge\ 1 - \frac{\delta}{2}.$$

*Proof.***(i) Recursion for the error accumulator and decomposition of the bad event.** Let $\widetilde{E}_n := E_{n-1} \setminus E_n = E_{n-1} \cap \{\mathsf{R}_n \ge s\}$ denote the event that the noise budget first crosses level s at time n, and set $\widetilde{\mathsf{R}}_n := \mathsf{R}_n\, \mathbf{1}_{E_{n-1}}$. Since the events $\{\widetilde{E}_k\}_{k\ge 0}$ are disjoint and $E_n = \bigcap_{k=0}^{n} \widetilde{E}_k^c$, it suffices to control $\sum_{k=0}^{n} \mathbb{P}(\widetilde{E}_k)$ via the recursion for $\widetilde{\mathsf{R}}_n$. Note that

$$\widetilde{\mathsf{R}}_n = \mathsf{R}_n\, \mathbf{1}_{E_{n-1}} = \widetilde{\mathsf{R}}_{n-1} - \mathsf{R}_{n-1}\mathbf{1}_{\widetilde{E}_{n-1}} + (\mathsf{R}_n - \mathsf{R}_{n-1})\, \mathbf{1}_{E_{n-1}}.$$

Taking expectations gives

$$\mathbb{E}[\widetilde{\mathsf{R}}_n] = \mathbb{E}[\widetilde{\mathsf{R}}_{n-1}] - \mathbb{E}\big[\mathsf{R}_{n-1}\mathbf{1}_{\widetilde{E}_{n-1}}\big] + \mathbb{E}\big[(\mathsf{R}_n - \mathsf{R}_{n-1})\, \mathbf{1}_{E_{n-1}}\big]. \tag{64}$$

Using $\mathsf{R}_n - \mathsf{R}_{n-1} = (\tilde{\mathsf{A}}_n^2 - \tilde{\mathsf{A}}_{n-1}^2) + (\tilde{\mathsf{B}}_n - \tilde{\mathsf{B}}_{n-1})$, $\tilde{\mathsf{A}}_n = \tilde{\mathsf{A}}_{n-1} + \eta_n(1 - \eta_n\tau^{-\frac{2}{\alpha}})\, \mathsf{A}_n\, \mathbf{1}_{\mathcal{E}_n}$, and $\tilde{\mathsf{B}}_n - \tilde{\mathsf{B}}_{n-1} = \frac{L_2}{2}\eta_n^2\|\mathbf{v}_n\|^2\, \mathbf{1}_{\mathcal{E}_n}$, we get

$$\mathsf{R}_n - \mathsf{R}_{n-1} = \eta_n^2(1 - \eta_n\tau^{-\frac{2}{\alpha}})^2\, \mathsf{A}_n^2\, \mathbf{1}_{\mathcal{E}_n} + 2\eta_n(1 - \eta_n\tau^{-\frac{2}{\alpha}})\, \mathsf{A}_n\, \mathbf{1}_{\mathcal{E}_n}\tilde{\mathsf{A}}_{n-1} + \frac{L_2}{2}\eta_n^2\|\mathbf{v}_n\|^2\, \mathbf{1}_{\mathcal{E}_n}. \tag{65}$$

By taking expectations, we use the following bounds for the terms in (65): For the first term in the right hand-side of (65), the Cauchy-Schwartz inequality for $\mathsf{A}_k := -\langle \nabla F(x_k),\, \mathbf{v}_k - \nabla F(x_k)\rangle$ yields

$$\begin{aligned}
\mathbb{E}[\mathsf{A}_n^2\, \mathbf{1}_{\mathcal{E}_n}] &\le \mathbb{E}[\|\nabla F(x_n)\|^2 \cdot \|\mathbf{v}_n - \nabla F(x_n)\|^2 \mathbf{1}_{\mathcal{E}_n}] \\
&\overset{(a)}{\le} L_1^2 \cdot \mathbb{E}[\|\mathbf{v}_n - \nabla F(x_n)\|^2 \mathbf{1}_{\mathcal{E}_n}] \\
&\overset{(b)}{\le} \frac{L_1^2\sigma^2}{(\lfloor n/S \rfloor S + 1)^{\beta}} + \frac{4L_1^2\tilde{L}^2 \cdot [n - \lfloor n/S \rfloor S]}{(\lfloor n/S \rfloor S + 1)^{\beta}} \cdot \max_{\lfloor n/S \rfloor S + 1 \le k \le n} \mathbb{E}[\|x_k - x_{k-1}\|^2] \\
&\overset{(c)}{\le} \frac{L_1^2\sigma^2}{(\lfloor n/S \rfloor S + 1)^{\beta}} + \frac{4L_1^2\tilde{L}^2 \cdot [n - \lfloor n/S \rfloor S]}{(\lfloor n/S \rfloor S + 1)^{\beta}} \cdot (2R^2 + 2(\mathrm{diam}(\mathcal{M}))^2),
\end{aligned} \tag{66}$$

where (a) follows from the local L_1-Lipschitz continuity of F, and (b) applies (54) to bound $\mathbb{E}\big[\|\mathbf{v}_n - \nabla F(x_n)\|^2\big]$ from above. (c) follows from the compactness of $\mathcal{M}$ and the fact that $x_k \in \mathcal{U}$ for all $k = \lfloor n/S \rfloor S, \ldots, n$. The second term in (65) is not tightly controlled by a direct Cauchy–Schwarz bound. Instead, we exploit the SARAH recursion:

(ii) Sharp control of the cross term $\mathbb{E}[\mathsf{A}_n\tilde{\mathsf{A}}_{n-1}\mathbf{1}_{\mathcal{E}_n}]$. Let $e_t := \mathbf{v}_t - \nabla F(x_t)$ and recall $\mathsf{A}_t = -\langle \nabla F(x_t), e_t\rangle$ and that $\tilde{\mathsf{A}}_{n-1}$ is $\mathcal{F}_{n-1}$–measurable by construction. On inner steps, the SARAH error obeys

$$\begin{aligned}
e_t \ &= \ e_{t-1} \ + \ \Delta_t, \\
&\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}] = 0, \\
&\mathbb{E}\big[\|\Delta_t\|^2 \mid \mathcal{F}_{t-1}\big] \ \le \ \frac{4\tilde{L}^2}{n_g^t}\, \|x_t - x_{t-1}\|^2.
\end{aligned} \tag{67}$$

At (mini-batch) restarts, $\mathbb{E}[\|e_t\|^2 \mid \mathcal{F}_{t-1}] \le \sigma^2/n_g^t$. Using the law of total expectation and the fact that $\mathbb{E}[\Delta_n \mid \mathcal{F}_{n-1}] = 0$, we get

$$\mathbb{E}\big[\mathsf{A}_n\, \tilde{\mathsf{A}}_{n-1}\, \mathbf{1}_{\mathcal{E}_n}\big] \ = \ \mathbb{E}\Big[-\big\langle \nabla F(x_n),\, e_{n-1}\big\rangle\, \tilde{\mathsf{A}}_{n-1}\, \mathbf{1}_{\mathcal{E}_{n-1}}\Big].$$

By Cauchy–Schwarz and local boundedness of the gradient on $\mathcal{E}_{n-1}$ (from (18), $\|\nabla F(x)\| \le L_1$ on the tube),

$$\mathbb{E}\big[\mathsf{A}_n\,\tilde{\mathsf{A}}_{n-1}\,\mathbf{1}_{\mathcal{E}_n}\big] \;\le\; L_1\Big(\mathbb{E}\big[\|e_{n-1}\|^2\,\mathbf{1}_{\mathcal{E}_{n-1}}\big]\Big)^{1/2}\Big(\mathbb{E}\big[\tilde{\mathsf{A}}_{n-1}^2\,\mathbf{1}_{\mathcal{E}_{n-1}}\big]\Big)^{1/2}. \tag{68}$$

Let $t_0 := \lfloor n/S\rfloor S$ be the last restart index. Within the current epoch, the batch size is $n_g^t \ge (\lfloor t/S\rfloor S+1)^\beta$. Unrolling e_t from t to t_0 and taking expectations yields

$$\mathbb{E}\big[\|e_{n-1}\|^2\,\mathbf{1}_{\mathcal{E}_{n-1}}\big] \;\le\; \frac{\sigma^2}{(\lfloor n/S\rfloor S+1)^\beta} \;+\; \frac{4\tilde{L}^{\,2}}{(\lfloor n/S\rfloor S+1)^\beta}\sum_{j=t_0+1}^{n-1}\mathbb{E}\big[\|x_j-x_{j-1}\|^2\,\mathbf{1}_{\mathcal{E}_{j-1}}\big].$$

By using the identity $\|x_j-x_{j-1}\|^2=\eta_{j-1}^2\|\mathbf{v}_{j-1}\|^2$ and the local boundedness $\mathbb{E}[\|\mathbf{v}_j\|^2\,\mathbf{1}_{\mathcal{E}_j}]\le C_v$ for a constant C_v depending only on the local neighborhood[15], we obtain

$$\mathbb{E}\big[\|e_{n-1}\|^2\,\mathbf{1}_{\mathcal{E}_{n-1}}\big] \;\le\; \frac{\sigma^2}{(\lfloor n/S\rfloor S+1)^\beta} \;+\; \frac{4\tilde{L}^{\,2}\,C_v}{(\lfloor n/S\rfloor S+1)^\beta}\sum_{j=t_0+1}^{n-1}\eta_{j-1}^2. \tag{69}$$

Since the step-size schedule satisfies $\sum_{k\ge0}\eta_k^2<\infty$ and is non-increasing, the tail $\sum_{j=t_0+1}^{n-1}\eta_{j-1}^2$ is upper bounded by a constant $H\eta_0^2$ *independent of* n *and* S. Hence

$$\mathbb{E}\big[\|e_{n-1}\|^2\,\mathbf{1}_{\mathcal{E}_{n-1}}\big] \;\le\; \frac{\sigma^2+4\,\tilde{L}^{\,2}\,C_v\,H\eta_0^2}{(\lfloor n/S\rfloor S+1)^\beta}. \tag{70}$$

To bound $\mathbb{E}[\tilde{\mathsf{A}}_{n-1}^2\mathbf{1}_{\mathcal{E}_n}]$, we use the definition of $\tilde{\mathsf{A}}_n$ in (59) as follows:

$$\begin{aligned}
\mathbb{E}[\tilde{\mathsf{A}}_{n-1}^2\mathbf{1}_{\mathcal{E}_n}] &\le \sum_{j=0}^{n-1}\left[\prod_{i=j+1}^{n-1}\Big(1-\tfrac{\eta_i}{\tau}\Big)^2\right]\eta_j^2\mathbb{E}[\mathsf{A}_j^2\mathbf{1}_{\mathcal{E}_n}]\\
&\overset{(a)}{\le} L_1^2\sum_{j=0}^{n-1}\left[\prod_{i=j+1}^{n-1}\Big(1-\tfrac{\eta_i}{\tau}\Big)^2\right]\eta_j^2\times\left\{\frac{\sigma^2}{(\lfloor j/S\rfloor S+1)^\beta}+\frac{4\tilde{L}^2\cdot[j-\lfloor j/S\rfloor S]}{(\lfloor j/S\rfloor S+1)^\beta}\cdot(2R^2+2(\mathrm{diam}(\mathcal{M}))^2)\right\}\\
&\le L_1^2\cdot\Big\{\sigma^2+4\tilde{L}^2\cdot(2R^2+2(\mathrm{diam}(\mathcal{M}))^2)\Big\}\times\sum_{j=0}^{n-1}\left[\prod_{i=j+1}^{n-1}\Big(1-\tfrac{\eta_i}{\tau}\Big)^2\right]\eta_j^2\\
&\overset{(b)}{\le} L_1^2\cdot\Big\{\sigma^2+4\tilde{L}^2\cdot(2R^2+2(\mathrm{diam}(\mathcal{M}))^2)\Big\}\cdot\frac{\eta_0^2C}{n^\eta},
\end{aligned}\tag{71}$$

where (a) comes from the bound on $\mathbb{E}[\mathsf{A}_j^2\mathbf{1}_{\mathcal{E}_n}]$ in (66) and (b) follows from the bound on $S_n := \sum_{j=0}^{n-1}\left[\prod_{i=j+1}^{n-1}\Big(1-\frac{\eta_i}{\tau}\Big)^2\right]\eta_j^2$ derived in Lemma 20 with $\eta_n=\eta_0(n+1)^{-\eta}$ and some constant $C>0$.
Plugging (70) and (71) into (68) gives

$$\mathbb{E}\big[\mathsf{A}_n\,\tilde{\mathsf{A}}_{n-1}\,\mathbf{1}_{\mathcal{E}_n}\big] \;\le\; L_1^2\sqrt{\frac{\sigma^2+4\tilde{L}^{\,2}\,C_v\,H\eta_0^2}{(\lfloor n/S\rfloor S+1)^\beta}}\times\sqrt{\Big\{\sigma^2+c_0\,\tilde{L}^{\,2}\big(R^2+\mathrm{diam}(\mathcal{M})^2\big)\Big\}\cdot\frac{\eta_0^2\,C}{n^\eta}}.$$

Finally, for the last term in the right hand-side of (65), by using (66), we get

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{v}_n\|^2\,\mathbf{1}_{\mathcal{E}_n}]\\
&\le 2\mathbb{E}[\|\nabla F(x_n)\|^2]+2\mathbb{E}[\|\mathbf{v}_n-\nabla F(x_n)\|^2]\\
&\le 2L_1^2+2\left\{\frac{\sigma^2}{(\lfloor n/S\rfloor S+1)^\beta}+\frac{4\tilde{L}^2[n-\lfloor n/S\rfloor S]}{(\lfloor n/S\rfloor S+1)^\beta}\cdot(2R^2+2(\mathrm{diam}(\mathcal{M}))^2)\right\},
\end{aligned}\tag{72}$$

[15] Using $\|a+b\|^2\le 2\|a\|^2+2\|b\|^2$, we obtain $\mathbb{E}\big[\|\mathbf{v}_j\|^2\big] \le 2\,\mathbb{E}\big[\|\mathbf{v}_j-\nabla F(\mathbf{x}_j)\|^2\big]+2\,\mathbb{E}\big[\|\nabla F(\mathbf{x}_j)\|^2\big] \le 2\big(\sigma^2+4\tilde{L}^2R^2\big)+2L_1^2$, where the second inequality uses (53) and the Lipschitz bounds for F.

Combining (65) with all the upper bounds computed above for each term on the right-hand side yields the following upper bound for $\mathbb{E}[\mathsf{R}_n]$.

$$\begin{aligned}
\mathbb{E}[\mathsf{R}_n] &\le \mathbb{E}[\mathsf{R}_{n-1}] + \eta_n^2(1-\eta_n\tau^{-\frac{2}{\alpha}})^2\,\mathbb{E}[\mathsf{A}_n^2\,\mathbf{1}_{\mathcal{E}_n}] + 2\eta_n(1-\eta_n\tau^{-\frac{2}{\alpha}})\,\mathbb{E}[\mathsf{A}_n\,\mathbf{1}_{\mathcal{E}_n}\tilde{\mathsf{A}}_{n-1}] + \frac{L_2}{2}\eta_n^2\mathbb{E}[\|\mathbf{v}_n\|^2\,\mathbf{1}_{\mathcal{E}_n}] \\
&\le \mathbb{E}[\mathsf{R}_{n-1}] + \eta_n^2(1-\eta_n\tau^{-\frac{2}{\alpha}})^2 \times L_1^2\left\{\frac{\sigma^2}{(\lfloor n/S\rfloor S+1)^\beta} + \frac{4\tilde{L}^2\cdot[n-\lfloor n/S\rfloor S]}{(\lfloor n/S\rfloor S+1)^\beta}\cdot(2R^2+2(\mathrm{diam}(\mathcal{M}))^2)\right\} \\
&+ 2\eta_n(1-\eta_n\tau^{-\frac{2}{\alpha}})\cdot L_1^2\sqrt{\frac{\sigma^2+4\tilde{L}^2\,C_v\,H\eta_0^2}{(\lfloor n/S\rfloor S+1)^\beta}} \times \sqrt{\left\{\sigma^2+c_0\,\tilde{L}^2\big(R^2+\mathrm{diam}(\mathcal{M})^2\big)\right\}\frac{\eta_0^2\,C}{n^\eta}} \\
&+ L_1L_2\eta_n^2 + \frac{L_2}{2}\eta_n^2\left\{\frac{\sigma^2}{(\lfloor n/S\rfloor S+1)^\beta} + \frac{4\tilde{L}^2[n-\lfloor n/S\rfloor S]}{(\lfloor n/S\rfloor S+1)^\beta}\cdot(2R^2+2(\mathrm{diam}(\mathcal{M}))^2)\right\} \\
&\le \mathbb{E}[\mathsf{R}_{n-1}] + \eta_0\mathsf{C}_1\cdot n^{-3\eta/2}(\lfloor n/S\rfloor S+1)^{-\beta/2} \times \left[L_2L_1^2+L_2\sigma^2+L_2\tilde{L}^2(R^2+(\mathrm{diam}(\mathcal{M}))^2)\right],
\end{aligned} \tag{73}$$

where in the last step we used $\eta_n = \eta_0/(n+1)^\eta$ and we grouped each term decaying with n in the dominant term on the right hand side of (73), multiplied by constant $\mathsf{C}_1 > 0$ independent of the problem parameters.

(iii) Final summation and tail bound. Now by plugging the bound in (73) for $\mathbb{E}[\mathsf{R}_n - \mathsf{R}_{n-1}]$ into (64), we get

$$\mathbb{E}[\widetilde{\mathsf{R}}_n] \le \mathbb{E}[\widetilde{\mathsf{R}}_{n-1}] - \mathbb{E}\big[\mathsf{R}_{n-1}\mathbf{1}_{\widetilde{E}_{n-1}}\big] + \eta_0\mathsf{C}_1\cdot L_2\Big[L_1^2+\sigma^2+\tilde{L}^2(R^2+(\mathrm{diam}(\mathcal{M}))^2)\Big]. \tag{74}$$

Recall that $\widetilde{E}_n := E_{n-1}\setminus E_n = \{\mathsf{R}_{n-1} < s,\ \mathsf{R}_n \ge s\}$. By using the fact that $\mathsf{R}_{n-1} \ge s$ on $\widetilde{E}_{n-1}$ and Markov's inequality,

$$\begin{aligned}
\mathbb{P}(\widetilde{E}_{n-1}) &= \mathbb{P}\big(E_{n-1}\cap\{\mathsf{R}_n \ge s\}\big) = \mathbb{E}\big[\mathbf{1}_{E_{n-1}}\mathbf{1}_{\{\mathsf{R}_n > s\}}\big] \\
&\le \frac{1}{s}\,\mathbb{E}\big[\mathbf{1}_{E_{n-1}}\mathsf{R}_n\big] = \frac{\mathbb{E}[\widetilde{\mathsf{R}}_n]}{s}.
\end{aligned} \tag{75}$$

On the other hand, (74) yields the recursive bound

$$s\,\mathbb{P}(\widetilde{E}_{n-1}) \;\le\; \mathbb{E}[\widetilde{\mathsf{R}}_{n-1}] - \mathbb{E}[\widetilde{\mathsf{R}}_n] + \eta_0\mathsf{C}_1\cdot L_2\left[L_1^2+\sigma^2+\tilde{L}^2(R^2+(\mathrm{diam}(\mathcal{M}))^2)\right] \times n^{-3\eta/2}(\lfloor n/S\rfloor S+1)^{-\beta/2}. \tag{76}$$

Summing this inequality from $k=1$ to $n+1$ and using $\widetilde{\mathsf{R}}_0 = 0$ and $\widetilde{\mathsf{R}}_n \ge 0$ gives

$$s\sum_{k=1}^{n+1}\mathbb{P}(\widetilde{E}_{k-1}) \le \eta_0\mathsf{C}_1\cdot L_2\left[L_1^2+\sigma^2+\tilde{L}^2(R^2+(\mathrm{diam}(\mathcal{M}))^2)\right] \times \sum_{k=1}^{n+1}k^{-3\eta/2}(\lfloor k/S\rfloor S+1)^{-\beta/2}. \tag{77}$$

By the step-size condition (63), we get

$$\sum_{k=0}^{n}\mathbb{P}(\widetilde{E}_k) < \frac{\delta}{2}.$$

Since the events $\{\widetilde{E}_k\}_{k\ge0}$ are disjoint and $E_n = \bigcap_{k=0}^n \widetilde{E}_k^c$, we obtain

$$\mathbb{P}(E_n) \;=\; 1-\mathbb{P}\left(\bigcup_{k=0}^{n}\widetilde{E}_k\right) \;=\; 1-\sum_{k=0}^{n}\mathbb{P}(\widetilde{E}_k) \;\ge\; 1-\frac{\delta}{2}.$$

□

Lemma 14. *Suppose η_0 is small enough that*

$$\eta_0^2\cdot\frac{4\mathsf{C}_2\cdot[L_1^2+\sigma^2+\tilde{L}^2(R^2+(\mathrm{diam}(\mathcal{M}))^2)]}{R^2} \times \sum_{k=0}^{\infty}\frac{1}{(k+1)^{2\eta}} \;\le\; \frac{\delta}{2}. \tag{78}$$

Then $\mathbb{P}(C_n) \ge 1-\delta/2$, where $C_n := \{\|x_{k+1}-x_k\| \le R/2,\ \forall\, k \le n\}$.

*Proof.*By Lemma 12, we have $\mathbb{P}(E_n) \geq 1-\delta/2$. By (72) in the proof of Lemma 12, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{v}_n\|^2 \mathbf{1}_{\mathcal{E}_n}] &\leq 2L_1^2 + 2\left\{\frac{\sigma^2}{(\lfloor n/S\rfloor S+1)^\beta} + \frac{4\tilde{L}^2[n-\lfloor n/S\rfloor S]}{(\lfloor n/S\rfloor S+1)^\beta}\cdot(2R^2+2(\mathrm{diam}(\mathcal{M}))^2)\right\} \\ &\leq \mathsf{C}_2\cdot[L_1^2+\sigma^2+\tilde{L}^2(R^2+(\mathrm{diam}(\mathcal{M}))^2)], \end{aligned} \tag{79}$$

where C_2 is the constant independent of the parameters of the problem. Moreover, by the additional step-size assumption and Markov's inequality,

$$\begin{aligned}\mathbb{P}(C_n) &= \mathbb{P}\Big(\forall\, k\leq n:\ \|x_{k+1}-x_k\|\leq \tfrac{R}{2}\Big) \\ &\geq\ 1-\sum_{k=0}^{n}\mathbb{P}\big(\|x_{k+1}-x_k\|>\tfrac{R}{2}\big) \qquad (80)\\ &= 1-\sum_{k=0}^{n}\mathbb{P}\Big(\|\mathbf{v}_k\|>\tfrac{R}{2\eta_k}\Big)\ \geq\ 1-\sum_{k=0}^{n}\frac{\mathbb{E}[\|\mathbf{v}_k\|^2]}{(R/(2\eta_k))^2} \\ &\geq 1-\frac{4\mathsf{C}_2\cdot[L_1^2+\sigma^2+\tilde{L}^2(R^2+(\mathrm{diam}(\mathcal{M}))^2)]}{R^2}\cdot\sum_{k=1}^{n}\eta_k^2 \\ &\geq 1-\frac{\delta}{2}. \qquad (81)\end{aligned}$$

□

From lemmas 13 and 14, we have

$$\begin{aligned}&\mathbb{P}(E_n)\geq 1-\delta/2,\quad \mathbb{P}(C_n)\geq 1-\delta/2 \\ &\implies\quad \mathbb{P}(E_n\cap C_n)\geq 1-\delta,\end{aligned} \tag{82}$$

which implies the following lemma:

Lemma 15. *Under the conditions in Lemmas 12 to 14, we have* $\mathbb{P}[\hat{E}_n]\geq 1-\delta$.

On $E_n\cap C_n$, Lemma 12 gives $D_{k+1}<s$ for all $k\leq n$, hence $x_{k+1}\notin\mathcal{R}$; since $x_{k+1}\in\mathbb{B}_2(\mathcal{M};3R/4)$ by C_n, we must have $x_{k+1}\in\mathbb{B}_2(\mathcal{M};R/2)=\mathcal{U}$. Therefore

$$E_n\cap C_n\ \subset\ \mathcal{E}_{n+1}.$$

Starting from $x_0\in\mathcal{U}_0$ (so $D_0\leq s/2$), an induction yields $\mathbb{P}(\mathcal{E}_n)\geq 1-\delta$ for all n.

G.3 Proof of optimal local convergence rate of SARAH

Up to this point, we have shown that the SARAH iterates remain within a neighborhood of the connected component $\mathcal{M}$ of local minimizers with high probability; concretely, $\mathbb{P}(\mathcal{E}_n)\geq 1-\delta$ for all $n\geq 1$. We now turn to deriving the optimal local convergence rate of SARAH by exploiting the dynamics induced by the local α-PŁ property.

Lemma 16. *Let $f:\mathbb{R}^d\to\mathbb{R}$ be L-smooth. For any $x_t\in\mathbb{R}^d$, step size $\eta>0$, and vector $\mathbf{v}_t\in\mathbb{R}^d$, define $x_{t+1}=x_t-\eta\mathbf{v}_t$. Then*

$$\begin{aligned}f(x_{t+1})\ &\leq\ f(x_t)-\frac{\eta}{2}\big\|\nabla f(x_t)\big\|^2 \\ &-\Big(\frac{1}{2\eta}-\frac{L}{2}\Big)\big\|x_{t+1}-x_t\big\|^2+\frac{\eta}{2}\big\|\mathbf{v}_t-\nabla f(x_t)\big\|^2.\end{aligned} \tag{83}$$

Proof. By L-smoothness, for any y,

$$f(y)\leq f(x_t)+\langle\nabla f(x_t),y-x_t\rangle+\frac{L}{2}\|y-x_t\|^2.$$

Take $y = x_{t+1} = x_t - \eta \mathbf{v}_t$ and write

$$\langle \nabla f(x_t), x_{t+1} - x_t \rangle = \langle \nabla f(x_t) - \mathbf{v}_t,\, x_{t+1} - x_t \rangle + \langle \mathbf{v}_t,\, x_{t+1} - x_t \rangle.$$

Let $\bar{x}_{t+1} := x_t - \eta \nabla f(x_t)$. Since $x_{t+1} - x_t = -\eta \mathbf{v}_t$,

$$\begin{aligned} &\langle \mathbf{v}_t, x_{t+1} - x_t \rangle = -\eta \|\mathbf{v}_t\|^2, \\ &\langle \nabla f(x_t) - \mathbf{v}_t, x_{t+1} - x_t \rangle = -\frac{1}{\eta} \langle x_{t+1} - \bar{x}_{t+1},\, x_{t+1} - x_t \rangle. \end{aligned} \tag{84}$$

Using the three-point identity

$$\langle \mathbf{u} - \mathbf{v},\, \mathbf{u} - \mathbf{w} \rangle = \tfrac{1}{2}\big(\|\mathbf{u} - \mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 \big)$$

with $\mathbf{u} = x_{t+1}$, $\mathbf{v} = \bar{x}_{t+1}$, $\mathbf{w} = x_t$, we obtain

$$\begin{aligned} &-\frac{1}{\eta} \langle x_{t+1} - \bar{x}_{t+1},\, x_{t+1} - x_t \rangle = \\ &-\frac{1}{2\eta} \|x_{t+1} - x_t\|^2 + \frac{\eta}{2} \|\nabla f(x_t)\|^2 - \frac{\eta}{2} \|\mathbf{v}_t - \nabla f(x_t)\|^2. \end{aligned}$$

Substituting these relations into the smoothness bound and grouping terms gives

$$\begin{aligned} &f(x_{t+1}) \le f(x_t) - \frac{\eta}{2} \|\nabla f(x_t)\|^2 \\ &- \Big(\frac{1}{2\eta} - \frac{L}{2} \Big) \|x_{t+1} - x_t\|^2 + \frac{\eta}{2} \|\mathbf{v}_t - \nabla f(x_t)\|^2, \end{aligned}$$

as claimed. □

Lemma 17. *For every $F \in \mathcal{F}_\alpha$, given $\{x_t\}_{t=1}^T \in \mathcal{N}(\mathcal{M})$, $l = F(x^*)$ for all $x^* \in \mathcal{M}$, we have the following recursion inequality for $\delta_t = \mathbb{E}[F(x_t) - l]$:*

$$\delta_{t+1} \le \delta_t - \frac{\eta_t}{2\tau^{\frac{2}{\alpha}}} \delta_t^{\frac{2}{\alpha}} + \eta_t \mathbb{E}[\|\mathbf{v}_t - \nabla F(x_t)\|^2]. \tag{85}$$

*Proof.*By applying Lemma 16 to F and $\eta_t \le \frac{1}{2L_2}$, we have

$$F(x_{t+1}) \le F(x_t) - \frac{\eta_t}{2} \|\nabla F(x_t)\|^2 + \frac{\eta_t}{2} \|\mathbf{v}_t - \nabla F(x_t)\|^2. \tag{86}$$

Using Assumption 2, we have

$$F(x_{t+1}) \le F(x_t) - \frac{\eta_t}{2\tau^{\frac{2}{\alpha}}} (F(x_t) - l)^{\frac{2}{\alpha}} + \frac{\eta_t}{2} \|\mathbf{v}_t - \nabla F(x_t)\|^2. \tag{87}$$

Let us define $\delta_t := \mathbb{E}[F(x_t)] - l$. By taking expectation of both sides of (87) and using Jensen's inequality ($\mathbb{E}[x^{2/\alpha}] \ge (\mathbb{E}[x])^{2/\alpha}$ for $\alpha \in [1, 2]$), we have

$$\delta_{t+1} \le \delta_t - \frac{\eta_t}{2\tau^{\frac{2}{\alpha}}} \delta_t^{\frac{2}{\alpha}} + \frac{\eta_t}{2} \mathbb{E}[\|\mathbf{v}_t - \nabla F(x_t)\|^2]. \tag{88}$$

□

Lemma 18. *If $\mathbb{E}[\|\mathbf{v}_t - \nabla F(x_t)\|^2] = C_0 t^{-\beta}$, $\eta_t = \eta_0 t^{-\eta}$, and $\eta < 1$, then*

$$\delta_t = \mathcal{O}\left(t^{-\min\{\frac{\alpha\beta}{2}, \frac{\alpha(1-\eta)}{2-\alpha}\}} \right). \tag{89}$$

*Proof.*Set $\gamma := \min\left\{\frac{\alpha\beta}{2}, \frac{\alpha(1-\eta)}{2-\alpha}\right\}$ and define $B_k := (k+1)^\gamma \delta_k$ for $k \geq 0$. We show that $B_k = \mathcal{O}(1)$ for $k \geq 1$, which implies $\delta_k = \mathcal{O}((k+1)^{-\gamma})$.

$$\begin{aligned}
B_{k+1} &\leq (k+2)^\gamma \delta_k + (k+2)^\gamma \frac{\eta_k}{2}\mathbb{E}[\|\mathbf{v}_k - \nabla F(x_k)\|^2] - (k+2)^\gamma \frac{\eta_k}{2\tau^{\frac{2}{\alpha}}}\delta_k^{\frac{2}{\alpha}}. \\
&= \left(\frac{k+2}{k+1}\right)^\gamma \left[B_k + C_0\eta_0(k+1)^\gamma(k+1)^{-\eta-\beta} - (k+1)^{\gamma-\frac{2}{\alpha}\gamma-\eta}\cdot\frac{\eta_0 B_k^{\frac{2}{\alpha}}}{2\tau^{\frac{2}{\alpha}}}\right] \qquad (90) \\
&= B_k + \left[\left(1+\frac{1}{k+1}\right)^\gamma - 1\right]B_k + \left(\frac{k+2}{k+1}\right)^\gamma \left[C_0\eta_0(k+1)^{\gamma-\eta-\beta} - (k+1)^{\gamma-\frac{2}{\alpha}\gamma-\eta}\cdot\frac{\eta_0 B_k^{\frac{2}{\alpha}}}{2\tau^{\frac{2}{\alpha}}}\right] \qquad (91)
\end{aligned}$$

where (90) is from $\eta_k = \eta_0(k+1)^{-\eta}$.
Note that for any $k \in \mathbb{N}\cup\{0\}$ we have

$$(k+2)^\gamma - (k+1)^\gamma = (k+1)^\gamma[(1+(k+1)^{-1})^\gamma - 1] \leq c_\gamma(k+1)^{\gamma-1} \qquad (92)$$

where $c_\gamma = \gamma 2^{\gamma-1}$ and the last inequality follows from

$$(1+a)^\gamma - 1 = \int_1^{1+a}\gamma x^{\gamma-1}dx \leq \gamma\cdot(1+a-1)\cdot(1+a)^{\gamma-1} \leq \gamma 2^{\gamma-1}a \qquad (93)$$

for $a = (k+1)^{-1}$. Hence

$$\begin{aligned}
B_{k+1} - B_k &\leq c_\gamma(k+1)^{-1}B_k + 2^\gamma\left[C_0\eta_0(k+1)^{\gamma-\eta-\beta} - (k+1)^{\gamma-\frac{2}{\alpha}\gamma-\eta}\cdot\frac{\eta_0 B_k^{\frac{2}{\alpha}}}{2\tau^{\frac{2}{\alpha}}}\right] \\
&= (k+1)^{\gamma-\frac{2}{\alpha}\gamma-\eta}\left(c_\gamma(k+1)^{-(\gamma-\frac{2}{\alpha}\gamma-\eta+1)}B_k + 2^\gamma\left[C_0\eta_0(k+1)^{-(-\frac{2}{\alpha}\gamma+\beta)} - \frac{\eta_0 B_k^{\frac{2}{\alpha}}}{2\tau^{\frac{2}{\alpha}}}\right]\right) \\
&\leq (k+1)^{-1}\left(c_\gamma B_k + 2^\gamma\left[C_0\eta_0 - \frac{\eta_0 B_k^{\frac{2}{\alpha}}}{2\tau^{\frac{2}{\alpha}}}\right]\right), \qquad (94)
\end{aligned}$$

where (94) uses that $\gamma \leq \alpha\beta/2$ and $\gamma \leq \alpha(1-\eta)/(2-\alpha)$, i.e., $-2\gamma/\alpha + \beta \geq 0$ and $\gamma - 2\gamma/\alpha - \eta + 1 \geq 0$. To upper bound (94), we use the following lemma.

Lemma 19. *Let $F(B) := A_0 B - A_1 B^{2/\alpha} + A_2$ where $A_0 > 0$, $A_1 > 0$, $A_2 \geq 0$, and $1 \leq \alpha < 2$. Then for $B \geq \max\{A_2/A_0, (2A_0/A_1)^{\alpha/(2-\alpha)}\}$, $F(B) \leq 0$ and for all $B \geq 0$, we have $F(B) \leq A_2 + (\alpha/2)^{\alpha/(2-\alpha)}\cdot(2-\alpha)/2\cdot A_0^{2/(2-\alpha)}A_1^{-\alpha/(2-\alpha)}$.*

Define

$$A_0 := c_\gamma, \qquad A_1 := 2^{\gamma-1}\eta_0\tau^{-2/\alpha}, \qquad A_2 := 2^\gamma C_0\eta_0,$$

and set $M := \max\left\{A_2/A_0, (2A_0/A_1)^{\alpha/(2-\alpha)}\right\}$ and $M' := A_2 + (\alpha/2)^{\alpha/(2-\alpha)}\cdot(2-\alpha)/2\cdot A_0^{2/(2-\alpha)}A_1^{-\alpha/(2-\alpha)}$. Then (94) implies

$$B_{k+1} \leq B_k + (A_0 B_k - A_1 B_k^{\frac{2}{\alpha}} + A_2)/(k+1). \qquad (95)$$

We show that $B_t \leq \max\{B_0, M\} + M'/t$ for $t \geq 1$ by induction, which concludes the proof. For the base case, $B_1 \leq B_0 + M'$ by (95) and Lemma 19. For the induction step, assume that $B_k \leq \max\{B_0, M\} + M'/k$. If $B_k \leq M$, then by (95) and Lemma 19, we have $B_{k+1} \leq B_k + M'/(k+1) \leq M + M'/(k+1)$. If $B_k \geq M$, then $(A_0 B_k - A_1 B_k^{\frac{2}{\alpha}} + A_2) \leq 0$ by Lemma 19, and (95) gives $B_{k+1} \leq B_k \leq \max\{B_0, M\} + M'/k$. □

*Proof.*For $B \geq \max\{A_2/A_0, (2A_0/A_1)^{\alpha/(2-\alpha)}\}$, we have

$$F(B) = A_0 B(1 - A_1 A_0^{-1} B^{2/\alpha - 1}) + A_2 \leq -A_0 B + A_2 \leq 0.$$

Note that $\max_{B\geq 0} F(B)$ is attained at $B_* \geq 0$ where $F'(B_*) = A_0 - (2/\alpha)\cdot A_1 B_*^{2/\alpha-1} = 0$. This implies $B_* = (\alpha A_0/(2A_1))^{\alpha/(2-\alpha)}$. Consequently,

$$F(B) \leq \max_{B\geq 0} F(B) = \left(\frac{\alpha}{2}\right)^{\alpha/(2-\alpha)} \cdot \frac{2-\alpha}{2} \cdot \frac{A_0^{2/(2-\alpha)}}{A_1^{\alpha/(2-\alpha)}} + A_2.$$

□

Theorem 3.2. *Fix $\alpha \in [1,2)$ and $\delta \in (0,1)$, and let $\mathcal{M}$ be a connected component of local minima of F with level $l = F(y)$ for all $y \in \mathcal{M}$. Assume $F \in \mathcal{F}_\alpha$ and choose $R > 0$ so that $\mathbb{B}_2(\mathcal{M}; R) \subseteq \mathcal{N}(\mathcal{M})$ (cf. Assumption 2). Let the oracle be batch-smooth, $O \in \mathsf{O}_\sigma^{\tilde{L}}$, and run Algorithm 1 from x_0 with step sizes $\eta_t = \eta_0(t+1)^{-\frac{\alpha}{2}-(2-\alpha)\mathsf{x}}$ and batch-sizes $n_g^t = (t+1)^{1-2\mathsf{x}}$, for an arbitrary small value $\mathsf{x} > 0$. Let $\{x_t\}_{t\geq 1}$ be the iterates and set $\hat{x} := x_T$. Then the following holds:*

- *There exists $s > 0$ such that, with $\mathcal{U} := \{x : \mathrm{dist}(x, \mathcal{M}) \leq R/2\}$, $\mathcal{U}_0 := \{x : \mathrm{dist}(x, \mathcal{M}) < R/2,\ F(x) - l \leq s/2\}$, and $\eta_0 = \mathcal{O}(\min\{\sqrt{s}, s^{-1}\delta, R\sqrt{\delta}\})$, if $x_0 \in \mathcal{U}_0$, the event $\mathcal{E}_T(\mathcal{U}) := \{\ x_t \in \mathcal{U}$ for all $t = 1, \ldots, T\ \}$ occurs with probability at least $1-\delta$.*
- *Let $N := \mathbb{E}\Big[\sum_{t=1}^T n_g^t \mid \mathcal{E}_T(\mathcal{U})\Big]$ denote the expected total number of oracle queries used up to iteration T given $\mathcal{E}_T(\mathcal{U})$. Then $\mathbb{E}[F(\hat{x}) - l \mid \mathcal{E}_T(\mathcal{U})] \leq \varepsilon$, with $N = \mathcal{O}\big(\varepsilon^{-2/\alpha}\big)$.*

*Proof of Theorem 3.2.*According to $\gamma = \min\{\frac{\alpha\beta}{2}, \frac{\alpha(1-\eta)}{2-\alpha}\}$ in Lemma 18, for $1 \leq \alpha < 2$, we set if $\eta = \frac{\alpha}{2} + (2-\alpha)\mathsf{x}$, $\beta = 1 - 2\mathsf{x}$, and then $\gamma = \frac{\alpha}{2} - \alpha\mathsf{x}$. For example, when $\alpha = 1$, we set $\eta = \frac{1}{2} + \mathsf{x}$, $\beta = 1 - 2\mathsf{x}$, and then $\gamma = \frac{1}{2} - \mathsf{x}$. The first part of the claim is the result of Lemma 15. We first check the conditions in Lemmas 12 to 14 for the choices of η and β in the following: Note that for in Lemma 12, the auxiliary variable $q = 3 - \alpha$, we get

$$\sum_{n=1}^{\infty} \frac{1}{(n+1)^{\frac{2q-2}{2-\alpha}}} = \sum_{n=1}^{\infty} \frac{1}{(n+1)^2} \leq \frac{\pi^2}{6} < \infty.$$

Then the condition in Lemma 12, reduces to

$$\eta_0 < \mathsf{UB}_1 := \left(\frac{3s}{\pi^2 \tilde{c}}\right)^{1/2}. \tag{96}$$

In Lemma 13, since $3\eta/2 + \beta/2 > 5/4$, we have

$$\sum_{k=1}^{\infty} k^{-3\eta/2}(\lfloor k/S \rfloor S + 1)^{-\beta/2} \leq \sum_{t=1}^{\infty} \frac{1}{t^{\frac{5}{4}}} = \zeta(5/4) < \infty. \tag{97}$$

where $\zeta(\cdot)$ is the Riemann's zeta function. Then the condition in Lemma 13 is restated as follows:

$$\eta_0 \leq \mathsf{UB}_2 := \delta(2s\mathsf{C}_1 \cdot L_2)^{-1} \times \left[L_1^2 + \sigma^2 + \tilde{L}^2(R^2 + (\mathrm{diam}(\mathcal{M}))^2)\right]^{-1} \cdot (\zeta(5/4))^{-1}, \tag{98}$$

In Lemma 14, since $\eta = \frac{\alpha}{2} + (2-\alpha)\mathsf{x} > \frac{1}{2}$,

$$\sum_{k=0}^{\infty} \frac{1}{(k+1)^{2\eta}} = \zeta(\alpha + 2(2-\alpha)\mathsf{x}) < \infty.$$

Then the condition in Lemma 14 can be restated as follows:

$$\eta_0 \ \leq\ \mathsf{UB}_3 := \sqrt{R^2\delta} \cdot 8^{-\frac{1}{2}}(\mathsf{C}_2 \cdot \zeta(2\eta))^{-\frac{1}{2}} \times [L_1^2 + \sigma^2 + \tilde{L}^2(R^2 + (\mathrm{diam}(\mathcal{M}))^2)]^{-\frac{1}{2}}. \tag{99}$$

In conclusion, we get

$$\eta_0 \leq \min\left\{\mathsf{UB}_1, \mathsf{UB}_2, \mathsf{UB}_3\right\}. \tag{100}$$

From Lemma 17, we have for $\delta_t = \mathbb{E}[F(x_t) - l]$:

$$\delta_{t+1} \leq \delta_t - \frac{\eta_t}{2\tau^{\frac{2}{\alpha}}}\delta_t^{\frac{2}{\alpha}} + \eta_t \mathbb{E}[\|\mathbf{v}_t - \nabla F(x_t)\|^2]. \tag{101}$$

Now, suppose that on block $(k-1)S < t \leq kS$, $k \geq 1$, we target the variance level $\varepsilon_k := (kS)^{-\beta}$ in the choice of n_g^t. Then, from Lemma 8, for all $t \geq 0$,

$$\delta_{t+1} \leq \delta_t - \frac{\eta_t}{2\tau^{\frac{2}{\alpha}}}\delta_t^{\frac{2}{\alpha}} + \eta_t \frac{C_g}{(\lceil t/S \rceil S)^{\beta}}, \tag{102}$$

and based on Lemma 18, if $\mathbb{E}[\|\mathbf{v}_t - \nabla F(x_t)\|^2] = C_0 t^{-\beta}$, $\eta_t = \eta_0(t+1)^{-\eta}$, and $\eta < 1$, then

$$\delta_t = \mathcal{O}(t^{-\gamma}), \tag{103}$$

where $\gamma = \min\{\frac{\alpha\beta}{2}, \frac{\alpha(1-\eta)}{2-\alpha}\}$. After $T = \mathcal{O}(\varepsilon^{-\frac{1}{\gamma}})$ iterations, $\delta_T \leq \varepsilon$. Furthermore, according to Lemma 16, we have: for $\eta_t \leq \frac{1}{2L_2}$,

$$\|x_{t+1} - x_t\|^2 \leq 4\eta_t[F(x_t) - F(x_{t+1})] + 2\eta_t^2\|\mathbf{v}_t - \nabla F(x_t)\|^2. \tag{104}$$

Define $\Delta_t := \mathbb{E}[F(x_t) - F(x_{t+1})]$. Taking expectation and summing from $t = (k-1)S+1$ to $t = kS$, we obtain

$$\begin{aligned}
\sum_{t=(k-1)S+1}^{kS} \mathbb{E}[\|x_{t+1} - x_t\|^2] &\leq 4 \sum_{t=(k-1)S+1}^{kS} \eta_t \Delta_t + 2 \sum_{t=(k-1)S+1}^{kS} \eta_t^2 \mathbb{E}[\|\mathbf{v}_t - \nabla F(x_t)\|^2] \\
&\leq 4\eta_{(k-1)S+1} \sum_{t=(k-1)S+1}^{kS} \Delta_t + 2(kS)^{-\beta} \sum_{t=(k-1)S+1}^{kS} \eta_t^2 \\
&\leq 4\eta_{(k-1)S+1}\delta_{(k-1)S+1} + 2(kS)^{-\beta} \sum_{t=(k-1)S+1}^{kS} \eta_t^2 \\
&\leq \mathcal{O}\big(\eta_0 S^{-\eta-\gamma} k^{-\eta-\gamma}\big) + \mathcal{O}\big(\eta_0^2 S^{1-2\eta-\beta} k^{-2\eta-\beta}\big),
\end{aligned} \tag{105}$$

where we used that η_t is decreasing, $\sum_{t=(k-1)S+1}^{kS} \Delta_t = \mathbb{E}[F(x_{(k-1)S+1}) - F(x_{kS+1})] \leq \delta_{(k-1)S+1}$, $\delta_t = \mathcal{O}(t^{-\gamma})$, and

$$\sum_{t=(k-1)S+1}^{kS} \frac{1}{(t+1)^{2\eta}} = \Theta\big(S^{1-2\eta}k^{-2\eta}\big),$$

for $\eta > 1/2$. Hence, according to Lemma 8, to obtain the variance bound

$$\mathbb{E}[\|\nabla F(x_t) - \mathbf{g}_t\|^2] \leq \frac{1}{(kS)^{\beta}}$$

on block k, it is enough that the expected total number of stochastic gradient samples up to time T satisfies

$$\sum_{\text{mod }(t,S)=0} \frac{2\sigma^2}{t^{-\beta}} + \sum_{\text{mod }(t,S)\neq 0} \frac{4 \cdot \tilde{L}^2 S \mathbb{E}[\|x_{t+1} - x_t\|^2]}{(\lceil t/S \rceil S)^{-\beta}}. \tag{106}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} n_g^t\right] &\leq \sum_{k=1}^{\lceil T/S \rceil} 2\sigma^2 (kS)^{\beta} + \sum_{k=1}^{\lceil T/S \rceil} \sum_{t=(k-1)S+1}^{kS} 4 \cdot \tilde{L}^2 S (kS)^{\beta} \mathbb{E}[\|x_{t+1} - x_t\|^2] \\
&\leq \mathcal{O}(\sigma^2 T^{\beta+1} S^{-1}) \\
&\quad + \mathcal{O}\left(\tilde{L}^2 \eta_0 S^{\beta+1-\eta-\gamma} \sum_{k=1}^{\lceil T/S \rceil} k^{\beta-\eta-\gamma}\right) \\
&\quad + \mathcal{O}\left(\tilde{L}^2 \eta_0^2 S^{2-2\eta} \sum_{k=1}^{\lceil T/S \rceil} k^{-2\eta}\right).
\end{aligned}$$

If $S = \lfloor T/q \rfloor$ where q is a fixed positive integer, then $\lceil T/S \rceil = \mathcal{O}(1)$ and, since $\eta > 1/2$, both sums in (107) are bounded by constants. Hence the expected total number of samples is

$$\mathcal{O}\left(\sigma^2 T^\beta + \tilde{L}^2 \eta_0 T^{\beta+1-\eta-\gamma} + \tilde{L}^2 \eta_0^2 T^{2-2\eta}\right)$$

For the choice

$$\eta = \frac{\alpha}{2} + (2-\alpha)\mathsf{x}, \qquad \beta = 1 - 2\mathsf{x}, \qquad \gamma = \frac{\alpha}{2} - \alpha\mathsf{x},$$

we have

$$\beta + 1 - \eta - \gamma = (2-\alpha)(1-2\mathsf{x}) \le \beta, \qquad 2 - 2\eta = (2-\alpha)(1-2\mathsf{x}) \le \beta. \tag{107}$$

Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T} n_g^t\right] = \mathcal{O}(T^\beta).$$

Since $T = \mathcal{O}(\varepsilon^{-1/\gamma})$ and $\beta/\gamma = 2/\alpha$, this gives

$$\mathbb{E}\left[\sum_{t=1}^{T} n_g^t\right] = \mathcal{O}(\varepsilon^{-2/\alpha}).$$

Finally, because $\mathbb{P}(\mathcal{E}_T(\mathcal{U})) \ge 1 - \delta$, the conditional sample complexity satisfies

$$N = \mathbb{E}\left[\sum_{t=1}^{T} n_g^t \mid \mathcal{E}_T(\mathcal{U})\right] \le (1-\delta)^{-1}\mathbb{E}\left[\sum_{t=1}^{T} n_g^t\right] = \mathcal{O}(\varepsilon^{-2/\alpha}),$$

which proves the claimed oracle upper bound. □

G.4 Supplementary lemmas for Section 3

Lemma 20. *Let*

$$S_n \;:=\; \sum_{j=1}^{n-1}\left[\prod_{I=j+1}^{n-1}\left(1-\frac{\eta_I}{\tau}\right)^2\right]\eta_j^2, \qquad \eta_k = \frac{\eta_0}{(k+1)^\eta},$$

with parameters $\eta \in (\frac{1}{2}, 1)$ and $\tau, \eta_0 > 0$ (empty products are 1). Then there exist constants $c, C > 0$ and $N \in \mathbb{N}$, depending only on (η, τ), such that for all $n \ge N$,

$$\eta_0^2 c\, n^{-\eta} \;\le\; S_n \;\le\; \eta_0^2 C\, n^{-\eta}.$$

In particular, $S_n = \Theta(n^{-\eta})$.

Proof. We use $(1-x) \le e^{-x}$ and $(1-x)^2 \ge 1 - 2x$ for $x \in [0,1]$, and for nonnegative $\{y_i\}$ with $\sum_i y_i \le 1$, $\prod_i (1-y_i) \ge 1 - \sum_i y_i$.

Head–tail split. Pick K so large that $\eta_I/\tau \le \frac{1}{2}$ for all $I \ge K$. Write

$$S_n = \sum_{j=1}^{K-1}\Big(\cdots\Big)\eta_j^2 \;+\; \sum_{j=K}^{n-1}\Big(\cdots\Big)\eta_j^2 =: S_n^{\text{head}} + S_n^{\text{tail}}.$$

For fixed $j \le K-1$,

$$\prod_{I=j+1}^{n-1}\left(1-\tfrac{\eta_I}{\tau}\right)^2 \;\le\; \exp\Big(-\tfrac{2}{\tau}\sum_{I=K}^{n-1}\eta_I\Big) \;\le\; \exp(-c\,n^{1-\eta}),$$

since $\sum_{I=K}^{n-1}\eta_I \asymp n^{1-\eta}$ for $\eta \in (0,1)$. Hence $S_n^{\text{head}} = O(e^{-c\,n^{1-\eta}}) = o(n^{-\eta})$.

For $j \geq K$, $(1-x) \leq e^{-x}$ gives

$$\prod_{I=j+1}^{n-1} \Big(1 - \tfrac{\eta_I}{\tau}\Big)^2 \;\leq\; \exp\Big(-\tfrac{2}{\tau} \sum_{I=j+1}^{n-1} \eta_I\Big).$$

Let $r := n-1-j \in \{0,1,\dots\}$. Monotonicity of (η_k) yields

$$\sum_{I=j+1}^{n-1} \eta_I \;\geq\; r\,\eta_{n-1} \;=\; r\,\frac{\eta_0}{n^\eta},$$

so with $\lambda := \frac{2\eta_0}{\tau}$,

$$\prod_{I=j+1}^{n-1} \Big(1 - \tfrac{\eta_I}{\tau}\Big)^2 \;\leq\; e^{-\lambda r/n^\eta}.$$

Thus

$$S_n^{\text{tail}} \;\leq\; \sum_{r=0}^{n-1} e^{-\lambda r/n^\eta}\; \eta_{n-1-r}^2.$$

Split the sum between $r \leq n/2$ and $r > n/2$. If $0 \leq r \leq n/2$, then $\eta_{n-1-r}^2 \leq \eta_0^2/(n/2)^{2\eta} = 2^{2\eta}\eta_0^2\, n^{-2\eta}$. Hence

$$\sum_{r=0}^{\lfloor n/2 \rfloor} e^{-\lambda r/n^\eta} \eta_{n-1-r}^2 \leq 2^{2\eta}\eta_0^2\, n^{-2\eta} \sum_{r\geq 0} e^{-\lambda r/n^\eta} = 2^{2\eta}\eta_0^2\, n^{-2\eta}\, \frac{1}{1-e^{-\lambda/n^\eta}}.$$

For large n, $1-e^{-x} \geq x/2$ when $x = \lambda/n^\eta \in (0,1]$, so this is $\leq \dfrac{2^{2\eta+1}\eta_0^2}{\lambda}\, n^{-\eta} = (2^{2\eta}\eta_0\tau)\, n^{-\eta}$. If $r > n/2$, then $e^{-\lambda r/n^\eta} \leq e^{-(\lambda/2)n^{1-\eta}}$ and $\eta_{n-1-r}^2 \leq \eta_0^2$, so the total is $O\big(n\, e^{-(\lambda/2)n^{1-\eta}}\big) = o(n^{-\eta})$. Therefore $S_n^{\text{tail}} \leq C\, n^{-\eta}$ for some $C > 0$, and hence $S_n \leq C\, n^{-\eta}$.

To give a lower bound, set

$$\gamma := \frac{\tau}{2^{\eta+3}\eta_0}, \qquad m := \lfloor \gamma n^\eta \rfloor.$$

For large n, $1 \leq m \leq n/2$. For any $j \in \{n-m, \dots, n-1\}$ we have

$$\sum_{I=j+1}^{n-1} \frac{2\eta_I}{\tau} \;\leq\; \sum_{I=n-m}^{n-1} \frac{2\eta_I}{\tau} \;\leq\; \frac{2m}{\tau}\,\eta_{n-m} \;\leq\; \frac{2^{\eta+1}\eta_0\gamma}{\tau} \;=\; \frac{1}{2},$$

using $n-m+1 \geq n/2$. Since $(1-x)^2 \geq 1-2x$ and $\prod_i (1-y_i) \geq 1 - \sum_i y_i$ when the sum is ≤ 1,

$$\prod_{I=j+1}^{n-1} \Big(1 - \tfrac{\eta_I}{\tau}\Big)^2 \;\geq\; \prod_{I=n-m}^{n-1} \Big(1 - \tfrac{2\eta_I}{\tau}\Big) \;\geq\; 1 - \sum_{I=n-m}^{n-1} \tfrac{2\eta_I}{\tau} \;\geq\; \tfrac{1}{2}.$$

Therefore

$$S_n \;\geq\; \frac{1}{2} \sum_{j=n-m}^{n-1} \eta_j^2 \;\geq\; \frac{1}{2}\, m\, \eta_{n-1}^2 \;\geq\; \Big(\frac{\gamma\eta_0^2}{2}\Big) n^{-\eta}.$$

Setting $c := \gamma\eta_0^2/2$ gives the claim. □

H PROOFS OF SECTION 4

Theorem 4.1. *For the family of domain sets $\mathbb{S}_R$, there exists a function F in the function class $\mathcal{F}^{\mathcal{X}}_{\alpha,\tau,\varepsilon}$, an oracle O in the family of oracles O^G, and $\alpha \in (1,2]$ and $\varepsilon \leq \min\{((\alpha-1)/\alpha)^\alpha \tau, 1\}$, such that the number of oracle queries for any stochastic first-order algorithm $\mathcal{A} \in \mathcal{A}_m$ required to output an estimate $\hat{x}$ satisfying $F(\hat{x}) - F^* \leq \varepsilon$ with probability at least $1-\delta$ is*

$$\Omega\left(\frac{G^2 \tau^{\frac{2}{\alpha}} \log\left(\frac{(\alpha-1)R}{2\alpha\varepsilon^{\frac{\alpha-1}{\alpha}}\tau^{\frac{1}{\alpha}}}\right)}{\varepsilon^{\frac{2}{\alpha}}}\right). \tag{108}$$

Proof of Theorem 4.1. We prove the lower bound by a reduction to the noisy binary search (NBS) problem. Herein, we consider the following: Assume that N sorted elements $\{a_1, \ldots, a_N\}$ are given and we want to insert a new element u using the queries of the form "Is $u > a_j$?". The oracle answers this query correctly with probability $1/2 + p$ for some fixed $p \in [0, 1/2)$. Let j^* be the unique index such that $a_{j^*} \leq u < a_{j^*+1}$. It is well known (see [Feige et al., 1994, Karp and Kleinberg, 2007]) that we need at least $\Omega\left(p^{-2} \log N\right)$ queries on average in order to identify j^*.

Reduction scheme: We will construct a stochastic optimization problem with the given parameters $(\bar{L}, \tau, \alpha)$, such that if there exists an algorithm that solves it (with a constant probability) after T first-order stochastic queries to the oracle O^G, then it can be used to identify j^* in NBS problem (with the same probability) using at most $2T$ queries.

First, at each iteration t, we define a random variable $Z_{t,j} \in \{-1, 1\}$ for every $1 \leq j \leq N$ as follows:

$$\mathbb{P}[Z_{t,j} = 1] = \begin{cases} \frac{1}{2} + p & j > j^*, \\ \frac{1}{2} - p & j \leq j^*. \end{cases} \tag{109}$$

$Z_{t,j}$ is the answer of the NBS oracle to query "Is $u > a_j$?" at the iteration t.

In the reduction scheme, we assume that function F has a one-dimensional domain $\mathcal{X}$. The diameter of this domain is $\sup_{x,y \in \mathcal{X}} |x - y| = R$, and without loss of generality, we assume that $\mathcal{X} = [0, R]$. We first divide the interval $[0, R]$ into N equal sub-intervals of length R/N each, and consider the element a_j as the smallest point in the j-th interval.

NBS oracle: At each iteration, NBS oracle is queried at a point $x \in \mathcal{X}$ and its response is $(Z_{t,j}, Z_{t,j+1})$, for $x \in [a_j, a_{j+1})$.

Stochastic first-order oracle: Using the noisy binary pairs $(Z_{t,j}, Z_{t,j+1})$ from NBS oracle queried at $x \in [a_j, a_{j+1})$, the output of this oracle at point x is constructed as follows:

$$\begin{aligned} &f'(x, Z_{t,j}, Z_{t,j+1}) \\ &\qquad = \frac{G}{2}\left(1 - g_j(x)\right) Z_{t,j} + \frac{G}{2}\left(1 + g_j(x)\right) Z_{t,j+1}, \end{aligned} \tag{110}$$

where $G > 0$ is some constant and

$$g_j(x) = \frac{\left|x - \frac{R}{2N} - a_j\right|^{\frac{1}{\alpha-1}} \cdot \operatorname{sgn}\left(x - \frac{R}{2N} - a_j\right)}{\left(\frac{R}{2N}\right)^{\frac{1}{\alpha-1}}} \mathbf{1}_{[a_j, a_{j+1})}(x). \tag{111}$$

Note that $\mathbb{E}[f'(x, Z_{t,j}, Z_{t,j+1})] = F'(x)$ and

$$|f'(x, Z_{t,j}, Z_{t,j+1})| = \begin{cases} G & \text{if } Z_{t,j} = Z_{t,j+1}, \\ G|g_j(x)| & \text{if } Z_{t,j} \neq Z_{t,j+1}. \end{cases}$$

Hence, $|f'(x, Z_{t,j}, Z_{t,j+1})| \leq G$. Taking expectation of $f'(x, Z_{t,j}, Z_{t,j+1})$, we obtain

$$\begin{aligned} F'(x) = \mathbb{E}[f'(x, Z_{t,j}, Z_{t,j+1})] &= pG\, \mathbf{1}_{[a_{j^*+1}, R]}(x) \\ &\quad - pG\, \mathbf{1}_{[0, a_{j^*})}(x) + pG g_{j^*}(x) \mathbf{1}_{[a_{j^*}, a_{j^*+1})}(x). \end{aligned} \tag{112}$$

Integrating $F'(x)$ with respect to x, we get

$$\begin{aligned} F(x) &= pG\left(x - a_{j^*+1}\right) \mathbf{1}_{[a_{j^*+1},\, R]}(x) \\ &\quad + pG\left(-x + a_{j^*}\right) \mathbf{1}_{[0,\, a_{j^*})}(x) \\ &\quad + \Bigg[pG\, \frac{\alpha - 1}{\alpha} \frac{\left|x - \frac{R}{2N} - a_{j^*}\right|^{\frac{\alpha}{\alpha-1}}}{\left(\frac{R}{2N}\right)^{\frac{1}{\alpha-1}}} \\ &\qquad\qquad - pG\, \frac{\alpha - 1}{2\alpha} \frac{R}{N}\Bigg] \mathbf{1}_{[a_{j^*},\, a_{j^*+1})}(x). \end{aligned} \tag{113}$$

Note that by construction, $\min_{x\in\mathcal{X}} F(x) = -pG(\alpha-1)R/(2\alpha N)$ and $a_{j^*} + R/(2N) =_{x\in\mathcal{X}} F(x)$. Moreover, function F given by (113) is convex and its domain is bounded ($\mathcal{X} = [0, R]$). In Lemma 21, we show that if

$$\tau \geq \frac{\alpha-1}{\alpha}\frac{R}{2N}(pG)^{1-\alpha}, \tag{114}$$

then F satisfies the local $(\alpha, \tau, R/N)$-gradient-dominance (Assumption 6). In our reduction, we need to show that if the output of a stochastic first-order method $\hat{x}$ satisfies $F(\hat{x}) - F^* \leq \varepsilon$, then j^* is identified (in other words, $\hat{x} \in [a_{j^*}, a_{j^*+1})$). If

$$pG\frac{\alpha-1}{2\alpha}\frac{R}{N} \geq 2\varepsilon, \tag{115}$$

we get $F(x) - F^* > \varepsilon$ for every $x \notin [a_{j^*}, a_{j^*+1})$. Indeed from the definition of the function (113), for every $x \notin [a_{j^*}, a_{j^*+1})$, we have

$$F(x) - F^* \geq pG\frac{\alpha-1}{2\alpha}\frac{R}{N}$$

and if $pG(\alpha-1)/(2\alpha)R/N \geq 2\varepsilon$, we get $F(x) - F^* > \varepsilon$.

We pick

$$p = \frac{2\varepsilon^{1/\alpha}}{G\tau^{1/\alpha}}, \quad N = \frac{(\alpha-1)R}{(2\alpha)\varepsilon^{(\alpha-1)/\alpha}\tau^{1/\alpha}}. \tag{116}$$

Subsequently, with these chosen values for p and N, the inequalities (114) and (115) are met for every $\varepsilon \leq 1$. For $\varepsilon \leq ((\alpha-1)/\alpha)^\alpha\tau$, we have: $R/N = 2\alpha\varepsilon^{(\alpha-1)/\alpha}\tau^{1/\alpha}(\alpha-1)^{-1} \geq \varepsilon$, and therefore, every local $(\alpha, \tau, R/N)$-gradient-dominated function is also a local $(\alpha, \tau, \varepsilon)$-gradient-dominated function. Consequently, $\mathcal{F}^{\mathcal{X}}_{\alpha,\tau,R/N} \subseteq \mathcal{F}^{\mathcal{X}}_{\alpha,\tau,\varepsilon}$, and as a result, $F \in \mathcal{F}^{\mathcal{X}}_{\alpha,\tau,\varepsilon}$.
Thus, for $(\mathcal{F}^{\mathcal{X}}_{\alpha,\tau,\varepsilon}, \mathsf{O}^G)$, any stochastic first-order algorithm that converges to an ε-minimizer can be used to identify j^* in a NBS problem for appropriately chosen p and N as in (116). Therefore, the probability-based minimax oracle complexity $\mathsf{T}_\varepsilon(\mathcal{F}^{\mathcal{X}}_{\alpha,\tau,\varepsilon}, \mathsf{O}^G)$ can be lower bounded by $\Omega\left(p^{-2}\log N\right)$. For every $\varepsilon \leq \min\{((\alpha-1)/\alpha)^\alpha\tau, 1\}$,

$$\mathsf{T}_\varepsilon(\mathcal{F}^{\mathcal{X}}_{\alpha,\tau,\varepsilon}, \mathsf{O}^G) = \Omega\left(\frac{G^2\tau^{\frac{2}{\alpha}}\log\left(\frac{(\alpha-1)R}{2\alpha\varepsilon^{\frac{\alpha-1}{\alpha}}\tau^{\frac{1}{\alpha}}}\right)}{\varepsilon^{\frac{2}{\alpha}}}\right).$$

□

Lemma 21. *Function F defined in (113) satisfies $(\alpha, \tau, R/N)$-gradient-dominance for $\tau \geq (\alpha-1)R(pG)^{1-\alpha}/(2\alpha N)$.*

Proof. For $x \in [a_{j^*}, a_{j^*+1})$, $F(x) - \min_{x\in\mathcal{X}} F(x) \leq \tau|F'(x)|^\alpha$ is equivalent to have

$$\begin{aligned} F(x) - \min_{x\in\mathcal{X}} F(x) &= pG\frac{\alpha-1}{\alpha}\frac{|x - \frac{R}{2N} - a_{j^*}|^{\frac{\alpha}{\alpha-1}}}{\left(\frac{R}{2N}\right)^{\frac{1}{\alpha-1}}} \\ &\leq \tau\left|pG\frac{|x - \frac{R}{2N} - a_j|^{\frac{1}{\alpha-1}}\cdot\operatorname{sgn}(x - \frac{R}{2N} - a_j)}{\left(\frac{R}{2N}\right)^{\frac{1}{\alpha-1}}}\right|^\alpha. \end{aligned} \tag{117}$$

If $\tau \geq (\alpha-1)R(pG)^{1-\alpha}/(2\alpha N)$, we get $F(x) - \min_{x\in\mathcal{X}} F(x) \leq \tau|F'(x)|^\alpha$. □

Remark 9. *In Appendix H.1, we consider the ϕ-Kurdyka-Łojasiewicz (KL) inequality [Yang and Lin, 2018] and recall the definition of function ϕ and ϕ-KL inequality. For the class of convex functions satisfying the ϕ-KL inequality with oracle O^G and domain sets $\mathbb{S}_R$, we derive the lower bound $\Omega\left(G^2(\phi'(\varepsilon))^2\log\left(R/(2\phi(\varepsilon))\right)\right)$. In this setting, the upper bound $T = \mathcal{O}\left(G^2(\phi(\varepsilon))^2\log(1/\varepsilon)/\varepsilon^2\right)$ from Yang and Lin [2018, Corollary 14] is larger than our lower bound by a multiplicative factor of $\mathcal{O}\left(\left((\phi(\varepsilon)/(\varepsilon\phi'(\varepsilon))\right)^2\cdot\log(1/\varepsilon)/\log(R/2\phi(\varepsilon))\right)$. It is noteworthy that this factor becomes a constant for $\phi(s) = C\cdot s^{1-1/\alpha}$ for $\alpha > 1$ and some constant $C > 0$. It would be interesting to characterize the minimax oracle complexity of first-order methods for achieving a global-optimum point of a convex bounded domain function that satisfies ϕ-KL inequality for other choices of function ϕ.*

H.1 Proof of Remark 9

Assumption 7. *Consider a continuous concave function $\phi : [0, \zeta) \to \mathbb{R}^+$ such that (i) $\phi(0) = 0$; (ii) ϕ is continuous on $(0, \zeta)$; (iii) and for all $s \in (0, \zeta)$, $\phi'(s) > 0$. Function $f(x)$ satisfies the ϕ-Kurdyka-Łojasiewicz (ϕ-KL) property at $\bar{x}$ if there exist $\zeta \in (0, \infty]$, a neighborhood $U_{\bar{x}}$ of $\bar{x}$ and for all $x \in U_{\bar{x}} \cap \{x : f(\bar{x}) < f(x) < f(\bar{x}) + \zeta\}$, the following inequality holds*

$$\phi'(f(x) - f(\bar{x})) \cdot \|\partial f(x)\|_2 \geq 1, \tag{118}$$

where $\|\partial f(x)\|_2 := \min_{\mathbf{g} \in \partial f(x)} \|\mathbf{g}\|_2$.

Stochastic first-order oracle: Using the noisy binary pairs $(Z_{t,j}, Z_{t,j+1})$ from NBS oracle which is queried at $x \in [a_j, a_{j+1})$, the output of this oracle at point x is constructed as follows:

$$\begin{aligned} f'(x, Z_{t,j}, Z_{t,j+1}) = \frac{G}{2}\,(1 - g_j(x))\, Z_{t,j} \\ + \frac{G}{2}\,(1 + g_j(x))\, Z_{t,j+1}, \end{aligned} \tag{119}$$

where G is some constant and

$$\begin{aligned} &g_j(x) = \\ &\quad \frac{\psi'(|x - \frac{R}{2N} - a_j|) \cdot \operatorname{sgn}(x - \frac{R}{2N} - a_j)}{\psi'(\frac{R}{2N})} \cdot \mathbf{1}_{[a_j, a_{j+1})}(x), \end{aligned} \tag{120}$$

where $\psi \equiv \phi^{-1}$ and then $\psi : [0, \infty) \to [0, \infty)$ is a continuous convex function such that $\psi(0) = 0$, $\psi'(x) > 0$ for $x \in \mathbb{R}^+$. Note that

$$|f'(x, Z_{t,j}, Z_{t,j+1})| = \begin{cases} G & \text{if } Z_{t,j} = Z_{t,j+1}, \\ G|g_j(x)| & \text{if } Z_{t,j} \neq Z_{t,j+1}. \end{cases}$$

Hence, $|f'(x, Z_{t,j}, Z_{t,j+1})| \leq G$. Taking expectation of $f'(x, Z_{t,j}, Z_{t,j+1})$, we obtain

$$\begin{aligned} F'(x) &= \mathbb{E}[f'(x, Z_{t,j}, Z_{t,j+1})] \\ &= pG\, \mathbf{1}_{[a_{j^*+1},\, R]}(x) \\ &\quad - pG\, \mathbf{1}_{[0,\, a_{j^*})}(x) + pG\, g_{j^*}(x)\, \mathbf{1}_{[a_{j^*},\, a_{j^*+1})}(x). \end{aligned} \tag{121}$$

Integrating $F'(x)$, we have

$$\begin{aligned} F(x) &= pG\,(x - a_{j^*+1})\, \mathbf{1}_{[a_{j^*+1},\, R]}(x) \\ &\quad + pG\,(-x + a_{j^*})\, \mathbf{1}_{[0,\, a_{j^*})}(x) \\ &\quad + \left[pG\, \frac{\psi\big(\big|x - \frac{B}{2N} - a_{j^*}\big|\big)}{\psi'\big(\frac{R}{2N}\big)} - pG\, \frac{\psi\big(\frac{R}{2N}\big)}{\psi'\big(\frac{R}{2N}\big)} \right] \mathbf{1}_{[a_{j^*},\, a_{j^*+1})}(x) \end{aligned} \tag{122}$$

Note that by construction, $\min_{x \in \mathcal{X}} F(x) = pG\psi(R/2N)/\psi'(R/2N)$ and $a_{j^*} + R/(2N) =_{x \in \mathcal{X}} F(x)$. Function F is convex and its domain is bounded ($\mathcal{X} = [0, R]$). From Lemma 22, if

$$pG \geq \psi'(R/2N), \tag{123}$$

then F satisfies ϕ-KL property (Assumption 7) in the interval $U_{a_{j^*+R/2N}} = [a_{j^*}, a_{j^*+1})$. In the reduction, we need to show that if the output of a stochastic first-order method $\hat{x}$ satisfies $F(\hat{x}) - F^* \leq \varepsilon$, then j^* is identified (more precisely, $\hat{x} \in [a_{j^*}, a_{j^*+1})$). If

$$pG \frac{\psi(R/2N)}{\psi'(R/2N)} \geq \varepsilon, \tag{124}$$

for every $x \notin [a_{j^*}, a_{j^*+1})$, we get $F(x) - F^* \geq \varepsilon$. Indeed from the definition of the function (122), for every $x \notin [a_{j^*}, a_{j^*+1})$, we have

$$F(x) - F^* \geq pG \frac{\psi(R/2N)}{\psi'(R/2N)},$$

and if $pG\psi(R/2N)(\psi'(R/2N))^{-1} > \varepsilon$, we get $F(x) - F^* > \varepsilon$.

Let $p = (G\phi'(\varepsilon))^{-1}$ and $N = R(2\phi(\varepsilon))^{-1}$. Then both conditions (123) and (124) hold with equality. Therefore, the minimax oracle complexity in this case, can be lower bounded by $\Omega\left(p^2 \log N\right)$ which is

$$\Omega\left(G^2(\phi'(\varepsilon))^2 \log\left(\frac{R}{2\phi(\varepsilon)}\right)\right). \tag{125}$$

Lemma 22. *Function F defined in* (122) *satisfies ϕ-KL property when $pG \geq \psi'(B/2N)$.*

Proof. By using $\phi \equiv \psi^{-1}$ and the condition $pG \geq \psi'(R/2N)$, we have

$$\begin{aligned}
&\phi'(F(x) - \min_{x\in\mathcal{X}} F(x)) = (\psi^{-1})'(F(x) - \min_{x\in\mathcal{X}} F(x)) \\
&= (\psi^{-1})'\left(\psi(|x - \frac{R}{2N} - a_j|)\right) = \frac{1}{\psi'(|x - \frac{R}{2N} - a_j|)} \\
&\geq \frac{\psi'(\frac{R}{2N})}{pG \cdot \psi'(|x - \frac{R}{2N} - a_j|)} = \frac{1}{|F'(x)|}
\end{aligned} \tag{126}$$

□

I COMPARISON BETWEEN THEOREM 4.1 AND FOSTER ET AL. (2019)

Regarding Theorem 4.1, we used a similar approach (reduction to NBS problem) as [Foster et al., 2019]. Foster et al. [2019] use the reduction to NBS problem in order to derive a complexity lower bound for stochastic first-order methods converging to the approximate first-order stationary point in expectation $\mathbb{E}[\|\nabla F(\hat{x})\|] \leq \varepsilon$ over the class of convex smooth functions. The differences of our lower bound proof with Theorem 4.1 are:

- Foster et al. [2019] derived their lower bound to find the average first-order stationary point while we are using this approach to derive the lower bound to find the approximate minimizer in average, i.e., $\mathbb{E}[F(\hat{x})] - F^* \leq \varepsilon$. For convex objective functions, the complexity of finding approximate stationary points is different from the complexity of finding approximate minimizers. For example, [Foster et al., 2019] showed that while SGD is (worst-case) optimal for stochastic convex optimization for finding approximate minimizer, it appears to be far from optimal for finding near-stationary points (a version of SGD3 [Allen-Zhu, 2018] is optimal in this case).
- The gradient estimator in [Foster et al., 2019] is $f'(x, Z_{t,j}, Z_{t,j+1}) = -2\varepsilon\,\mathbf{1}_{(-\infty,0)}(x) + 2\varepsilon\,\mathbf{1}_{[R,\infty)}(x) + \sum_{j=1}^{N-1} \mathbf{1}_{[a_j,a_{j+1})}(x)\big(h_j(x)Z_{t,j+1} + (1-h_j(x))Z_{t,j}\big)$. where $h_j := (x - a_j)(R/N)^{-1}$. A naive approach to extend their construction to the case where the function satisfies local $(\alpha, \tau, \varepsilon)$-PŁ property (Assumption 6) is the straightforward replacement of $h_j(x)$ with $|x - a_j|^{1/(\alpha-1)}\mathrm{sgn}(x - a_j)(R/n)^{-1/(\alpha-1)}$. A drawback of this construction is that if the minimum of $f(x)$ is close to a_{j^*} and when the approximate minimizer of the function is in $[a_{j^*-1}, a_{j^*})$, then $[a_{j^*}, a_{j^*+1})$ is not identified and the reduction to NBS problem does not work. The solution is to use a version of $f'(x, Z_{t,j}, Z_{t,j+1})$ in (110) which has the following two properties: 1) the function satisfying local $(\alpha, \tau, \varepsilon)$-PŁ 2) Finding the approximate minimizer of this function uniquely identifies the interval $[a_{j^*}, a_{j^*+1})$.

J EXPERIMENT DETAILS: RELU AND DICTIONARY LEARNING

This appendix details the problem and protocols used in our two empirical studies in Section 5: (1) binary classification with a ReLU network, and (2) dictionary learning.

Code and dataset instructions for reproducing all our experiments are available at `GitHub`.

Our theory shows that the optimal local oracle complexity takes the form

$$\mathrm{poly}\big(\tau(\alpha)\big)\,\varepsilon^{-2/\alpha},$$

where $\tau(\alpha)$ is the local α–PŁ constant around the target component $\mathcal{M}$ and ε is the target accuracy. Thus, larger α improves the ε–exponent ($2/\alpha$ decreases) but may increase $\tau(\alpha)$; the best choice balances these two effects.

Empirically, we estimate

$$\tau(\alpha) \;:=\; \sup_{x\in\mathbb{B}_2(\hat{x};r)} \frac{F(x)-F(\hat{x})}{\|\nabla F(x)\|^{\alpha}},$$

and typically observe that $\tau(\alpha)$ is finite for all $\alpha \in [1,2]$ and fairly flat near $\alpha = 1$. (i) For very small ε, the term $\varepsilon^{-2/\alpha}$ dominates, so taking α closer to 2 is advantageous. (ii) For moderate ε, a flatter (often smaller) $\tau(\alpha)$ near $\alpha = 1$ can yield the smaller overall bound. This explains why we care about $\alpha \in [1,2]$ and motivates tuning α to minimize the predicted cost.

Let $\hat{x}$ be the terminal iterate returned by full-batch gradient descent (FB-GD). In all reported runs we have $\|\nabla F(\hat{x})\| \le 10^{-5}$, so $\hat{x}$ is an *almost critical point*. Fix a radius $r > 0$ and consider the ball $\mathbb{B}_2(\hat{x};r) := \{x : \|x-\hat{x}\|_2 \le r\}$ in parameter space. For $\alpha \in [1,2]$, define

$$\mathsf{R}_\alpha(x) := \frac{F(x)-F(\hat{x})}{\|\nabla F(x)\|^{\alpha}}, \qquad \tau(\alpha) = \sup_{x\in\mathbb{B}_2(\hat{x};r)} \mathsf{R}_\alpha(x),$$

so that $F(x) - F(\hat{x}) \le \tau(\alpha)\,\|\nabla F(x)\|^{\alpha}$ on $\mathbb{B}_2(\hat{x};r)$. Because $\hat{x}$ is only an almost critical point, these plots should be interpreted as empirical envelopes around $\hat{x}$ rather than a verification of Assumption 2 for an isolated component of local minimizers. Empirically, we approximate the supremum by sampling m points $\{x_i\}_{i=1}^m \subset \mathbb{B}_2(\hat{x};r)$ (uniform in direction, radius drawn with density proportional to r^{d-1}) and compute

$$\widehat{\tau}(\alpha) = \max_{1\le i\le m} \frac{F(x_i)-F(\hat{x})}{\|\nabla F(x_i)\|^{\alpha}}.$$

In Figure 2 we use $m = 10^5$ probes for each radius $r \in \{0.5, 0.05, 0.005\}$; in Figure 3 we use $(r,m) = (0.35, 450)$ for MNIST and $(r,m) = (0.40, 600)$ for WDBC. In the plots, we also report the baseline $10\widehat{\tau}(1)$ and highlight in green the set $\{\alpha \in [1,2] : \widehat{\tau}(\alpha) \le 10\widehat{\tau}(1)\}$, which visualizes the near-flat regime around $\alpha = 1$.

J.1 Binary classification with a one-hidden-layer ReLU network

Data. Two-dimensional synthetic binary classification with four Gaussian blobs (means at $(\pm 1.5, \pm 1.5)$, isotropic covariance $0.35I$); labels form an XOR pattern.

Model and loss. A 1-hidden-layer ReLU network with m hidden units: $x \mapsto a = \mathrm{ReLU}(xW_1 + b_1) \mapsto \ell = a^\top W_2 + b_2 \mapsto p = \sigma(\ell)$. Loss F is the average binary cross-entropy, optionally with ℓ_2 weight decay $\frac{\lambda}{2}(\|W_1\|_F^2 + \|W_2\|_2^2)$.

Training. FB-GD with fixed step size; stop either at a gradient-norm threshold or a max-iteration budget; set $\hat{x}$ to the final iterate.

Neighborhood sampling and estimation. Form $\mathbb{B}_2(\hat{x};R)$ in *parameter space* and draw m probing parameters. Compute gaps $\Delta_i := F(x_i) - F(\hat{x})$ and gradient norms $g_i := \|\nabla F(x_i)\|$. Evaluate $\widehat{\tau}(\alpha)$ on a grid $\alpha \in [1,2]$, plot $\widehat{\tau}(\alpha)$ together with $10\widehat{\tau}(1)$, and highlight the set $\{\alpha : \widehat{\tau}(\alpha) \le 10\widehat{\tau}(1)\}$.

Figure 2 reports the curves $\widehat{\tau}(\alpha)$ for radii $r \in \{0.5,\ 0.05,\ 0.005\}$. By definition, $\widehat{\tau}(\alpha)$ is a (sampled) supremum over the ball $B_2(x^\star;r)$, hence it is *monotone in* r. The radius r also selects the *local geometry* being used. Smaller r isolates the immediate neighborhood of $x^\star$, where the landscape is closer to a single regime and may admit a *smaller constant* and a *wider near-flat regime* around $\alpha = 1$. As $r \downarrow 0$, the curve reflects the most local behavior around $x^\star$ and the set $\{\alpha : \widehat{\tau}(\alpha) \le 10\widehat{\tau}(1)\}$ typically expands.

Real-data binary classification. We additionally repeat the same procedure on two standard real datasets: MNIST (digits 0 vs. 1) and the UCI Breast Cancer Wisconsin Diagnostic dataset (WDBC; malignant vs. benign). For both datasets, we fit the same one-hidden-layer ReLU network by full-batch gradient descent to obtain an almost critical point $\hat{x}$, then probe a small ball $\mathbb{B}_2(\hat{x};r)$ in parameter space by uniform sampling and compute $\widehat{\tau}(\alpha)$ over $\alpha \in [1,2]$. Figure 3 shows the resulting envelopes.

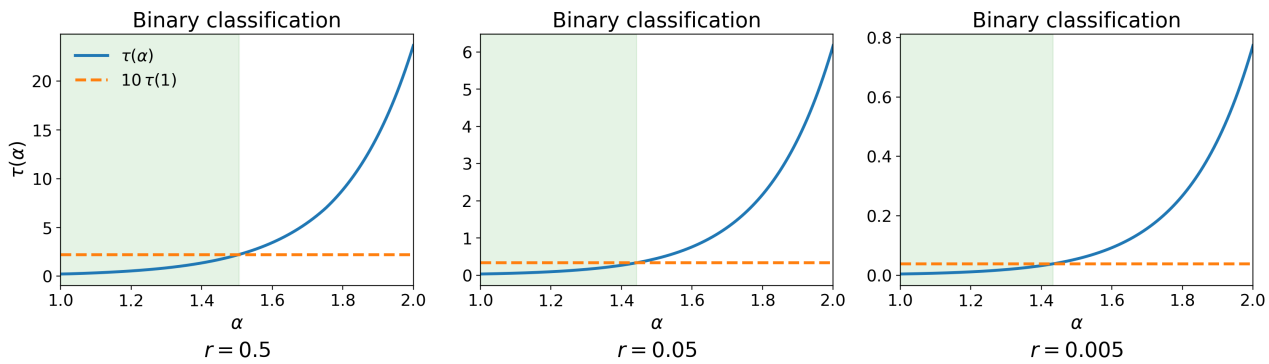

Figure 2: **Synthetic ReLU classification (XOR).** Weight decay $\lambda = 0$. Variation of $\tau(\alpha)$ over $\alpha \in [1, 2]$ for three values of r. The dashed line is $10\tau(1)$ and the green region is $\{\alpha : \tau(\alpha) \leq 10\tau(1)\}$. Ranges: $r = 0.5$, $\tau(\alpha) : [1, 2] \to [0.221, 23.67]$; $r = 0.05$, $\tau(\alpha) : [1, 2] \to [0.0336, 6.17]$; $r = 0.005$, $\tau(\alpha) : [1, 2] \to [0.00382, 0.773]$.

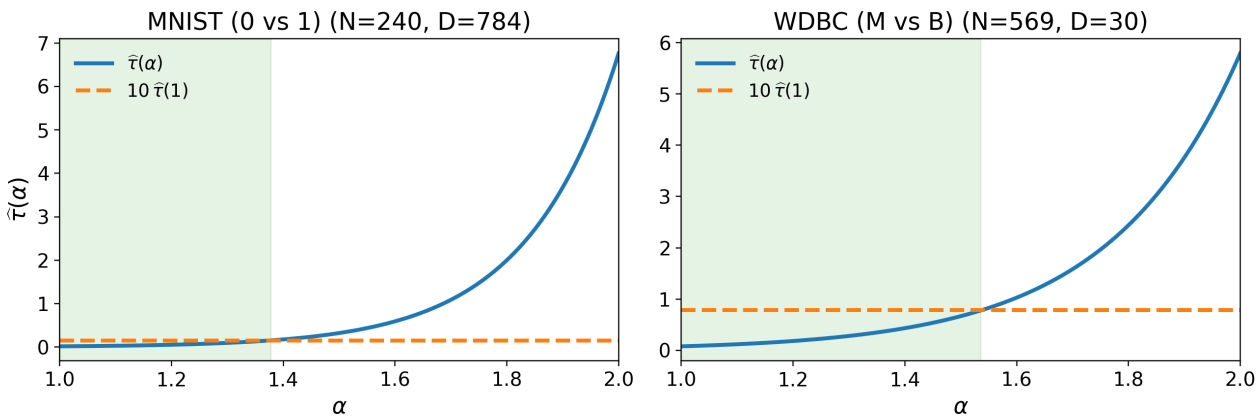

Figure 3: **Real-data ReLU classification.** Weight decay $\lambda = 10^{-4}$. Estimated envelopes $\alpha \mapsto \widehat{\tau}(\alpha)$ for MNIST (0 vs. 1) and WDBC (malignant vs. benign). The dashed line is $10\widehat{\tau}(1)$ and the green region is $\{\alpha : \widehat{\tau}(\alpha) \leq 10\widehat{\tau}(1)\}$. Ranges of $\widehat{\tau}(\alpha)$ over $\alpha \in [1, 2]$: MNIST [0.015, 6.76], WDBC [0.078, 5.79].

J.2 Dictionary learning with smooth sparsity

Data. Generate a ground-truth dictionary $D_0 \in \mathbb{R}^{p\times k}$ with unit-norm columns; draw sparse codes $A_0 \in \mathbb{R}^{k\times n}$ with s nonzeros per column; form $X = D_0A_0 + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2 I)$.

Objective. We minimize

$$F(D, A) = \frac{1}{2n}\|DA - X\|_F^2 \;+\; \lambda_1 \sum_{ij} \phi_\delta(A_{ij}) + \;\frac{\lambda_2}{2}\|A\|_F^2 \;+\; \frac{\mu}{4}\sum_{j=1}^{k}\left(\|d_j\|_2^2 - 1\right)^2, \tag{127}$$

where ϕ_δ is the pseudo-Huber penalty $\phi_\delta(a) = \delta^2(\sqrt{1 + (a/\delta)^2} - 1)$ with derivative $\phi'_\delta(a) = a/\sqrt{1 + (a/\delta)^2}$. The last term softly enforces unit-norm atoms and removes the $D-A$ scaling ambiguity.

Training. FB-GD with backtracking (Armijo) on (D, A) until a small gradient norm is reached; set $\hat{x} = (\hat{D}, \hat{A})$.

Neighborhood sampling and estimation. Sample $\mathbb{B}_2(\hat{x}; R)$ in the *joint* (D, A) parameter space by perturbing $\hat{D}$ and $\hat{A}$ together. In the main-body dictionary learning experiment (Figure 1, bottom) we use $R = 0.1$

and $m = 2500$ probe points. Compute Δ_i and g_i as above and evaluate $\widehat{\tau}(\alpha)$ on $\alpha \in [1, 2]$.

K ADDITIONAL EXPERIMENTS: POLYNOMIAL MATRIX FACTORIZATION AND DEEP LINEAR FACTORIZATION

In both experiments we validate Assumption 2 (local α–PŁ) by estimating, on a bounded neighborhood of an almost–critical point,

$$R_\alpha(\theta) \;:=\; \frac{F(\theta) - F(\theta^\star)}{\|\nabla F(\theta)\|^\alpha}, \qquad \tau(\alpha) \;:=\; \sup_{\theta \in B_2(\theta^\star; r)} R_\alpha(\theta), \quad \alpha \in [1, 2],$$

where $\theta^\star$ is the terminal iterate of full-batch gradient descent (FB-GD), hence an *almost* critical point ($\|\nabla F(\theta^\star)\|$ small). In the plots below, we visualize the envelope $\alpha \mapsto \tau(\alpha)$ and highlight in green the range $\{\alpha : \tau(\alpha) \le 10\tau(1)\}$ to make the "flat" regime around $\alpha = 1$ directly comparable across setups.

K.1 Polynomial Matrix Factorization

Problem. Given $X \in \mathbb{R}^{p \times n}$ and a target rank r, factorize $X \approx UV$ with

$$F(U, V) \;=\; \frac{1}{2n}\,\|UV - X\|_F^2 \;+\; \frac{\lambda}{2}\big(\|U\|_F^2 + \|V\|_F^2\big), \quad U \in \mathbb{R}^{p \times r},\ V \in \mathbb{R}^{r \times n}.$$

This is a *polynomial* loss, thus fits our tame/definable setting.

Synthetic setup. We generate $X = U_0 V_0 + \xi$ with $U_0 \in \mathbb{R}^{p \times r}$, $V_0 \in \mathbb{R}^{r \times n}$ i.i.d. Gaussian and $\xi \sim \mathcal{N}(0, \sigma^2 I)$; optionally, we inject a small fraction of *spikes* (large entries) to test robustness.[16] We run FB-GD with backtracking to obtain $\theta^\star = (U^\star, V^\star)$ and estimate $\widehat{\tau}(\alpha)$ by probing a neighborhood around $\theta^\star$ (radius $R = 0.05$, $m = 1500$ probes).

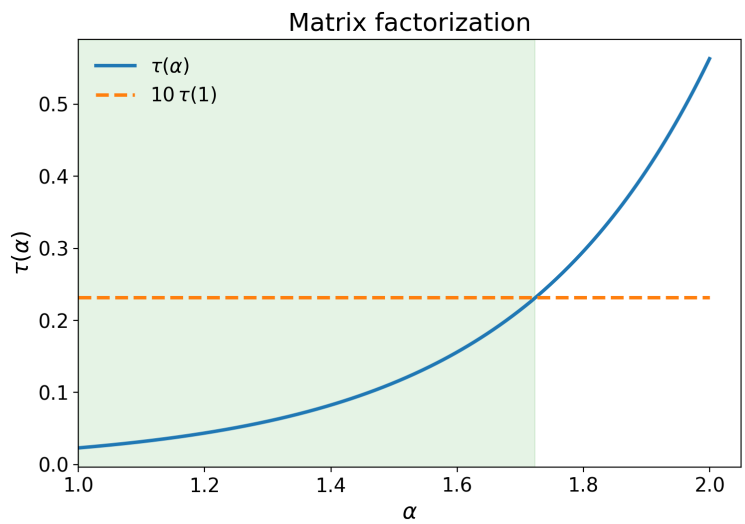

Figure 4: **Synthetic polynomial MF.** $\lambda = 10^{-3}$, $R = 0.05$, $m = 1500$. Variation of $\widehat{\tau}(\alpha)$ for $\alpha \in [1, 2]$ (synthetic $X = U_0 V_0 + \xi$). Range of $\widehat{\tau}(\alpha)$ over $\alpha \in [1, 2]$: [0.023, 0.563].

Explanation. In Figure 4 we plot the empirical envelope $\widehat{\tau}(\alpha) = \max_i \Delta_i / g_i^\alpha$ on a fixed-radius neighborhood around the terminal iterate $\theta^\star$. The curve is relatively flat for α close to 1 and then increases as α approaches 2, indicating that (at this radius) near-linear domination in $\|\nabla F\|$ provides a smaller constant than near-quadratic domination.

[16] We use a spike fraction s and a multiplier κ relative to $\mathrm{std}(X)$; both are reported with the plots.

Setup and estimation. We fit by FB-GD with backtracking to obtain an almost critical point $\theta^\star = (U^\star, V^\star)$, then probe a ball $B_2(\theta^\star; R)$ by uniform sampling. For each probe θ_i we compute the optimality gap $\Delta_i := F(\theta_i) - F(\theta^\star)$ and gradient norm $g_i := \|\nabla F(\theta_i)\|$. For a grid $\alpha \in [1, 2]$ we estimate

$$\widehat{\tau}(\alpha) \;=\; \max_i \frac{\Delta_i}{g_i^\alpha}.$$

We report the curve $\alpha \mapsto \widehat{\tau}(\alpha)$ and highlight $\{\alpha : \widehat{\tau}(\alpha) \le 10\widehat{\tau}(1)\}$ (for the real-data run below: $R = 0.006$, $m = 600$ probes in each subplot).

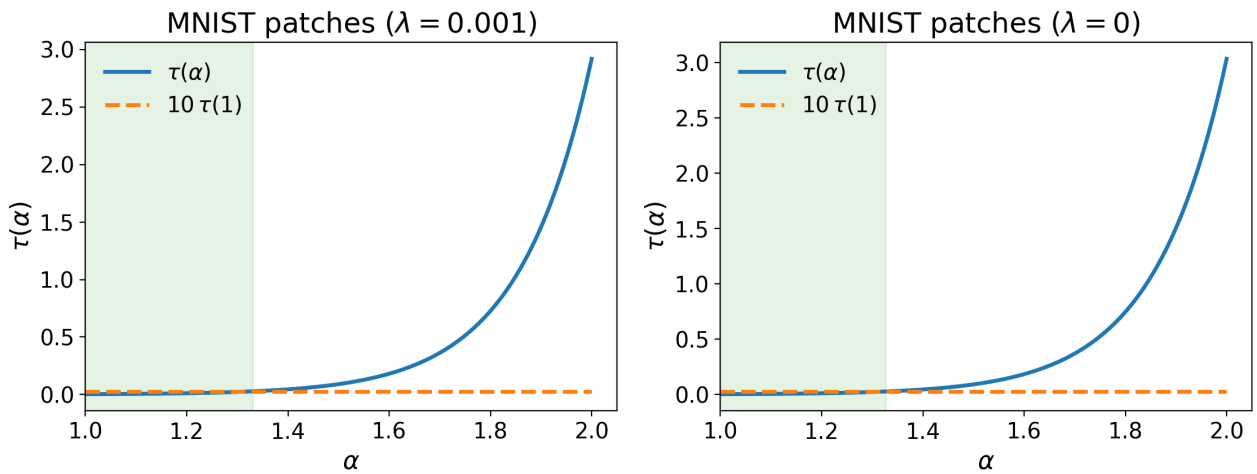

Figure 5: **Real-data polynomial MF (MNIST patches).** $R = 0.006$, $m = 600$. Estimated envelopes $\alpha \mapsto \tau(\alpha)$ for two settings: regularized ($\lambda = 10^{-3}$; range [0.0028, 2.919]) and unregularized ($\lambda = 0$; range [0.0028, 3.034]). The green region is $\{\alpha : \tau(\alpha) \le 10\tau(1)\}$, mirroring the synthetic plots.

Explanation. In Figure 5 the two subplots compare how ridge regularization affects the local envelope around an almost-critical point fitted on MNIST patches. Both regularized and unregularized runs exhibit a broad "flat" regime near $\alpha = 1$ (green band), while $\tau(\alpha)$ grows with α and is largest near $\alpha = 2$. The similarity of the curves here suggests that, at this radius, the local PL-type behavior is driven primarily by the data-fit term rather than the ridge term.

K.2 Deep Linear Matrix Factorization

Problem. Approximate $X \in \mathbb{R}^{p\times n}$ by a product of L linear factors

$$X \approx W_0 W_1 \cdots W_{L-1}, \qquad F(W_0, \ldots, W_{L-1}) \;=\; \frac{1}{2n}\left\|W_0 \cdots W_{L-1} - X\right\|_F^2 + \frac{\lambda}{2}\sum_{\ell=0}^{L-1}\|W_\ell\|_F^2,$$

with $W_0 \in \mathbb{R}^{p\times r_1}$, $W_\ell \in \mathbb{R}^{r_\ell \times r_{\ell+1}}$, and $W_{L-1} \in \mathbb{R}^{r_{L-1}\times n}$. This loss is a *polynomial of degree* $2L$ in the parameters.

Setup. We train by FB-GD (with backtracking) to obtain $\theta^\star = (W_0^\star, \ldots, W_{L-1}^\star)$ and probe $B_2(\theta^\star; R)$ as above, producing pairs (Δ_i, g_i). We compute the empirical envelope

$$\widehat{\tau}(\alpha) \;=\; \max_i \frac{\Delta_i}{g_i^\alpha}, \qquad \alpha \in [1, 2].$$

Synthetic setup. We also report the same envelope estimation for a synthetic instance (low-rank ground truth plus Gaussian noise), using $R = 0.002$ and $m = 2000$ probes:

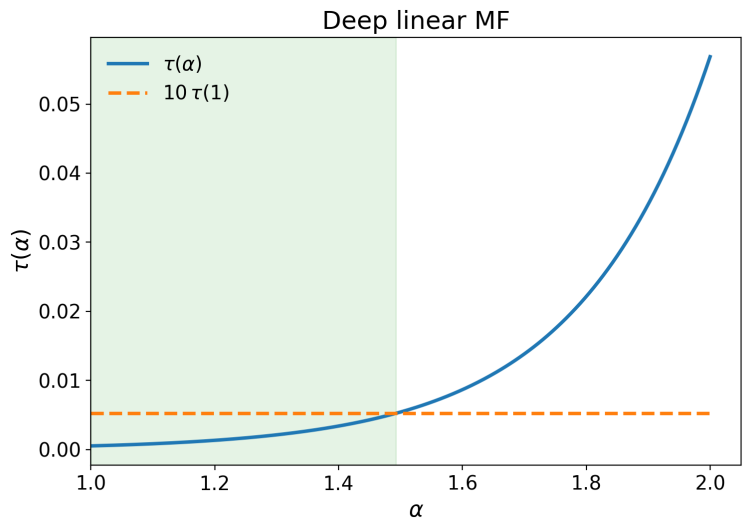

Figure 6: **Synthetic deep linear MF.** $\lambda = 10^{-3}$, $R = 0.002$, $m = 2000$, $L = 4$. Variation of $\widehat{\tau}(\alpha)$ over $\alpha \in [1, 2]$ (synthetic X). Range of $\widehat{\tau}(\alpha)$ over $\alpha \in [1, 2]$: $[5.3 \times 10^{-4}, 0.0569]$.

Explanation. In Figure 6, on a synthetic instance, the estimated envelope stays small for α close to 1 and increases as $\alpha \uparrow 2$. This reflects the fact that, even for polynomial objectives, the best local domination exponent can be strictly less than 2 on a fixed neighborhood, and the constant for $\alpha = 2$ can be much larger than for $\alpha \approx 1$.

Real-data deep linear MF. We repeat the same procedure on a real dataset by extracting 8×8 patches from MNIST training images and treating them as columns of X. We compare a regularized run ($\lambda > 0$) to an unregularized run with $\lambda = 0$, using $R = 0.001$ and $m = 700$ probes in each subplot.

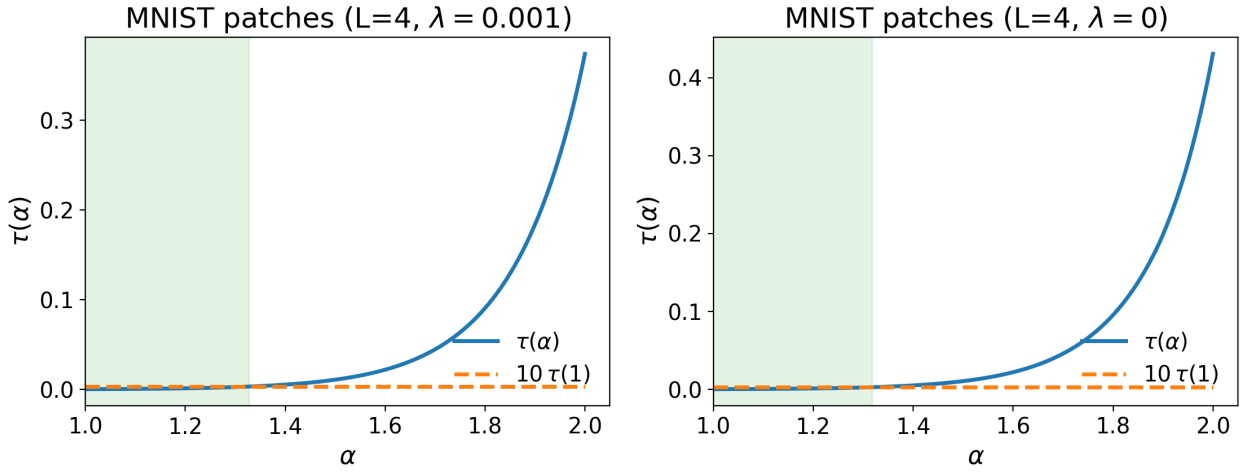

Figure 7: **Real-data deep linear MF (MNIST patches).** $R = 0.001$, $m = 700$, $L = 4$. Estimated envelopes $\alpha \mapsto \tau(\alpha)$ for a regularized run ($\lambda = 10^{-3}$; range $[3.3 \times 10^{-4}, 0.374]$) and an unregularized run ($\lambda = 0$; range $[2.9 \times 10^{-4}, 0.431]$). The green region is $\{\alpha : \tau(\alpha) \leq 10\tau(1)\}$, matching the synthetic plot convention.

Explanation. In Figure 7 we repeat the envelope computation on MNIST patches. The green region highlights the near-$\alpha = 1$ regime where $\tau(\alpha) \leq 10\tau(1)$ and the curve is comparatively flat, while the envelope grows

sharply as α approaches 2. The $\lambda = 0$ run has a slightly larger maximum envelope over $\alpha \in [1, 2]$, consistent with regularization improving local conditioning in the deep parameterization.

Degenerate critical point (unregularized). For depth $L \geq 3$ and $\lambda = 0$, the trivial point $\theta^\star = 0$ is a stationary point of deep linear MF, and the local geometry can be much flatter than quadratic. We estimate the envelope around $\theta^\star = 0$ on the same MNIST-patch matrix using $R = 0.005$ and $m = 8000$ probe points:

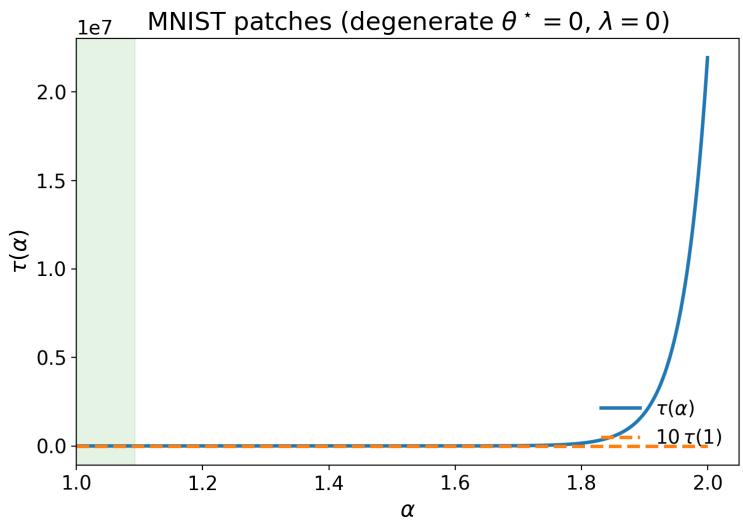

Figure 8: **Real-data deep linear MF around the degenerate point $\theta^\star = 0$ (MNIST patches).** $\lambda = 0$, $R = 0.005$, $m = 8000$, $L = 4$. Estimated envelope $\alpha \mapsto \tau(\alpha)$ around $\theta^\star = 0$. Range of $\tau(\alpha)$ over $\alpha \in [1, 2]$: $[3.9 \times 10^{-4}, 2.2 \times 10^{7}]$.

Explanation. In Figure 8 we intentionally probe around a *degenerate* stationary point. The green region marks the small "flat" regime near $\alpha = 1$ (where $\tau(\alpha) \leq 10\tau(1)$), but the envelope then grows extremely rapidly, yielding a very large $\tau(2)$. This illustrates that, without regularization, deep linear MF can have regions where a quadratic ($\alpha = 2$) PL-type inequality holds only with an enormous constant (or effectively fails to be informative), even though smaller exponents around $\alpha \simeq 1$ can still yield reasonable constants on the same neighborhood.