Tight Heavy-Tailed Complexity Bounds for Optimizing Polyak–Łojasiewicz Objectives

Anonymous authors
Paper under double-blind review

Abstract

We study optimizing L-smooth and μ-Polyak–Łojasiewicz (PL) objectives with unbiased stochastic gradients under α-heavy-tailed noise ($1 < \alpha \leq 2$). In the unrestricted **high-dimensional** setting, we establish the noise-adaptive lower bound for attaining ϵ-suboptimal function value. Under the generalized mirror-PL condition, we propose a centered-clipped mirror-descent method that achieves a high-probability upper bound matching the lower bound up to logarithmic factors. We further provide the improved tight stochastic complexity in **fixed** dimensions.

1 Introduction

The Polyak–Łojasiewicz inequality provides an intermediate model between strongly convex and fully nonconvex optimization. For a differentiable objective f, it requires $\|\nabla f(x)\|^2 \geq 2\mu\big(f(x) - f^\star\big)$ for some $\mu > 0$, without requiring f itself to be convex. Nevertheless, when f is also L-smooth, this inequality is sufficient for gradient descent to converge globally at a linear rate governed by the condition number $\kappa = L/\mu$ (Polyak, 1963a; Karimi et al., 2016a). The PL framework therefore captures a broad class of objectives that retain a global optimization guarantee even though their geometry may be nonconvex (Hardt & Ma, 2016; Li et al., 2018; Liu et al., 2022; Allen-Zhu et al., 2019). Understanding the oracle complexity of PL optimization is consequently important both for identifying the limits of first-order algorithms and for explaining which parts of strongly convex optimization theory continue to hold under the weaker gradient-dominance assumption.

In stochastic optimization, however, the geometry of the objective is only one source of difficulty. Classical analyses often assume bounded variance or sub-Gaussian gradient noise, while empirical studies of stochastic gradients in modern learning problems indicate that their distributions can be strongly non-Gaussian and heavy-tailed (Şimşekli et al., 2019). A natural model assumes conditional unbiasedness and a finite centered α-moment for each stochastic gradient G_t:

$$\mathbb{E}[G_t \mid \mathscr{F}_{t-1}] = \nabla f(x_t), \qquad \mathbb{E}[\|G_t - \nabla f(x_t)\|^\alpha \mid \mathscr{F}_{t-1}] \leq \sigma^\alpha, \qquad 1 < \alpha \leq 2.$$

For $\alpha < 2$, this assumption permits infinite variance; empirical averages need not exhibit Gaussian concentration, and rare but arbitrarily large oracle errors cannot be neglected. Gradient clipping, normalization, and robust mean estimation have consequently become central tools for obtaining high-probability guarantees under heavy-tailed noise (Cutkosky & Mehta, 2021a; Sadiev et al., 2023; Puchkin et al., 2024). Yet clipping a stochastic gradient around the origin can introduce a bias depending on $\|\nabla f(x)\|$, which may be unbounded on an unbounded domain. This makes it nontrivial to obtain a high-probability theory that is simultaneously robust to heavy tails, compatible with mirror geometry, and free of bounded-gradient and bounded-domain assumptions.

A second challenge concerns optimality. In the deterministic high-dimensional setting, smooth PL optimization has the lower-bound scale $\Omega(\kappa \log(\Delta_0/\epsilon))$, where Δ_0 bounds the initial objective gap. This shows that the general PL class does not admit the $\sqrt{\kappa}$ acceleration available under strong convexity (Yue et al., 2023a). Stochastic lower bounds based on probabilistic zero-chains and random rotations provide a powerful way to hide sequential information from arbitrary first-order algorithms (Arjevani et al., 2023a). For *heavy-tailed*

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

noise, however, a complete lower bound must capture two different obstructions at once: the deterministic optimization transient and the statistical cost of estimating a mean under only a finite α-moment. Moreover, it should be *noise-adaptive*: when $\sigma = 0$, the stochastic term vanishes, but the lower bound should reduce to the correct exact-gradient complexity rather than becoming trivial.

The role of dimension creates a third difficulty. Randomly rotated zero-chains require a dimension that grows with the number of hidden directions, whereas low-dimensional optimization may permit geometric search strategies unavailable in high dimension. Gradient-flow trapping, for example, exploits separating curves or hypersurfaces to locate approximate stationary points in fixed dimension (Bubeck & Mikulincer, 2020a). Extending this idea to the present stochastic setting is delicate because the algorithm has access only to noisy gradients, rather than exact function values on an entire separating surface. It is therefore not clear a priori whether the condition-number factor in the high-dimensional stochastic complexity remains unavoidable when d is fixed, or how the answer depends jointly on the dimension and the tail exponent.

Motivated by these issues, this paper asks the following question:

What is the optimal first-order complexity of smooth PL optimization as a function of κ, σ, ϵ, the moment order α, and the dimension d, when the guarantee must hold with high probability and remain informative in the noiseless limit?

1.1 Our Contributions

Our results separate the roles of geometry, noise, and dimension. For brevity, write $\theta_\alpha := \alpha/[2(\alpha-1)]$ and $p_\alpha := \alpha/(\alpha-1)$; Section 2 gives the full notation and query model.

- **A noise-adaptive high-dimensional lower bound.** For arbitrary adaptive randomized gradient-only algorithms, sufficiently large κ, and sufficiently small ϵ/Δ_0, we establish $\Omega\big(\kappa\log(\Delta_0/\epsilon) + \kappa(\sigma^2/(\mu\epsilon))^{\theta_\alpha}\big)$ in a sufficiently large finite dimension. The two terms capture the deterministic transient and the heavy-tailed estimation cost. In particular, the bound recovers $\Omega(\kappa\log(\Delta_0/\epsilon))$ when $\sigma = 0$ (Theorem 3.1).
- **A matching high-probability upper bound.** For a nonempty closed convex domain, possibly unbounded, we combine centered clipping with mirror descent. Under mirror-PL, the method returns an ϵ-suboptimal point with probability at least $1-\delta$ using $\widetilde{O}\big(\kappa\log(\Delta_0/\epsilon) + \kappa(\sigma^2/(\mu\epsilon))^{\theta_\alpha}\big)$ stochastic-gradient calls (Theorem 4.3). No bounded-gradient or sub-Gaussian assumption is needed. On the unconstrained Euclidean specialization, this upper bound matches the lower bound up to logarithmic factors.
- **Improved complexity in fixed dimensions.** For every fixed $d \geq 1$ and $\kappa \geq 6$, we prove the lower bound $\Omega(\log(\Delta_0/\epsilon) + (\sigma^2/(\mu\epsilon))^{\theta_\alpha})$ in the stated accuracy range (Theorem 5.1). Robust stochastic bisection matches this bound up to logarithmic factors for $d = 1$. Gradient-flow trapping gives upper bounds $\widetilde{O}(\kappa^{2/3}\log(\Delta_0/\epsilon) + (\sigma^2/(\mu\epsilon))^{\theta_\alpha})$ for $d = 2$ and $\widetilde{O}(\kappa\log(\Delta_0/\epsilon) + (\sigma^2/(\mu\epsilon))^{\theta_\alpha})$ for $d = 3$. Thus the stochastic term is matched in both dimensions. For fixed $d > 3$, the same stochastic characterization holds when $p_\alpha \geq d-1$; the deterministic term becomes $\widetilde{O}(\kappa^{(d-1)/2}\log(\Delta_0/\epsilon))$. Theorem 5.2 also states the remaining gap when $p_\alpha < d-1$.
- **Ordinary PL versus constrained mirror-PL.** We construct a smooth objective satisfying ordinary PL on an unbounded convex domain for which exact mirror descent remains at a nonoptimal boundary point (Proposition 4.2). This shows why constrained mirror descent requires a condition on feasible descent, rather than the full gradient norm. In the unconstrained Euclidean case, mirror-PL reduces to ordinary PL.

Comparison with Masiha et al. (2026). That work studies lower bounds under a sublevel geometric PL condition with exponent $\nu \in [1,2)$, namely $f - f^\star \leq \tau\|\nabla f\|^\nu$, while retaining global smoothness and using bounded-variance oracles. The geometric exponent ν is distinct from the noise-moment order α used throughout our paper. Our results concern the

Regime	Lower bound	High-prob. upper bound	Comparison
Growing d	$\Omega\big(\kappa A_\epsilon + \kappa Q_\epsilon\big)$	$\widetilde{O}\big(\kappa A_\epsilon + \kappa Q_\epsilon\big)$	Matched
$d = 1$	$\Omega\big(A_\epsilon + Q_\epsilon\big)$	$\widetilde{O}\big(A_\epsilon + Q_\epsilon\big)$	Matched
$d = 2$	$\Omega\big(A_\epsilon + Q_\epsilon\big)$	$\widetilde{O}\big(\kappa^{2/3} A_\epsilon + Q_\epsilon\big)$	Stochastic term matched
$d = 3$	$\Omega\big(A_\epsilon + Q_\epsilon\big)$	$\widetilde{O}\big(\kappa A_\epsilon + Q_\epsilon\big)$	Stochastic term matched
Fixed $d > 3$, $p_\alpha \geq d - 1$	$\Omega\big(A_\epsilon + Q_\epsilon\big)$	$\widetilde{O}\big(\kappa^{(d-1)/2} A_\epsilon + Q_\epsilon\big)$	Stochastic term matched

Table 1: Complexity bounds in the stated condition-number and accuracy ranges; matching is up to logarithmic factors. Upper bounds hold with probability at least $1 - \delta$, and lower bounds hold in expectation and at constant confidence. In the last row, one may take the minimum with $\widetilde{O}(\kappa[A_\epsilon + Q_\epsilon])$. The quantities A_ϵ, Q_ϵ, and p_α are defined in (1)–(2).

classical global PL geometry and finite centered α-moments, together with high-probability algorithms and fixed-dimensional refinements. Appendix A compares the stated exponent ranges and explains the distinction between formal endpoint substitution and a theorem valid at that endpoint.

1.2 Related Work

Polyak–Łojasiewicz (PL) condition. The PL condition originates in the work of Polyak (1963b) and Lojasiewicz (1963). It relaxes strong convexity and has been used in several optimization settings, including minimax problems (Guo et al., 2020; 2023). Other related conditions include error bounds (Luo & Tseng, 1993), restricted secant inequalities (Zhang & Yin, 2013), quadratic growth (Anitescu, 2000), essential strong convexity (Liu et al., 2014), and weak strong convexity (Necoara et al., 2019); see Guille-Escuret et al. (2021) for a comparison. Gradient descent converges linearly under PL (Polyak, 1963b), and Yue et al. (2023b) established a tight deterministic lower bound. Karimi et al. (2016b) extended the framework to constrained optimization through proximal PL. In the stochastic setting, Khaled & Richtárik (2020) and Li et al. (2021) obtained expected oracle-complexity bounds with stochastic term $O(L\sigma^2/(\mu^2\epsilon))$, where σ^2 bounds the noise variance.

Heavy-tailed stochastic optimization. Empirical evidence of non-Gaussian stochastic gradients motivates analysis beyond finite-variance and sub-Gaussian models (Şimşekli et al., 2019). Gradient clipping yields high-probability guarantees for smooth convex and strongly convex optimization, including accelerated methods, without light-tail assumptions (Gorbunov et al., 2020a). Plain SGD has also been analyzed under specialized state-dependent infinite-variance models for strongly convex objectives (Wang et al., 2021). For stochastic convex optimization with a finite moment strictly larger than one, Vural et al. (2022a) established minimax-optimal mirror-descent bounds without explicit clipping or normalization. For smooth nonconvex objectives, Cutkosky & Mehta (2021a) combined clipping, momentum, and normalized steps, while Liu et al. (2023b) showed how additional sample-function structure can improve on unstructured-oracle rates. Sadiev et al. (2023) developed high-probability results under bounded centered α-moments for nonconvex, PL, convex, and strongly convex optimization, as well as variational inequalities. Further developments include general Bregman geometries (Fatkhullin & He, 2024), clipping stochastic gradient differences (Gorbunov et al., 2024), smoothed median-of-means estimation (Puchkin et al., 2024), and clipped adaptive methods (Chezhegov et al., 2025). Fradin et al. (2026) established lower and upper bounds under generalized stochastic smoothness and developed double-clipped methods with high-probability guarantees.

Lower bounds. Classical deterministic first-order lower bounds address convex optimization (Nesterov, 2013) and nonconvex stationarity (Carmon et al., 2020). Stochastic lower bounds were developed for convex objectives by Agarwal et al. (2009); Raginsky & Rakhlin

(2011), extended to heavy-tailed noise by Vural et al. (2022b), and established for nonconvex stationarity by Arjevani et al. (2023b). Saad et al. (2025) used divergence decomposition for structured nonconvex classes, including objectives satisfying a restricted secant inequality. Other settings include finite-sum optimization (Fang et al., 2018; Zhou & Gu, 2019) and zeroth-order oracles (Duchi et al., 2015). For PL-type objectives, Yue et al. (2023b) treated the deterministic case, Bai et al. (2024) studied finite-sum complexity, and Masiha et al. (2026) considered the sublevel geometric condition discussed above. Our lower bounds instead retain classical global PL geometry while accounting for both finite-moment noise and the noiseless transient.

High-probability convergence. Nemirovski et al. (2009) analyzed stochastic mirror descent under sub-Gaussian noise. Gorbunov et al. (2020b; 2021) weakened the noise assumptions in smooth and nonsmooth settings. Beyond convexity, Li & Orabona (2020); Cutkosky & Mehta (2021b) studied high-probability guarantees for momentum-based methods, with the latter using clipping for heavy-tailed noise. Liu et al. (2023a) developed further stochastic mirror-descent analyses. Adaptive methods have been studied with AdaGrad stepsizes (Kavis et al., 2022; Attia & Koren, 2023) and with Adam (Hong & Lin, 2024); Liu et al. (2024) treated clipping for nonsmooth nonconvex optimization under heavy-tailed noise. Our focus is the joint dependence on PL geometry, the noise moment, and dimension, with bounds that also retain the noiseless optimization cost.

2 Preliminaries

Notation. All norms and inner products are Euclidean, and $\mathrm{Lip}(h)$ denotes the Lipschitz constant of a map h. Throughout, $L, \mu, \Delta_0 > 0$, $\sigma \geq 0$, and $1 < \alpha \leq 2$. Constants c_α, C_α depend only on α; subscripts d or (d, α) indicate analogous dependence. Unnamed constants may change between bounds. The notation $\widetilde{O}$ and $\widetilde{\Theta}$ suppresses logarithmic factors; their constants may depend on α and, in fixed-dimensional results, on d. Unless a base is specified, log is the natural logarithm. We use

$$\theta_\alpha := \frac{\alpha}{2(\alpha-1)}, \qquad \beta_\alpha := \frac{2(\alpha-1)}{\alpha} = \frac{1}{\theta_\alpha}, \qquad p_\alpha := \frac{\alpha}{\alpha-1} = 2\theta_\alpha, \tag{1}$$

and, for noise level σ, PL parameter μ, initial-gap bound Δ_0, and target accuracy ϵ, define

$$Q_\epsilon := \left(\frac{\sigma^2}{\mu\epsilon}\right)^{\theta_\alpha}, \qquad A_\epsilon := \log\frac{\Delta_0}{\epsilon}. \tag{2}$$

Problem setup. We consider $\min_{x\in\mathcal{X}} f(x)$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is nonempty and convex, but need not be bounded, and f is continuously differentiable and possibly nonconvex. Write $f^\star := \inf_{x\in\mathcal{X}} f(x) > -\infty$ and $\mathcal{X}^\star := \arg\min_{x\in\mathcal{X}} f(x)$. Our smoothness and geometric assumptions are as follows.

Definition 2.1 (L-smoothness)**.** The function f is L-smooth if, for all $x, y \in \mathcal{X}$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|. \tag{3}$$

Definition 2.2 (μ-PL condition)**.** The function f satisfies the μ-Polyak–Łojasiewicz (PL) condition if

$$\|\nabla f(x)\|^2 \geq 2\mu\big(f(x) - f^\star\big) \qquad (x \in \mathcal{X}). \tag{4}$$

Remark 2.3. When $\mathcal{X} = \mathbb{R}^d$, L-smoothness also gives $\|\nabla f(x)\|^2 \leq 2L(f(x) - f^\star)$ by the descent inequality. For an L-smooth, μ-PL objective, write $\kappa := L/\mu$.

Oracle model and adaptive queries. At physical oracle call $t \geq 1$, an adaptive randomized algorithm chooses x_t and receives $G_t = G(x_t, \xi_t)$. The pre-call history $\mathscr{F}_{t-1}$ includes previous queries and responses, the algorithm's randomness used so far, and any randomness used to choose x_t. Thus x_t is $\mathscr{F}_{t-1}$-measurable before the current oracle noise is drawn. Write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathscr{F}_{t-1}]$.

Definition 2.4 (Heavy-tailed stochastic-gradient oracle)**.** For $1 < \alpha \le 2$ and $\sigma \ge 0$, the oracle satisfies, almost surely,

$$\mathbb{E}_t[G_t] = \nabla f(x_t), \qquad \mathbb{E}_t[\|G_t - \nabla f(x_t)\|^\alpha] \le \sigma^\alpha. \tag{5}$$

We denote the class of such oracles by $\mathcal{O}_{\alpha,\sigma}(f)$, with the batch convention below.

Remark 2.5. Samples requested together at a fixed, adaptively chosen point are conditionally independent given the history before that request, and each satisfies (5). No independence across adaptively chosen query locations is assumed. Each sample counts as one physical oracle call.

Function classes and performance criteria. Let $\mathcal{F}_d(L, \mu, \Delta_0)$ be the class of globally L-smooth, globally μ-PL functions on $\mathbb{R}^d$ with $f(0) - f^\star \le \Delta_0$. Translation allows us to use $x_0 = 0$ for lower bounds; the algorithms may start at any specified x_0 satisfying the same gap bound. An algorithm A using at most T physical oracle calls may return any measurable output $\widehat{x}_T$, not necessarily a queried point. The minimax expected risk is

$$\mathsf{R}_T^{(d)} := \inf_{\mathsf{A}} \sup_{f \in \mathcal{F}_d(L,\mu,\Delta_0)} \sup_{G \in \mathcal{O}_{\alpha,\sigma}(f)} \mathbb{E}_{\mathsf{A},G}[f(\widehat{x}_T) - f^\star]. \tag{6}$$

For high-probability accuracy, define separately

$$\mathsf{T}_{\epsilon,\delta}^{(d)} := \inf\left\{ T \in \mathbb{N}_0 : \ \exists \mathsf{A}, \sup_{f \in \mathcal{F}_d(L,\mu,\Delta_0)} \sup_{G \in \mathcal{O}_{\alpha,\sigma}(f)} \mathbb{P}_{\mathsf{A},G}\big(f(\widehat{x}_T) - f^\star > \epsilon\big) \le \delta \right\}. \tag{7}$$

Both criteria include algorithmic and oracle randomness; dependence on the problem parameters is suppressed. The dimension-growing criterion takes $\sup_{d \ge 1} \mathsf{R}_T^{(d)}$. Only oracle calls are counted, not arithmetic, geometric-median computation, or exact mirror subproblems. In the algorithms, t indexes optimization updates, each of which may require several physical calls; N denotes their total number.

3 Lower Bound in the High-Dimensional Setting

We first establish a lower bound that separates the cost of deterministic optimization from the cost imposed by heavy-tailed noise. The hard dimension may grow with the target accuracy and noise level.

Theorem 3.1. *Fix $1 < \alpha \le 2$. There are constants $c_{0,\alpha} > 0$ and $\kappa_{\mathrm{hd}} \ge 1$ such that the following holds whenever $\kappa \ge \kappa_{\mathrm{hd}}$ and $0 < \epsilon \le c_{0,\alpha}\Delta_0$. For every $d \ge d_{\mathrm{hd}}$, where the threshold in (166) satisfies $d_{\mathrm{hd}} = \widetilde{O}_\alpha(\kappa^2[\Delta_0/\epsilon + Q_\epsilon])$, any adaptive randomized gradient-only algorithm guaranteeing $\mathbb{E}[f(\widehat{x}_T) - f^\star] \le \epsilon$ for all $f \in \mathcal{F}_d(L, \mu, \Delta_0)$ and $G \in \mathcal{O}_{\alpha,\sigma}(f)$ requires at least*

$$\Omega\left(\kappa \log \frac{\Delta_0}{\epsilon} + \kappa \left(\frac{\sigma^2}{\mu\epsilon} \right)^{\theta_\alpha} \right). \tag{8}$$

Hard Instance: Idea. Two different obstacles produce the two terms. For the stochastic term, a chain hides one new direction at a time; revealing that direction rarely, but amplifying its value when revealed, preserves unbiasedness while exhausting the available moment budget (Lemma B.3). The waiting time is paid across a chain of width proportional to κ. For the noiseless term, geometrically decreasing layer amplitudes keep the unresolved tail small enough that adding layers does not worsen the PL constant (Lemmas B.12–B.13). Compactification and random rotation prevent arbitrarily large or adaptive queries from bypassing either chain. The maximum of these two lower bounds controls their sum up to a constant; the exact-gradient family remains legal even when a positive noise level is announced. The complete proof is in Appendix B.5, using the constructions in Appendices B.3 and B.4.

4 High-Probability Upper Bound

We next combine centered clipping with mirror descent to obtain a high-probability guarantee on a closed convex domain. The geometric assumption must control feasible descent, not merely the full gradient.

4.1 Mirror Descent and the Mirror-PL Condition

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be nonempty, closed, and convex, and let f be L-smooth on $\mathcal{X}$. The mirror map ψ is differentiable on an open convex neighborhood of $\mathcal{X}$ and 1-strongly convex in the Euclidean norm. Its Bregman divergence satisfies

$$D_\psi(z, x) := \psi(z) - \psi(x) - \langle \nabla\psi(x), z - x\rangle \geq \tfrac{1}{2}\|z - x\|^2 \qquad (x, z \in \mathcal{X}). \tag{9}$$

For $v \in \mathbb{R}^d$ define

$$Q_x(z; v) := \langle v, z - x\rangle + LD_\psi(z, x), \qquad z \in \mathcal{X}. \tag{10}$$

As a function of z, this model is coercive and L-strongly convex, so its minimizer exists and is unique even when $\mathcal{X}$ is unbounded.

Definition 4.1 (Mirror-PL Condition)**.** Define

$$\mathcal{G}^2_{\psi,L}(x) := -2L \min_{z\in\mathcal{X}} Q_x(z; \nabla f(x)). \tag{11}$$

The function satisfies the μ-mirror-PL condition if

$$\frac{1}{2}\mathcal{G}^2_{\psi,L}(x) \geq \mu\big(f(x) - f^\star\big), \qquad x \in \mathcal{X}. \tag{12}$$

If $\mathcal{X} = \mathbb{R}^d$ and $\psi(x) = \|x\|^2/2$, then $\mathcal{G}^2_{\psi,L}(x) = \|\nabla f(x)\|^2$, so (12) is exactly (4). For Euclidean projected gradient descent, the projected-PL condition based on $G_L(x) = L(x - \Pi_\mathcal{X}(x - \nabla f(x)/L))$ implies mirror-PL, because projection optimality yields $\mathcal{G}^2_{\psi,L}(x) \geq \|G_L(x)\|^2$.

In the unconstrained case, the PL inequality controls an available descent direction. Under constraints, the negative gradient can instead point outside the feasible region. The next proposition shows why ordinary PL alone does not guarantee convergence of projected or mirror descent.

Proposition 4.2. *There exist a closed convex unbounded set $\mathcal{X} \subset \mathbb{R}^2$, a globally 3-smooth function $f : \mathbb{R}^2 \to \mathbb{R}$, and a constant $\mu = 8/9$ such that*

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu\big(f(x) - \inf_{z\in\mathcal{X}} f(z)\big) \qquad \textit{for every } x \in \mathcal{X}, \tag{13}$$

but exact projected or mirror descent can remain forever at a nonoptimal point of $\mathcal{X}$.

The construction and proof are given in Appendix C.

4.2 Centered Gradient Clipping

For $v \in \mathbb{R}^d$ and radius $r > 0$, define the radial clipping operator

$$\mathrm{clip}_r(v) := \begin{cases} v, & \|v\| \leq r, \\ r\, v/\|v\|, & \|v\| > r. \end{cases} \tag{14}$$

Clipping a raw stochastic gradient around the origin can introduce a bias that depends on $\|\nabla f(x)\|$, which can be arbitrarily large on an unbounded domain. We avoid this dependence by first estimating the gradient with a geometric-median pilot and then clipping a second collection of block means around that pilot, using samples independent of the pilot batch.

At a fixed query point x, divide $2Kb$ fresh samples into $2K$ blocks and form

$$Z_j := \frac{1}{b}\sum_{\ell=1}^{b} G(x, \xi_{j,\ell}), \qquad j = 1, \ldots, 2K. \tag{15}$$

Choose a geometric median of the first K block means as the pilot,

$$a \in \arg\min_{u\in\mathbb{R}^d} \sum_{j=1}^{K} \|u - Z_j\|, \tag{16}$$

Algorithm 1 Centered-Clipped Mirror Descent (CC-MD)

Require: Initial point $x_0 \in \mathcal{X}$, constants $L, \mu > 0$, mirror map ψ, number of stages S, stage length J, number of blocks per half K, block sizes $\{b_s\}_{s=0}^{S-1}$, clipping thresholds $\{r_s\}_{s=0}^{S-1}$.

if $\mu > L$ **then return** x_0.
for $s = 0, \dots, S-1$ **do**
  for $t = sJ, \dots, (s+1)J - 1$ **do**
    Query $2Kb_s$ fresh samples $G_{t,k,\ell} = G(x_t, \xi_{t,k,\ell})$, $k = 1, \dots, 2K$, $\ell = 1, \dots, b_s$
    Form the block means $Z_{t,k} \leftarrow \dfrac{1}{b_s} \displaystyle\sum_{\ell=1}^{b_s} G_{t,k,\ell}, \quad k = 1, \dots, 2K.$
    Compute a geometric-median pilot from the first K blocks:
    $a_t \in \arg\min_{u \in \mathbb{R}^d} \displaystyle\sum_{k=1}^{K} \|u - Z_{t,k}\|.$
    Clip the remaining K block means around the pilot:
    $Y_{t,k} \leftarrow a_t + \mathrm{clip}_{r_s}(Z_{t,K+k} - a_t), \quad k = 1, \dots, K.$
    $\widehat{g}_t \leftarrow \dfrac{1}{K} \displaystyle\sum_{k=1}^{K} Y_{t,k}.$
    Perform the mirror-descent update:
    $x_{t+1} \leftarrow \arg\min_{z \in \mathcal{X}} \{\langle \widehat{g}_t, z - x_t \rangle + L D_\psi(z, x_t)\}.$
  end for
end for
return $x_{\mathrm{out}} \leftarrow x_{JS}$.

and average the remaining block means after clipping around a:

$$\widehat{g}(x) := \frac{1}{K} \sum_{j=1}^{K} Y_j, \quad \text{where} \quad Y_j := a + \mathrm{clip}_r(Z_{K+j} - a). \tag{17}$$

Using $\widehat{g}(x)$ in the mirror update gives Centered-Clipped Mirror Descent (**CC-MD**), presented in Algorithm 1. Appendix C.1 gives the estimator parameters and its conditional accuracy guarantee.

4.3 Convergence Result

The next theorem combines the robust estimator with the mirror-PL contraction. The parameter choices and oracle counts are explicit in the appendix.

Theorem 4.3. *Under the domain and mirror-map assumptions of Section 4.1, let f be L-smooth and μ-mirror-PL, with $f(x_0) - f^\star \le \Delta_0$. Suppose the oracle satisfies Definition 2.4. For every $\epsilon \in (0, \Delta_0)$ and $\delta \in (0, 1/2)$, return $x_{\mathrm{out}} = x_0$ with zero queries if $\mu > L$; otherwise run CC-MD with the parameter choices in Appendix C.2. The output satisfies $\mathbb{P}\big(f(x_{\mathrm{out}}) - f^\star \le \epsilon\big) \ge 1 - \delta$. For $\mu \le L$, the number of stochastic-gradient calls is at most*

$$\widetilde{O}\left(\left(\kappa \left\lceil \log_2 \frac{\Delta_0}{\epsilon} \right\rceil + \kappa \left(\frac{\sigma^2}{\mu\epsilon}\right)^{\theta_\alpha}\right) \log\left(\frac{16\kappa}{\delta} \left\lceil \log_2 \frac{\Delta_0}{\epsilon} \right\rceil\right)\right)$$

Upper Bound: Idea. The estimator must suppress outliers without truncating a possibly large true gradient. Block averages provide weak concentration, and a geometric-median pilot supplies a reliable center; clipping fresh blocks around this center yields accuracy independent of the gradient magnitude (Lemma C.4). The deterministic mirror step then converts squared estimation error into an additive objective-gap error (Lemma C.5). At gap scale Δ, an error of order $\sqrt{\mu\Delta}$ permits a constant-factor improvement after $O(\kappa)$ steps. Successive stages halve the gap: their deterministic costs add, whereas their increasing sampling costs are dominated by the finest scale. A conditional union bound makes every stage valid simultaneously. The full proof and parameter choices are in Appendix C.2.

Remark 4.4. The related guarantees discussed above cover constrained mirror descent under sub-Gaussian noise (Fatkhullin & He, 2024) and unconstrained PL optimization under finite centered α-moments (Sadiev et al., 2023). Here centered clipping yields a high-probability guarantee under the α-heavy tailed model on closed convex, possibly unbounded domains.

5 Fixed-Dimensional Refinements

We now fix the dimension and work with the unconstrained class $\mathcal{F}_d(L, \mu, \Delta_0)$. The high-dimensional chains no longer apply directly; different information-hiding and geometric-search mechanisms yield sharper bounds.

Theorem 5.1. *Fix $d \geq 1$, $1 < \alpha \leq 2$, $\sigma \geq 0$, and $\kappa \geq 6$. There is a constant $c_0 > 0$ such that, for $0 < \epsilon \leq c_0\Delta_0$, any adaptive randomized gradient-only algorithm guaranteeing $\mathbb{E}[f(\widehat{x}_T) - f^\star] \leq \epsilon$ for all $f \in \mathcal{F}_d(L, \mu, \Delta_0)$ and $G \in \mathcal{O}_{\alpha,\sigma}(f)$ requires at least*

$$\Omega\left(\log\frac{\Delta_0}{\epsilon} + \left(\frac{\sigma^2}{\mu\epsilon}\right)^{\theta_\alpha}\right). \tag{18}$$

Hard Instance: Idea. Fixed dimension requires information hiding without a long chain of orthogonal directions. For the stochastic term, two shifted quadratics produce identical oracle answers until a rare informative sample appears; finite-moment calibration makes the waiting time equal to the claimed statistical scale (Lemma D.3). Conditional on no revelation, the average squared distance to the two minimizers stays large, which directly gives an expected-error lower bound. For the noiseless term, a recursively nested interval hides the minimizer, and each exact-gradient query reveals only a constant expected number of location bits (Lemma D.5). Both constructions occupy one coordinate and embed into any fixed dimension. Taking their maximum combines the two obstacles without imposing a condition-number multiplier on the noise term. The complete proof is in Appendix D.2.

Remark (Origin of the κ multiplier). The stochastic barrier κQ_ϵ in Theorem 3.1 couples mean estimation with sequential exploration of $m = \Theta(\kappa)$ hidden directions. In the genuinely noisy regime, the next direction is revealed with probability at most $q = \Theta_\alpha(Q_\epsilon^{-1})$, forcing $\Omega(m/q)$ queries; random rotation prevents shortcuts (Appendix B.3). The exact chain likewise uses $\Theta(\kappa)$ directions per accuracy layer (Appendix B.4). Thus these multipliers reflect the high-dimensional geometry of the constructions, not mean estimation alone. Indeed, Theorem 5.2 removes the polynomial κ multiplier from the stochastic term when $p_\alpha \geq d - 1$, including $d = 1, 2, 3$. The deterministic gaps for $d \geq 2$ and the complementary stochastic regime remain unresolved.

For the upper bounds, we use geometric search tailored to the dimension. Both algorithms access only stochastic gradients and are evaluated using $\mathsf{T}^{(d)}_{\epsilon,\delta}$ from (7).

Robust stochastic bisection ($d = 1$). The PL condition makes the minimizer set an interval, with the derivative negative to its left and positive to its right. Starting from a bracket obtained from Δ_0, we estimate the derivative at its midpoint. A reliable sign selects the half containing a minimizer; a small estimate certifies accuracy through PL. Algorithm 2 gives the pseudocode, and Appendix D.2 proves the guarantee.

Gradient-flow trapping ($d \geq 2$). Maintain a compact box R and an interior pivot x with the boundary certificate $f(y) > f(x) - c\|y - x\|$ for all $y \in \partial R$. If every gradient in R exceeded c in norm, the normalized descent flow would reach the boundary with a larger decrease, a contradiction. Hence R contains a point with gradient norm at most c (Lemma D.11). Given $f(x) - f^\star \leq \Delta$, set $\eta = \sqrt{\mu\Delta}$ and center a cube of side $32\Delta/\eta$ at x, with $c = \eta/8$. The gap bound certifies its boundary without any queries. Two parallel cuts along a longest side either retain the pivot or replace it by a certified lower-value grid point. Each refinement reduces volume by at least $1/6$ while controlling the increase in c (Lemma D.12). Reconstruction tolerances and mesh widths decrease with the cut scale, keeping the accumulated slope increase below $\eta/8$; a strict interpolation margin extends the certificate from grid vertices to the entire new boundary.

Relative values on the cuts are reconstructed by stochastic line integrals. A dyadic tree shares coarse-segment estimates among grid points, rather than independently estimating each full path from the pivot (Lemma D.9). After $O_d(\log(2+\kappa))$ refinements, $c \le \eta/4$ and $\mathrm{diam}(R) \le \eta/(4L)$; smoothness makes the pivot $\eta/2$-stationary, and PL halves the gap. The flow is only a proof device; no function-value oracle is queried. See Algorithm 3 and Appendices D.3–D.4 for reconstruction, error budgets, and the complete stagewise method. Combining these guarantees with the lower bounds gives the following phase diagram.

Below, we provide the upper bound for any fixed $d \ge 1$.

Theorem 5.2. *Fix $d \ge 1$, $1 < \alpha \le 2$, $\sigma \ge 0$, and $\kappa \ge 6$. For a sufficiently small constant $c_0 > 0$, let $0 < \epsilon \le c_0 \Delta_0$ and $0 < \delta \le 1/4$. Suppressing logarithmic factors in κ, Δ_0/ϵ, and $1/\delta$, we have*

$$d = 1: \ \mathsf{T}^{(1)}_{\epsilon,\delta} = \widetilde{\Theta}\left[\log \frac{\Delta_0}{\epsilon} + Q_\epsilon\right], \tag{19}$$

$$d = 2: \ \Omega\left[\log \frac{\Delta_0}{\epsilon} + Q_\epsilon\right] \le \mathsf{T}^{(2)}_{\epsilon,1/4} \le \widetilde{O}\left[\kappa^{2/3} \log \frac{\Delta_0}{\epsilon} + Q_\epsilon\right], \tag{20}$$

$$d = 3: \ \Omega\left[\log \frac{\Delta_0}{\epsilon} + Q_\epsilon\right] \le \mathsf{T}^{(3)}_{\epsilon,1/4} \le \widetilde{O}\left[\kappa \log \frac{\Delta_0}{\epsilon} + Q_\epsilon\right], \tag{21}$$

$$d > 3, p_\alpha \ge d-1: \ \Omega\left[\log \frac{\Delta_0}{\epsilon} + Q_\epsilon\right] \le \mathsf{T}^{(d)}_{\epsilon,1/4} \le \widetilde{O}\left[\kappa^{(d-1)/2} \log \frac{\Delta_0}{\epsilon} + Q_\epsilon\right], \tag{22}$$

$$d > 3, p_\alpha < d-1: \ \Omega\left[\log \frac{\Delta_0}{\epsilon} + Q_\epsilon\right] \le \mathsf{T}^{(d)}_{\epsilon,1/4} \le \widetilde{O}\left[\kappa^{\frac{d-1}{2}} \log \frac{\Delta_0}{\epsilon} + \kappa^{\frac{d-1-p_\alpha}{2}} Q_\epsilon\right]. \tag{23}$$

Upper Bound: Idea. In one dimension, the derivative's sign identifies which side of a query contains the minimizer interval, so a robust sign estimate supports bisection (Lemma D.4). In higher dimensions, boundary values certify that a box traps a descending gradient trajectory; shrinking such a box eventually makes its pivot nearly stationary (Lemmas D.11–D.12). Function values are unavailable, so shared stochastic line integrals reconstruct relative values on each cutting surface. Across dyadic levels, the growth in the number of surface points competes with the decay in integration length. The resulting noise sum is bounded up to logarithms precisely when $p_\alpha \ge d-1$ (Lemma D.9); otherwise its finest levels produce the extra factor. PL converts stationarity to accuracy, and Theorem D.2 supplies the lower terms. The full derivation is in Appendix D.5, in the proof of Corollary D.13.

Thus the full complexity is matched up to logarithmic factors in dimension one. In dimensions two and three, and in fixed dimension $d > 3$ when $p_\alpha \ge d-1$, the stochastic term Q_ϵ is matched; the condition-number dependence of the deterministic term remains unresolved.

6 Conclusion and Future Directions

We establish noise-adaptive complexity bounds for smooth PL optimization with conditionally unbiased stochastic gradients satisfying only a finite centered α-moment. In growing dimension, the lower bound separates the exact-gradient transient from the heavy-tailed estimation cost and remains informative when $\sigma = 0$. Under mirror-PL, centered-clipped mirror descent matches this bound up to logarithmic factors with high probability, including on closed convex, unbounded domains. Fixed dimension permits sharper geometric search: robust stochastic bisection characterizes the one-dimensional complexity up to logarithmic factors, while gradient-flow trapping matches the stochastic term for $d = 2, 3$ and for $d > 3$ whenever $p_\alpha \ge d-1$.

The remaining questions concern the deterministic condition-number dependence in dimensions $d \ge 2$ and the stochastic complexity when $p_\alpha < d-1$. In particular, it remains to determine whether the extra surface-reconstruction factor is necessary or can be removed by a different search strategy. Developing methods that do not require prior knowledge of the noise and PL parameters is another natural direction.

AI use statement

In this work, we used generative AI tools for polishing the grammar and reorganizing the paper's structure. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

All assumptions and theoretical claims are stated in the paper, with complete proofs provided in the appendix.

References

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

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pp. 242–252. PMLR, 2019.

Mihai Anitescu. Degenerate nonlinear programming with a quadratic growth condition. *SIAM Journal on Optimization*, 10(4):1116–1135, 2000.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199:165–214, 2023a. Preliminary version: arXiv:1912.02365.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023b.

Amit Attia and Tomer Koren. Sgd with adagrad stepsizes: Full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *International Conference on Machine Learning*, pp. 1147–1171. PMLR, 2023.

Yunyan Bai, Yuxing Liu, and Luo Luo. On the complexity of finite-sum smooth optimization under the polyak-{\L} ojasiewicz condition. *arXiv preprint arXiv:2402.02569*, 2024.

Sébastien Bubeck and Dan Mikulincer. How to trap a gradient flow. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 940–960. PMLR, 2020a.

Sébastien Bubeck and Dan Mikulincer. How to trap a gradient flow. In *Conference on Learning Theory*, pp. 940–960. PMLR, 2020b.

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

Savelii Chezhegov, Yaroslav Klyukin, Andrei Semenov, Aleksandr Beznosikov, Alexander Gasnikov, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. Clipping improves Adam-norm and AdaGrad-norm when the noise is heavy-tailed. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10269–10333. PMLR, 2025. URL https://proceedings.mlr.press/v267/chezhegov25a.html.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4883–4895, 2021a.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34: 4883–4895, 2021b.

John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

Ilyas Fatkhullin and Niao He. Taming nonconvex stochastic mirror descent with general bregman divergence. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3493–3501. PMLR, 2024. URL https://proceedings.mlr.press/v238/fatkhullin24a.html.

Adrien Fradin, Abdurakhmon Sadiev, Laurent Condat, and Peter Richtárik. Tight lower bounds and optimal algorithms for stochastic nonconvex optimization with heavy-tailed noise. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 2548–2556. PMLR, 2026. URL https://proceedings.mlr.press/v300/fradin26a.html.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems*, volume 33, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/abd1c782880cc59759f4112fda0b8f98-Abstract.html.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020b.

Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov. Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2106.05958*, pp. 1, 2021.

Eduard Gorbunov, Abdurakhmon Sadiev, Marina Danilova, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15951–16070. PMLR, 2024. URL https://proceedings.mlr.press/v235/gorbunov24a.html.

Charles Guille-Escuret, Manuela Girotti, Baptiste Goujaud, and Ioannis Mitliagkas. A study of condition numbers for first-order optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1261–1269. PMLR, 2021.

Zhishuai Guo, Yan Yan, Zhuoning Yuan, and Tianbao Yang. Fast objective & duality gap convergence for non-convex strongly-concave min-max problems with pl condition. *arXiv preprint arXiv:2006.06889*, 2020.

Zhishuai Guo, Yan Yan, Zhuoning Yuan, and Tianbao Yang. Fast objective & duality gap convergence for non-convex strongly-concave min-max problems with pl condition. *Journal of Machine Learning Research*, 24(148):1–63, 2023.

Peter Hall and Christopher C. Heyde. *Martingale Limit Theory and Its Application*. Probability and Mathematical Statistics. Academic Press, New York, 1980. ISBN 978-0-12-319350-6.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

Yusu Hong and Junhong Lin. On convergence of adam for stochastic optimization under relaxed assumptions. *Advances in Neural Information Processing Systems*, 37:10827–10877, 2024.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak–łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016a.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016b.

Ali Kavis, Kfir Yehuda Levy, and Volkan Cevher. High probability bounds for a class of nonconvex algorithms with adagrad stepsize. *arXiv preprint arXiv:2204.02833*, 2022.

Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive sgd with momentum. In *Workshop on Beyond First Order Methods in ML Systems at ICML'20*, 2020.

Xiaoyu Li, Zhenxun Zhuang, and Francesco Orabona. A second look at exponential and cosine step sizes: Simplicity, adaptivity, and performance. In *International Conference on Machine Learning*, pp. 6553–6564. PMLR, 2021.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pp. 2–47. PMLR, 2018.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Ji Liu, Steve Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. In *International Conference on Machine Learning*, pp. 469–477. PMLR, 2014.

Langqi Liu, Yibo Wang, and Lijun Zhang. High-probability bound for non-smooth non-convex stochastic optimization with heavy tails. In *Forty-first International Conference on Machine Learning*, 2024.

Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Nguyen. High probability convergence of stochastic gradient methods. In *International Conference on Machine Learning*, pp. 21884–21914. PMLR, 2023a.

Zijian Liu, Jiawei Zhang, and Zhengyuan Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2266–2290. PMLR, 2023b. URL https://proceedings.mlr.press/v195/liu23c.html.

Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5): 1145–1190, 2019. doi: 10.1007/s10208-019-09427-x. URL https://doi.org/10.1007/s10208-019-09427-x.

Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

Saeed Masiha, Negar Kiyavash, and Patrick Thiran. Sharp first-order lower bounds under sublevel α-polyak–łojasiewicz conditions, 2026. URL https://arxiv.org/abs/2606.28278.

Stanislav Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015. doi: 10.3150/14-BEJ645. URL https://doi.org/10.3150/14-BEJ645.

Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical programming*, 175(1):69–107, 2019.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Iosif Pinelis. Best possible bounds of the von bahr–esseen type. *Annals of Functional Analysis*, 6(4):1–29, 2015. doi: 10.15352/afa/06-4-1. URL https://doi.org/10.15352/afa/06-4-1.

Boris T. Polyak. Gradient methods for minimizing functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963a.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963b.

Nikolay Puchkin, Eduard Gorbunov, Nikita Kutuzov, and Alexander Gasnikov. Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 856–864. PMLR, 2024.

Maxim Raginsky and Alexander Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Transactions on Information Theory*, 57(10): 7036–7056, 2011.

El Mehdi Saad, Weicheng Lee, and Francesco Orabona. New lower bounds for stochastic non-convex optimization through divergence decomposition. *Proceedings of Machine Learning Research vol*, 291:1–54, 2025.

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: The case of unbounded variance. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29563–29648. PMLR, 2023. URL https://proceedings.mlr.press/v202/sadiev23a.html.

Umut Şimşekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5827–5837. PMLR, 2019.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303, 1965. doi: 10.1214/aoms/1177700291. URL https://doi.org/10.1214/aoms/1177700291.

Nuri Mert Vural, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A. Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 65–102. PMLR, 2022a. URL https://proceedings.mlr.press/v178/vural22a.html.

Nuri Mert Vural, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Conference on Learning Theory*, pp. 65–102. PMLR, 2022b.

Hongjian Wang, Mert Gürbüzbalaban, Lingjiong Zhu, Umut Simsekli, and Murat A. Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18866–18877, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/9cdf26568d166bc6793ef8da5afa0846-Abstract.html.

Andrew Chi-Chih Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science*, pp. 222–227. IEEE, 1977. doi: 10.1109/SFCS.1977.24. URL https://doi.org/10.1109/SFCS.1977.24.

Pengyun Yue, Cong Fang, and Zhouchen Lin. On the lower bound of minimizing polyak–łojasiewicz functions. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2948–2968. PMLR, 2023a. URL https://proceedings.mlr.press/v195/yue23a.html.

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

Hui Zhang and Wotao Yin. Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv:1303.4645*, 2013.

Dongruo Zhou and Quanquan Gu. Lower bounds for smooth nonconvex finite-sum optimization. In *International Conference on Machine Learning*, pp. 7574–7583. PMLR, 2019.

APPENDIX

Notation and scope of the proofs. We use the conventions of Section 2 throughout: $f^\star$ and $\mathcal{X}^\star$ denote the optimal value and minimizer set, and θ_α, β_α, p_α, Q_ϵ, and A_ϵ are defined in (1)–(2). The filtration $\mathscr{F}_{t-1}$ records the full pre-call history; $\mathcal{F}_d(L, \mu, \Delta_0)$ denotes a function class, not a filtration. The expected risk $\mathsf{R}_T^{(d)}$ and the high-probability query complexity $\mathsf{T}_{\epsilon,\delta}^{(d)}$ are distinct criteria, defined in (6) and (7). All conditional oracle statements use this full history, with conditional independence only for samples requested together at a fixed location.

The confidence parameter is δ; spatial mesh widths are h or h_t. A prescribed-budget output is denoted by $\widehat{x}_T$, and an algorithmic output by x_{out}. Construction-specific indices and parameters are introduced locally. In particular, the layer count in Appendix B and the block count in Appendix C are different uses of K. Constants such as $C_d, C_\alpha, C_{d,\alpha}$ may change between bounds; named construction constants and algorithm parameters do not. Only physical oracle calls are counted, under the convention in Section 2.

A Comparison with Masiha et al. (2026)

We compare with the statements in version 2 of Masiha et al. (2026). Write $\nu \in [1, 2)$ for their geometric exponent, to distinguish it from the noise-moment order α and the geometric decay factor γ used below. Their model imposes $f - f^\star \le \tau\|\nabla f\|^\nu$ on the initial sublevel set while retaining global smoothness. Their Theorems 2.1, 3.1, and 4.1 respectively establish triviality under a global inequality with $\nu < 2$, a deterministic lower bound of order $L\tau^{2/\nu}\epsilon^{-(2-\nu)/\nu}$, and a bounded-variance lower bound adding $L\sigma^2\tau^{4/\nu}\epsilon^{-(4-\nu)/\nu}$. Those theorems do not state a separate globally PL endpoint result at $\nu = 2$.

Our geometric assumption is instead the global quadratic PL inequality, while the centered noise has only a finite α-moment. The finite-variance specialization of our stochastic term is $\kappa\sigma^2/(\mu\epsilon)$. Substituting $\nu = 2$ and $\tau = (2\mu)^{-1}$ formally into the polynomial noise expression above gives $L\sigma^2/(4\mu^2\epsilon)$, the same algebraic scale. This algebraic comparison does not extend their theorems beyond the stated exponent range. Our proofs separately establish the global-PL heavy-tail bound, its exact-gradient transient, the mirror-PL upper bound, and the fixed-dimensional refinements.

B Proofs for Section 3

The lower bound separates two obstructions: learning the next hidden direction under heavy-tailed noise, and traversing many accuracy scales even with exact gradients. We first develop each normalized construction and then choose its parameters when matching the prescribed L, μ, σ, and ϵ. The two resulting ambient dimensions are compared only when the constructions are combined in the final subsection.

B.1 Main ideas and proof sketch

The lower bound has two logically independent hard families. The first is the Bernoulli-thinned heavy-tailed chain developed in the next sections. The second is an exact-gradient geometrically weighted chain. Because an exact oracle has centered α-moment zero, it is legal for every announced noise level $\sigma \ge 0$. The noise-adaptive lower bound is therefore at least the maximum of the two component risks, which gives the additive noise-adaptive statement.

For the stochastic obstruction, take a chain whose width is proportional to κ. Its next hidden gradient coordinate is returned rarely and enlarged on revelation so that the oracle remains unbiased. The moment constraint sets the waiting time per coordinate, while radial compactification makes the objective globally smooth and globally PL. A random rotation prevents an adaptive method from bypassing the chain by guessing hidden directions.

The exact-gradient transient needs a different construction. Concatenating unweighted chains would worsen the PL constant with the number of layers. Instead, we decrease the coordinate amplitudes geometrically: the squared mass of the unresolved tail stays proportional to the layer width times the first unresolved squared amplitude. This permits logarithmically many layers without changing the order of the condition number. A guard that vanishes near the minimizer and a rotation argument with nonuniform thresholds preserve the last layer's small objective gap.

B.2 Bernoulli-thinned robust zero-chain and global PL compactification

We first construct the scalar activation, coupling, and barrier functions. Their local bounds will control both the information structure and the global geometry of the chain.

Define

$$\hat{\Theta}(t) = \begin{cases} 0, & t \le \frac{1}{2}, \\ 3(2t-1)^2 - 2(2t-1)^3, & \frac{1}{2} < t < 1, \\ 1, & t \ge 1, \end{cases} \tag{24}$$

$$\phi(t) = 1 - \exp\left(-\frac{(t-1)^2}{2}\right), \tag{25}$$

and

$$B(t) = \begin{cases} 2 + \frac{1}{2}(t+\frac{1}{2})^2, & t \le -\frac{1}{2}, \\ 2, & -\frac{1}{2} < t \le \frac{1}{2}, \\ 2(1-\hat{\Theta}(t)), & \frac{1}{2} < t < 1, \\ \frac{1}{2}(t-1)^2, & t \ge 1. \end{cases} \tag{26}$$

All three functions are continuously differentiable. Their derivatives are Lipschitz, although their second derivatives need not agree at the break points.

Lemma B.1. *For every $t \in \mathbb{R}$,*

$$0 \le \hat{\Theta}(t) \le 1, \qquad \left\|\hat{\Theta}'\right\|_\infty \le 3, \qquad \mathrm{Lip}(\hat{\Theta}') \le 24, \tag{27}$$

$$0 \le \phi(t) \le 1, \qquad \|\phi'\|_\infty = e^{-1/2} < 1, \qquad \mathrm{Lip}(\phi') \le 1, \tag{28}$$

$$\mathrm{Lip}(B') \le 48. \tag{29}$$

In addition,

$$B(t) \le \frac{1}{2}t^2 + 2, \qquad |B'(t) - t| \le 7, \qquad tB'(t) \ge \frac{1}{2}t^2 - \frac{13}{2}. \tag{30}$$

Proof. On $(1/2, 1)$ put $u = 2t - 1 \in (0, 1)$. Then

$$\hat{\Theta}(t) = 3u^2 - 2u^3, \qquad \hat{\Theta}'(t) = 12u(1-u), \qquad \hat{\Theta}''(t) = 24(1-2u).$$

This proves (27), including the endpoint matching. Furthermore,

$$\phi'(t) = (t-1)e^{-(t-1)^2/2}, \qquad \phi''(t) = \big(1 - (t-1)^2\big)e^{-(t-1)^2/2}.$$

The first expression has maximum absolute value $e^{-1/2}$, and the second has absolute value at most 1, proving (28).

For B, the derivative equals

$$B'(t) = \begin{cases} t + \frac{1}{2}, & t < -\frac{1}{2}, \\ 0, & -\frac{1}{2} < t < \frac{1}{2}, \\ -2\hat{\Theta}'(t), & \frac{1}{2} < t < 1, \\ t - 1, & t > 1. \end{cases}$$

The largest absolute slope of these pieces is 48, and the derivative is continuous at the break points, giving (29).

The three inequalities in (30) follow by checking the same four intervals. For completeness, on $t \le -1/2$ one has $B(t) = 2 + \frac{1}{2}(t+1/2)^2 \le 2 + t^2/2$ and $B'(t) - t = 1/2$; on $[-1/2, 1/2]$ they are immediate; on $(1/2, 1)$, one uses $0 \le B \le 2$, $-6 \le B' \le 0$, and $t \le 1$; and on $[1, \infty)$ one uses $B(t) = \frac{1}{2}(t-1)^2$ and $B'(t) = t - 1$. In particular, on $(1/2, 1)$, $tB'(t) \ge -6 \ge t^2/2 - 13/2$. This proves all claims. □

We now combine the scalar functions into a chain. Its support property is the information constraint that will survive Bernoulli thinning.

For $N \ge 1$, define

$$F_N(x) := \phi(x_1) + \sum_{i=1}^{N-1} \hat{\Theta}(x_i)\phi(x_{i+1}) + \sum_{i=1}^{N} B(x_i), \qquad x \in \mathbb{R}^N. \tag{31}$$

For $a \ge 0$, define the progress index

$$\mathrm{prog}_a(x) := \max\big(\{i : |x_i| > a\} \cup \{0\}\big). \tag{32}$$

Lemma B.2. *For every $N \ge 1$:*

(i) $F_N(x) \ge 0$ for all x, $F_N(\mathbf{1}_N) = 0$, and $F_N(0) < \frac{5}{2}N$.

(ii) F_N is globally 79-smooth.

(iii) If $k = \mathrm{prog}_{1/4}(x)$, then

$$\nabla_j F_N(x) = 0 \qquad \text{for every } j \ge k+2. \tag{33}$$

If $k < N$, then

$$|\nabla_{k+1} F_N(x)| \le 1. \tag{34}$$

(iv) If $N = 2m$ and $\mathrm{prog}_{1/4}(x) < m$, then

$$F_{2m}(x) \ge 2m. \tag{35}$$

Proof. Nonnegativity follows from the nonnegativity of $\hat{\Theta}, \phi, B$. At $x = \mathbf{1}_N$, one has $\phi(1) = B(1) = 0$, hence $F_N(\mathbf{1}_N) = 0$. At the origin, all coupling terms vanish because $\hat{\Theta}(0) = 0$, so

$$F_N(0) = \phi(0) + 2N < \frac{5}{2}N.$$

For an interior coordinate $2 \le i \le N-1$,

$$\nabla_i F_N(x) = \hat{\Theta}(x_{i-1})\phi'(x_i) + \hat{\Theta}'(x_i)\phi(x_{i+1}) + B'(x_i). \tag{36}$$

At $i = 1$, the incoming term is $\phi'(x_1)$; at $i = N$, omit the outgoing term. Omit nonexistent neighbors throughout, including when $N = 1$. By lemma B.1, for arbitrary x, y,

$$|\nabla_i F_N(x) - \nabla_i F_N(y)| \le 3\,|x_{i-1} - y_{i-1}| + 73\,|x_i - y_i| + 3\,|x_{i+1} - y_{i+1}|\,.$$

Indeed, the first product contributes at most $3\,|x_{i-1} - y_{i-1}| + |x_i - y_i|$, the second contributes at most $24\,|x_i - y_i| + 3\,|x_{i+1} - y_{i+1}|$, and B' contributes at most $48\,|x_i - y_i|$. The corresponding nonnegative tridiagonal coefficient matrix has maximum row sum and maximum column sum at most 79, hence its Euclidean operator norm is at most $\sqrt{\|A\|_1 \|A\|_\infty} \le 79$. Thus F_N is 79-smooth.

Let $k = \mathrm{prog}_{1/4}(x)$. If $j \ge k+2$, then $|x_{j-1}|, |x_j|, |x_{j+1}| \le 1/4$. Consequently $\hat{\Theta}(x_{j-1}) = 0$, $\hat{\Theta}'(x_j) = 0$, and $B'(x_j) = 0$, which proves (33). For $j = k+1$ with $k \ge 1$, all terms in (36) vanish except possibly $\hat{\Theta}(x_k)\phi'(x_{k+1})$. When $k = 0$, the only possibly nonzero term is instead the leading term $\phi'(x_1)$. In both cases, $|\nabla_{k+1} F_N(x)| \le 1$, proving (34).

Finally, if $N = 2m$ and $\mathrm{prog}_{1/4}(x) < m$, then each coordinate $i = m+1, \dots, 2m$ lies in $[-1/4, 1/4]$. Hence $B(x_i) = 2$, and all other terms are nonnegative. Thus

$$F_{2m}(x) \ge \sum_{i=m+1}^{2m} B(x_i) = 2m.$$

□

Fix $m \geq 1$ and $q \in (0, 1]$. Let $z \sim \text{Bernoulli}(q)$. For a query $x \in \mathbb{R}^{2m}$, put $k = \text{prog}_{1/4}(x)$ and define

$$[g_{m,q}(x, z)]_i := \nabla_i F_{2m}(x) \left[1 + \mathbf{1}_{\{i>k\}} \left(\frac{z}{q} - 1\right)\right]. \tag{37}$$

Already exposed coordinates are exact, whereas every hidden coordinate is multiplied by z/q. By the zero-chain property, at most the single coordinate $k + 1$ can be both hidden and nonzero.

Lemma B.3. *For every $x \in \mathbb{R}^{2m}$,*

$$\mathbb{E}_z[g_{m,q}(x, z)] = \nabla F_{2m}(x), \tag{38}$$

$$\mathbb{E}_z\|g_{m,q}(x, z) - \nabla F_{2m}(x)\|^\alpha \leq 2q^{1-\alpha}. \tag{39}$$

When $q = 1$, the left-hand side of (39) is in fact zero. Moreover,

$$\text{prog}_0(g_{m,q}(x, z)) \leq \text{prog}_{1/4}(x) + 1 \tag{40}$$

always, and

$$\mathbb{P}_z\Big(\text{prog}_0(g_{m,q}(x, z)) > \text{prog}_{1/4}(x)\Big) \leq q. \tag{41}$$

Proof. For $i \leq k$, the multiplier in (37) is one; for $i > k$, it is z/q, whose expectation is one. This proves unbiasedness. Only coordinate $k + 1$ can fluctuate, and its true magnitude is at most one. Therefore

$$\begin{aligned}
&\mathbb{E}_z\|g_{m,q}(x, z) - \nabla F_{2m}(x)\|^\alpha \\
&\quad\leq q\left(\frac{1}{q} - 1\right)^\alpha + (1 - q) = (1 - q)^\alpha q^{1-\alpha} + (1 - q) \\
&\quad\leq q^{1-\alpha} + 1 \leq 2q^{1-\alpha},
\end{aligned}$$

where the last inequality uses $q^{1-\alpha} \geq 1$. For $q = 1$, the multiplier is identically one and the error is zero. The support statement follows from the deterministic zero-chain property. Progress beyond the current coordinate can occur only when $z = 1$, which has probability q. □

The zero-chain support rule alone does not prevent unbounded queries from bypassing the hidden-coordinate sequence. We therefore compactify the effective queries and add a quadratic guard. The construction below preserves the information constraint and verifies global smoothness and PL; its Jacobian estimates also control the oracle's centered moment.

Fix

$$\eta := \frac{1}{20}, \qquad R_m := 16\sqrt{m}. \tag{42}$$

Define $h_m : [0, \infty) \to [0, 2R_m)$ by

$$h_m(r) = \begin{cases} r, & 0 \leq r \leq R_m, \\ R_m + R_m\left(1 - e^{-(r-R_m)/R_m}\right), & r > R_m, \end{cases} \tag{43}$$

and define $\rho_m : \mathbb{R}^d \to \mathbb{R}^d$ by

$$\rho_m(x) = \begin{cases} \dfrac{h_m(\|x\|)}{\|x\|} x, & x \neq 0, \\ 0, & x = 0. \end{cases} \tag{44}$$

Lemma B.4. *The map ρ_m is continuously differentiable and satisfies*

$$\rho_m(x) = x \quad \text{if } \|x\| \leq R_m, \qquad \|\rho_m(x)\| < 2R_m, \tag{45}$$

$$\|J\rho_m(x)\|_{\text{op}} \leq 1, \qquad \text{Lip}(J\rho_m) \leq \frac{16}{R_m}. \tag{46}$$

In addition, if $\|x\| \geq 16R_m$, then

$$\|J\rho_m(x)\|_{\text{op}} \leq \frac{2R_m}{\|x\|}. \tag{47}$$

Proof. For $x \neq 0$, put $r = \|x\|$, $e = x/r$, and $a(r) = h_m(r)/r$. A direct differentiation gives

$$J\rho_m(x) = a(r)I + \big(h'_m(r) - a(r)\big)ee^\top. \tag{48}$$

Thus the radial eigenvalue is $h'_m(r)$ and every tangential eigenvalue is $a(r)$. For $r \leq R_m$, both equal 1. For $r > R_m$, $0 < h'_m(r) = e^{-(r-R_m)/R_m} \leq 1$. Also $h_m(R_m) = R_m$ and $h'_m \leq 1$, so $h_m(r) \leq r$ and hence $0 < a(r) \leq 1$. This proves the first part of (46); (45) is immediate from (43).

For $r > R_m$, one has $h_m(r) \leq 2R_m$, $|h''_m(r)| \leq R_m^{-1}$,

$$|a'(r)| = \left|\frac{h'_m(r)r - h_m(r)}{r^2}\right| \leq \frac{3}{R_m}, \qquad |(h'_m - a)'(r)| \leq \frac{4}{R_m}.$$

Moreover, the differential of $ee^\top$ has operator norm at most $2/r$. Differentiating (48) therefore gives, on $r > R_m$,

$$\|D(J\rho_m)(x)\|_{\mathrm{op}} \leq \frac{3}{R_m} + \frac{4}{R_m} + \frac{2\,|h'_m(r) - a(r)|}{r} \leq \frac{13}{R_m}.$$

The Jacobian is continuous at $r = R_m$. Integrating this derivative along line segments, splitting a segment at its intersections with the sphere $r = R_m$ if necessary, yields the stated safe bound $16/R_m$.

Finally, for $r \geq 16R_m$, $a(r) < 2R_m/r$. Also $h'_m(r) = e^{-(r-R_m)/R_m} \leq 2R_m/r$. Equation (48) then proves (47). □

We can now define the compactified family and its oracle. The following estimates verify smoothness and admissibility before we turn to global PL.

Let $U \in \mathbb{R}^{d\times 2m}$ have orthonormal columns. Define

$$\widehat{F}_{m,U}(x) := F_{2m}\big(U^\top \rho_m(x)\big) + \frac{\eta}{2}\|x\|^2, \qquad x \in \mathbb{R}^d, \tag{49}$$

and the associated stochastic oracle

$$\widehat{G}_{m,q,U}(x,z) := J\rho_m(x)^\top U g_{m,q}\big(U^\top \rho_m(x), z\big) + \eta x. \tag{50}$$

Lemma B.5. *For all $m \geq 1$, all orthonormal U, and every $q \in (0,1]$,*

$$\mathrm{Lip}(\nabla \widehat{F}_{m,U}) \leq L_b := 128, \tag{51}$$

$$\widehat{F}_{m,U}(0) - \widehat{F}^\star_{m,U} < 5m, \tag{52}$$

$$\mathbb{E}_z[\widehat{G}_{m,q,U}(x,z)] = \nabla \widehat{F}_{m,U}(x), \tag{53}$$

$$\mathbb{E}_z\|\widehat{G}_{m,q,U}(x,z) - \nabla \widehat{F}_{m,U}(x)\|^\alpha \leq 2q^{1-\alpha}. \tag{54}$$

For $q = 1$, the stochastic error in the last display is identically zero.

Proof. Write $N = 2m$ and separate

$$H_N(y) := \phi(y_1) + \sum_{i=1}^{N-1} \hat{\Theta}(y_i)\phi(y_{i+1}), \qquad F_N(y) = H_N(y) + \sum_{i=1}^{N} B(y_i).$$

Each coordinate of ∇H_N has absolute value at most 4, so

$$\|\nabla H_N(y)\| \leq 4\sqrt{N}. \tag{55}$$

By the scalar estimates proved above,

$$\|\nabla F_N(y)\| \leq \|y\| + 11\sqrt{N}. \tag{56}$$

If $y = U^\top \rho_m(x)$, then $\|y\| < 2R_m$. Since $R_m = 8\sqrt{2}\sqrt{N} > 11\sqrt{N}$,

$$\|\nabla F_N(y)\| \leq 3R_m. \tag{57}$$

Using the chain smoothness and the radial clipping estimates, for arbitrary x, x' the gradient of the composed chain obeys

$$\begin{aligned}&\|J\rho_m(x)^\top U\nabla F_N(U^\top\rho_m(x)) - J\rho_m(x')^\top U\nabla F_N(U^\top\rho_m(x'))\|\\&\quad\le 79\|x-x'\| + \frac{16}{R_m}(3R_m)\|x-x'\| \le 127\|x-x'\|.\end{aligned}$$

Adding the gradient ηx proves (51). At the origin, $\widehat{F}_{m,U}(0) = F_{2m}(0) < 5m$, whereas $\widehat{F}^\star_{m,U} \ge 0$, proving (52). Unbiasedness follows from Lemma B.3 and the chain rule. Finally, U is an isometry on $\mathbb{R}^{2m}$ and $\|J\rho_m(x)\|_{\mathrm{op}} \le 1$, so

$$\begin{aligned}&\mathbb{E}_z\|\widehat{G}_{m,q,U}(x,z) - \nabla\widehat{F}_{m,U}(x)\|^\alpha\\&\quad\le \mathbb{E}_z\|g_{m,q}(U^\top\rho_m(x),z) - \nabla F_{2m}(U^\top\rho_m(x))\|^\alpha \le 2q^{1-\alpha}.\end{aligned}$$

The $q = 1$ assertion is immediate. □

It remains to establish global PL, which is not implied by the zero-chain property. We first analyze a regularized chain without radial composition.

Define, for general N,

$$\Gamma_N(y) := F_N(y) + \frac{\eta}{2}\|y\|^2. \tag{58}$$

Lemma B.6. *For every $N \ge 1$ and every $y \in \mathbb{R}^N$,*

$$\Gamma_N(y) - \Gamma_N^\star \le C_\Gamma N\|\nabla\Gamma_N(y)\|^2, \qquad C_\Gamma := 20000. \tag{59}$$

Proof. Let $r = \|y\|$. From (55) and (30),

$$\Gamma_N(y) \le \frac{21}{40}r^2 + 3N, \tag{60}$$

$$\langle\nabla\Gamma_N(y), y\rangle \ge \frac{11}{20}r^2 - 4\sqrt{N}\,r - \frac{13}{2}N. \tag{61}$$

Indeed, $0 \le H_N \le N$, $\sum_i B(y_i) \le r^2/2 + 2N$, $\langle\nabla H_N(y), y\rangle \ge -4\sqrt{N}\,r$, and $\sum_i y_i B'(y_i) \ge r^2/2 - 13N/2$.

We next isolate a box containing every critical point. Put

$$\mathcal{B}_N := [7/8, 9/8]^N.$$

On this box,

$$\hat\Theta \ge \frac{27}{32}, \quad \left|\hat\Theta'\right| \le \frac{9}{4}, \quad \left|\hat\Theta''\right| \le 24, \quad \phi'' \ge \frac{1}{2}, \quad \phi \le \frac{1}{128}, \quad |\phi'| \le \frac{1}{8}. \tag{62}$$

At points of twice differentiability, every diagonal entry of $\nabla^2 F_N$ is at least $79/64$, while every off-diagonal entry has absolute value at most $9/32$. To verify the diagonal bound, note that a typical diagonal entry is

$$\hat\Theta(y_{i-1})\phi''(y_i) + \hat\Theta''(y_i)\phi(y_{i+1}) + B''(y_i).$$

At $i = 1$, the incoming term is $\phi''(y_1)$; at $i = N$, the outgoing term is omitted. If $y_i \ge 1$, the last two terms sum to 1; if $y_i < 1$, then $\hat\Theta''(y_i) \le -12$ and $B''(y_i) = -2\hat\Theta''(y_i)$, so their sum is $(-\hat\Theta''(y_i))(2 - \phi(y_{i+1})) > 23$. The off-diagonal term is $\hat\Theta'(y_i)\phi'(y_{i+1})$. Gershgorin's theorem therefore gives

$$\nabla^2 F_N \succeq \frac{1}{2}I \quad \text{almost everywhere on } \mathcal{B}_N. \tag{63}$$

For line segments contained in a breakpoint hyperplane, approximate them by generic segments and pass to the limit using gradient continuity. Thus the Lipschitz continuity of ∇F_N and integration along line segments imply that F_N, and hence Γ_N, is $1/2$-strongly convex on $\mathcal{B}_N$.

We now prove that the gradient is uniformly nonzero outside the box. Let $y \notin \mathcal{B}_N$, and let j be the first index for which $y_j \notin [7/8, 9/8]$. Set $a = 1$ if $j = 1$ and $a = \hat{\Theta}(y_{j-1})$ otherwise. Then $a \geq 27/32$. With $s = \phi(y_{j+1}) \in [0, 1]$ when $j < N$ and $s = 0$ when $j = N$, one has

$$\nabla_j \Gamma_N(y) = a\phi'(y_j) + s\hat{\Theta}'(y_j) + B'(y_j) + \eta y_j. \tag{64}$$

Direct substitution of the scalar formulas gives

$$\nabla_j \Gamma_N(y) \begin{cases} \geq 1/8, & y_j > 9/8, \\ \leq -9/160, & 1/2 < y_j < 7/8, \\ \leq -3/40, & -1/2 \leq y_j \leq 1/2, \\ \leq -1/8, & y_j < -1/2. \end{cases} \tag{65}$$

For $1/2 < y_j < 7/8$, put $u = 1 - y_j \in (1/8, 1/2)$. Since $s\hat{\Theta}' + B' = (s-2)\hat{\Theta}' \leq -\hat{\Theta}' \leq 0$ and $ue^{-u^2/2} \geq (1/8)e^{-1/128}$,

$$a\phi'(y_j) + \eta y_j \leq -\frac{27}{256}e^{-1/128} + \frac{7}{160} < -\frac{9}{160}.$$

On $[-1/2, 1/2]$, the last two chain terms vanish. Here $(1 - y_j)e^{-(1-y_j)^2/2} \geq (1/2)e^{-1/8}$, and therefore $a\phi'(y_j) + \eta y_j \leq -(27/64)e^{-1/8} + 1/40 < -3/40$. On $(-\infty, -1/2)$, split at -1: for $y_j \leq -1$, $(1+\eta)y_j + 1/2 \leq -11/20$, and for $-1 < y_j < -1/2$, $a\phi'(y_j) \leq -(27/32)2e^{-2}$ while $(1+\eta)y_j + 1/2 \leq -1/40$. Thus all constants in (65) are valid. In particular,

$$y \notin \mathcal{B}_N \quad \Longrightarrow \quad \|\nabla \Gamma_N(y)\| \geq \frac{9}{160}. \tag{66}$$

The function Γ_N is coercive, so it has a minimizer. By (66), every minimizer lies in $\mathcal{B}_N$. Fix one minimizer $y^\star$.

If $y \in \mathcal{B}_N$, strong convexity on the box gives

$$\Gamma_N(y) - \Gamma_N(y^\star) \leq \langle \nabla \Gamma_N(y), y - y^\star \rangle - \frac{1}{4}\|y - y^\star\|^2 \leq \|\nabla \Gamma_N(y)\|^2.$$

Suppose next that $y \notin \mathcal{B}_N$ and $r \leq 9\sqrt{N}$. Since $\Gamma_N^\star \geq 0$, (60) and (66) imply

$$\Gamma_N(y) - \Gamma_N^\star \leq \frac{1821}{40}N \leq 20000N\|\nabla \Gamma_N(y)\|^2.$$

Finally, suppose $r > 9\sqrt{N}$. The right-hand side of (61) is at least $r^2/40$; this follows by writing $r = s\sqrt{N}$, observing that $\frac{21}{40}s^2 - 4s - 13/2 \geq 0$ for $s \geq 9$. Hence $\|\nabla \Gamma_N(y)\| \geq r/40$. Also (60) and $N < r^2/81$ give

$$\Gamma_N(y) - \Gamma_N^\star \leq \left(\frac{21}{40} + \frac{3}{81}\right) r^2 \leq 900\|\nabla \Gamma_N(y)\|^2.$$

Combining the three regions proves (59). □

The auxiliary PL estimate can now be transferred through radial clipping. Inside the clipping ball the transfer is exact; outside it a radial or tangential gradient component prevents spurious near-stationarity.

Lemma B.7. *For every $m \geq 1$, every $d \geq 2m$, every orthonormal $U \in \mathbb{R}^{d \times 2m}$, and every $x \in \mathbb{R}^d$,*

$$\widehat{F}_{m,U}(x) - \widehat{F}_{m,U}^\star \leq C_{\mathrm{PL}} m \left\|\nabla \widehat{F}_{m,U}(x)\right\|^2, \qquad C_{\mathrm{PL}} := 40000. \tag{67}$$

Proof. Write $N = 2m$. We first identify the minimum value. For any x, let $y = U^\top \rho_m(x)$. Since $\|y\| \leq \|\rho_m(x)\| \leq \|x\|$,

$$\widehat{F}_{m,U}(x) \geq F_N(y) + \frac{\eta}{2}\|y\|^2 = \Gamma_N(y) \geq \Gamma_N^\star. \tag{68}$$

Conversely, a minimizer $y^\star$ of Γ_N satisfies

$$\frac{\eta}{2}\|y^\star\|^2 \le \Gamma_N(y^\star) \le \Gamma_N(0) = F_N(0) < 5m.$$

Thus $\|y^\star\|^2 < 200m < R_m^2$. Consequently $\rho_m(Uy^\star) = Uy^\star$, and $\widehat{F}_{m,U}(Uy^\star) = \Gamma_N(y^\star)$. Therefore

$$\widehat{F}^\star_{m,U} = \Gamma^\star_N. \tag{69}$$

First consider $\|x\| \le R_m$. Decompose

$$y = U^\top x, \qquad w = x - Uy, \qquad U^\top w = 0.$$

Since $\rho_m(x) = x$,

$$\widehat{F}_{m,U}(x) - \widehat{F}^\star_{m,U} = \Gamma_N(y) - \Gamma^\star_N + \frac{\eta}{2}\|w\|^2, \tag{70}$$

$$\left\|\nabla \widehat{F}_{m,U}(x)\right\|^2 = \|\nabla\Gamma_N(y)\|^2 + \eta^2\|w\|^2. \tag{71}$$

By lemma B.6, $N = 2m$, and $1/(2\eta) = 10$, these identities imply (67) inside the ball.

It remains to consider $r := \|x\| > R_m$. Put

$$e = x/r, \qquad h = h_m(r), \qquad a = h/r, \qquad z = U^\top e, \qquad y = hz,$$
$$g = \nabla F_N(y), \qquad \beta = \langle g, z\rangle.$$

Using (48), the radial and tangential parts of the gradient are orthogonal and satisfy

$$\nabla \widehat{F}_{m,U}(x) = \big(h_m'(r)\beta + \eta r\big)e + a\big(Ug - \beta e\big), \tag{72}$$

so

$$\left\|\nabla \widehat{F}_{m,U}(x)\right\|^2 = \big(h_m'(r)\beta + \eta r\big)^2 + a^2\big(\|g\|^2 - \beta^2\big). \tag{73}$$

If $\beta \ge -\eta r/2$, then, because $0 \le h_m'(r) \le 1$, the radial term in (72) has magnitude at least $\eta r/2$. Suppose instead that $\beta < -\eta r/2$. Then $\langle g, y\rangle = h\beta < 0$. The analogue of (61) without the regularizing term gives

$$\langle \nabla F_N(y), y\rangle \ge \frac{1}{2}\|y\|^2 - 4\sqrt{N}\|y\| - \frac{13}{2}N. \tag{74}$$

Therefore

$$\|y\| < (4 + \sqrt{29})\sqrt{N}.$$

Since $h \ge R_m = 8\sqrt{2}\sqrt{N}$, it follows that $\|z\|^2 < 7/10$. Hence

$$\beta^2 \le \|g\|^2\|z\|^2 \le \frac{7}{10}\|g\|^2, \qquad \|g\|^2 - \beta^2 \ge \frac{3}{7}\beta^2.$$

Also $a|\beta| > \eta h/2 \ge \eta R_m/2$. In either case,

$$\left\|\nabla \widehat{F}_{m,U}(x)\right\|^2 \ge \frac{3\eta^2 R_m^2}{28}. \tag{75}$$

If $R_m < r \le 16R_m$, then (30) and $0 \le H_N \le N$ give

$$F_N(y) \le \frac{1}{2}\|y\|^2 + 3N \le 2R_m^2 + 3N.$$

Because $\widehat{F}^\star_{m,U} \ge 0$,

$$\begin{aligned}\widehat{F}_{m,U}(x) - \widehat{F}^\star_{m,U} &\le 2R_m^2 + 3N + \frac{\eta}{2}(16R_m)^2 \\ &\le 31448\left\|\nabla \widehat{F}_{m,U}(x)\right\|^2 \le C_{\mathrm{PL}} m\left\|\nabla \widehat{F}_{m,U}(x)\right\|^2.\end{aligned}$$

The numerical inequality uses $R_m^2 = 128N$ and (75).

Finally, suppose $r > 16R_m$. By (57), $\|g\| \le 3R_m$, and by (47),

$$\left\|J\rho_m(x)^\top Ug\right\| \le \frac{6R_m^2}{r} \le \frac{\eta r}{2},$$

where the last inequality follows from $r^2 > 256R_m^2$ and $\eta = 1/20$. Thus $\left\|\nabla \widehat{F}_{m,U}(x)\right\| \ge \eta r/2 = r/40$. Moreover,

$$\widehat{F}_{m,U}(x) - \widehat{F}^\star_{m,U} \le 2R_m^2 + 3N + \frac{r^2}{40} = 259N + \frac{r^2}{40} \le 53\left\|\nabla \widehat{F}_{m,U}(x)\right\|^2.$$

This completes all regions and proves (67). □

B.3 Random rotations, scaling, and the stochastic component

This section proves the information-theoretic part of the construction. The argument is a self-contained form of the probabilistic zero-chain and random-rotation technique used in stochastic first-order lower bounds; see, for example, Arjevani et al. (2023b).

The only concentration input for the rotation argument is the following elementary spherical tail estimate.

Lemma B.8. *Let v be uniform on the unit sphere in an r-dimensional Euclidean space, where $r \ge 4$, and let a be a fixed unit vector. For every $s \in (0,1)$,*

$$\mathbb{P}\big(|\langle v, a\rangle| \ge s\big) \le 4\exp\left(-\frac{rs^2}{16}\right). \tag{76}$$

Proof. By rotational invariance, take $a = e_1$ and write $v = Z/\|Z\|$ with $Z \sim N(0, I_r)$. If $|v_1| \ge s$, then either $|Z_1| \ge s\sqrt{r/2}$ or $\|Z\|^2 \le r/2$. The Gaussian tail gives

$$\mathbb{P}\big(|Z_1| \ge s\sqrt{r/2}\big) \le 2e^{-rs^2/4}.$$

For the second event, Markov's inequality applied to $e^{-\|Z\|^2/2}$ gives

$$\mathbb{P}(\|Z\|^2 \le r/2) \le e^{r/4}\mathbb{E}e^{-\|Z\|^2/2} = e^{r/4}2^{-r/2} \le e^{-r/16}.$$

Since $s < 1$, the sum is at most $4e^{-rs^2/16}$. □

We next combine this concentration bound with the one-coordinate revelation rule. The lemma permits coefficients to depend on all current projections, so its proof keeps track of the enlarged projection history.

We state the lemma only in the form needed here.

Lemma B.9. *Let $D \ge M \ge 5$, $q \in (0,1]$, $\delta \in (0, 1/2)$, and $R \ge 1$. Suppose a measurable stochastic oracle $g : \mathbb{R}^D \times \mathcal{Z} \to \mathbb{R}^D$ satisfies, for all x and every history,*

$$\mathrm{prog}_0(g(x,z)) \le \mathrm{prog}_{1/4}(x) + 1 \quad \textit{almost surely}, \tag{77}$$

$$\mathbb{P}\big(\mathrm{prog}_0(g(x,z)) > \mathrm{prog}_{1/4}(x) \mid x, \textit{history}\big) \le q. \tag{78}$$

Let $U \in \mathbb{R}^{d\times D}$ be Haar-uniform on the Stiefel manifold of orthonormal D-frames, independently of the algorithm and the fresh, independent oracle seeds $z_1, z_2, \ldots$. An arbitrary adaptive randomized algorithm makes queries $y_t \in \mathbb{R}^d$ with $\|y_t\| \le R$ and receives $Ug(U^\top y_t, z_t)$.

For the universal constant $C_{\mathrm{rot,s}} = 4096$, if

$$d \ge 2D + C_{\mathrm{rot,s}}R^2\frac{M}{q}\log\left(\frac{16DM}{q\delta}\right), \tag{79}$$

then, for every positive integer

$$t \le \frac{M - \log(2/\delta)}{2q}, \tag{80}$$

$$\mathbb{P}\left(\mathrm{prog}_{1/4}(U^\top y_s) < M \textit{ for every } s \le t\right) \ge 1 - \delta. \tag{81}$$

The probability includes U, the oracle randomness, and the internal randomness of the algorithm.

Proof. We use deferred decisions for the columns $u_1, \dots, u_D$ of U. Put

$$\pi_s := \max_{i \le s} \mathrm{prog}_{1/4}(U^\top y_i), \qquad \gamma_s := \max_{i \le s} \mathrm{prog}_0(g(U^\top y_i, z_i)), \qquad \gamma_0 := 0.$$

Let $\mathcal{V}_s$ be the event $\mathrm{prog}_{1/4}(U^\top y_s) \le \gamma_{s-1}$. Define the two stopping times

$$\tau_V := \inf\{s \ge 1 : \mathcal{V}_s^c\}, \qquad \tau_\gamma := \inf\{s \ge 1 : \gamma_s \ge M\},$$

with the convention $\inf \varnothing = \infty$. If $\pi_t \ge M$, then either $\tau_\gamma \le t$ occurs before τ_V, or $\tau_V \le t$ occurs no later than τ_γ. Indeed, consider the first query whose progress reaches M: if no leakage has occurred, its progress is bounded by the previously revealed support, while otherwise the leakage stopping time has already occurred. Hence

$$\mathbb{P}(\pi_t \ge M) \le \mathbb{P}(\tau_\gamma \le t, \tau_\gamma < \tau_V) + \mathbb{P}(\tau_V \le t, \tau_V \le \tau_\gamma). \tag{82}$$

We first control the progress-before-leakage event. Let r_{A} denote the algorithm's random seed and use the filtration

$$\mathscr{G}_i := \sigma(r_{\mathsf{A}}, U, z_1, \dots, z_i).$$

The query y_i, the event $\mathcal{V}_i$, and γ_{i-1} are $\mathscr{G}_{i-1}$-measurable. Before either stopping time, the support rule (77) implies that $\gamma_i - \gamma_{i-1} \in \{0, 1\}$, and (78) bounds the conditional probability of the value 1 by q. Define the stopped increments

$$\iota_i := (\gamma_i - \gamma_{i-1})\mathbf{1}_{\{i < \tau_V,\, \gamma_{i-1} < M\}}.$$

Then

$$\mathbb{E}[e^{\iota_i} \mid \mathscr{G}_{i-1}] \le 1 - q + qe \le e^{2q}.$$

If $\tau_\gamma \le t$ and $\tau_\gamma < \tau_V$, then $\sum_{i=1}^t \iota_i \ge M$. Markov's inequality and iterated conditional expectations therefore give

$$\mathbb{P}(\tau_\gamma \le t, \tau_\gamma < \tau_V) \le \mathbb{P}\left(\sum_{i=1}^t \iota_i \ge M\right) \le e^{2qt - M}. \tag{83}$$

Under (80), this is at most $\delta/2$.

It remains to control leakage before the support reaches M, namely the second event in (82). Fix a horizon t satisfying (80). For each column index j and query index i, apply Gram–Schmidt to y_i after the vectors $u_1, \dots, u_{j-1}, y_1, \dots, y_{i-1}$. If the residual is nonzero, denote its normalized direction by $w_{j,i}$; otherwise put $w_{j,i} = 0$. Thus the nonzero vectors $w_{j,1}, \dots, w_{j,i}$ are orthonormal and orthogonal to $u_1, \dots, u_{j-1}$.

We record the deferred-decisions step precisely. Let r_{A} denote the algorithm's random seed and define

$$\mathscr{U}_{j-1}^{i-1} := \sigma\big(r_{\mathsf{A}}, z_1, \dots, z_{i-1}, u_1, \dots, u_{j-1}, U^\top y_1, \dots, U^\top y_{i-1}\big). \tag{84}$$

The event $\{j > \gamma_{i-1}\}$ is measurable with respect to this σ-field. On that event, every past oracle response can be reconstructed from the information in (84): the vector $g(U^\top y_s, z_s)$ is known and has support below j, so

$$U g(U^\top y_s, z_s) = \sum_{\ell < j} [g(U^\top y_s, z_s)]_\ell u_\ell.$$

It follows recursively that the next query y_i, and hence $w_{j,i}$, is fixed conditional on $\mathscr{U}_{j-1}^{i-1}$ and $\{j > \gamma_{i-1}\}$.

Put

$$\mathcal{S}_{j,i-1} := \mathrm{span}\{u_1, \dots, u_{j-1}, y_1, \dots, y_{i-1}\}, \qquad P_{j,i-1} := P_{\mathcal{S}_{j,i-1}^\perp}.$$

Conditional on the same information, the normalized residual

$$\widehat{u}_j := \frac{P_{j,i-1} u_j}{\|P_{j,i-1} u_j\|} \tag{85}$$

is uniform on the unit sphere of $\mathcal{S}_{j,i-1}^\perp$ whenever the residual is nonzero. If the residual is zero, then $\langle u_j, w_{j,i} \rangle = 0$ and the tail estimate below is automatic, so no normalized residual

is needed. Here is a complete invariance argument. Let Q be any orthogonal transformation that fixes $\mathcal{S}_{j,i-1}$ pointwise. Left multiplication $U \mapsto QU$ preserves Haar measure. It fixes $u_1, \dots, u_{j-1}$ and every previous query, and it also preserves every conditioned projection because $(QU)^\top y_s = U^\top Q^\top y_s = U^\top y_s$. It therefore preserves both the σ-field in (84) and the event $\{j > \gamma_{i-1}\}$. The corresponding regular conditional law of $P_{j,i-1}u_j$ is consequently invariant under every orthogonal map of $\mathcal{S}_{j,i-1}^\perp$. Its norm is fixed by the conditioned projections and the identity $\|u_j\| = 1$; hence its direction has the unique rotation-invariant probability law on the sphere, namely the uniform law. Since $\dim(\mathcal{S}_{j,i-1}^\perp) \geq d - i - j + 2$, the safe lower bound $d - i - j$ may be used below.

Consequently, whenever $j > \gamma_{i-1}$ and no previous leakage has occurred, $w_{j,i} \in \mathcal{S}_{j,i-1}^\perp$ and $|\langle u_j, w_{j,i}\rangle| \leq |\langle \widehat{u}_j, w_{j,i}\rangle|$. Thus lemma B.8 gives

$$\mathbb{P}\left(|\langle u_j, w_{j,i}\rangle| \geq \frac{1}{4R\sqrt{t}} \;\middle|\; \text{past} \right) \leq 4 \exp\left(-\frac{d-i-j}{256R^2 t} \right). \tag{86}$$

Here conditioning on more information than the algorithm observes is legitimate and only strengthens the deferred-decisions argument.

Suppose all the inner products in (86) are below their threshold for a fixed j and all $i \leq s$. Since u_j is orthogonal to $u_1, \dots, u_{j-1}$, the Gram–Schmidt expansion of y_s and Cauchy–Schwarz give

$$\begin{aligned} |\langle u_j, y_s\rangle| &\leq \left(\sum_{i=1}^{s} |\langle u_j, w_{j,i}\rangle|^2 \right)^{1/2} \left(\sum_{i=1}^{s} |\langle w_{j,i}, y_s\rangle|^2 \right)^{1/2} \\ &< \frac{\sqrt{s}}{4R\sqrt{t}} \|y_s\| \leq \frac{1}{4}. \end{aligned} \tag{87}$$

At the stopping time τ_V on the event $\{\tau_V \leq t, \tau_V \leq \tau_\gamma\}$, one has $\gamma_{\tau_V - 1} < M$, and a failure of one of the events in (86) must occur for some $i \leq \tau_V$ and some $j \leq D$. A union bound therefore yields

$$\mathbb{P}(\tau_V \leq t, \tau_V \leq \tau_\gamma) \leq 4tD \exp\left(-\frac{d-t-D}{256R^2 t} \right). \tag{88}$$

Let $A = \log(16DM/(q\delta))$. Since $t \leq M/(2q)$, the dimension requirement (79) implies $d - t - D \geq \frac{1}{2} C_{\mathrm{rot,s}} R^2 (M/q) A$ and hence

$$\frac{d-t-D}{256R^2 t} \geq 16A.$$

Substituting this into (88), and using $4tD \leq 2DM/q$, shows that the leakage probability is at most $\delta/2$. Together with (82) and (83), this proves (81). □

We now combine the geometry and information arguments.

Proposition B.10. *There is a universal constant $c_{\mathrm{n}} > 0$ such that the following holds. Let $m \geq 5$, $q \in (0, 1]$, and define*

$$d_{m,q} := \left\lceil 4m + 2^{22} \frac{m^2}{q} \log\left(\frac{128m^2}{q} \right) \right\rceil. \tag{89}$$

Let U be Haar-uniform on the Stiefel manifold of $d_{m,q} \times 2m$ orthonormal matrices. For every adaptive randomized algorithm using the oracle (50), whenever $T \leq c_{\mathrm{n}} m/q$,

$$\mathbb{P}_{U,z,\mathsf{A}}\left(\widehat{F}_{m,U}(\widehat{x}_T) - \widehat{F}^\star_{m,U} \geq (2-\eta)m \right) \geq \frac{3}{4}, \tag{90}$$

$$\mathbb{E}_{U,z,\mathsf{A}}\left[\widehat{F}_{m,U}(\widehat{x}_T) - \widehat{F}^\star_{m,U} \right] \geq m. \tag{91}$$

One may take $c_{\mathrm{n}} = 1/16$.

Proof. Given an algorithm querying the clipped oracle at x, simulate it using the strictly stronger rotated-chain oracle as follows. Query $y = \rho_m(x)$, which satisfies $\|y\| < 2R_m$, receive $Ug_{m,q}(U^\top y, z)$, and return to the original algorithm

$$J\rho_m(x)^\top U g_{m,q}(U^\top y, z) + \eta x.$$

This is exactly (50). Thus it suffices to lower-bound the stronger experiment.

Apply lemma B.9 with

$$D = 2m, \qquad M = m, \qquad R = 2R_m = 32\sqrt{m}, \qquad \delta = \frac{1}{4}.$$

The dimension in (89) is larger than the requirement (79). The output point may be arbitrary and need not have been queried. Append $y_{T+1} = \rho_m(\widehat{x}_T)$ as one additional query and ignore its response. If $T \le m/(16q)$ and $m \ge 5$, then

$$T + 1 \le \frac{m - \log 8}{2q},$$

so lemma B.9 implies, with probability at least 3/4,

$$\mathrm{prog}_{1/4}\big(U^\top \rho_m(\widehat{x}_T)\big) < m. \tag{92}$$

On this event, (35) gives

$$F_{2m}\big(U^\top \rho_m(\widehat{x}_T)\big) \ge 2m.$$

On the other hand, because $\sqrt{2m} < R_m$,

$$\widehat{F}^\star_{m,U} \le \widehat{F}_{m,U}(U\mathbf{1}_{2m}) = \frac{\eta}{2}\,\|U\mathbf{1}_{2m}\|^2 = \eta m.$$

The quadratic term at the output is nonnegative, so on (92),

$$\widehat{F}_{m,U}(\widehat{x}_T) - \widehat{F}^\star_{m,U} \ge (2 - \eta)m.$$

The event (92) has probability at least 3/4, so the preceding inequality proves (90). The gap is always nonnegative; hence its expectation is at least $\frac{3}{4}(2-\eta)m > m$, proving (91). □

Remark B.11 (A common fixed hard rotation). Averaging over U is only an existence argument. For a fixed randomized algorithm define

$$p_{\mathsf{A}}(U) = \mathbb{P}_{z,\mathsf{A}}\Big(\widehat{F}_{m,U}(\widehat{x}_T) - \widehat{F}^\star_{m,U} \ge (2-\eta)m\Big).$$

Proposition B.10 gives $\mathbb{E}_U p_{\mathsf{A}}(U) \ge 3/4$. Hence there is a fixed U with $p_{\mathsf{A}}(U) \ge 3/4$. For this same U, nonnegativity of the gap implies

$$\mathbb{E}_{z,\mathsf{A}}[\widehat{F}_{m,U}(\widehat{x}_T) - \widehat{F}^\star_{m,U}] \ge \tfrac{3}{4}(2-\eta)m > m.$$

Thus one rotation is simultaneously hard in probability and expectation. Selecting separate maximizers of the two criteria would not justify this conclusion. The instance and oracle are fixed before the algorithm is run.

The normalized instance is now ready to be calibrated. Its smoothness and PL constants determine the chain width through

$$A_0 := C_{\mathrm{PL}} L_b = 40000 \cdot 128, \qquad \kappa_0 := 12A_0. \tag{93}$$

Assume that $\kappa = L/\mu \ge \kappa_0$ and set

$$m_{\mathrm{s}} := \left\lfloor \frac{\kappa}{2A_0} \right\rfloor, \qquad m := m_{\mathrm{s}}. \tag{94}$$

Then $m \ge 5$ and

$$\frac{\kappa}{3A_0} \le m \le \frac{\kappa}{2A_0}. \tag{95}$$

For $0 < \epsilon \le \Delta_0/10$, choose the objective scale to make the unrevealed tail exceed the target accuracy, and the spatial scale to match L:

$$\lambda := \frac{2\epsilon}{m}, \qquad s^2 := \frac{L}{\lambda L_b}. \tag{96}$$

For a fixed rotation U, define

$$f_U(x) := \lambda \widehat{F}_{m,U}(sx), \tag{97}$$

and scale the oracle consistently:

$$G_U(x,z) := \lambda s\, \widehat{G}_{m,q,U}(sx,z). \tag{98}$$

First we verify the geometric requirements after scaling.

By (51),

$$\mathrm{Lip}(\nabla f_U) \le \lambda s^2 L_b = L. \tag{99}$$

By the global PL estimate for $\widehat{F}_{m,U}$,

$$\begin{aligned} f_U(x) - f_U^\star &\le \lambda C_{\mathrm{PL}} m \|\nabla \widehat{F}_{m,U}(sx)\|^2 \\ &= \frac{C_{\mathrm{PL}} m}{\lambda s^2}\|\nabla f_U(x)\|^2 = \frac{C_{\mathrm{PL}} m L_b}{L}\|\nabla f_U(x)\|^2 \le \frac{1}{2\mu}\|\nabla f_U(x)\|^2. \end{aligned}$$

Thus f_U is globally μ-PL. Moreover,

$$f_U(0) - f_U^\star < 5\lambda m = 10\epsilon. \tag{100}$$

Next we calibrate the revelation probability to the announced noise moment.

For $q < 1$, Lemma B.5 gives

$$\begin{aligned} &\mathbb{E}\|G_U(x,z) - \nabla f_U(x)\|^\alpha \\ &\quad \le 2(\lambda s)^\alpha q^{1-\alpha} = 2\left(\frac{\lambda L}{L_b}\right)^{\alpha/2} q^{1-\alpha}. \end{aligned} \tag{101}$$

Define

$$q_\star := \left[\frac{2(\lambda L/L_b)^{\alpha/2}}{\sigma^\alpha}\right]^{1/(\alpha-1)} = 2^{1/(\alpha-1)}\left(\frac{\lambda L}{L_b\sigma^2}\right)^{\theta_\alpha}, \qquad q := \min\{1, q_\star\} = q_{\mathrm{s}} \quad (\sigma > 0). \tag{102}$$

For $\sigma = 0$, set $q_{\mathrm{s}} = q = 1$ directly and omit $q_\star$; the oracle is exact. For $\sigma > 0$, if $q_\star \le 1$, (101) is at most σ^α. If $q_\star > 1$, then $q = 1$ and the oracle is exact, so its centered moment is zero.

The normalized lower bound transfers exactly through scaling: the objective gap is multiplied by λ, and a scaled algorithm can be simulated on the normalized instance. Thus, for $T \le c_{\mathrm{n}} m/q$, there is one fixed rotation U such that

$$\mathbb{P}\big(f_U(\widehat{x}_T) - f_U^\star \ge 2(2-\eta)\epsilon\big) \ge \frac{3}{4}, \tag{103}$$

$$\mathbb{E}[f_U(\widehat{x}_T) - f_U^\star] \ge 2\epsilon. \tag{104}$$

Suppose first that $\sigma > 0$ and $q_\star \le 1$.

Using (96),

$$\begin{aligned} \frac{m}{q} &= 2^{-1/(\alpha-1)} m \left(\frac{L_b\sigma^2}{\lambda L}\right)^{\theta_\alpha} \\ &= 2^{-1/(\alpha-1)} m \left(\frac{m L_b \sigma^2}{2\epsilon L}\right)^{\theta_\alpha}. \end{aligned} \tag{105}$$

Since $mL_b/L \geq 1/(3C_{\mathrm{PL}}\mu)$ and $m \geq \kappa/(3A_0)$,

$$\frac{m}{q} \geq \frac{2^{-1/(\alpha-1)}}{3A_0}\left(\frac{1}{6C_{\mathrm{PL}}}\right)^{\theta_\alpha}\kappa\left(\frac{\sigma^2}{\mu\epsilon}\right)^{\theta_\alpha}. \tag{106}$$

For the remaining positive-noise case $q_\star > 1$, the calibrated oracle is exact but still gives the required barrier.

The inequality $q_\star > 1$ implies

$$\sigma^2 < 2^{2/\alpha}\frac{\lambda L}{L_b} = 2^{1+2/\alpha}\frac{\epsilon L}{mL_b}.$$

Consequently,

$$\left(\frac{\sigma^2}{\mu\epsilon}\right)^{\theta_\alpha} < \left(2^{1+2/\alpha}\frac{\kappa}{mL_b}\right)^{\theta_\alpha}. \tag{107}$$

Because $m = \lfloor \kappa/(2A_0)\rfloor$ and $m \geq 5$, $\kappa < (12/5)A_0 m$. Hence the right-hand side of (107) is bounded by a constant depending only on α. Also $\kappa \leq (12/5)A_0 m$. Therefore

$$\kappa\left(\frac{\sigma^2}{\mu\epsilon}\right)^{\theta_\alpha} \leq C_\alpha m. \tag{108}$$

The order-m exact-gradient barrier thus dominates the desired lower bound.

Combining the two cases, there is $c_{\mathrm{large},\alpha} > 0$ such that, for $\kappa \geq \kappa_0$, every

$$T \leq c_{\mathrm{large},\alpha}\kappa\left(\frac{\sigma^2}{\mu\epsilon}\right)^{\theta_\alpha} \tag{109}$$

leaves expected error at least 2ϵ and gap at least 2ϵ with probability at least $3/4$ on some fixed valid instance. The initial-gap condition follows from (100), since $10\epsilon \leq \Delta_0$ in the stated accuracy range.

With the width and revelation probability now fixed, the rotation argument requires the ambient dimension

$$d_{\mathrm{s}} := d_{m_{\mathrm{s}},q_{\mathrm{s}}} = \left\lceil 4m_{\mathrm{s}} + 2^{22}\frac{m_{\mathrm{s}}^2}{q_{\mathrm{s}}}\log\left(\frac{128m_{\mathrm{s}}^2}{q_{\mathrm{s}}}\right)\right\rceil. \tag{110}$$

This is precisely (89) at the calibrated parameters. It completes the stochastic construction, including its finite hard dimension.

B.4 Geometrically weighted exact chain and the noiseless component

The stochastic construction is deliberately tuned to the noise scale. Setting $\sigma = 0$ in its final formula only produces the trivial value zero; it does not by itself prove the exact-gradient transient $\Omega(\kappa\log(\Delta_0/\epsilon))$. We now build a separate exact-gradient hard family. Since an exact oracle belongs to $\mathcal{O}_{\alpha,\sigma}(f)$ for every $\sigma \geq 0$, this family can be combined with the stochastic family by taking the maximum of the two minimax obstructions.

The new ingredient is a geometrically weighted robust zero-chain. It has K layers, each of width $m = \Theta(\kappa)$. The layer amplitudes decay geometrically, so the last unresolved layer has value comparable to the prescribed accuracy. Crucially, the total squared weight after any first bad coordinate is only $O(m)$ times that coordinate's squared weight. This keeps the global PL coefficient at $O(m)$ although the total chain length is mK.

We now specify the amplitudes and the weighted chain explicitly.

Fix

$$\gamma := \frac{7}{8}, \qquad r_\gamma := \gamma^2 = \frac{49}{64}. \tag{111}$$

Let $m, K \geq 1$ be integers, put $N := mK$, and define

$$a_i := \gamma^{\lfloor (i-1)/m\rfloor}, \qquad u_i := \frac{x_i}{a_i}, \qquad i = 1,\dots,N. \tag{112}$$

Thus each layer has m equal amplitudes, and successive layers differ by the factor γ. Using the scalar functions from (24)–(26), define

$$\mathsf{W}_{m,K}(x) := a_1^2\phi(u_1) + \sum_{i=1}^{N-1} a_{i+1}^2\hat{\Theta}(u_i)\phi(u_{i+1}) + \sum_{i=1}^{N} a_i^2 B(u_i). \tag{113}$$

For the nonuniform threshold vector $\boldsymbol{a} = (a_1, \dots, a_N)$, put

$$\mathrm{prog}_{\boldsymbol{a}}(x) := \max\left(\left\{i : |x_i| > \frac{a_i}{4}\right\} \cup \{0\}\right). \tag{114}$$

Finally, define the tail squared-weight mass

$$S_j := \sum_{i=j}^{N} a_i^2, \qquad j = 1, \dots, N. \tag{115}$$

Lemma B.12 (Geometric tail mass)**.** *For every j,*

$$S_j \le \frac{m}{1-\gamma^2} a_j^2 = \frac{64}{15} m a_j^2 < 5 m a_j^2. \tag{116}$$

In particular, $S_1 < 5m$.

Proof. Let j belong to layer ℓ, so $a_j = \gamma^\ell$. There are at most m remaining entries in that layer and exactly m entries in each later layer. Therefore

$$S_j \le m\gamma^{2\ell} \sum_{q=0}^{\infty} \gamma^{2q} = \frac{m}{1-\gamma^2} a_j^2.$$

Since $1 - \gamma^2 = 15/64$, the numerical statements follow. □

The tail-mass estimate is what prevents the PL coefficient from growing with the number of layers. We next verify all properties of the weighted chain together, so that the information barrier and its geometry refer to the same function.

Lemma B.13. *There are universal constants*

$$C_{\mathrm{init}} := 11, \qquad L_{\mathrm{W}} := 79, \qquad C_{\mathrm{W}} := 10^5, \tag{117}$$

such that, for every $m, K \ge 1$:

(i) $\mathsf{W}_{m,K} \ge 0$, $\mathsf{W}_{m,K}(\boldsymbol{a}) = 0$, *and*

$$\mathsf{W}_{m,K}(0) \le C_{\mathrm{init}} m. \tag{118}$$

(ii) $\mathsf{W}_{m,K}$ *is globally L_{W}-smooth.*

(iii) If $k = \mathrm{prog}_{\boldsymbol{a}}(x)$, then

$$\nabla_j \mathsf{W}_{m,K}(x) = 0 \qquad (j \ge k+2), \tag{119}$$

and, when $k < N$,

$$|\nabla_{k+1} \mathsf{W}_{m,K}(x)| \le a_{k+1}. \tag{120}$$

(iv) If $\mathrm{prog}_{\boldsymbol{a}}(x) \le N - m$, then

$$\mathsf{W}_{m,K}(x) \ge 2m\gamma^{2(K-1)}. \tag{121}$$

(v) The global gradient-dominance inequality

$$\mathsf{W}_{m,K}(x) \le C_{\mathrm{W}} m \, \|\nabla \mathsf{W}_{m,K}(x)\|^2 \qquad \text{for every } x \in \mathbb{R}^N \tag{122}$$

holds. In particular, its coefficient depends on the layer width m but not on the number of layers K.

Proof. Nonnegativity and $\mathsf{W}_{m,K}(\boldsymbol{a}) = 0$ follow exactly as for the unweighted chain. At the origin all coupling terms vanish, so

$$\mathsf{W}_{m,K}(0) = \phi(0) + 2S_1 < 1 + 10m \le 11m$$

for $m \ge 1$, proving (i).

For $i = 1, \dots, N$, define

$$A_i := \begin{cases} 1, & i = 1, \\ \hat{\Theta}(u_{i-1}), & i \ge 2, \end{cases} \qquad r_i := \frac{a_{i+1}}{a_i} \in \{1, \gamma\} \quad (i < N).$$

A direct differentiation gives

$$\nabla_i \mathsf{W}_{m,K}(x) = a_i \left[A_i \phi'(u_i) + B'(u_i) + \mathbf{1}_{\{i<N\}} r_i^2 \hat{\Theta}'(u_i)\phi(u_{i+1}) \right]. \tag{123}$$

The dependence of this coordinate on x_{i-1}, x_i, x_{i+1} has respective Lipschitz coefficients at most $3, 73, 3$. Omit nonexistent neighbors and, when $i = N$, the entire outgoing term. All adjacent amplitude ratios are at most one, and the scalar estimates are those of Lemma B.1. The same tridiagonal row/column-sum argument used in Lemma B.2 therefore proves global 79-smoothness.

Let $k = \mathrm{prog}_{\boldsymbol{a}}(x)$. If $j \ge k + 2$, then $|u_{j-1}|, |u_j|, |u_{j+1}| \le 1/4$; hence every term in (123) vanishes. For $j = k + 1$, only the incoming term can remain, and its magnitude is at most a_{k+1}. This proves (iii). If the progress is at most $N - m$, the last m normalized coordinates lie in $[-1/4, 1/4]$, where $B = 2$, and their common amplitude is γ^{K-1}. This proves (iv).

It remains to establish the global PL inequality. We divide the proof into four steps.

Step 1: local strong convexity.

Let

$$\mathcal{B}_{\boldsymbol{a}} := \prod_{i=1}^{N} \left[\frac{7}{8} a_i, \frac{9}{8} a_i \right]. \tag{124}$$

Using the explicit formulas for the scalar derivatives (and $e^{-x} \ge 1 - x$), on $[7/8, 9/8]$,

$$\hat{\Theta} \ge \frac{27}{32}, \quad \left|\hat{\Theta}'\right| \le \frac{9}{4}, \quad \phi \le \frac{1}{128}, \quad |\phi'| \le \frac{1}{8}, \quad \phi'' \ge \frac{31}{32}. \tag{125}$$

At points of twice differentiability, the diagonal Hessian entry is

$$A_i \phi''(u_i) + B''(u_i) + \mathbf{1}_{\{i<N\}} r_i^2 \hat{\Theta}''(u_i)\phi(u_{i+1}),$$

and the $(i, i+1)$ entry is $r_i \hat{\Theta}'(u_i)\phi'(u_{i+1})$. If $u_i \ge 1$, the diagonal is at least $1 + (27/32)(31/32) > 7/4$. If $u_i < 1$, then $\hat{\Theta}''(u_i) \le -12$ and $B''(u_i) = -2\hat{\Theta}''(u_i)$, so the last two terms are

$$(-\hat{\Theta}''(u_i))\big(2 - r_i^2 \phi(u_{i+1})\big) > 23.$$

Every off-diagonal entry has magnitude at most $9/32$, and each row has at most two such entries. Gershgorin's theorem gives

$$\nabla^2 \mathsf{W}_{m,K} \succeq I \quad \text{a.e. on } \mathcal{B}_{\boldsymbol{a}}. \tag{126}$$

Since the gradient is Lipschitz, integration along line segments shows that $\mathsf{W}_{m,K}$ is 1-strongly convex on this box. Its zero $\boldsymbol{a}$ lies in the box. Hence, for $x \in \mathcal{B}_{\boldsymbol{a}}$,

$$\mathsf{W}_{m,K}(x) \le \frac{1}{2} \left\| \nabla \mathsf{W}_{m,K}(x) \right\|^2. \tag{127}$$

Step 2: the first bad coordinate.

Let $x \notin \mathcal{B}_{\boldsymbol{a}}$ and let j be the first coordinate with $u_j \notin [7/8, 9/8]$. Then $A_j \ge 27/32$. Put $c_j = r_j^2 \phi(u_{j+1}) \in [0, 1]$ if $j < N$, and $c_j = 0$ if $j = N$. Dividing (123) by a_j gives

$$D_j := \frac{\nabla_j \mathsf{W}_{m,K}(x)}{a_j} = A_j \phi'(u_j) + B'(u_j) + c_j \hat{\Theta}'(u_j). \tag{128}$$

We claim

$$|D_j| \geq \frac{3}{32}. \tag{129}$$

If $u_j > 9/8$, then $\hat{\Theta}'(u_j) = 0$, $B'(u_j) = u_j - 1 \geq 1/8$, and $\phi'(u_j) \geq 0$. If $1/2 < u_j < 7/8$, then $B' + c_j\hat{\Theta}' = (c_j - 2)\hat{\Theta}' \leq 0$, while

$$-A_j\phi'(u_j) \geq \frac{27}{32}\frac{1}{8}e^{-1/128} > \frac{3}{32}.$$

If $-1/2 \leq u_j \leq 1/2$, then $B' = \hat{\Theta}' = 0$ and $-\phi'(u_j) = (1-u_j)e^{-(1-u_j)^2/2} \geq 3/8$, so the same conclusion holds. Finally, if $u_j \leq -1/2$, then either $u_j \leq -1$, in which case $B'(u_j) \leq -1/2$, or $-1 < u_j \leq -1/2$, in which case $-\phi'(u_j) \geq 2e^{-2} > 1/4$; all remaining terms are nonpositive. This proves (129), and therefore

$$|\nabla_j \mathsf{W}_{m,K}(x)| \geq \frac{3}{32}a_j. \tag{130}$$

Step 3: prefix and tail control.

When $j \geq 2$, let $P_j(\cdot; x_j)$ consist of all terms in (113) involving coordinates $1, \ldots, j-1$, including the boundary coupling $a_j^2\hat{\Theta}(u_{j-1})\phi(u_j)$. For $j = 1$, put $P_1 = 0$. For fixed x_j, this prefix is 1-strongly convex on $\prod_{i<j}[7a_i/8, 9a_i/8]$. The preceding Hessian proof still applies: for the final prefix coordinate, $\phi(u_j) \leq 1$ gives $B'' + r^2\hat{\Theta}''\phi(u_j) \geq -\hat{\Theta}'' \geq 12$ whenever that coordinate is below one. Moreover, on every lower face the corresponding partial derivative is negative, and on every upper face it is positive, by the same calculation as in (128). Thus its constrained minimizer is interior. Strong convexity then gives

$$P_j(x_{<j}; x_j) \leq P_j(\boldsymbol{a}_{<j}; x_j) + \frac{1}{2}\left\|\nabla_{<j}\mathsf{W}_{m,K}(x)\right\|^2 \leq a_j^2 + \frac{1}{2}\left\|\nabla_{<j}\mathsf{W}_{m,K}(x)\right\|^2. \tag{131}$$

Define the tail function, with $A_j = 1$ for $j = 1$ and $A_j = \hat{\Theta}(u_{j-1})$ otherwise, by

$$Q_j(x_{j:N}) := A_j a_j^2\phi(u_j) + \sum_{i=j}^{N-1} a_{i+1}^2\hat{\Theta}(u_i)\phi(u_{i+1}) + \sum_{i=j}^{N} a_i^2 B(u_i). \tag{132}$$

Then

$$\nabla Q_j(x_{j:N}) = \nabla_{j:N}\mathsf{W}_{m,K}(x). \tag{133}$$

Let $r = \|x_{j:N}\|$. The non-B part of Q_j is at most S_j and its gradient norm is at most $4\sqrt{S_j}$. Lemma B.1 therefore gives

$$Q_j(x_{j:N}) \leq \frac{1}{2}r^2 + 3S_j, \tag{134}$$

$$\langle \nabla Q_j(x_{j:N}), x_{j:N}\rangle \geq \frac{1}{2}r^2 - 4\sqrt{S_j}\,r - \frac{13}{2}S_j. \tag{135}$$

If $r \geq 18\sqrt{S_j}$, the right-hand side of (135) is at least $r^2/4$, so $\|\nabla Q_j\| \geq r/4$, and (134) gives $Q_j \leq 9\|\nabla Q_j\|^2$. If $r < 18\sqrt{S_j}$, then $Q_j < 165S_j$. Hence, in all cases,

$$Q_j(x_{j:N}) \leq 9\left\|\nabla_{j:N}\mathsf{W}_{m,K}(x)\right\|^2 + 165S_j. \tag{136}$$

Step 4: global assembly.

For $j \geq 2$, the boundary coupling into j is counted in both P_j and Q_j; it is nonnegative. Thus $\mathsf{W}_{m,K} \leq P_j + Q_j$ (with equality when $j = 1$). Combining (131), (136), Lemma B.12, and (130),

$$\begin{aligned}
\mathsf{W}_{m,K}(x) &\leq \frac{1}{2}\left\|\nabla_{<j}\mathsf{W}_{m,K}(x)\right\|^2 + 9\left\|\nabla_{j:N}\mathsf{W}_{m,K}(x)\right\|^2 + a_j^2 + 165S_j \\
&\leq \frac{1}{2}\left\|\nabla_{<j}\mathsf{W}_{m,K}(x)\right\|^2 + 9\left\|\nabla_{j:N}\mathsf{W}_{m,K}(x)\right\|^2 + (1+825m)\frac{1024}{9}\left|\nabla_j\mathsf{W}_{m,K}(x)\right|^2 \\
&\leq 10^5 m\left\|\nabla\mathsf{W}_{m,K}(x)\right\|^2.
\end{aligned}$$

Together with (127), this proves (v) and the lemma. □

The weighted chain is already globally PL. The next construction bounds all effective queries without lifting its zero minimum.

The random-rotation argument requires all effective queries to have bounded norm. The quadratic guard used in the stochastic construction is unsuitable here because it would assign order-m value to the geometrically weighted minimizer and overwhelm the final layer. We instead use a radial guard that vanishes on a ball containing the entire minimizer.

Set

$$R := 256\sqrt{m}, \qquad \eta_{\mathrm{g}} := 16. \tag{137}$$

Define h_R and ρ_R by

$$h_R(r) := \begin{cases} r, & 0 \le r \le R, \\ R + R(1 - e^{-(r-R)/R}), & r > R, \end{cases} \qquad \rho_R(x) := \begin{cases} h_R(\|x\|)x/\|x\|, & x \neq 0, \\ 0, & x = 0. \end{cases} \tag{138}$$

The proof of Lemma B.4, with R_m replaced by R, gives

$$\|\rho_R(x)\| < 2R, \qquad \|J\rho_R(x)\|_{\mathrm{op}} \le 1, \qquad \mathrm{Lip}(J\rho_R) \le \frac{16}{R}. \tag{139}$$

Let

$$Q_R(x) := \frac{\eta_{\mathrm{g}}}{2}\left(\|x\| - \frac{R}{2}\right)_+^2. \tag{140}$$

This is η_{g}-smooth because it equals $\eta_{\mathrm{g}}\,\mathrm{dist}(x, \overline{B}(0, R/2))^2/2$ and the gradient of half the squared distance to a closed convex set is 1-Lipschitz.

For $d \ge N$ and an orthonormal frame $U \in \mathbb{R}^{d\times N}$, define

$$\widehat{\mathsf{W}}_{m,K,U}(x) := \mathsf{W}_{m,K}(U^\top \rho_R(x)) + Q_R(x). \tag{141}$$

Lemma B.14. *Let*

$$L_{\mathrm{det}} := 144, \qquad C_{\mathrm{det}} := 2C_{\mathrm{W}} = 2 \cdot 10^5. \tag{142}$$

Then, for every m, K, d, U as above,

$$\mathrm{Lip}(\nabla\widehat{\mathsf{W}}_{m,K,U}) \le L_{\mathrm{det}}, \tag{143}$$

$$\widehat{\mathsf{W}}_{m,K,U}(x) \le C_{\mathrm{det}} m \left\|\nabla\widehat{\mathsf{W}}_{m,K,U}(x)\right\|^2 \qquad \text{for every } x \in \mathbb{R}^d. \tag{144}$$

Moreover,

$$\widehat{\mathsf{W}}^\star_{m,K,U} = 0, \qquad \widehat{\mathsf{W}}_{m,K,U}(0) \le C_{\mathrm{init}} m. \tag{145}$$

Proof. By Lemma B.1 and the coordinate formula (123),

$$\|\nabla\mathsf{W}_{m,K}(y)\| \le \|y\| + 11\sqrt{S_1}. \tag{146}$$

Indeed, the non-B part has gradient norm at most $4\sqrt{S_1}$, while $|B'(u_i) - u_i| \le 7$. On the range $\|y\| < 2R$ this is at most $3R$. The chain-rule estimate, Lemma B.13, and (139) therefore give

$$\mathrm{Lip}\big(\nabla[\mathsf{W}_{m,K}(U^\top\rho_R(\cdot))]\big) \le 79 + \frac{16}{R}(3R) = 127.$$

Adding the 16-smooth guard proves (143).

Since $\|\boldsymbol{a}\|^2 = S_1 < 5m < R^2/4$, the point $U\boldsymbol{a}$ lies inside the zero region of the guard, where ρ_R is the identity. Both summands in (141) are nonnegative and vanish there. This proves (145); the initial bound follows from Lemma B.13.

We prove (144) in three radial regions. If $\|x\| \le R/2$, then $\rho_R(x) = x$ and $Q_R(x) = 0$. Hence

$$\widehat{\mathsf{W}}(x) = \mathsf{W}(U^\top x), \qquad \left\|\nabla\widehat{\mathsf{W}}(x)\right\| = \left\|\nabla\mathsf{W}(U^\top x)\right\|,$$

and Lemma B.13 applies.

Suppose $R/2 < r := \|x\| \le R$. Put

$$y = U^\top x, \qquad g = \nabla \mathsf{W}_{m,K}(y), \qquad e = x/r, \qquad b = \eta_{\mathrm{g}}(r - R/2).$$

Then

$$\nabla \widehat{\mathsf{W}}(x) = Ug + be. \tag{147}$$

The weighted analogue of (74), obtained from (30), is

$$\langle \nabla \mathsf{W}(y), y \rangle \ge \frac{1}{2}\|y\|^2 - 4\sqrt{S_1}\,\|y\| - \frac{13}{2}S_1. \tag{148}$$

Because $R/4 \ge 18\sqrt{S_1}$, if $\|y\| \ge R/4$ then $\langle g, y\rangle \ge 0$, and (147) gives $\left\|\nabla \widehat{\mathsf{W}}\right\|^2 \ge \|g\|^2 + b^2$. If $\|y\| < R/4$, set $z = U^\top e$. Since $r > R/2$, $\|z\| < 1/2$, and the orthogonal decomposition of e gives

$$\begin{aligned}\|Ug + be\|^2 &= \|g + bz\|^2 + b^2(1 - \|z\|^2) \\ &\ge \frac{1}{2}\|g\|^2 + b^2(1 - 2\|z\|^2) \ge \frac{1}{2}(\|g\|^2 + b^2).\end{aligned}$$

In both cases,

$$\left\|\nabla \widehat{\mathsf{W}}(x)\right\|^2 \ge \frac{1}{2}(\|g\|^2 + b^2). \tag{149}$$

Now $\mathsf{W}(y) \le C_{\mathrm{W}} m \|g\|^2$ and $Q_R(x) = b^2/(2\eta_{\mathrm{g}})$, so (144) follows in the annulus with coefficient $2C_{\mathrm{W}}m$.

Finally, let $r > R$. On the clipped range, (146) gives $\left\|\nabla \mathsf{W}(U^\top \rho_R(x))\right\| \le 3R$, and $\|J\rho_R\| \le 1$. The guard gradient has magnitude $b = \eta_{\mathrm{g}}(r - R/2) \ge 8R$. Therefore

$$\left\|\nabla \widehat{\mathsf{W}}(x)\right\| \ge b - 3R \ge 5R, \qquad \left\|\nabla \widehat{\mathsf{W}}(x)\right\| \ge \frac{5}{8}b. \tag{150}$$

Also, by $B(t) \le t^2/2 + 2$ and nonnegativity of the coupling terms,

$$\mathsf{W}(U^\top \rho_R(x)) \le \frac{1}{2}\left\|U^\top \rho_R(x)\right\|^2 + 3S_1 < 3R^2. \tag{151}$$

Equations (150) and (151) control both this term and $Q_R = b^2/(2\eta_{\mathrm{g}})$ by a universal multiple of the squared gradient, which is smaller than $C_{\mathrm{det}} m$ for $m \ge 1$. This completes the proof. □

The decreasing amplitudes require a rotation lemma with nonuniform activation thresholds. The proof below adapts the earlier projection-history argument to the smallest threshold.

Lemma B.15. *Let $D \ge M \ge 2$, $R_0 \ge 1$, $\delta \in (0, 1/2)$, and $0 < \underline{a} \le a_j \le 1$ for $j = 1, \dots, D$, and define $\mathrm{prog}_{\boldsymbol{a}}$ as in (114). Suppose a measurable deterministic map $h : \mathbb{R}^D \to \mathbb{R}^D$ satisfies*

$$\mathrm{prog}_0(h(z)) \le \mathrm{prog}_{\boldsymbol{a}}(z) + 1 \qquad \text{for every } z. \tag{152}$$

Let $U \in \mathbb{R}^{d\times D}$ be Haar-uniform on the Stiefel manifold and independent of an arbitrary randomized adaptive algorithm. The algorithm makes an integer number $1 \le t < M$ of queries y_i satisfying $\|y_i\| \le R_0$ and receives $Uh(U^\top y_i)$. For the universal constant $C_{\mathrm{rot,d}} = 1024$, if

$$d \ge 2D + C_{\mathrm{rot,d}} \frac{R_0^2 t}{\underline{a}^2} \log\left(\frac{8Dt}{\delta}\right), \tag{153}$$

then

$$\mathbb{P}\left(\mathrm{prog}_{\boldsymbol{a}}(U^\top y_i) < M \text{ for all } i \le t\right) \ge 1 - \delta. \tag{154}$$

The probability includes the random frame and the algorithm's internal randomness.

Proof. Condition first on the algorithm's random seed. Let

$$\gamma_i := \max_{s \le i} \mathrm{prog}_0(h(U^\top y_s)), \qquad \gamma_0 := 0.$$

Call query i nonleaking if $\mathrm{prog}_{\boldsymbol{a}}(U^\top y_i) \le \gamma_{i-1}$. On a nonleaking history, (152) implies $\gamma_i \le \gamma_{i-1} + 1$, and hence $\gamma_i \le i$.

We bound the probability of the first leakage by deferred decisions. For column u_j of U and query y_i, apply Gram–Schmidt to y_i after $u_1, \dots, u_{j-1}, y_1, \dots, y_{i-1}$; call the normalized residual $w_{j,i}$, or zero if the residual vanishes. On the event $j > \gamma_{i-1}$, all previous responses use only columns below j and can be reconstructed from the previous projections and those columns. Thus, conditional on this enlarged history, y_i and $w_{j,i}$ are fixed, while the normalized residual of u_j orthogonal to

$$\operatorname{span}\{u_1, \dots, u_{j-1}, y_1, \dots, y_{i-1}\}$$

is uniform on the sphere in that orthogonal complement. This is exactly the Haar-invariance argument given in the proof of Lemma B.9; the remaining dimension is at least $d - i - j$.

Lemma B.8 therefore yields, conditionally on every nonleaking past,

$$\mathbb{P}\left(|\langle u_j, w_{j,i} \rangle| \geq \frac{\underline{a}}{4R_0\sqrt{t}} \;\middle|\; \text{past} \right) \leq 4 \exp\left[-\frac{(d-i-j)\underline{a}^2}{256 R_0^2 t} \right]. \tag{155}$$

If none of these events occurs for a fixed j up to query s, orthogonality and Cauchy–Schwarz give

$$|\langle u_j, y_s \rangle| \leq \left(\sum_{i=1}^{s} |\langle u_j, w_{j,i} \rangle|^2 \right)^{1/2} \|y_s\| < \frac{\underline{a}}{4} \leq \frac{a_j}{4}.$$

With $C_{\mathrm{rot,d}} = 1024$, the exponent in (155) is at least $4 \log(8Dt/\delta)$. A union bound over $i \leq t$ and $j \leq D$ therefore makes the probability of any leakage at most $4Dt(\delta/(8Dt))^4 \leq \delta$. On the complementary event, $\gamma_i \leq i < M$ and every query has progress at most $\gamma_{i-1} < M$. This proves (154). Averaging back over the algorithm's random seed completes the proof. □

We now combine the weighted barrier, compactification, and rotation estimate, and scale them to the prescribed smoothness and PL constants.

Theorem B.16. *There are universal constants $\kappa_{\mathrm{det}} \geq 1$, $c_{\mathrm{det}} > 0$, and $c_{0,\mathrm{det}} > 0$ such that the following holds. Let $\kappa = L/\mu \geq \kappa_{\mathrm{det}}$ and $0 < \epsilon \leq c_{0,\mathrm{det}} \Delta_0$. For every adaptive randomized gradient-only method using fewer than*

$$c_{\mathrm{det}} \kappa \log \frac{\Delta_0}{\epsilon} \tag{156}$$

exact-gradient calls, there are a finite dimension d and a fixed globally L-smooth, globally μ-PL function with initial gap at most Δ_0 such that

$$\mathbb{P}\big(f(\widehat{x}_T) - f^\star \geq 4\epsilon\big) \geq \frac{3}{4}. \tag{157}$$

In particular, its expected error is larger than ϵ. The proof gives an explicit admissible dimension in (163).

Proof. Let

$$m_{\mathrm{d}} := \left\lfloor \frac{\kappa}{4 C_{\mathrm{det}} L_{\mathrm{det}}} \right\rfloor, \qquad m := m_{\mathrm{d}}. \tag{158}$$

Take $\kappa_{\mathrm{det}} := 24 C_{\mathrm{det}} L_{\mathrm{det}}$. Then $m \geq 5$ and $m \geq \kappa/(8 C_{\mathrm{det}} L_{\mathrm{det}})$. Put

$$\lambda := \frac{\Delta_0}{C_{\mathrm{init}} m}, \qquad s^2 := \frac{L}{\lambda L_{\mathrm{det}}}. \tag{159}$$

Let

$$K := 2 + \left\lfloor \frac{\log(\Delta_0/(16 C_{\mathrm{init}} \epsilon))}{\log(1/r_\gamma)} \right\rfloor. \tag{160}$$

Take $c_{0,\mathrm{det}} := (16 C_{\mathrm{init}})^{-2}$. Then the logarithm is nonnegative and, writing $B = \log(16 C_{\mathrm{init}})$, $\log(\Delta_0/\epsilon) \geq 2B$. Since $1 + \lfloor z \rfloor \geq z$,

$$K - 1 \geq \frac{1}{2 \log(1/r_\gamma)} \log \frac{\Delta_0}{\epsilon} =: c_K \log \frac{\Delta_0}{\epsilon}. \tag{161}$$

For an orthonormal U, define

$$f_U(x) := \lambda \widehat{\mathsf{W}}_{m,K,U}(sx). \tag{162}$$

By Lemma B.14, this function is L-smooth. Furthermore,

$$\begin{aligned} f_U(x) - f_U^\star &\le \lambda C_{\mathrm{det}} m \left\| \nabla \widehat{\mathsf{W}}_{m,K,U}(sx) \right\|^2 \\ &= \frac{C_{\mathrm{det}} m L_{\mathrm{det}}}{L} \left\| \nabla f_U(x) \right\|^2 \le \frac{1}{4\mu} \left\| \nabla f_U(x) \right\|^2 \le \frac{1}{2\mu} \left\| \nabla f_U(x) \right\|^2 . \end{aligned}$$

Thus it is globally μ-PL, and $f_U(0) - f_U^\star \le \lambda C_{\mathrm{init}} m = \Delta_0$.

The calibrated chain has $N = mK$ coordinates; set $H := N - m = m(K-1)$ for the number that must be traversed before its final layer. Recall that $R = 256\sqrt{m}$ is the clipping radius. The scaled exact gradient can be simulated from the rotated-chain response. At an external query x, send $y = \rho_R(sx)$ to the oracle returning $U \nabla \mathsf{W}_{m,K}(U^\top y)$, and compute

$$\lambda s \left[J\rho_R(sx)^\top U \nabla \mathsf{W}_{m,K}(U^\top \rho_R(sx)) + \nabla Q_R(sx) \right] = \nabla f_U(x).$$

Thus every effective query has norm below $2R$. Append the algorithm's output as one final effective query whose response is ignored. Apply Lemma B.15 with

$$D = N, \quad M = H + 1, \quad R_0 = 2R, \quad \underline{a} = a_N = \gamma^{K-1}, \quad \delta = \frac{1}{4}.$$

For $T \le H/4$, the appended output gives at most $t_{\mathrm{d}} := \lfloor H/4 \rfloor + 1$ effective queries. In particular, $t_{\mathrm{d}} < H + 1 = M$. This determines an admissible ambient dimension:

$$d_{\mathrm{d}} := \left\lceil 2D + C_{\mathrm{rot,d}} \frac{(2R)^2 t_{\mathrm{d}}}{\underline{a}^2} \log(32 D t_{\mathrm{d}}) \right\rceil . \tag{163}$$

It satisfies (153), so with probability at least $3/4$,

$$\mathrm{prog}_{\boldsymbol{a}}\big(U^\top \rho_R(s\widehat{x}_T)\big) \le N - m.$$

The compactified guard is nonnegative, so Lemma B.13 gives on this event

$$f_U(\widehat{x}_T) - f_U^\star \ge 2\lambda m \gamma^{2(K-1)}. \tag{164}$$

If $x_\epsilon = \log(\Delta_0/(16 C_{\mathrm{init}}\epsilon))/\log(1/r_\gamma)$, then $K - 1 = 1 + \lfloor x_\epsilon \rfloor \le 1 + x_\epsilon$ and, because $r_\gamma < 1$,

$$r_\gamma^{K-1} \ge r_\gamma^{1+x_\epsilon} = 16 C_{\mathrm{init}} r_\gamma \frac{\epsilon}{\Delta_0}.$$

Consequently,

$$f_U(\widehat{x}_T) - f_U^\star \ge 32 r_\gamma \epsilon > 4\epsilon. \tag{165}$$

Averaging the probability over Haar U and then using $\sup_U p(U) \ge \mathbb{E}_U p(U)$ selects one fixed rotation for the given randomized algorithm.

Finally, (158) and (161) imply $H \ge c\kappa \log(\Delta_0/\epsilon)$. Choosing c_{det} below one quarter of this universal constant proves the theorem. □

B.5 Completion of the high-dimensional lower bound

All ingredients are now available. We combine the two families at the minimax level rather than adding their objectives, then verify the finite hard dimension and its extension to larger prescribed dimensions.

Proof of Theorem 3.1. Choose $\kappa_{\mathrm{hd}} := \max\{\kappa_0, \kappa_{\mathrm{det}}\}$ and $c_{0,\alpha} := \min\{1/10, c_{0,\mathrm{det}}\}$. These choices enforce the parameter ranges of both constructions. To place either hard instance in one prescribed dimension, set

$$d_{\mathrm{hd}} := \max\{d_{\mathrm{s}}, d_{\mathrm{d}}\}. \tag{166}$$

Let

$$A := \kappa \log \frac{\Delta_0}{\epsilon}, \qquad B := \kappa \left(\frac{\sigma^2}{\mu\epsilon} \right)^{\theta_\alpha}.$$

The stochastic construction above provides a constant $c_{\mathrm{s},\alpha} > 0$ such that every $T < c_{\mathrm{s},\alpha} B$ fails on a fixed legal instance in dimension d_{s}. The exact-gradient construction provides a universal $c_{\mathrm{d}} > 0$ such that every $T < c_{\mathrm{d}} A$ fails on a fixed legal instance in dimension d_{d}. The latter oracle has centered α-moment zero and is therefore legal for every announced $\sigma \geq 0$.

Choose

$$c_{1,\alpha} := \frac{1}{4} \min\{c_{\mathrm{s},\alpha}, c_{\mathrm{d}}\}.$$

If $T < c_{1,\alpha}(A + B)$, then $T < \frac{1}{2} \min\{c_{\mathrm{s},\alpha}, c_{\mathrm{d}}\} \max\{A, B\}$. If $A \geq B$, the exact-gradient family applies; if $B > A$, the stochastic family applies. In either case the selected instance has dimension at most d_{hd}, expected gap larger than ϵ, and the stated constant-probability error. This proves (8).

It remains to simplify the dimension. Since $m_{\mathrm{s}} = \Theta(\kappa)$ and, when $\sigma > 0$,

$$q_{\mathrm{s}}^{-1} \leq C_\alpha \max\left\{1, \left(\frac{\sigma^2}{\mu\epsilon}\right)^{\theta_\alpha}\right\},$$

equation (110) gives

$$d_{\mathrm{s}} = \widetilde{O}_\alpha \left[\kappa^2 \left(1 + \left(\frac{\sigma^2}{\mu\epsilon}\right)^{\theta_\alpha}\right)\right]. \tag{167}$$

For the exact chain, $m_{\mathrm{d}} = \Theta(\kappa)$, $K = \Theta(\log(\Delta_0/\epsilon))$, and $t_{\mathrm{d}} = O(m_{\mathrm{d}} K)$. The choice of K gives

$$r_\gamma^{K-1} \geq 16 C_{\mathrm{init}} r_\gamma \frac{\epsilon}{\Delta_0}, \qquad \underline{a}^{-2} \leq \frac{\Delta_0}{16 C_{\mathrm{init}} r_\gamma \epsilon}.$$

Using $R^2 = 2^{16} m_{\mathrm{d}}$ in (163),

$$d_{\mathrm{d}} = \widetilde{O}\left(\kappa^2 \frac{\Delta_0}{\epsilon}\right). \tag{168}$$

Because Δ_0/ϵ is bounded below by a universal constant in the theorem's range, combining these dimension estimates gives

$$d_{\mathrm{hd}} = \widetilde{O}_\alpha\left(\kappa^2 \left[\frac{\Delta_0}{\epsilon} + Q_\epsilon\right]\right). \tag{169}$$

To obtain the prescribed-dimensional formulation in Theorem 3.1, let $d \geq d_{\mathrm{hd}}$ and let a selected hard instance have dimension $d_0 \leq d_{\mathrm{hd}}$. Extend it by

$$\widetilde{f}(y, z) = f(y) + \frac{\mu}{2}\|z\|^2, \qquad \widetilde{G}(y, z) = (G(y), \mu z), \qquad (y, z) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d - d_0}.$$

This extension is L-smooth and μ-PL, has the same initial gap and centered moment bound, and reveals no new information about the hard part. Any algorithm on the extension can be simulated with one original oracle call per query. Its gap is at least the original-part gap. Thus the lower bound holds in every dimension at or above the explicit threshold. □

The delicate points are the following. Bernoulli thinning has centered moment $O(q^{1-\alpha})$, not $O(q^{-\alpha})$. Under the scaling $f(x) = \lambda \widehat{F}(sx)$, the oracle noise is multiplied by $\lambda s = \sqrt{\lambda L / L_b}$, which is why the exponent is θ_α. The case $q_{\mathrm{s}} = 1$ is not discarded: the exact order-m_{s} chain already dominates the desired stochastic expression there. The weighted chain's PL coefficient is independent of its number of layers because $S_j \leq 5 m_{\mathrm{d}} a_j^2$ while $|\nabla_j \mathsf{W}| \gtrsim a_j$. Finally, Haar randomization is only an existence device: averaging over U and then taking a supremum selects one fixed hard rotation for the given randomized algorithm.

C Proofs for Section 4

We first resolve the distinction between ordinary PL and feasible descent, then analyze the robust gradient estimator, and finally combine the two through a deterministic contraction. Throughout this section $\mathcal{X}$ is nonempty, closed, and convex. The mirror map is differentiable

on an open convex neighborhood of $\mathcal{X}$ and is 1-strongly convex. Its Bregman divergence therefore satisfies (9). The model $Q_x(z;v) = \langle v, z-x\rangle + LD_\psi(z,x)$ is continuous, coercive, and L-strongly convex as a function of $z \in \mathcal{X}$. Consequently its minimizer exists and is unique, even on an unbounded domain. The mirror-PL assumption is the feasible-decrement inequality (12), not ordinary gradient-norm PL on a constrained set.

Proof of Proposition 4.2. Let

$$\mathcal{X} := \mathbb{R} \times [0,\infty), \qquad u(s) := 1 - \cos s, \qquad q(s) := u(s) - u(s)^2,$$

and define

$$f(s,y) := q(s) + (y+1)^2. \tag{170}$$

Because $u(s) \in [0,2]$, one has $-2 \le q(s) \le 1/4$. Moreover,

$$q'(s) = \sin s\,(2\cos s - 1), \qquad q''(s) = 4\cos^2 s - \cos s - 2.$$

The quadratic polynomial $4c^2 - c - 2$ lies in $[-33/16, 3]$ for $c \in [-1,1]$. Hence $|q''(s)| \le 3$, and the Hessian of f is the diagonal matrix $\mathrm{diag}(q''(s), 2)$. Thus f is globally 3-smooth.

The minimum over $\mathcal{X}$ is attained when $y = 0$ and $q(s) = -2$, so $f^\star := \inf_{\mathcal{X}} f = -1$. Put $a = (y+1)^2 \ge 1$. Then

$$f(s,y) - f^\star = q(s) + a + 1 \le a + \frac{5}{4},$$

whereas

$$\frac{1}{2}\,\|\nabla f(s,y)\|^2 = \frac{1}{2}\,|q'(s)|^2 + 2(y+1)^2 \ge 2a.$$

Since $2a \ge \frac{8}{9}(a + 5/4)$ for every $a \ge 1$, inequality (13) holds with $\mu = 8/9$.

Now take $x_0 = (0,0)$. Then $\nabla f(x_0) = (0,2)$ and $f(x_0) - f^\star = 2$. Let ψ be any differentiable strongly convex mirror map, and let $\eta_{\mathrm{md}} > 0$ be arbitrary. For every $z = (z_1, z_2) \in \mathcal{X}$,

$$\langle \nabla f(x_0), z - x_0\rangle + \frac{1}{\eta_{\mathrm{md}}} D_\psi(z, x_0) = 2z_2 + \frac{1}{\eta_{\mathrm{md}}} D_\psi(z, x_0) \ge 0,$$

with equality at $z = x_0$. Thus x_0 is the unique mirror-descent update. The same argument repeats at every iteration, so the method never reaches a minimizer. This remains true for an exact oracle, completing the proof. □

Remark C.1. The obstruction is not a technical artifact: a constrained upper bound must control a feasible descent measure rather than the full gradient norm. We therefore use Definition 4.1. In the unconstrained Euclidean setting it is exactly the ordinary PL condition used in the lower bound, so the two parts of the manuscript match on their common domain.

C.1 Centered clipping under a finite α-moment

This subsection separates three tasks: controlling the moment of a block mean, bounding the number of unreliable blocks, and converting those bounds into a translation-invariant gradient estimate. Clipping is centered at an estimated mean, so none of the following bounds requires a bounded true gradient. The geometric-median principle is classical (Minsker, 2015; Lugosi & Mendelson, 2019); the elementary conditional argument below makes the needed constants explicit.

Lemma C.2. *Let $\zeta_1, \dots, \zeta_b$ be independent random vectors in a Euclidean space with $\mathbb{E}\zeta_i = 0$ and $\mathbb{E}\|\zeta_i\|^\alpha \le \sigma^\alpha$. Then*

$$\mathbb{E}\left\|\frac{1}{b}\sum_{i=1}^{b}\zeta_i\right\|^\alpha \le \frac{2^\alpha \sigma^\alpha}{b^{\alpha-1}}. \tag{171}$$

The same statement holds conditionally on any sigma-field with respect to which the vectors are conditionally independent and centered.

Proof. Introduce an independent copy $\zeta_1', \dots, \zeta_b'$ and independent Rademacher signs ε_i. Jensen's inequality and conditional symmetry give

$$\begin{aligned}\mathbb{E}\left\|\sum_i \zeta_i\right\|^\alpha &\le \mathbb{E}\left\|\sum_i (\zeta_i - \zeta_i')\right\|^\alpha \\ &= \mathbb{E}\left\|\sum_i \varepsilon_i(\zeta_i - \zeta_i')\right\|^\alpha \\ &\le 2^{\alpha-1}\mathbb{E}\left\|\sum_i \varepsilon_i\zeta_i\right\|^\alpha + 2^{\alpha-1}\mathbb{E}\left\|\sum_i \varepsilon_i\zeta_i'\right\|^\alpha \\ &= 2^\alpha \mathbb{E}\left\|\sum_i \varepsilon_i\zeta_i\right\|^\alpha.\end{aligned}$$

Conditioning on $\zeta_1, \dots, \zeta_b$ and using $\alpha/2 \le 1$,

$$\begin{aligned}\mathbb{E}_\varepsilon\left\|\sum_i \varepsilon_i\zeta_i\right\|^\alpha &\le \left(\mathbb{E}_\varepsilon\left\|\sum_i \varepsilon_i\zeta_i\right\|^2\right)^{\alpha/2} \\ &= \left(\sum_i \|\zeta_i\|^2\right)^{\alpha/2} \le \sum_i \|\zeta_i\|^\alpha.\end{aligned}$$

Taking expectations yields $\mathbb{E}\|\sum_i \zeta_i\|^\alpha \le 2^\alpha b\sigma^\alpha$. Division by b^α proves the result. Every step is valid after conditioning, proving the conditional version. □

Fix a query x, target accuracy $r > 0$, and failure probability $\rho \in (0, 1/2)$. Define

$$B_\alpha := (16 \cdot 8^\alpha)^{1/(\alpha-1)}, \qquad b := \max\left\{1, \left\lceil B_\alpha \left(\frac{\sigma}{r}\right)^{\alpha/(\alpha-1)} \right\rceil\right\}, \qquad K := \left\lceil 16 \log \frac{4}{\rho} \right\rceil. \tag{172}$$

Draw $2Kb$ fresh samples at the fixed point x, divide them into $2K$ blocks, and write

$$Z_j := \frac{1}{b}\sum_{\ell=1}^{b} G(x, \xi_{j,\ell}), \qquad j = 1, \dots, 2K. \tag{173}$$

Choose a geometric median of the first half,

$$a \in \arg\min_{u \in \mathbb{R}^d} \sum_{j=1}^{K} \|u - Z_j\|, \tag{174}$$

and define

$$Y_j := a + \mathrm{clip}_r(Z_{K+j} - a), \qquad \widehat{g}(x) := \frac{1}{K}\sum_{j=1}^{K} Y_j. \tag{175}$$

The block-moment estimate gives only a constant success probability for one block. We therefore need a concentration statement that remains valid when its indicators are merely adapted rather than independent.

Lemma C.3. *Let $I_1, \dots, I_K \in \{0, 1\}$ be adapted to a filtration and suppose $\mathbb{E}[I_j \mid \mathscr{F}_{j-1}] \le p$. Then, for $a > p$,*

$$\mathbb{P}\left(\sum_{j=1}^{K} I_j \ge aK\right) \le \exp[-2K(a-p)^2]. \tag{176}$$

In particular, for $p = 1/16$ and $a = 1/4$, the right-hand side is $\exp(-9K/128)$.

Proof. Put $D_j = I_j - \mathbb{E}[I_j \mid \mathscr{F}_{j-1}]$. Conditional Hoeffding's lemma gives $\mathbb{E}[e^{\lambda D_j} \mid \mathscr{F}_{j-1}] \le e^{\lambda^2/8}$. Iteration and Markov's inequality imply

$$\mathbb{P}\left(\sum_{j=1}^{K} D_j \ge u\right) \le \exp(-2u^2/K).$$

If $\sum I_j \ge aK$, then $\sum D_j \ge (a-p)K$, proving the claim. □

We can now combine the block and counting estimates. The pilot controls the clipping center, and the second collection limits the contribution of all remaining outliers.

Lemma C.4. *Under (5), for every adaptively selected x, every $r > 0$, and $\rho \in (0, 1/2)$, the estimator above satisfies*

$$\mathbb{P}\left(\|\widehat{g}(x) - \nabla f(x)\| \le r \mid \textit{history before the estimator}\right) \ge 1 - \rho. \tag{177}$$

It uses at most

$$C_{\mathrm{est},\alpha}\left[1 + \left(\frac{\sigma}{r}\right)^{\alpha/(\alpha-1)}\right]\log\frac{4}{\rho} \tag{178}$$

stochastic-gradient calls, where one may take $C_{\mathrm{est},\alpha} = 34(1 + B_\alpha)$.

Proof. Condition on the full history before the estimator and put $m_x = \nabla f(x)$. Lemma C.2 and the definition of b give, conditionally at the beginning of every block,

$$\mathbb{E}\|Z_j - m_x\|^\alpha \le \frac{2^\alpha \sigma^\alpha}{b^{\alpha-1}}.$$

Consequently,

$$\mathbb{P}\left(\|Z_j - m_x\| > \frac{r}{4} \;\middle|\; \text{past blocks}\right) \le \frac{2^\alpha 4^\alpha \sigma^\alpha}{b^{\alpha-1} r^\alpha} = \frac{8^\alpha \sigma^\alpha}{b^{\alpha-1} r^\alpha} \le \frac{1}{16}. \tag{179}$$

The last inequality also holds when $b = 1$: in that case $B_\alpha(\sigma/r)^{\alpha/(\alpha-1)} \le 1$, which is equivalent to the required bound.

Call a block good if $\|Z_j - m_x\| \le r/4$. Applying Lemma C.3 separately to the pilot and clipping collections shows that, with conditional probability at least $1 - \rho/2$, at least $3K/4$ blocks are good in each collection. We work on this event.

Let G, B be the good and bad pilot indices and put $d_0 = \|a - m_x\|$. For $j \in G$,

$$\|a - Z_j\| - \|m_x - Z_j\| \ge d_0 - r/2,$$

whereas for $j \in B$ the same difference is at least $-d_0$. Optimality of the geometric median gives

$$0 \ge (|G| - |B|)d_0 - |G|r/2.$$

Since $|G| \ge 3K/4$ and $|B| \le K/4$,

$$\|a - m_x\| \le \frac{3r}{4}. \tag{180}$$

For a good block in the second collection, $\|Z_{K+j} - a\| \le r$, so clipping leaves it unchanged and $\|Y_j - m_x\| \le r/4$. For a bad block, $\|Y_j - m_x\| \le r + 3r/4 = 7r/4$. At most $K/4$ blocks are bad; hence

$$\|\widehat{g}(x) - m_x\| \le \frac{3}{4}\frac{r}{4} + \frac{1}{4}\frac{7r}{4} = \frac{5r}{8} < r.$$

This proves (177). Finally, $b \le 1 + B_\alpha(\sigma/r)^{\alpha/(\alpha-1)}$ and $K \le 17\log(4/\rho)$, which gives (178). □

C.2 Inexact mirror contraction and the stagewise method

The preceding estimator supplies a deterministic error tolerance on a simultaneous high-probability event. We now quantify how that error perturbs the exact mirror step; no probabilistic argument is needed in this part. For L-smooth f, (9) implies $f(z) \le f(x) + Q_x(z; \nabla f(x))$. Hence $\mathcal{G}^2_{\psi,L}(x) \le 2L(f(x) - f^\star)$. Together with mirror-PL, this shows $\mu \le L$ whenever the objective has a nonoptimal feasible point. If $\mu > L$, every feasible point is optimal and returning x_0 without any oracle call suffices. This is the first branch of Algorithm 1 and Theorem 4.3. The positive-query bound below concerns only the remaining case $\kappa = L/\mu \ge 1$.

Given $x \in \mathcal{X}$ and an approximate gradient $\widehat{g} = \nabla f(x) + e$, define

$$x_+ := \arg\min_{z \in \mathcal{X}} Q_x(z; \widehat{g}). \tag{181}$$

Let

$$z_+ := \arg\min_{z \in \mathcal{X}} Q_x(z; \nabla f(x)) \tag{182}$$

be the corresponding exact step.

Lemma C.5. *Under 1-strong convexity of ψ, L-smoothness of $f(x)$, and (12),*

$$f(x_+) - f^\star \le \left(1 - \frac{\mu}{L}\right)(f(x) - f^\star) + \frac{1}{L}\|e\|^2. \tag{183}$$

Proof. The variational inequalities for (181) and (182) are

$$\langle \widehat{g} + L(\nabla\psi(x_+) - \nabla\psi(x)), z_+ - x_+ \rangle \ge 0, \tag{184}$$
$$\langle \nabla f(x) + L(\nabla\psi(z_+) - \nabla\psi(x)), x_+ - z_+ \rangle \ge 0. \tag{185}$$

Adding them and using

$$\langle \nabla\psi(x_+) - \nabla\psi(z_+), x_+ - z_+ \rangle \ge \|x_+ - z_+\|^2$$

yields

$$L\|x_+ - z_+\|^2 \le \langle e, z_+ - x_+ \rangle \le \|e\|\,\|x_+ - z_+\|. \tag{186}$$

Consequently,

$$\|x_+ - z_+\| \le \frac{\|e\|}{L}. \tag{187}$$

Optimality of x_+ for the perturbed model gives $Q_x(x_+; \widehat{g}) \le Q_x(z_+; \widehat{g})$. Therefore

$$\begin{aligned} Q_x(x_+; \nabla f(x)) &= Q_x(x_+; \widehat{g}) - \langle e, x_+ - x \rangle \\ &\le Q_x(z_+; \widehat{g}) - \langle e, x_+ - x \rangle \\ &= Q_x(z_+; \nabla f(x)) + \langle e, z_+ - x_+ \rangle \\ &\le Q_x(z_+; \nabla f(x)) + \frac{1}{L}\|e\|^2, \end{aligned} \tag{188}$$

where the last step uses (187). Smoothness and 1-strong convexity of ψ imply

$$\begin{aligned} f(x_+) &\le f(x) + \langle \nabla f(x), x_+ - x \rangle + \frac{L}{2}\|x_+ - x\|^2 \\ &\le f(x) + Q_x(x_+; \nabla f(x)). \end{aligned}$$

Combining this with (188) and the definition (11),

$$\begin{aligned} f(x_+) - f^\star &\le f(x) - f^\star - \frac{1}{2L}\mathcal{G}^2_{\psi,L}(x) + \frac{1}{L}\|e\|^2 \\ &\le \left(1 - \frac{\mu}{L}\right)(f(x) - f^\star) + \frac{1}{L}\|e\|^2, \end{aligned}$$

where the last inequality is (12). □

If $\mu > L$, Algorithm 1 returns x_0 without querying the oracle. Otherwise $\kappa \geq 1$; for $\epsilon \in (0, \Delta_0)$ and $\delta \in (0, 1/2)$, define

$$S := \left\lceil \log_2 \frac{\Delta_0}{\epsilon} \right\rceil, \qquad J := \lceil \kappa \log 4 \rceil, \qquad \Delta_s := \frac{\Delta_0}{2^s}, \qquad r_s := \frac{1}{2}\sqrt{\mu \Delta_s}. \tag{189}$$

At every one of the J iterations of stage s, call the estimator of Subsection C.1 at the current point with accuracy r_s and conditional failure probability

$$\rho := \frac{\delta}{JS}, \tag{190}$$

then update

$$x_{t+1} := \arg\min_{z \in \mathcal{X}} \left\{ \langle \widehat{g}_t, z - x_t \rangle + L D_\psi(z, x_t) \right\}. \tag{191}$$

Return the last iterate after stage $S - 1$. In (172), use $r = r_s$ to obtain b_s and $\rho = \delta/(JS)$ to obtain the common K. Together with (189), these are all parameters of Algorithm 1. For $\mu \leq L$, we establish the following explicit version of the sample bound in Theorem 4.3; for $\mu > L$, $N = 0$:

$$N \leq C_{0,\alpha} \kappa \left[\left\lceil \log_2 \frac{\Delta_0}{\epsilon} \right\rceil + Q_\epsilon \right] \log \left(\frac{16\kappa}{\delta} \left\lceil \log_2 \frac{\Delta_0}{\epsilon} \right\rceil \right). \tag{192}$$

Proof of Theorem 4.3. The case $\mu > L$ follows from the feasible-decrement argument at the start of this subsection: every feasible point is optimal, so the returned x_0 is exact and costs no queries. Suppose henceforth that $\mu \leq L$. There are JS robust-estimator calls. The conditional guarantee in Lemma C.4 and a union bound imply that, with probability at least $1 - \delta$,

$$\|\widehat{g}_t - \nabla f(x_t)\| \leq r_s \tag{193}$$

at every update of every stage. Work pathwise on this event; all probabilistic estimates have already been established.

We prove inductively that the objective gap at the beginning of stage s is at most Δ_s. Write $\Delta_{s,j} := f(x_{sJ+j}) - f^\star$ for the gap after j updates in stage s. Lemma C.5 gives

$$\Delta_{s,j+1} \leq \left(1 - \frac{1}{\kappa}\right) \Delta_{s,j} + \frac{r_s^2}{L}. \tag{194}$$

After J iterations,

$$\Delta_{s,J} \leq e^{-J/\kappa} \Delta_s + \frac{r_s^2}{\mu} \leq \frac{\Delta_s}{4} + \frac{\Delta_s}{4} = \Delta_{s+1}. \tag{195}$$

Thus $f(x_{\text{out}}) - f^\star \leq \Delta_S \leq \epsilon$.

It remains to count oracle calls. By Lemma C.4, one estimator call in stage s costs at most

$$C_{\text{est},\alpha} \left[1 + \left(\frac{\sigma}{r_s} \right)^{\alpha/(\alpha-1)} \right] \log \left(\frac{4JS}{\delta} \right) = C_{\text{est},\alpha} \left[1 + 2^{2\theta_\alpha} \left(\frac{\sigma^2}{\mu \Delta_s} \right)^{\theta_\alpha} \right] \log \left(\frac{4JS}{\delta} \right).$$

Hence

$$N \leq C_{\text{est},\alpha} J \log \left(\frac{4JS}{\delta} \right) \left[S + 2^{2\theta_\alpha} \left(\frac{\sigma^2}{\mu} \right)^{\theta_\alpha} \sum_{s=0}^{S-1} \Delta_s^{-\theta_\alpha} \right]. \tag{196}$$

The geometric sum satisfies

$$\sum_{s=0}^{S-1} \Delta_s^{-\theta_\alpha} = \frac{\Delta_S^{-\theta_\alpha} - \Delta_0^{-\theta_\alpha}}{2^{\theta_\alpha} - 1} < \frac{\Delta_S^{-\theta_\alpha}}{2^{\theta_\alpha} - 1} < \frac{1}{1 - 2^{-\theta_\alpha}} \epsilon^{-\theta_\alpha}. \tag{197}$$

Because $J \leq 3\kappa$ and $4JS \leq 12\kappa S$, all factors depending only on α can be absorbed into $C_{0,\alpha}$, yielding (192). □

D Proofs for Section 5

We work with $f \in \mathcal{F}_d(L, \mu, \Delta_0)$, the oracle model (5), and a fixed dimension d. Theorem D.1 states the explicit bounds underlying Theorem 5.2; Theorem D.2 supplies the expected-risk and confidence lower bounds. The algorithms below are specified before their analysis. Bisection uses the one-dimensional ordering of minimizers, whereas gradient-flow trapping uses cutting surfaces and reconstructs relative values from gradients only. Their subroutines share the inputs G, L, σ, α and the exponent p_α from (1). In bisection, M is the number of blocks per half of the estimator; it plays the role of K in Appendix C.1. For reproducible parameter choices fix

$$\mathfrak{c}_d = 2^{16} d^2 \quad (d \geq 2), \qquad C_{\mathrm{rep},\alpha} = \max\{3, 32^{1/(\alpha-1)}\}. \tag{198}$$

These constants are deliberately conservative; the complexity statements absorb dimension- and moment-dependent factors. Finite minimizations use a fixed lexicographic tie-breaking rule. A geometric median may be chosen as the minimum-norm element of its nonempty compact convex minimizer set. The upper-bound algorithms handle the trivial constant-objective case by returning x_0 when the supplied parameters satisfy $\mu > L$.

Algorithm 2 Robust Stochastic Bisection

Require: A globally L-smooth, μ-PL function $f : \mathbb{R} \to \mathbb{R}$ and a conditionally unbiased stochastic-gradient oracle G with centered α-moment bounded by σ^α, where $\alpha \in (1, 2]$.
Require: $x_0 \in \mathbb{R}$ with $f(x_0) - f^\star \leq \Delta_0$; $\epsilon \in (0, \Delta_0)$ and $\delta \in (0, 1/2)$.
Ensure: x_{out} such that $\mathbb{P}(f(x_{\mathrm{out}}) - f^\star \leq \epsilon) \geq 1 - \delta$.

$\mathrm{clip}_r(v) := \mathrm{sgn}(v) \min\{|v|, r\}$, $\log_2^+(z) := \max\{0, \log_2 z\}$.

1: **if** $\mu > L$ **then return** x_0.
2: $\kappa \leftarrow \dfrac{L}{\mu}$, $D_0 \leftarrow \sqrt{\dfrac{2\Delta_0}{\mu}}$, $r_\epsilon \leftarrow \dfrac{1}{4}\sqrt{2\mu\epsilon}$.
3: $K_\epsilon \leftarrow 1 + \left\lceil \dfrac{1}{2} \log_2^+ \left(\dfrac{\kappa \Delta_0}{\epsilon} \right) \right\rceil$, $\rho \leftarrow \dfrac{\delta}{K_\epsilon}$.
4: $B_\alpha \leftarrow (16 \cdot 8^\alpha)^{1/(\alpha-1)}$, $M \leftarrow \left\lceil 16 \log \dfrac{4}{\rho} \right\rceil$.
5: $b \leftarrow \max\left\{ 1, \left\lceil B_\alpha \left(\dfrac{\sigma}{r_\epsilon} \right)^{\alpha/(\alpha-1)} \right\rceil \right\}$.
6: $(\ell_0, u_0) \leftarrow (x_0 - D_0, x_0 + D_0)$.
7: **for** $k = 0, \ldots, K_\epsilon - 1$ **do**
8: $\quad c_k \leftarrow (\ell_k + u_k)/2$.
9: $\quad$ Draw $2Mb$ fresh, conditionally independent oracle samples $G_{j,q} = G(c_k, \xi_{j,q})$, $j = 1, \ldots, 2M$, $q = 1, \ldots, b$.
10: $\quad Z_j \leftarrow \dfrac{1}{b} \sum_{q=1}^{b} G_{j,q}$, $j = 1, \ldots, 2M$.
11: $\quad a_k \in \arg\min_{a \in \mathbb{R}} \sum_{j=1}^{M} |a - Z_j|$.
12: $\quad \widehat{g}_k \leftarrow a_k + \dfrac{1}{M} \sum_{j=1}^{M} \mathrm{clip}_{r_\epsilon}(Z_{M+j} - a_k)$.
13: $\quad$ **if** $|\widehat{g}_k| \leq 2r_\epsilon$ **then**
14: $\quad\quad$ **return** c_k.
15: $\quad$ **else if** $\widehat{g}_k > 2r_\epsilon$ **then**
16: $\quad\quad (\ell_{k+1}, u_{k+1}) \leftarrow (\ell_k, c_k)$.
17: $\quad$ **else**
18: $\quad\quad (\ell_{k+1}, u_{k+1}) \leftarrow (c_k, u_k)$.
19: $\quad$ **end if**
20: **end for**
21: **return** $(\ell_{K_\epsilon} + u_{K_\epsilon})/2$.

Algorithm 3 Stagewise Gradient-Flow Trapping

Require: A fixed dimension $d \geq 2$; a globally L-smooth, μ-PL function $f : \mathbb{R}^d \to \mathbb{R}$; a conditionally unbiased stochastic-gradient oracle G with centered α-moment bounded by σ^α.
Require: x_0 with $f(x_0) - f^\star \leq \Delta_0$; $\alpha \in (1, 2]$, $\epsilon \in (0, \Delta_0)$, and $\delta \in (0, 1/2)$.
Ensure: x_{out} such that $\mathbb{P}(f(x_{\mathrm{out}}) - f^\star \leq \epsilon) \geq 1 - \delta$.

if $\mu > L$ **then return** x_0.
$\kappa \leftarrow \dfrac{L}{\mu}, \qquad S \leftarrow \left\lceil \log_2 \dfrac{\Delta_0}{\epsilon} \right\rceil, \qquad J \leftarrow \mathfrak{c}_d \left\lceil \log(2 + \kappa) \right\rceil.$
$x \leftarrow x_0, \quad \rho \leftarrow \delta/(2SJ).$
for $s = 0, \ldots, S - 1$ **do**
  $\Delta_s \leftarrow \dfrac{\Delta_0}{2^s}, \qquad \eta \leftarrow \sqrt{\mu \Delta_s}, \qquad c \leftarrow \dfrac{\eta}{8}, \qquad a \leftarrow \dfrac{\eta}{\mathfrak{c}_d J}.$
  $H_0 \leftarrow \dfrac{2\Delta_s}{c}, \qquad R \leftarrow \prod_{j=1}^{d} [x_j - H_0, x_j + H_0].$
  for $t = 0, \ldots, J - 1$ **do**
    Write $R = \prod_{j=1}^{d} [\ell_j, u_j]$ and choose $i \in \arg\max_{1 \leq j \leq d} (u_j - \ell_j)$.
    $s_t \leftarrow u_i - \ell_i$.
    if $s_t \leq \eta/(4L\sqrt{d})$ **then**
      break.
    end if
    $\tau_t \leftarrow \dfrac{a s_t}{\mathfrak{c}_d}, \qquad h_t \leftarrow \sqrt{\dfrac{a s_t}{\mathfrak{c}_d L}}, \qquad n_t \leftarrow 2^{\lceil \log_2(\mathfrak{c}_d s_t / h_t) \rceil}.$
    Define the oriented coordinate $\zeta(z) = z_i - \ell_i$ if $x_i \geq (\ell_i + u_i)/2$, and $\zeta(z) = u_i - z_i$ otherwise.
    $E \leftarrow \{z \in R : \zeta(z) = s_t/6\}, \qquad F \leftarrow \{z \in R : \zeta(z) = s_t/3\}.$
    for all $H \in \{E, F\}$ **do**
      $(\mathcal{N}_H, \widehat{D}_H) \leftarrow \textsc{CutValues}(H, x, n_t, \tau_t, \rho)$.
    end for
    Let $\widehat{D}(z) = \widehat{D}_H(z)$ for $z \in \mathcal{N}_H$, $H \in \{E, F\}$.
    $A \leftarrow \left\{ z \in \mathcal{N}_E \cup \mathcal{N}_F : \widehat{D}(z) \leq -c\|z - x\| - \tau_t \right\}.$
    if $A = \varnothing$ **then**
      $R \leftarrow R \cap \{z : \zeta(z) \geq s_t/3\}.$ ▷ Keep the current pivot
    else
      $z_\star \in \arg\min_{z \in A} \widehat{D}(z).$
      if $z_\star \in \mathcal{N}_F$ **then**
        $R \leftarrow R \cap \{z : \zeta(z) \geq s_t/6\}.$
      else
        $R \leftarrow R \cap \{z : \zeta(z) \leq s_t/3\}.$
      end if
      $x \leftarrow z_\star$.
    end if
    $c \leftarrow c + a$.
  end for
end for
return x.

Algorithm 4 Gradient-Only Potential Reconstruction on a Cutting Surface

Require: A k-dimensional cutting hyperrectangle H; a reference pivot x within distance $\sqrt{k+1}r$ of H, where all side lengths of H lie in $[r/3, r]$; a dyadic resolution $n = 2^M$, $M \in \mathbb{N}_0$; accuracy $\tau > 0$ and failure probability $\rho \in (0, 1/2)$.

function CutValues(H, x, n, τ, ρ)
  $M \leftarrow \log_2 n$.
  Construct nested Cartesian grids $\mathcal{N}_0, \dots, \mathcal{N}_M$ on H, using 2^ℓ equal subintervals per free coordinate at level ℓ.
  $\mathcal{E}_0 \leftarrow \{(x, z) : z \in \mathcal{N}_0\}$. ▷ Connect the coarsest grid to the pivot
  for $\ell = 1, \dots, M$ **do**
    for all $z \in \mathcal{N}_\ell \setminus \mathcal{N}_{\ell-1}$ **do**
      $\pi(z) \in \arg\min_{y \in \mathcal{N}_{\ell-1}} \|z - y\|$.
    end for
    $\mathcal{E}_\ell \leftarrow \{(\pi(z), z) : z \in \mathcal{N}_\ell \setminus \mathcal{N}_{\ell-1}\}$.
  end for
  $\mathcal{E} \leftarrow \bigcup_{\ell=0}^{M} \mathcal{E}_\ell, \qquad \rho_{\text{edge}} \leftarrow \frac{\rho}{|\mathcal{E}|}$.
  $A_{\text{det}} \leftarrow \sum_{j=0}^{M} 2^{(3k-4)j/5}, \qquad A_{\text{noise}} \leftarrow \sum_{j=0}^{M} 2^{(k-p_\alpha)j/(p_\alpha+1)}$.
  for $\ell = 0, \dots, M$ **do**
    $u_\ell \leftarrow \frac{\tau\, 2^{(3k-4)\ell/5}}{2A_{\text{det}}}, \qquad v_\ell \leftarrow \frac{\tau\, 2^{(k-p_\alpha)\ell/(p_\alpha+1)}}{2A_{\text{noise}}}$.
    for all $e = (a_e, b_e) \in \mathcal{E}_\ell$ **do**
      $\widehat{I}_e \leftarrow$ RobustLineIntegral$(a_e, b_e, u_\ell, v_\ell, \rho_{\text{edge}})$.
    end for
  end for
  for all $z \in \mathcal{N}_M$ **do**
    Let $P(x, z)$ be the directed tree path from x to z.
    $\widehat{D}(z) \leftarrow \sum_{e \in P(x,z)} \widehat{I}_e$.
  end for
  return $(\mathcal{N}_M, \widehat{D})$.
end function

Algorithm 5 Robust Stochastic Line Integral

Require: Endpoints $a, b \in \mathbb{R}^d$; quadrature and noise tolerances $u, v > 0$; failure probability $\rho \in (0, 1/2)$.

function RobustLineIntegral(a, b, u, v, ρ)
  $h \leftarrow \|b - a\|$.
  if $h = 0$ **then**
    return 0.
  end if
  $m \leftarrow \max\left\{1, \left\lceil C_{\text{rep},\alpha} \left[\left(\frac{Lh^2}{u}\right)^{2/3} + \left(\frac{\sigma h}{v}\right)^{p_\alpha} \right] \right\rceil \right\}$.
  $B \leftarrow \left\lceil 4 \log \frac{4}{\rho} \right\rceil$.
  for $r = 1, \dots, B$ **do**
    for $j = 1, \dots, m$ **do**
      Draw a fresh $U_{r,j} \sim \text{Unif}([(j-1)/m, j/m])$.
      Query the oracle at $a + U_{r,j}(b - a)$ with fresh randomness.
      $W_{r,j} \leftarrow \langle G(a + U_{r,j}(b-a), \xi_{r,j}),\, b - a \rangle$.
    end for
    $\widehat{I}^{(r)} \leftarrow \frac{1}{m} \sum_{j=1}^{m} W_{r,j}$.
  end for
  return median$\{\widehat{I}^{(1)}, \dots, \widehat{I}^{(B)}\}$.
end function

With the procedures specified, we state their guarantees explicitly before proving the lower and upper bounds separately.

Theorem D.1. *Let $d \geq 1$ be fixed, $1 < \alpha \leq 2$, $\sigma \geq 0$, $0 < \epsilon < \Delta_0$, and $0 < \delta < 1/2$. Assume $f \in \mathcal{F}_d(L, \mu, \Delta_0)$, $\kappa \geq 1$, and the oracle convention stated above. The algorithms may use a translated initial point with the same gap bound.*

(i) *In dimension $d = 1$, Algorithm 2 returns $f(x_{\mathrm{out}}) - f^\star \leq \epsilon$ with probability at least $1 - \delta$ using*

$$N \leq C_\alpha K_\epsilon \big[1 + Q_\epsilon\big] \log \frac{4K_\epsilon}{\delta}, \qquad K_\epsilon := 1 + \left\lceil \frac{1}{2} \log_2^+ \frac{\kappa \Delta_0}{\epsilon} \right\rceil. \tag{199}$$

(ii) *Let $d \geq 2$, put $k = d - 1$, and let*

$$J := \mathfrak{c}_d \left\lceil \log(2 + \kappa) \right\rceil, \qquad S := \left\lceil \log_2 \frac{\Delta_0}{\epsilon} \right\rceil, \tag{200}$$

where $\mathfrak{c}_d$ is fixed in (198). Define

$$\mathcal{D}_k(z) := \begin{cases} z^{2/3}, & k = 1, \\ z^{k/2}, & k \geq 2, \end{cases} \tag{201}$$

and

$$\mathcal{H}_{k,p}(\kappa, J) := \begin{cases} J^{p+1}, & k < p, \\ J^{p+1} \big[\log(2 + \kappa J)\big]^{p+1}, & k = p, \\ J^p (\kappa J)^{(k-p)/2}, & k > p. \end{cases} \tag{202}$$

Then Algorithm 3 uses

$$N \leq C_{d,\alpha} \Lambda_{d,\alpha} \left[S\, \mathcal{D}_k(\kappa J) + \mathcal{H}_{k,p_\alpha}(\kappa, J) Q_\epsilon \right] \tag{203}$$

stochastic-gradient calls, where

$$\Lambda_{d,\alpha} := \log \left(\frac{C_{d,\alpha} (S + 1)(J + 1)(2 + \kappa J)^{k/2}}{\delta} \right). \tag{204}$$

With probability at least $1 - \delta$, its output satisfies $f(x_{\mathrm{out}}) - f^\star \leq \epsilon$.

D.1 Main ideas and proof sketches

The proof has four independent ingredients.

1. *A universal one-coordinate heavy-tail obstruction.* A pair of shifted quadratics is coupled so that the two oracle laws are identical unless a rare event of probability q occurs. At a fixed mean shift, the centered α-moment scales as $q^{1-\alpha}$. Calibrating the shift to the moment budget makes the minimizer separation of order $\sigma q^{(\alpha-1)/\alpha}/\mu$. Taking $q \asymp (\mu\epsilon/\sigma^2)^{\theta_\alpha}$ produces the sharp mean-estimation scale Q_ϵ. Embedding the construction in the first coordinate works in every fixed dimension.
2. *One-dimensional localization.* For a one-dimensional PL function, the minimizer set is an interval, the derivative is negative to its left and positive to its right, and PL implies a quadratic-growth inequality. A robust estimate of the derivative sign therefore supports bisection. This removes every polynomial dependence on κ from the stochastic term.
3. *Gradient-flow traps in dimensions at least two.* A compact box whose boundary cannot be reached from its pivot while decreasing f faster than slope c must contain a point with gradient norm at most c. Two parallel cuts of the longest side shrink such a box by a constant volume factor. Approximate relative values on a fine net of the new cut preserve the trap after paying only $O_d(\tau/s + Lh^2/s + ch^2/s^2)$ in the slope parameter.

4. *Potential reconstruction from stochastic gradients.* Relative function values are line integrals of the gradient. Stratified randomized quadrature on a segment of length h has quadrature and oracle-noise errors of order $Lh^2/m^{3/2}$ and $\sigma h/m^{(\alpha-1)/\alpha}$, respectively, at constant confidence. A dyadic tree shares these line integrals among all points of a $(d-1)$-dimensional cut. The accumulated noise is bounded up to logarithms exactly when $p_\alpha \geq d-1$; this is the source of the phase transition in (260).

Appendix D.2 proves the two lower-bound mechanisms and the bisection guarantee. Appendices D.3 and D.4 analyze reconstruction and box refinement; Appendix D.5 then assembles the phase diagram.

D.2 Universal lower bounds and the complete one-dimensional theory

The lower bound uses two distinct information-hiding mechanisms. Rare-event observations conceal the sign of a shifted quadratic, whereas exact gradient queries reveal only a few bits of a nested minimizer location. We state the expected-risk and confidence consequences separately, then prove the bisection guarantee from one-dimensional PL geometry.

Theorem D.2. *Let $d \geq 1$ be fixed, $\kappa \geq 6$, $1 < \alpha \leq 2$, and $\sigma \geq 0$. There are positive constants c_α, C, c_0 such that, for every integer $T \geq 0$,*

$$\mathsf{R}_T^{(d)} \geq c_\alpha \left[\Delta_0 e^{-CT} + \min\left\{\Delta_0, \frac{\sigma^2}{\mu(T+1)^{\beta_\alpha}}\right\}\right]. \tag{205}$$

For every algorithm there is a fixed legal instance on which the gap is at least the same bracket times a positive constant with probability strictly larger than $1/4$. For $0 < \epsilon \leq c_0\Delta_0$ and $0 < \delta \leq 1/4$,

$$\mathsf{T}_{\epsilon,\delta}^{(d)} \geq c \log(\Delta_0/\epsilon). \tag{206}$$

If also $\sigma^2 \geq C_\alpha \mu\epsilon$, then

$$\mathsf{T}_{\epsilon,\delta}^{(d)} \geq c_\alpha Q_\epsilon \log\frac{1}{2\delta}. \tag{207}$$

At constant confidence $1/4$, combining the two bounds gives $c_\alpha[\log(\Delta_0/\epsilon) + Q_\epsilon]$ for all $\sigma \geq 0$.

The following two-point family supplies the stochastic part of this theorem.

Lemma D.3. *Fix $d \geq 1$ and $q \in (0, 1/2]$. Let $\vartheta \in \{-1, +1\}$ and set*

$$a := \frac{\sigma}{(2q)^{1/\alpha}}, \qquad v_\vartheta := -\frac{\vartheta q a}{\mu}, \qquad f_\vartheta(x) := \frac{\mu}{2}\|x - v_\vartheta e_1\|^2. \tag{208}$$

At every query x, let the oracle return

$$G_\vartheta(x) := \mu x + Z_\vartheta e_1, \qquad Z_\vartheta := \begin{cases} \vartheta a, & \text{with probability } q, \\ 0, & \text{with probability } 1-q, \end{cases} \tag{209}$$

using fresh conditionally independent draws. Then f_ϑ is μ-smooth and μ-PL,

$$\mathbb{E}[G_\vartheta(x)] = \nabla f_\vartheta(x), \qquad \mathbb{E}\|G_\vartheta(x) - \nabla f_\vartheta(x)\|^\alpha \leq \sigma^\alpha. \tag{210}$$

If an algorithm makes T calls and returns X, then under the uniform prior on ϑ,

$$\mathbb{P}\left(f_\vartheta(X) - f_\vartheta^\star \geq \frac{\sigma^2}{2^{1+2/\alpha}\mu} q^{\beta_\alpha}\right) \geq \frac{1}{2}(1-q)^T. \tag{211}$$

Writing $b_\alpha = 2^{-1-2/\alpha}$, the same experiment satisfies

$$\mathbb{E}[f_\vartheta(X) - f_\vartheta^\star] \geq b_\alpha \frac{\sigma^2}{\mu} q^{\beta_\alpha}(1-q)^T. \tag{212}$$

Proof. The mean of Z_ϑ is $\vartheta q a$, so

$$\mathbb{E}G_\vartheta(x) = \mu x + \vartheta q a e_1 = \mu(x - v_\vartheta e_1) = \nabla f_\vartheta(x).$$

The centered noise has only its first coordinate nonzero, and

$$\begin{aligned}\mathbb{E}|Z_\vartheta - \mathbb{E}Z_\vartheta|^\alpha &= a^\alpha\left[q(1-q)^\alpha + (1-q)q^\alpha\right] \\ &\le 2qa^\alpha = \sigma^\alpha.\end{aligned}$$

The quadratic identity $\|\nabla f_\vartheta(x)\|^2/2 = \mu(f_\vartheta(x) - f_\vartheta^\star)$ proves smoothness and PL.

Realize each draw as $Z_\vartheta = \vartheta a B_i$, where $B_i \sim \text{Bernoulli}(q)$ are independent of the sign and the algorithm's random seed. Let $E_T = \{B_1 = \cdots = B_T = 0\}$, so that $\mathbb{P}(E_T) = (1-q)^T$ even when $\sigma = 0$. On E_T, every response equals μx, independently of ϑ; hence the complete adaptive transcript, including the algorithm's internal randomness, has the same conditional law for the two signs. The two minimizers have first coordinates $\pm r$, where

$$r = \frac{qa}{\mu} = \frac{\sigma}{2^{1/\alpha}\mu} q^{(\alpha-1)/\alpha}.$$

For every realized output X, $\max\{|X_1 - r|, |X_1 + r|\} \ge r$. Conditional on E_T and averaged over the uniform sign, the probability of squared distance at least r^2 is therefore at least $1/2$. Multiplying this conditional probability by $\mathbb{P}(E_T) = (1-q)^T$ proves (211). For the expected gap, average over the two signs on E_T: $\big((X_1 - r)^2 + (X_1 + r)^2\big)/2 = X_1^2 + r^2 \ge r^2$. The other coordinates contribute nonnegative terms. Multiplying by $\mu/2$ and by $\mathbb{P}(E_T)$ proves (212). Averaging over the two signs selects a fixed sign for either criterion; no adversarial randomness remains in the resulting instance. □

Stochastic part of Theorem D.2. Assume first $\sigma > 0$ and put $b_\alpha = 2^{-1-2/\alpha}$. For the confidence bound, choose

$$q_\epsilon = \left(\frac{4\mu\epsilon}{b_\alpha \sigma^2}\right)^{\theta_\alpha}. \tag{213}$$

Then the gap threshold in (211) is exactly 4ϵ. The condition $\sigma^2 \ge (4 \cdot 2^{\beta_\alpha}/b_\alpha)\mu\epsilon$ ensures $q_\epsilon \le 1/2$. Its initial gap is $4\epsilon \le \Delta_0$ provided $\epsilon \le \Delta_0/4$. For $T < (2q_\epsilon)^{-1}\log(1/(2\delta))$, the inequality $\log(1-q) \ge -2q$ on $[0, 1/2]$ gives $\frac{1}{2}(1-q_\epsilon)^T > \delta$. Lemma D.3 therefore proves (207) on a fixed sign instance.

For a fixed budget instead choose

$$q_T = \min\left\{\frac{1}{2(T+1)}, \left(\frac{\mu\Delta_0}{b_\alpha\sigma^2}\right)^{\theta_\alpha}\right\}. \tag{214}$$

This choice has $0 < q_T \le 1/2$ and the initial gap is

$$g_T = \min\left\{\Delta_0, \frac{b_\alpha\sigma^2}{\mu[2(T+1)]^{\beta_\alpha}}\right\}.$$

Bernoulli's inequality yields $(1-q_T)^T \ge 1 - Tq_T > 1/2$. Consequently (212) gives expected gap greater than $g_T/2$, and (211) gives gap at least g_T with probability greater than $1/4$, each on a fixed legal instance. The initial gap is at most Δ_0 by construction. This proves the stochastic term of (205). When $\sigma = 0$ that term is zero and no rare-event construction is required. □

We turn from statistical indistinguishability to one-dimensional geometry. The following structural lemma explains why derivative signs suffice to localize a minimizer.

Lemma D.4. *Let $f : \mathbb{R} \to \mathbb{R}$ be globally L-smooth, bounded below, and μ-PL. Then its minimizer set $\mathcal{S} = \arg\min f$ is a nonempty closed interval (possibly a singleton, ray, or all of $\mathbb{R}$). Moreover,*

$$f'(x) < 0 \qquad (x < \inf\mathcal{S}), \qquad f'(x) > 0 \qquad (x > \sup\mathcal{S}), \tag{215}$$

$$f(x) - f^\star \ge \frac{\mu}{2}\operatorname{dist}(x, \mathcal{S})^2, \tag{216}$$

$$|f'(x)| \ge \mu \operatorname{dist}(x, \mathcal{S}), \tag{217}$$

$$f(x) - f^\star \le \frac{L}{2}\operatorname{dist}(x, \mathcal{S})^2. \tag{218}$$

Proof. By PL, every critical point is a global minimizer. We first prove that a minimizer exists. Otherwise f' never vanishes and, by continuity, has a constant sign. Suppose $f' > 0$; the other case is symmetric. Put $h = f - f^\star > 0$. PL yields $f'(x) \geq \sqrt{2\mu h(x)}$, hence

$$\frac{d}{dx}\sqrt{h(x)} = \frac{f'(x)}{2\sqrt{h(x)}} \geq \sqrt{\frac{\mu}{2}}.$$

Integrating from $x < x_0$ to x_0 gives $\sqrt{h(x_0)} - \sqrt{h(x)} \geq \sqrt{\mu/2}(x_0 - x)$, impossible as $x \to -\infty$.

The zero set of f' therefore equals the nonempty minimizer set $\mathcal{S}$. If $a, b \in \mathcal{S}$ and some $c \in (a, b)$ had $f(c) > f^\star$, then the maximum of f on $[a, b]$ would be attained at an interior nonoptimal critical point, a contradiction. Thus $\mathcal{S}$ is an interval. On each component of $\mathbb{R} \setminus \mathcal{S}$, f' has constant sign. To the right it cannot be negative, because then f would fall below $f^\star$; to the left it cannot be positive. This proves (215).

For $x > \sup \mathcal{S} =: s$, the same differential inequality now gives

$$\sqrt{f(x) - f^\star} \geq \sqrt{\frac{\mu}{2}}(x - s).$$

The argument to the left of $\mathcal{S}$ is symmetric. This proves (216). Combining it with PL gives (217). Finally, for $s = \Pi_{\mathcal{S}}(x)$, smoothness and $f'(s) = 0$ imply $f(x) - f^\star \leq L|x - s|^2/2$, proving (218). □

The geometry lemma supplies both an initial bracket and a stopping certificate. We use it to analyze Algorithm 2.

From (216) and the initial gap,

$$\operatorname{dist}(x_0, \mathcal{S}) \leq D_0 := \sqrt{\frac{2\Delta_0}{\mu}}. \tag{219}$$

Start with $I_0 = [x_0 - D_0, x_0 + D_0]$ and set

$$r_\epsilon := \frac{1}{4}\sqrt{2\mu\epsilon}. \tag{220}$$

At the midpoint c_k of the current interval, invoke the centered-clipping estimator of Lemma C.4 with accuracy r_ϵ and failure probability δ/K_ϵ. If $|\widehat{g}_k| \leq 2r_\epsilon$, return c_k. If $\widehat{g}_k > 2r_\epsilon$, retain the left half; if $\widehat{g}_k < -2r_\epsilon$, retain the right half. If no early termination occurs, return the midpoint after K_ϵ halvings.

Proof of Theorem D.1(i). By Lemma C.4 and a union bound, with probability at least $1 - \delta$ all estimates satisfy $|\widehat{g}_k - f'(c_k)| \leq r_\epsilon$. Work on this event. If the method stops early, then $|f'(c_k)| \leq 3r_\epsilon$, so PL gives

$$f(c_k) - f^\star \leq \frac{9r_\epsilon^2}{2\mu} = \frac{9}{16}\epsilon < \epsilon.$$

If $\widehat{g}_k > 2r_\epsilon$, then $f'(c_k) > 0$, and Lemma D.4 places every minimizer to the left of c_k. The negative case is symmetric. Hence every retained interval meets $\mathcal{S}$. After K_ϵ halvings, the final midpoint c satisfies

$$\operatorname{dist}(c, \mathcal{S}) \leq D_0 2^{-K_\epsilon},$$

and (218) yields

$$f(c) - f^\star \leq \frac{L}{2} D_0^2 4^{-K_\epsilon} = \kappa\Delta_0 4^{-K_\epsilon} \leq \epsilon.$$

One estimator call costs, by Lemma C.4,

$$C_\alpha \left[1 + \left(\frac{\sigma}{r_\epsilon}\right)^{p_\alpha}\right] \log\frac{4K_\epsilon}{\delta} \leq C_\alpha[1 + Q_\epsilon] \log\frac{4K_\epsilon}{\delta}.$$

Multiplication by K_ϵ proves (199). □

To complete the one-dimensional characterization, we still need the logarithmic noiseless transient. The rare-event family alone cannot provide it. We now give a fixed one-dimensional distribution of strongly convex functions for which each exact-gradient query reveals only a constant expected number of nested location bits.

For an interval $I = [a, a+s]$, define

$$I_0 := \left[a + \frac{s}{8}, a + \frac{3s}{8}\right], \qquad I_1 := \left[a + \frac{5s}{8}, a + \frac{7s}{8}\right], \qquad A(s) := \frac{6}{7}\mu s. \tag{221}$$

Fix a root interval $I_\varnothing = [-\ell/2, \ell/2]$. Given a word $\omega \in \{0,1\}^K$, define a continuous piecewise-affine function h_ω recursively. On a node interval $I = [a, a+s]$, its boundary values are $-A(s)$ and $A(s)$. If the next bit is b, use the recursively defined profile on I_b; on each component of $I \setminus I_b$, interpolate affinely between the adjacent parent and child boundary values. At depth K, interpolate affinely across the final interval. Outside the root interval, extend with slope μ.

Every affine slope can be computed explicitly. On a terminal interval of length s, the change from $-A(s)$ to $A(s)$ is $2A(s) = 12\mu s/7$, hence the slope is $12\mu/7$. If the left child is selected, the left connector has length $s/8$ and changes by $A(s) - A(s/4) = 9\mu s/14$, hence its slope is $36\mu/7$; the right connector has length $5s/8$ and the same change, hence slope $36\mu/35$. The two values are interchanged when the right child is selected. Outside the root the prescribed slope is μ. Therefore every affine slope belongs to

$$\left\{\mu, \frac{36}{35}\mu, \frac{12}{7}\mu, \frac{36}{7}\mu\right\}. \tag{222}$$

Let $x^\star_\omega$ be the midpoint of the final interval and define

$$f_\omega(x) := \int_{x^\star_\omega}^{x} h_\omega(t)\, dt. \tag{223}$$

Then f_ω is μ-strongly convex and $(36/7)\mu$-smooth, hence μ-PL.

Lemma D.5. *Let ω be uniform on $\{0,1\}^K$. For every adaptive randomized exact-gradient algorithm using T queries, if $K = 8T + 2$, then with probability at least $1/4$,*

$$f_\omega(\widehat{x}_T) - f^\star_\omega \geq \frac{\mu\ell^2}{128} 4^{-K}, \tag{224}$$

where ℓ is the length of the root interval.

Proof. Condition on the algorithm's internal random seed, independently of ω. This is the distributional argument underlying Yao's principle (Yao, 1977). Keep ω uniform. Give the algorithm a stronger oracle: in addition to $h_\omega(x)$, a query reveals the next unknown bit and continues revealing bits for as long as the query point lies in the selected child interval. This stronger response determines the exact derivative value, because once the query lies outside the selected child, all deeper modifications are confined to that child.

Conditional on the currently revealed prefix, the query is fixed and every unrevealed bit is independent and uniform. The first new bit is revealed for free. To reveal at least $j \geq 2$ new bits, the query must lie in the selected child at each of the preceding $j-1$ levels. The two children are disjoint, so each continuation has conditional probability at most $1/2$. Thus, if J_t is the number of new bits revealed at query t,

$$\mathbb{P}(J_t \geq j \mid \text{past}) \leq 2^{-(j-1)}, \qquad \mathbb{E}[J_t \mid \text{past}] \leq 2.$$

If R_T is the total revealed prefix length, then $\mathbb{E}R_T \leq 2T$, and Markov's inequality gives $\mathbb{P}(R_T \geq K/2) < 1/2$.

On the event $R_T < K/2$, the next bit is still uniform. If the current parent interval has length $s = \ell 4^{-R_T}$, then for every output x,

$$\max\{\mathrm{dist}(x, I_0), \mathrm{dist}(x, I_1)\} \geq \frac{s}{8}.$$

Hence, with conditional probability at least $1/2$, the actual minimizer lies at distance at least $s/8 \geq \ell 4^{-K/2}/8$ from the output. Strong convexity then gives (224). Multiplying the two conditional probabilities gives a probability strictly larger than $1/4$. The argument holds after conditioning on every internal random seed; averaging over that seed proves the assertion for randomized algorithms. □

Choose

$$\ell^2 := \frac{4\Delta_0}{3\mu}. \tag{225}$$

Since every f_ω is 6μ-smooth and $|x^\star_\omega| \leq \ell/2$, its initial gap at $x_0 = 0$ is at most $3\mu\ell^2/4 = \Delta_0$. Lemma D.5 therefore gives

$$\inf_{\mathsf{A}} \sup_{f \in \mathcal{F}_1(L,\mu,\Delta_0)} \mathbb{P}\left(f(\widehat{x}_T) - f^\star \geq c\Delta_0 e^{-CT}\right) \geq \frac{1}{4} \tag{226}$$

whenever $L/\mu \geq 6$. Equivalently,

$$\mathsf{T}^{(1)}_{\epsilon,1/4} \geq c \log \frac{\Delta_0}{\epsilon} \tag{227}$$

for $\epsilon \leq c_0\Delta_0$. Adding $\frac{\mu}{2}\sum_{j=2}^d x_j^2$ embeds this family into every fixed dimension without changing smoothness, PL, or oracle information. The strict probability bound in the proof of Lemma D.5 justifies (227) even at confidence $1/4$: choose the budget so that its gap threshold exceeds ϵ.

Completion of Theorem D.2 and proof of Theorem 5.1. With $K = 8T + 2$, (224) and (225) give a gap of at least $\Delta_0 \exp[-8(\log 4)T]/1536$ with probability greater than $1/4$. They therefore also give an expected gap greater than one quarter of that quantity. Together with the stochastic construction above and $\max\{u, v\} \geq (u + v)/2$, this proves (205). For the probability claim, choose whichever hard family has the larger threshold; this chooses one instance whose threshold controls half the sum and whose failure probability is strictly greater than $1/4$. The same deterministic construction proves (206). In the regime where (207) applies, combine the two complexity lower bounds by taking their maximum. In the complementary regime Q_ϵ is bounded by a constant depending only on α, and the deterministic term absorbs it. This proves the last claim of Theorem D.2.

For clarity, the expected-error statement of Theorem 5.1 follows directly, not by converting an upper confidence guarantee into an expectation. The exact family forces $T \geq c\log(\Delta_0/\epsilon)$ when ϵ/Δ_0 is a sufficiently small universal constant. If Q_ϵ is above the fixed threshold required for (213), use that choice of q_ϵ. For $T \leq 1/(4q_\epsilon)$, (212) and $(1 - q_\epsilon)^T \geq 1 - Tq_\epsilon \geq 3/4$ give expected gap at least $3\epsilon > \epsilon$. Thus $T \geq c_\alpha Q_\epsilon$ in this regime. Below this threshold, Q_ϵ is bounded and is absorbed by the deterministic logarithm. Taking the maximum of these two necessary budgets proves (18), the expected-error bound in the main text. □

D.3 Gradient-only potential reconstruction on cutting surfaces

The algorithms for $d \geq 2$ use no function-value oracle. Every required relative value is reconstructed from the identity

$$f(b) - f(a) = \int_0^1 \langle \nabla f(a + t(b - a)), b - a\rangle \, dt.$$

The moment estimate below is a martingale-difference analogue of the classical von Bahr–Esseen inequality in von Bahr & Esseen (1965). Sharp extensions of von Bahr–Esseen-type bounds to martingales and more general moment functions were developed by Pinelis (2015). We include a short self-contained proof because the precise conditional formulation needed for the subsequent adaptive line-integral estimator is particularly simple.

Lemma D.6. *Let $1 < \alpha \le 2$ and let $(X_i, \mathscr{F}_i)$ be real-valued martingale differences. Then*

$$\mathbb{E}\left|\sum_{i=1}^{m} X_i\right|^{\alpha} \le C_\alpha \sum_{i=1}^{m} \mathbb{E}|X_i|^{\alpha}. \tag{228}$$

Moreover,

$$\mathbb{E}\max_{j\le m}\left|\sum_{i=1}^{j} X_i\right|^{\alpha} \le C_\alpha \sum_{i=1}^{m} \mathbb{E}|X_i|^{\alpha}. \tag{229}$$

Proof. Let $r = \alpha - 1 \in (0, 1]$ and $H(u) = \operatorname{sgn}(u)|u|^r$. For equal signs, concavity gives $|H(u) - H(v)| \le |u - v|^r$; for opposite signs, $|u|^r + |v|^r \le 2^{1-r}(|u| + |v|)^r$. Thus H is r-Hölder with constant $2^{2-\alpha}$. Integrating the derivative of $|u + tv|^\alpha$ from $t = 0$ to 1 yields

$$|u + v|^\alpha \le |u|^\alpha + \alpha|u|^{\alpha-2}uv + 2^{2-\alpha}|v|^\alpha,$$

with the linear term zero at $u = 0$. Applying this inequality to the successive partial sums and taking conditional expectations eliminates the linear term. Iteration proves (228) with constant $2^{2-\alpha}$.

For completeness, write $M_j = \sum_{i\le j} X_i$ and $M^* = \max_{j\le m} |M_j|$. The first crossing time of $\lambda > 0$ and the submartingale property of $|M_j|$ imply $\lambda\mathbb{P}(M^* \ge \lambda) \le \mathbb{E}[|M_m|\mathbf{1}_{\{M^*\ge\lambda\}}]$. Integrating this inequality against $\alpha\lambda^{\alpha-2}$ gives

$$\mathbb{E}(M^*)^\alpha \le \frac{\alpha}{\alpha - 1}\mathbb{E}\big[|M_m|(M^*)^{\alpha-1}\big].$$

Hölder's inequality yields $\mathbb{E}(M^*)^\alpha \le (\alpha/(\alpha-1))^\alpha \mathbb{E}|M_m|^\alpha$. All moments are finite if the right-hand side of (228) is finite; otherwise the assertion is trivial. This proves (229), including a possible larger constant. Every argument remains valid after conditioning on an initial history. This is the usual maximal-inequality argument; see also Hall & Heyde (1980). □

The next estimator has two error sources with different scalings. Stratification controls the randomized quadrature error through a second moment, while the oracle error uses only the martingale α-moment estimate just proved.

Lemma D.7. *Let $a, b \in \mathbb{R}^d$ be measurable before this request, $h = \|b - a\|$, and let $u, v > 0$ and $\rho \in (0, 1/2)$. There is a gradient-only estimator $\widehat{I}_{a,b}$ such that*

$$\mathbb{P}\left(\left|\widehat{I}_{a,b} - (f(b) - f(a))\right| > u + v \;\middle|\; \textit{past}\right) \le \rho, \tag{230}$$

and the number of oracle calls is at most

$$C_\alpha\left[1 + \left(\frac{Lh^2}{u}\right)^{2/3} + \left(\frac{\sigma h}{v}\right)^{p_\alpha}\right]\log\frac{4}{\rho}. \tag{231}$$

Proof. Condition on the history preceding the request. If $a = b$, return zero without a query. Otherwise parameterize the segment by $\gamma(t) = a + t(b - a)$ and put $\varphi(t) = \langle\nabla f(\gamma(t)), b - a\rangle$. Then $|\varphi(t) - \varphi(s)| \le Lh^2|t - s|$.

Partition $[0, 1]$ into m equal intervals. In interval j, draw U_j uniformly, independently of the past, query the oracle at $\gamma(U_j)$, and form

$$\widehat{I}^{(1)} = \frac{1}{m}\sum_{j=1}^{m}\langle G(\gamma(U_j), \xi_j), b - a\rangle.$$

Write its error as $Q + N$. If $\overline{\varphi}_j = m\int_{(j-1)/m}^{j/m}\varphi(t)\,dt$, then the quadrature increments $m^{-1}(\varphi(U_j) - \overline{\varphi}_j)$ are centered and bounded by Lh^2/m^2. Therefore

$$\mathbb{E}Q^2 \le \frac{L^2h^4}{m^3}. \tag{232}$$

For the oracle noise $\zeta_j = G(\gamma(U_j), \xi_j) - \nabla f(\gamma(U_j))$, the increments $m^{-1}\langle \zeta_j, b-a\rangle$ are martingale differences with conditional α-moment at most $\sigma^\alpha h^\alpha / m^\alpha$. Hence Lemma D.6 gives

$$\mathbb{E}|N|^\alpha \le C_\alpha \frac{\sigma^\alpha h^\alpha}{m^{\alpha-1}}. \tag{233}$$

Choose

$$m = \max\left\{1, \left\lceil C_{\mathrm{rep},\alpha}\left[\left(\frac{Lh^2}{u}\right)^{2/3} + \left(\frac{\sigma h}{v}\right)^{p_\alpha}\right]\right\rceil\right\},$$

with $C_{\mathrm{rep},\alpha}$ from (198). In (233) one may use $2^{2-\alpha} \le 2$. Consequently Markov's inequality bounds $\mathbb{P}(|Q| > u \mid \text{past})$ and $\mathbb{P}(|N| > v \mid \text{past})$ by $1/16$ each. One replicate is accurate to $u+v$ with conditional probability at least $7/8$. Generate $B = \lceil 4\log(4/\rho)\rceil$ replicates sequentially, using fresh quadrature points and oracle samples, and return any median. Independence between replicate errors is unnecessary: each bad-replicate indicator has conditional mean at most $1/8$. Lemma C.3, with $p = 1/8$ and threshold $1/2$, bounds the probability of at least $B/2$ bad replicates by $\exp(-9B/32) \le \rho$. On the complementary event more than half of the replicates are accurate, so their median is accurate. Counting Bm oracle calls proves (231) and justifies Algorithm 5 under adaptive noise. □

A single line integral is not enough to reconstruct a cutting surface efficiently. We therefore share edge estimates along a nested dyadic tree.

Let H be a k-dimensional axis-aligned hyperrectangle with all side lengths between $r/3$ and r, and let $n = 2^M$. Its nested dyadic grids have at most $C_k 2^{k\ell}$ new points at level ℓ, and each new point can be joined to a parent at level $\ell-1$ by a segment of length at most $C_k r 2^{-\ell}$. The resulting tree has $O_k(n^k)$ edges and a root-to-leaf path uses at most one edge per level. The finest grid is a $C_k r/n$-net of H.

We use the following elementary allocation identity.

Lemma D.8. *For $q > 0$, weights $w_0, \dots, w_M > 0$, and $\tau > 0$,*

$$\inf_{x_\ell > 0:\, \sum x_\ell \le \tau} \sum_{\ell=0}^{M} w_\ell x_\ell^{-q} = \tau^{-q}\left(\sum_{\ell=0}^{M} w_\ell^{1/(q+1)}\right)^{q+1}. \tag{234}$$

Proof. The Lagrange equations give $x_\ell = c w_\ell^{1/(q+1)}$; the active constraint determines $c = \tau/(\sum w_j^{1/(q+1)})$. Substitution proves the identity. Convexity of $x \mapsto x^{-q}$ makes this stationary point the global minimizer. □

The allocation identity now balances accuracy across levels. Applying it separately to quadrature and oracle noise exposes the different dimension thresholds of the two costs.

Lemma D.9. *Let H be as above, let $z_0 \in H$, and let $\tau > 0$, $\rho \in (0, 1/2)$. All values $f(z) - f(z_0)$ on a $C_k r/n$-net of H can be estimated simultaneously to absolute error at most τ, with conditional failure probability at most ρ, using*

$$C_{k,\alpha} \log\left(\frac{C_k n^k (M+1)}{\rho}\right)\left[n^k + \left(\frac{Lr^2}{\tau}\right)^{2/3} \mathcal{A}_k(n) + \left(\frac{\sigma r}{\tau}\right)^{p_\alpha} \mathcal{C}_{k,p_\alpha}(n)\right] \tag{235}$$

oracle calls, where

$$\mathcal{A}_k(n) := \begin{cases} 1, & k = 1, \\ n^{k-4/3}, & k \ge 2, \end{cases} \tag{236}$$

and

$$\mathcal{C}_{k,p}(n) := \begin{cases} 1, & k < p, \\ (1+\log n)^{p+1}, & k = p, \\ n^{k-p}, & k > p. \end{cases} \tag{237}$$

An external reference z_0 is also allowed provided $\mathrm{dist}(z_0, H) \le \sqrt{k+1}\, r$. The necessary root edges then have length $O_k(r)$ and do not change the order of the bound. Without a bound on this distance, the root-edge cost must be counted separately.

Proof. Estimate every tree edge using Lemma D.7, assign each edge failure probability at most $\rho/(C_k n^k(M+1))$, and sum edge estimates along root-to-leaf paths. If every edge estimate succeeds and if the level-ℓ deterministic and stochastic error allowances are u_ℓ, v_ℓ, respectively, the uniform path error is at most $\sum_\ell (u_\ell + v_\ell)$. Allocate each sum at most $\tau/2$.

There are $O_k(2^{k\ell})$ level-ℓ edges of length $O_k(r2^{-\ell})$. The deterministic quadrature cost, before the common logarithmic factor, is therefore

$$C_k(Lr^2)^{2/3} \sum_{\ell=0}^{M} 2^{(k-4/3)\ell} u_\ell^{-2/3}.$$

Apply Lemma D.8 with $q = 2/3$. If $k = 1$, the resulting geometric sum is bounded; if $k \geq 2$, it is dominated by its last term and equals $O_k(n^{k-4/3})$ after taking the required power. This yields the second term in (235).

The stochastic cost is

$$C_{k,\alpha}(\sigma r)^{p_\alpha} \sum_{\ell=0}^{M} 2^{(k-p_\alpha)\ell} v_\ell^{-p_\alpha}.$$

Applying Lemma D.8 with $q = p_\alpha$ gives

$$C_{k,\alpha} \left(\frac{\sigma r}{\tau}\right)^{p_\alpha} \left[\sum_{\ell=0}^{M} 2^{(k-p_\alpha)\ell/(p_\alpha+1)}\right]^{p_\alpha+1},$$

which is exactly (237). Finally, every edge requires at least one sample, producing the n^k term. The union bound over edges and the common median-amplification factor give the logarithm in (235). A constant number of additional root edges is absorbed into the same estimates. □

D.4 Gradient-flow trapping and the upper bound for $d \geq 2$

The gradient-flow viewpoint originates in Bubeck & Mikulincer (2020b). Here we need a version that tolerates approximate relative values and uses no function-value oracle. We first prove that a boundary certificate traps a small gradient. We then show how two cuts preserve that certificate with a controlled increase in its slope, including a strict interpolation margin. The preceding reconstruction lemma will supply all values used in these cuts.

Definition D.10 (Gradient-flow trap)**.** Let $R \subset \mathbb{R}^d$ be a compact axis-aligned box with positive side lengths, $x \in \mathrm{int}(R)$, and $c \geq 0$. The pair (R, x) is a c-trap if

$$f(y) > f(x) - c\|y - x\| \qquad (y \in \partial R). \tag{238}$$

Lemma D.11 (Trap lemma)**.** *Every c-trap contains a point z with $\|\nabla f(z)\| \leq c$.*

Proof. Suppose instead that $\|\nabla f\| > c$ on R. Compactness gives $m_R := \min_R \|\nabla f\| > c \geq 0$. The vector field $-\nabla f/\|\nabla f\|$ is Lipschitz on a neighborhood of R, because the gradient is Lipschitz and bounded away from zero there. Starting at x, its unit-speed trajectory is therefore uniquely defined up to exit and can be continued as long as it remains in R. While inside,

$$\frac{d}{dt} f(z(t)) = -\|\nabla f(z(t))\| \leq -m_R < -c.$$

The trajectory cannot remain in R for all time: otherwise its objective value tends to $-\infty$, contrary to the minimum of the continuous function f on the compact box. At the finite first boundary time $\tau > 0$, let $y = z(\tau) \in \partial R$. Its length is τ, so

$$f(y) - f(x) \leq -m_R\tau < -c\tau \leq -c\|y - x\|,$$

contradicting (238). □

The trap lemma is existential, but it becomes algorithmic once the box is small. The next refinement rule shrinks volume by a fixed factor using only a finite set of approximate values; it also accounts explicitly for the error between grid vertices.

Lemma D.12. *Let $d \geq 2$ and let (R, x) be a c-trap whose aspect ratio is at most 3. Write $s > 0$ for its longest side. There are two parallel $(d-1)$-dimensional cross-sections E, F with the following property. Equip each cut with a Cartesian grid of coordinate mesh at most h, where $0 < h \leq s$, and use its standard simplicial triangulation. Suppose $\tau > 0$ and every grid vertex has an estimate satisfying*

$$|\widehat{D}(z) - (f(z) - f(x))| \leq \tau. \tag{239}$$

One can select $R_+ \subset R$ and $x_+ \in \operatorname{int}(R_+)$ such that

$$\operatorname{vol}(R_+) \leq \tfrac{5}{6} \operatorname{vol}(R), \qquad \operatorname{aspect}(R_+) \leq 3, \tag{240}$$

and (R_+, x_+) is a c_+-trap with

$$c_+ \leq c + 400d \left(\frac{\tau}{s} + \frac{Lh^2}{s} + \frac{ch^2}{s^2} \right). \tag{241}$$

Any larger nonnegative slope is also a valid certificate.

Proof. Permute and reflect coordinates so that $R = [0, s] \times R'$ and $x_1 \geq s/2$; every side of R' lies in $[s/3, s]$. Set $E = \{s/6\} \times R'$ and $F = \{s/3\} \times R'$. Call a grid vertex certified if $\widehat{D}(z) \leq -c\|z - x\| - \tau$, and let A contain all certified vertices of both cuts. If $A = \varnothing$, keep $R_+ = [s/3, s] \times R'$ and $x_+ = x$. Otherwise choose $z_\star \in \arg\min_{z \in A} \widehat{D}(z)$ and put $x_+ = z_\star$. If $z_\star \in F$, keep $[s/6, s] \times R'$; if $z_\star \in E$, keep $[0, s/3] \times R'$. This is the rule in Algorithm 3.

No certified vertex can lie on ∂R: the old trap inequality and (239) give $\widehat{D}(z) > -c\|z - x\| - \tau$ on that boundary. Hence the new pivot is interior, including in the unchanged-pivot case. The new side length is $2s/3$, $5s/6$, or $s/3$, and all other side lengths lie in $[s/3, s]$. This proves (240). If a new pivot is selected, certification implies

$$f(z_\star) - f(x) \leq -c\|z_\star - x\|. \tag{242}$$

For every retained point y of the old boundary, the triangle inequality therefore yields

$$f(y) - f(z_\star) > c(\|z_\star - x\| - \|y - x\|) \geq -c\|y - z_\star\|.$$

For an unchanged pivot the old-boundary certificate is immediate.

Let H be the newly introduced boundary cut. Its distance from x_+ is at least $s/6$. At a certified vertex $z \in H$, minimality of $\widehat{D}(z_\star)$ gives $f(z) - f(x_+) \geq -2\tau$. At a noncertified vertex, $f(z) - f(x) > -c\|z - x\| - 2\tau$; combining this with (242) and the triangle inequality gives the same uniform bound

$$f(z) - f(x_+) \geq -c\|z - x_+\| - 2\tau.$$

This bound also holds when $A = \varnothing$. Define

$$\gamma_{\mathrm{net}} = c + 18\tau/s, \qquad \Phi(y) = f(y) + \gamma_{\mathrm{net}}\|y - x_+\| \quad (y \in H). \tag{243}$$

Every grid vertex satisfies $\Phi(z) \geq f(x_+) + \tau$. The distance term has tangential Hessian norm at most $6/s$, so $\Phi|_H$ has Lipschitz gradient at most $L + 6\gamma_{\mathrm{net}}/s$.

For y in a simplex with vertices z_i and barycentric weights λ_i, smoothness applied at y and $\sum_i \lambda_i(z_i - y) = 0$ imply

$$\Phi(y) \geq \sum_i \lambda_i \Phi(z_i) - E_h, \qquad E_h = \frac{d-1}{2}(L + 6\gamma_{\mathrm{net}}/s)h^2.$$

Set $c_+ = \gamma_{\mathrm{net}} + 6E_h/s$. Since $\|y - x_+\| \geq s/6$,

$$f(y) + c_+\|y - x_+\| \geq \Phi(y) + E_h \geq f(x_+) + \tau > f(x_+).$$

Thus the full new face has the required strict trap inequality. Increasing c also preserves the old-boundary inequalities. Finally,

$$\begin{aligned} c_+ - c &= 18\tau/s + 3(d-1)Lh^2/s + 18(d-1)ch^2/s^2 + 324(d-1)\tau h^2/s^3 \\ &\leq 400d\left(\tau/s + Lh^2/s + ch^2/s^2\right), \end{aligned}$$

where $h \leq s$. This proves (241) with a fixed constant and preserves strictness even when some vertex estimates are at their error limits. □

We now allocate this slope increase across a finite sequence of refinements. The construction uses the explicit constants in (198); the resulting inequalities establish both termination and the claimed query count on one simultaneous success event.

Proof of Theorem D.1 (ii). Assume a current pivot x satisfies

$$f(x) - f^\star \le \Delta, \tag{244}$$

and put

$$\eta := \sqrt{\mu\Delta}. \tag{245}$$

If a new point has gradient norm at most η, PL gives gap at most $\Delta/2$.

Set $c_0 = \eta/8$ and let R_0 be the cube centered at x whose distance from x to every face is $H = 2\Delta/c_0$. For every boundary point y, $f(y) \ge f^\star \ge f(x) - \Delta > f(x) - c_0\|y - x\|$, so (R_0, x) is a c_0-trap. Its longest side is

$$s_0 = 2H = \frac{4\Delta}{c_0} = \frac{32\Delta}{\eta}. \tag{246}$$

Run Lemma D.12 repeatedly. Because each refinement shrinks volume by at least $5/6$ and keeps aspect ratio at most 3, after

$$J = \mathfrak{c}_d\lceil\log(2+\kappa)\rceil \tag{247}$$

steps the longest side is at most $\eta/(4L\sqrt{d})$, with the choice (198): indeed, the initial-to-final ratio is $O_d(L\Delta/\eta^2) = O_d(\kappa)$. More explicitly, $s_t \le 3(5/6)^{t/d}s_0$, so it suffices that $t \ge d\log(384\kappa\sqrt{d})/\log(6/5)$. For $d \ge 2$ and $\kappa \ge 1$, this is smaller than $\mathfrak{c}_d\lceil\log(2+\kappa)\rceil$: use $\log(384\kappa\sqrt{d}) \le (8+d)\log(2+\kappa)$ and $1/\log(6/5) < 6$. If the stopping length is reached earlier, we stop and use its last box and pivot as R_J, x_J; the same bound $c_J \le \eta/4$ applies. In the cost sums below, the index t runs only over refinements actually performed, whose number is at most J.

Allocate

$$a := \frac{\eta}{\mathfrak{c}_d J}, \qquad \tau_t := \frac{as_t}{\mathfrak{c}_d}, \qquad h_t^2 := \frac{as_t}{\mathfrak{c}_d L}, \tag{248}$$

where s_t is the current longest side. We maintain $c_t = \eta/8 + ta$. For $t \le J$ this is at most $\eta/8 + \eta/\mathfrak{c}_d \le \eta/4$. Before termination $Ls_t > \eta/(4\sqrt{d})$, so $c_t/(Ls_t) \le \sqrt{d}$ and $h_t^2/s_t^2 < 4\sqrt{d}/(\mathfrak{c}_d^2 J) < 1$. Using the explicit constant in Lemma D.12, the required slope increase is at most

$$400d\left(2 + \frac{c_t}{Ls_t}\right)\frac{a}{\mathfrak{c}_d} \le \frac{400d(2+\sqrt{d})}{\mathfrak{c}_d}a \le a.$$

Thus the algorithm's update $c_{t+1} = c_t + a$ preserves the trap by induction; no assumption about a later slope is used to prove an earlier step. Consequently

$$c_J \le \frac{\eta}{4}. \tag{249}$$

By Lemma D.11, the final box contains z with $\|\nabla f(z)\| \le \eta/4$. If x_J is the final pivot, then

$$\|\nabla f(x_J)\| \le \|\nabla f(z)\| + L\|x_J - z\| \le \frac{\eta}{4} + L\,\mathrm{diam}(R_J) \le \frac{\eta}{2} < \eta.$$

Thus one stage halves the objective gap.

At refinement t, each of the two cross-sections has dimension $k = d-1$, diameter $O_d(s_t)$, and target mesh h_t. Put

$$n_t := 2^{\lceil\log_2(\mathfrak{c}_d s_t/h_t)\rceil}, \qquad \chi_t := \frac{Ls_t}{a}. \tag{250}$$

Then $n_t^2 \asymp_d \chi_t$, while $Ls_t^2/\tau_t \asymp_d \chi_t$ and $\sigma s_t/\tau_t \asymp_d \sigma/a$. The root of each cut tree is connected to the pivot by a constant number of segments of length $O_d(s_t)$. Lemma D.9 therefore bounds one refinement by

$$C_{d,\alpha}\Lambda_{d,\alpha}\left[\mathcal{D}_k(\chi_t) + \left(\frac{\sigma}{a}\right)^{p_\alpha}\mathcal{C}_{k,p_\alpha}(n_t)\right]. \tag{251}$$

For $k = 1$, the exponent $2/3$ comes from stratified line integration; for $k \geq 2$, the n_t^k tree size and deterministic quadrature have the same order $\chi_t^{k/2}$.

The longest sides decrease geometrically up to dimension-dependent constants: from the volume reduction and bounded aspect ratio,

$$s_t \leq C_d(5/6)^{t/d} s_0. \tag{252}$$

Using $a = \eta/(\mathfrak{c}_d J)$, $s_0 = 32\Delta/\eta$, and $\eta^2 = \mu\Delta$, we obtain

$$\chi_0 \leq C_d \kappa J. \tag{253}$$

Consequently,

$$\sum_{\text{performed } t} \mathcal{D}_k(\chi_t) \leq C_d \mathcal{D}_k(\kappa J). \tag{254}$$

For the stochastic part, first note

$$\left(\frac{\sigma}{a}\right)^{p_\alpha} \leq C_{d,\alpha} J^{p_\alpha} \left(\frac{\sigma^2}{\mu\Delta}\right)^{\theta_\alpha}. \tag{255}$$

If $k < p_\alpha$, the surface factor in (237) is uniformly bounded and summing over the J refinements adds one factor J. If $k = p_\alpha$, it adds at most $J[\log(2 + \kappa J)]^{p_\alpha+1}$. If $k > p_\alpha$, the geometric decay in (252) gives

$$\sum_t n_t^{k-p_\alpha} \leq C_{d,\alpha}(\kappa J)^{(k-p_\alpha)/2}.$$

Combining these estimates proves that a gap-Δ stage costs at most

$$C_{d,\alpha}\Lambda_{d,\alpha}\left[\mathcal{D}_k(\kappa J) + \mathcal{H}_{k,p_\alpha}(\kappa, J)\left(\frac{\sigma^2}{\mu\Delta}\right)^{\theta_\alpha}\right]. \tag{256}$$

Every cutting-surface reference is the current pivot, lying inside the parent box. Its distance from a cut is at most s_t, so the external-reference condition in Lemma D.9 is satisfied. Assign each of the at most $2JS$ cutting-surface requests conditional failure probability $\delta/(2JS)$. Each reconstruction allocates this budget among its tree edges; the line-integral guarantee is conditional on the full history at each edge request. Iterated conditioning and a union bound therefore cover all adaptively selected cuts, without assuming that their errors are independent. The number of edges per cut is $O_d((\kappa J)^{k/2})$ and its depth is $O_d(\log(2 + \kappa J))$. Since $\log(2 + \kappa J) = O_d(J + 1)$, the logarithm in (204), enlarged by a dimension-dependent constant, bounds all median-amplification factors.

Now set $\Delta_s = \Delta_0/2^s$ and run $S = \lceil \log_2(\Delta_0/\epsilon) \rceil$ stages, assigning failure probability δ/S to each. The deterministic cost repeats S times, whereas

$$\sum_{s=0}^{S-1}\left(\frac{\sigma^2}{\mu\Delta_s}\right)^{\theta_\alpha} \leq C_\alpha Q_\epsilon. \tag{257}$$

A union bound makes all line-integral and trap certificates simultaneous with probability at least $1 - \delta$. Equations (256)–(257) prove Theorem D.1(ii). □

D.5 Assembly of the fixed-dimensional phase diagram

The remaining step is to compare the two lower-bound mechanisms with the explicit upper bounds. This requires keeping the confidence criterion separate from expected risk and retaining the condition-number restriction of the exact-gradient lower family. The following corollary records the precise scope of the main-text comparison, including its last regime.

Corollary D.13. *Let $d \geq 1$ be fixed, $\kappa \geq 6$, $1 < \alpha \leq 2$, $\sigma \geq 0$, $0 < \epsilon \leq c_0\Delta_0$, and $0 < \delta \leq 1/4$, with $c_0 > 0$ small enough. Then*

$$\mathsf{T}^{(1)}_{\epsilon,\delta} = \widetilde{\Theta}_\alpha\left[\log(\Delta_0/\epsilon) + Q_\epsilon\right], \tag{258}$$

where logarithmic factors in $\kappa, \Delta_0/\epsilon, 1/\delta$ are suppressed. For $d \geq 2$,

$$c_\alpha \left[\log(\Delta_0/\epsilon) + Q_\epsilon\right] \leq \mathsf{T}^{(d)}_{\epsilon,1/4} \leq \widetilde{O}_{d,\alpha}\left[K_d(\kappa)\log(\Delta_0/\epsilon) + \kappa^{(d-1-p_\alpha)_+/2} Q_\epsilon\right], \tag{259}$$

where $(r)_+ = \max\{r, 0\}$, $K_2(\kappa) = \kappa^{2/3}$, and $K_d(\kappa) = \kappa^{(d-1)/2}$ for $d \geq 3$. In particular, for $d > 3$,

$$p_\alpha \geq d-1 \quad \Longleftrightarrow \quad \alpha \leq (d-1)/(d-2), \tag{260}$$

and the upper and lower stochastic terms established here match up to logarithmic factors in this range. No necessity claim is made for the extra factor when $p_\alpha < d-1$.

Proof of Corollary D.13 and Theorem 5.2. For $d = 1$, combine Theorem D.1(i) with the constant-confidence consequence of Theorem D.2. The latter also applies to every $\delta \leq 1/4$, because the minimax complexity is nonincreasing in the allowed failure probability. The bisection bound introduces only the logarithmic factors displayed in (199). This proves (258).

For $d \geq 2$ put $k = d-1$. Both J and $\Lambda_{d,\alpha}$ in Theorem D.1(ii) are logarithmic factors for fixed d. If $k < p_\alpha$, $\mathcal{H}_{k,p_\alpha}$ contains only such factors. If $k = p_\alpha$, the dyadic level sum adds $[\log(2+\kappa J)]^{p_\alpha+1}$, still logarithmic. If $k > p_\alpha$, its polynomial contribution is $\kappa^{(k-p_\alpha)/2}$. Meanwhile $\mathcal{D}_1(\kappa J)$ contributes $\kappa^{2/3}$ and $\mathcal{D}_k(\kappa J)$ contributes $\kappa^{k/2}$ for $k \geq 2$. These observations give the upper side of (259); Theorem D.2 gives its lower side for every σ.

For $d = 2$, $k = 1 < p_\alpha$ throughout the allowed moment range. For $d = 3$, $k = 2 \leq p_\alpha$, with equality only at $\alpha = 2$. These are the rates in (20) and (21). For $d > 3$, multiplication by the positive quantity $\alpha - 1$ gives $p_\alpha \geq d-1$ if and only if $(d-2)\alpha \leq d-1$, proving (260). This also explains the rate in (22) and supplies the last row of Theorem 5.2, namely (23). □