DENSE WEAK HIDING: CLOSING COMPLEXITY GAPS IN NONCONVEX AND PL FINITE-SUM OPTIMIZATION UNDER INDIVIDUAL SMOOTHNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Under individual smoothness, the optimal incremental first-order oracle (IFO) complexity of nonconvex finite-sum optimization is still open. Known algorithms use $O(n + \sqrt{n}\,\Delta L_{\max}/\varepsilon^2)$ calls, but the best lower bounds miss a factor of $\sqrt{n}$ on the $\Delta L_{\max}/\varepsilon^2$ term. We prove the matching lower bound $\Omega(n + \sqrt{n}\,\Delta L_{\max}/\varepsilon^2)$. Thus PAGE and SPIDER are minimax optimal up to universal constants, under both individual and mean-squared smoothness. Unlike earlier bounds for linear-span algorithms, our result applies to the broader class of randomized IFO algorithms, allowing arbitrary choices of component indices and query points based on the complete preceding history.

Under the global Polyak–Łojasiewicz (PL) condition, we use a similar idea to obtain an $\Omega(n + \kappa_{\max}\sqrt{n}\log(\Delta/\varepsilon))$ lower bound for large $\kappa_{\max}$. We also note that existing PL lower bounds have focused only on the regime where κ is relatively large. To fill this gap, we study the regime $\kappa_{\max} < \sqrt{n}$ and obtain a new lower-bound rate, $\Omega(n + n\log(\Delta/\varepsilon)/(1 + \log(\sqrt{n}/\kappa_{\max})))$. The new lower bound motivates us to propose Restarted PAGE. Its upper bound matches the new rate for small $\kappa_{\max}$ and recovers the standard PAGE rate for large $\kappa_{\max}$, showing that our new lower bound is nearly tight.

Our new lower bounds are based on our proposed *dense weak hiding* construction. By spreading each hidden direction across all components, the construction makes every single IFO query weakly informative while preserving the direction in the full row average. Because unrevealed stages remain inactive even for arbitrary query points, an algorithm must spend many calls to expose a stage before it can make substantial progress. We choose the bias to balance this revelation cost against the number of stages permitted by individual smoothness, and this balance produces the missing $\sqrt{n}$ factor.

1 INTRODUCTION

Finite-sum methods optimize $F(x) = n^{-1}\sum_{i=1}^{n} f_i(x)$ by querying a fixed set of component functions. Unlike an online stochastic oracle, the same components remain available for repeated queries. Unlike full-gradient optimization, one call accesses only one component. The algorithm may choose its next component and query point from all preceding observations. Modern variance-reduced methods exploit this access pattern (Defazio et al., 2014; Fang et al., 2018; Li et al., 2021). To determine whether these methods are optimal, a lower bound must use the same finite-sum oracle model and smoothness assumption.

For nonconvex stationarity, PAGE and SPIDER use $O(n + \sqrt{n}\,\Delta L_{\mathrm{ms}}/\varepsilon^2)$ IFO calls under mean-squared smoothness (Fang et al., 2018; Li et al., 2021). Individual $L_{\max}$-smoothness implies the same condition with $L_{\mathrm{ms}} \le L_{\max}$, so this upper bound is $O(n + \sqrt{n}\,\Delta L_{\max}/\varepsilon^2)$ under individual smoothness. The known lower bounds do not match it. Zhou & Gu (2019) and Han et al. (2024) cover linear-span algorithm classes and give only $\Omega(n + \Delta L_{\max}/\varepsilon^2)$ under individual smoothness. Emmenegger et al. (2022) allow unrestricted queries but give only $\Omega(\Delta L_{\max}/\varepsilon^2)$. This leaves the first question. *Is the factor $\sqrt{n}$ unavoidable for randomized IFO algorithms that may choose component indices and query points using their full preceding history?*

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.

Closing this gap requires more than a rescaled version of an existing hard instance. The sparse-link construction of Zhou & Gu (2019) preserves the sequential structure but raises individual smoothness by a factor of $\sqrt{n}$. Like that of Han et al. (2024), its linear-span argument does not cover queries outside the generated span. Emmenegger et al. (2022) assign a separate block to each component. This allows arbitrary query points but weakens the average gradient by $n^{-1/2}$. A construction that avoids all three problems must distribute hidden information across components, preserve the signal in the finite-sum average without this $n^{-1/2}$ loss, and prevent unrestricted queries from activating later stages.

We introduce *dense weak hiding* to meet these requirements. At each stage, a hidden direction is encoded in a fixed $n \times D$ sign table. Each row is only weakly biased toward the direction, while the exact row average is $\rho\Theta^{(j)}$. A bounded radial map turns any query point into an alignment vector in the unit ball. A smooth gate satisfies $G = G' = 0$ below its threshold, so unopened links drop out of both the function value and the gradient. Revealing one stage requires $\Omega(\min\{n, \rho^{-2}\})$ distinct rows. The construction satisfies the gap and individual smoothness constraints with $\Theta(\Delta L_{\max}\rho/\varepsilon^2)$ stages. Their product is

$$\Omega\left(\frac{\Delta L_{\max}}{\varepsilon^2}\min\{n\rho, \rho^{-1}\}\right),$$

which is maximized at $\rho \asymp n^{-1/2}$. This gives the missing $\sqrt{n}$ factor and shows that PAGE and SPIDER are minimax optimal for the randomized IFO class considered here.

The global Polyak–Łojasiewicz (PL) condition poses a second minimax question. Standard PAGE gives $O((n + \kappa_{\mathrm{ms}}\sqrt{n})\log(\Delta/\varepsilon))$ (Li et al., 2021, Corollary 6), but two gaps remained. Under individual smoothness with unrestricted queries, no earlier result showed how the complexity depends on the condition number, and no matching rate was known for small κ. Yue et al. (2023) gave the first PL lower bound. It treats deterministic full-gradient methods and assumes a large objective condition number. Bai et al. (2024) obtain the $\kappa_{\mathrm{ms}}\sqrt{n}$ term for linear-span IFO algorithms, but only at large condition numbers. The version of their result that holds for all condition numbers gives only $\Omega(n)$. This raises the second question. *What is the minimax complexity under individual smoothness as $\kappa_{\max}$ moves across $\sqrt{n}$, and is the standard PAGE guarantee under mean-squared smoothness tight below this transition?*

We determine the complexity in both $\kappa_{\max}$ regimes. For $3 \le \kappa_{\max} \le \sqrt{n}$, a two-scale construction gives $\Omega(n + n\log(\Delta/\varepsilon)/[1 + \log(\sqrt{n}/\kappa_{\max})])$. Its average objective is globally strongly convex. For $\kappa_{\max} \ge \sqrt{n}$, a geometrically weighted construction gives $\Omega(n + \kappa_{\max}\sqrt{n}\log(\Delta/\varepsilon))$. These lower bounds hold under individual smoothness for randomized IFO algorithms and hence also give the mean-squared lower bounds. A similar scale appears in PAGE. One epoch costs $O(n)$ calls in expectation and shrinks the function value gap by $\Theta(\kappa_{\mathrm{ms}}/\sqrt{n})$. Restarted PAGE uses these epochs as restart periods. It gives the same logarithmic dependence and matches the lower bounds in both κ_{ms} regimes under mean-squared smoothness.

Contributions. Our main contributions are summarized as follows.

1. **Dense weak hiding.** We give a construction for fixed finite sums. Its information analysis covers randomized IFO algorithms with unrestricted query selection.
2. **Tight nonconvex complexity under individual smoothness.** We prove the matching $\Omega(n + \sqrt{n}\,\Delta L_{\max}/\varepsilon^2)$ lower bound.
3. **PL lower bounds in both $\kappa_{\max}$ regimes.** We prove the lower bounds under individual smoothness for every $\kappa_{\max} \ge 3$. The average objective of the small-$\kappa_{\max}$ hard instance is globally strongly convex.
4. **Restarted PAGE matching the lower bound.** Under mean-squared smoothness, Restarted PAGE improves the standard guarantee for $\kappa_{\mathrm{ms}} < \sqrt{n}$ and matches the lower bound in both κ_{ms} regimes.

2 Related Work

Nonconvex stationarity. Early work established oracle lower bounds for finding stationary points of smooth nonconvex full objectives (Carmon et al., 2020; 2021). Arjevani et al. (2023) extend

Table 1: Finite-sum bounds; lower bounds hold with constant success probability.

Nonconvex stationarity			
Result	**Oracle complexity**	**Algorithm class**	**Reference**
Upper	$O\big(n+\sqrt{n}\,\Delta L_{\mathrm{ms}}/\varepsilon^2\big)$	PAGE / SPIDER	Fang et al., 2018; Li et al., 2021
Lower	$\Omega\big(\Delta L_{\max}/\varepsilon^2\big)$	randomized methods; arbitrary queries	Emmenegger et al., 2022
Lower	$\Omega\big(n+\Delta L_{\max}/\varepsilon^2\big)$	linear-span IFO / PIFO	Zhou & Gu, 2019; Han et al., 2024
New lower	$\boldsymbol{\Omega\big(n+\sqrt{n}\,\Delta L_{\max}/\varepsilon^2\big)}$	**randomized IFO algorithms**	**Theorem 3.1**

Global PL condition			
Result	**Oracle complexity and range**	**Algorithm class**	**Reference**
Upper	$O\left((n+\kappa_{\mathrm{ms}}\sqrt{n})\log\dfrac{\Delta}{\varepsilon}\right),\quad \kappa_{\mathrm{ms}}\geq 1$	standard PAGE	Li et al., 2021, Cor. 6
Lower	$\Omega\left(\kappa_F\log\dfrac{\Delta}{\varepsilon}\right),\quad \kappa_F\gtrsim 10^6$	deterministic full-gradient	Yue et al., 2023
Lower	$\Omega\left(n+\kappa_{\mathrm{ms}}\sqrt{n}\log\dfrac{\Delta}{\varepsilon}\right),\quad \kappa_{\mathrm{ms}}\gtrsim 10^5\sqrt{n}$	linear-span IFO algorithms	Bai et al., 2024
New upper	$\begin{cases} O\left(n+\dfrac{n\log(\Delta/\varepsilon)}{1+\log(\sqrt{n}/\kappa_{\mathrm{ms}})}\right), & 1\leq\kappa_{\mathrm{ms}}\leq\sqrt{n}, \\ O\left(n+\kappa_{\mathrm{ms}}\sqrt{n}\log\dfrac{\Delta}{\varepsilon}\right), & \kappa_{\mathrm{ms}}\geq\sqrt{n}. \end{cases}$	**Restarted PAGE**	**Alg. 1; Prop. E.2**
New lower	$\begin{cases} \Omega\left(n+\dfrac{n\log(\Delta/\varepsilon)}{1+\log(\sqrt{n}/\kappa_{\max})}\right), & 3\leq\kappa_{\max}\leq\sqrt{n}, \\ \Omega\left(n+\kappa_{\max}\sqrt{n}\log\dfrac{\Delta}{\varepsilon}\right), & \kappa_{\max}\geq\sqrt{n}. \end{cases}$	**randomized IFO algorithms**	**Theorem 3.3**

zero-chain techniques to arbitrary randomized stochastic first-order methods. They prove sharp lower bounds under bounded variance and mean-squared smoothness, with fresh oracle randomness at each call. Concurrent work by Jin (2026) also uses weakly biased fresh signs under bounded noise. For finite-sum upper bounds, SPIDER and PAGE achieve $O(n+\sqrt{n}\,\Delta L_{\mathrm{ms}}/\varepsilon^2)$ under mean-squared smoothness (Fang et al., 2018; Li et al., 2021). Under individual smoothness, prior finite-sum lower bounds do not match this rate. Zhou & Gu (2019), Xie et al. (2019), and Han et al. (2024) track linear-span progress. Of these, Zhou & Gu (2019) and Han et al. (2024) obtain only $\Omega(n+\Delta L_{\max}/\varepsilon^2)$. The construction of Emmenegger et al. (2022) allows arbitrary queries but gives only $\Omega(\Delta L_{\max}/\varepsilon^2)$. All three bounds therefore miss the factor $\sqrt{n}$ multiplying $\Delta L_{\max}/\varepsilon^2$. Table 1 also records which algorithms each bound covers. Our lower bound recovers this factor without a linear-span restriction.

PL lower bounds. Yue et al. (2023) prove a lower bound for deterministic full-gradient methods that depends explicitly on the condition number, while Bai et al. (2024) obtain the $\kappa_{\mathrm{ms}}\sqrt{n}$ term for linear-span IFO algorithms. These results require the large condition numbers recorded in Table 1. The theorem of Bai et al. (2024) that covers all condition numbers gives only $\Omega(n)$. Two gaps therefore remained. Under individual smoothness, no lower bound showed any dependence on κ, and the small-κ regime was open for randomized IFO algorithms. Our lower bound covers every $\kappa_{\max}\geq 3$ for randomized IFO algorithms, and Restarted PAGE gives matching upper bounds under mean-squared smoothness. The exact parameter ranges in prior work are recorded in Appendix G.

Relation to strongly convex finite-sum lower bounds. The average objective in our small-κ construction is globally strongly convex, although its components need not be convex. Han et al. (2024, Theorem 66) obtain the same small-κ denominator $1+(\log(\sqrt{n}/\kappa))_+$ for a strongly convex objective under average smoothness in their linear-span PIFO class. Our result establishes this dependence under individual smoothness for randomized IFO algorithms with unrestricted query selection. Classical lower bounds for strongly convex finite sums with convex components give the rate $n+\sqrt{n\kappa_{\max}}\log(1/\varepsilon)$ (Agarwal & Bottou, 2015; Woodworth & Srebro, 2016; Lan & Zhou,

2018; Allen-Zhu, 2018). Those bounds assume convex components, which our hard instances need not have.

Proof techniques. Standard tools for lower bounds include zero-chains, resisting oracles, random rotations, and arguments that track how component queries reveal new coordinates (Nemirovski & Yudin, 1983; Nesterov, 2018; Agarwal & Bottou, 2015; Woodworth & Srebro, 2016; Carmon et al., 2020; 2021). The closest related construction is that of Zhou (2026), which studies deterministic first-order algorithms for full objectives with higher-order smoothness. That work lifts a scalar hard function from Carmon et al. (2020; 2021) to a block chain and transfers zero-respecting hardness by an orthogonal rotation. Here the oracle accesses a fixed finite sum, and the algorithm may randomize both its component choices and query points. An information argument therefore replaces the span condition. Fixed encoding tables hide the directions, and a mutual information bound for rows drawn without replacement controls what the replies reveal.

3 Setting and Main Results

For $d \geq 1$, write $F \in \mathcal{F}_{\mathrm{ind}}^{n,d}(\Delta, L_{\max})$ when a specified finite-sum decomposition satisfies

$$F : \mathbb{R}^d \to \mathbb{R}, \qquad F = \frac{1}{n}\sum_{i=1}^{n} f_i, \qquad F(0) - \inf_x F(x) \leq \Delta, \tag{1}$$

and every component satisfies

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_{\max}\|x - y\| \qquad \forall x, y. \tag{2}$$

Every hard component constructed in this paper is C^∞. The ambient dimension is arbitrary and finite, and the dimension of the hard instance may depend on the problem parameters and oracle budget. When d is omitted from a class notation, all finite dimensions are allowed. Appendix G.1 gives the formal definition.

We also write $F \in \mathcal{F}_{\mathrm{ms}}^{n,d}(\Delta, L_{\mathrm{ms}})$ when the same finite-sum and gap conditions hold and

$$\left(\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f_i(y)\|^2\right)^{1/2} \leq L_{\mathrm{ms}}\|x - y\| \qquad \forall x, y. \tag{3}$$

At the same numerical bound L, $\mathcal{F}_{\mathrm{ind}}^{n}(\Delta, L) \subseteq \mathcal{F}_{\mathrm{ms}}^{n}(\Delta, L)$. Following the literature on finite-sum lower bounds (Zhou & Gu, 2019; Emmenegger et al., 2022; Han et al., 2024), we call (2) *individual smoothness* (also called componentwise smoothness). Condition (3) is *mean-squared smoothness* (often called average smoothness), as used in the SPIDER/PAGE analyses (Fang et al., 2018; Li et al., 2021).

Incremental first-order oracle. We use the standard IFO model introduced by Agarwal & Bottou (2015) and later studied for randomized incremental gradient methods by Lan & Zhou (2018). Fix a dimension d. An exact IFO call selects $(i, x) \in [n] \times \mathbb{R}^d$ and returns

$$\mathcal{O}_i(x) = \big(f_i(x), \nabla f_i(x)\big). \tag{4}$$

Throughout the paper, *IFO* refers to this exact oracle returning both value and gradient.

Algorithmic scope. We consider randomized IFO algorithms with unrestricted query selection. At every round the algorithm may choose both the component index and the query point as arbitrary functions of its private randomness and the complete history of queries and responses. It may repeat indices, stop at a random time, and return a point that it never queried. We impose no linear-span or zero-respecting restriction, and the rule for choosing component indices may depend on the full history. This formulation follows standard conventions in oracle complexity (Carmon et al., 2020; Arjevani et al., 2023). Appendix G.1 gives the measurable model and the minimax definition over all dimensions used below.

For a class $\mathcal{F}$ and error criterion $\mathcal{E}$, write $\mathrm{Comp}_{\varepsilon,\delta}^{\mathrm{IFO}}(\mathcal{F}; \mathcal{E})$ for the smallest fixed budget that achieves error at most ε with probability at least $1 - \delta$, uniformly over dimensions and instances in the worst case. We use $\mathcal{E}_F^{\mathrm{nc}}(x) = \|\nabla F(x)\|$ and $\mathcal{E}_F^{\mathrm{PL}}(x) = F(x) - F^*$.

Theorem 3.1 (Nonconvex lower bound for randomized IFO algorithms)**.** *There are universal constants $n_0, c_0, p_0 > 0$ such that, for every $n \geq n_0$, all $L_{\max}, \Delta > 0$, every ε satisfying $0 < \varepsilon^2 \leq c_0 \Delta L_{\max}$, and every fixed $\delta \in (0, p_0)$,*

$$\mathrm{Comp}^{\mathrm{IFO}}_{\varepsilon,\delta}\left(\mathcal{F}^n_{\mathrm{ind}}(\Delta, L_{\max}); \mathcal{E}^{\mathrm{nc}}\right) = \Omega\left(n + \sqrt{n}\,\frac{\Delta L_{\max}}{\varepsilon^2}\right). \tag{5}$$

The implicit constant is universal.

Appendix F extends the same rate to expected call budgets in the worst case and to expected gradient norm.

To state both smoothness models compactly, let $\mathsf{s} \in \{\mathrm{ind}, \mathrm{ms}\}$ indicate the model and let L_{s} denote the corresponding smoothness bound.

Corollary 3.2 (Tight IFO complexity for nonconvex stationarity)**.** *For every $n \geq n_0$, $L_{\mathsf{s}}, \Delta > 0$, any fixed $\delta \in (0, p_0)$, and $\mathsf{s} \in \{\mathrm{ind}, \mathrm{ms}\}$, the minimax IFO complexity is*

$$\mathrm{Comp}^{\mathrm{IFO}}_{\varepsilon,\delta}\left(\mathcal{F}^n_{\mathsf{s}}(\Delta, L_{\mathsf{s}}); \mathcal{E}^{\mathrm{nc}}\right) = \Theta\left(n + \sqrt{n}\,\frac{\Delta L_{\mathsf{s}}}{\varepsilon^2}\right) \tag{6}$$

whenever $0 < \varepsilon^2 \leq c_0 \Delta L_{\mathsf{s}}$. The upper bound follows from PAGE and SPIDER (Fang et al., 2018; Li et al., 2021), since individual $L_{\max}$-smoothness implies mean-squared smoothness with $L_{\mathrm{ms}} \leq L_{\max}$. Theorem 3.1 gives the lower bound for $\mathsf{s} = \mathrm{ind}$, and the same hard instances satisfy the mean-squared condition at the same numerical bound.

For the PL result, use the corresponding smoothness constant and condition number:

$$(L_{\mathsf{s}}, \kappa_{\mathsf{s}}) = \begin{cases} (L_{\max}, \kappa_{\max}), & \mathsf{s} = \mathrm{ind}, \\ (L_{\mathrm{ms}}, \kappa_{\mathrm{ms}}), & \mathsf{s} = \mathrm{ms}, \end{cases} \qquad \kappa_{\max} := \frac{L_{\max}}{\mu}, \quad \kappa_{\mathrm{ms}} := \frac{L_{\mathrm{ms}}}{\mu}. \tag{7}$$

Following the standard PL convention (Polyak, 1963; Karimi et al., 2016), define $\mathcal{F}^{n,d}_{\mathsf{s},\mathrm{PL}}(\Delta, L_{\mathsf{s}}, \mu)$ as the subclass of $\mathcal{F}^{n,d}_{\mathsf{s}}(\Delta, L_{\mathsf{s}})$ for which $F^* > -\infty$ and

$$\|\nabla F(x)\|^2 \geq 2\mu(F(x) - F^*) \qquad \forall x. \tag{8}$$

Here $F^* = \inf_x F(x)$. When d is omitted, the same convention over finite dimensions applies.

Theorem 3.3 (Tight PL complexity under both smoothness assumptions)**.** *There is a universal constant $c_\varepsilon > 0$ such that, for every integer $n \geq n_0$, every $\mathsf{s} \in \{\mathrm{ind}, \mathrm{ms}\}$, all $L_{\mathsf{s}}, \mu, \Delta > 0$ with $\kappa_{\mathsf{s}} \geq 3$, every ε satisfying $0 < \varepsilon \leq c_\varepsilon \Delta$, and every fixed $\delta \in (0, p_0)$,*

$$\mathrm{Comp}^{\mathrm{IFO}}_{\varepsilon,\delta}\left(\mathcal{F}^n_{\mathsf{s},\mathrm{PL}}(\Delta, L_{\mathsf{s}}, \mu); \mathcal{E}^{\mathrm{PL}}\right) = \begin{cases} \Theta\left(n + \dfrac{n \log(\Delta/\varepsilon)}{1 + \log(\sqrt{n}/\kappa_{\mathsf{s}})}\right), & 3 \leq \kappa_{\mathsf{s}} \leq \sqrt{n}, \\ \Theta\left(n + \kappa_{\mathsf{s}}\sqrt{n}\log\dfrac{\Delta}{\varepsilon}\right), & \kappa_{\mathsf{s}} \geq \sqrt{n} \end{cases}. \tag{9}$$

The hard instances are individually $L_{\max}$-smooth and also satisfy mean-squared smoothness with bound $L_{\max}$. In the small-$\kappa_{\max}$ branch, their average objective is globally strongly convex. By contrast, the average-smooth hard instances of Zhou & Gu (2019) and Han et al. (2024) have component smoothness of order $\sqrt{n}\,L_{\mathrm{ms}}$. Rewriting their bounds in terms of $L_{\max}$ therefore loses the factor $\sqrt{n}$. Section 7 gives the matching Restarted PAGE upper bound under mean-squared smoothness in terms of κ_{ms}.

4 Dense Weak Hiding

Dense weak hiding combines a fixed encoding table with a bounded radial map and a smooth gate. The table preserves the hidden signal in the average, and the radial map turns any query point into an alignment vector in the unit ball. Because the gate satisfies $G = G' = 0$ below its threshold, unopened links contribute neither value nor gradient. Figure 1 shows these ingredients and the resulting bias tradeoff.

Fixed encoding table. Fix a bias $\rho \in (0, 1/8]$ on the parity-compatible grid specified in Appendix A. At each stage j, independently draw $\Theta^{(j)} \sim \mathrm{Unif}\{\pm 1\}^D$ and sample one $n \times D$ sign table $S^{(j)} = (S^{(j)}_{i\ell})$ whose columns satisfy

$$\sum_{i=1}^{n} S^{(j)}_{i\ell} = n\rho\Theta^{(j)}_{\ell}. \tag{10}$$

Component i stores row $S^{(j)}_{i:} \in \{\pm 1\}^D$. The table is fixed before the interaction, so querying the same component returns the same row. Any one row is only weakly biased toward $\Theta^{(j)}$, but the column constraint gives the exact identity $n^{-1}\sum_i S^{(j)}_{i:} = \rho\Theta^{(j)}$. The average therefore preserves the full direction after rescaling by $1/\rho$. Appendix B proves that choosing indices from previous replies still requires many distinct rows.

Bounded radial map. Following the bounded radial map of Carmon et al. (2020) and its adaptation to finite sums with unrestricted queries by Emmenegger et al. (2022), define

$$\Psi(x) = \frac{x}{\sqrt{1 + \|x\|^2}}, \qquad u_\theta = \frac{\theta}{\sqrt{D}}. \tag{11}$$

For this local calculation, suppress the stage index. Given a direction θ and a row S_i, define

$$q_\theta(x) = \langle u_\theta, \Psi(x)\rangle, \qquad \widetilde{q}_i(x) = \frac{1}{\rho}\left\langle \frac{S_i}{\sqrt{D}}, \Psi(x)\right\rangle. \tag{12}$$

Then $n^{-1}\sum_i \widetilde{q}_i = q_\theta$ exactly. Since $\|\Psi(x)\| < 1$, even an arbitrarily large query point produces a bounded alignment vector. Appendix A shows that the derivative bounds for Ψ do not depend on D, while those for $\widetilde{q}_i$ scale as $O(1/\rho)$.

Smooth threshold gate. Fix $0 < a < b < 1$ and a nondecreasing $G \in C^\infty(\mathbb{R})$ satisfying

$$G(s) = G'(s) = 0 \quad (s \le a), \qquad G(s) = 1, \quad G'(s) = 0 \quad (s \ge b). \tag{13}$$

Below a, a link and all of its value and gradient contributions vanish exactly. Thus unopened links are invisible in an IFO reply. Bounded derivatives of G and the bounds for Ψ control each component Hessian. Appendix A gives the explicit choices. Appendix B makes the coupling across stages precise.

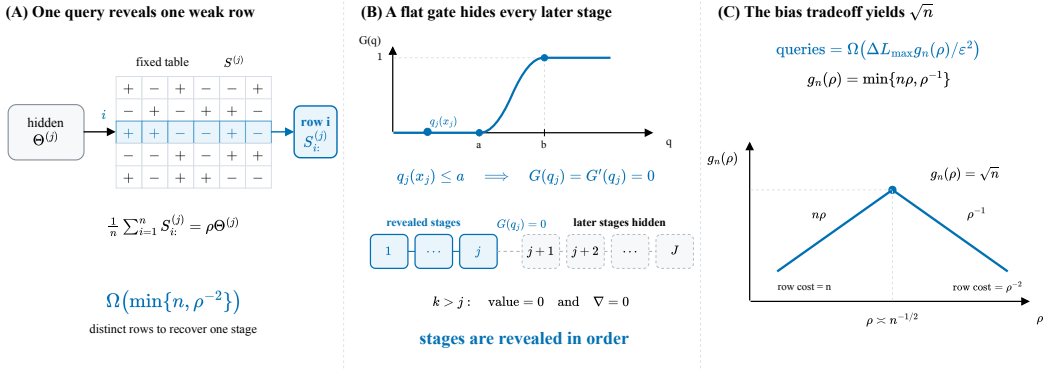

Figure 1: Dense weak hiding. (A) A fixed sign table encodes each stage direction; the blue outline marks the row accessed when component i is queried, and averaging all rows gives $\rho\Theta^{(j)}$. (B) Below a, the link following stage j is zero in value and gradient, so every later stage $k > j$ remains hidden. (C) The product of the row cost per stage and the number of stages is proportional to $g_n(\rho)$, which is maximized at $\rho \asymp n^{-1/2}$ and yields $\sqrt{n}$.

Origin of the $\sqrt{n}$ factor. The one-stage information bound in Section 5 shows that completing one stage requires

$$r_\rho = \Omega\big(\min\{n, \rho^{-2}\}\big) \tag{14}$$

distinct rows, even when their indices are chosen from all previous responses. The nonconvex scaling in Appendix C can take $T_\rho = \Theta(\Delta L_{\max}\rho/\varepsilon^2)$ stages and still preserve the gap and individual smoothness bounds. Therefore

$$r_\rho T_\rho = \Omega\left(\frac{\Delta L_{\max}}{\varepsilon^2}\min\{n\rho, \rho^{-1}\}\right). \tag{15}$$

The two branches balance at $n\rho = \rho^{-1}$, hence $\rho \asymp n^{-1/2}$ and the product is $\Omega(\sqrt{n}\,\Delta L_{\max}/\varepsilon^2)$. Proposition C.6 in Appendix C formalizes this tradeoff for every admissible bias, using the one-stage row cost, sequential revelation, and the gradient bound for the final stage. Appendix A handles the discrete choice of ρ.

5 Information Cost of Revealing Hidden Stages

To align above the gate threshold with a direction that has not been revealed, an algorithm must see many distinct rows. This remains true even when row indices depend on the preceding history. The row costs then add across successive stages. Until stage j becomes current, its hidden pair $(\Theta^{(j)}, S^{(j)})$ is independent of the preceding history.

5.1 One-stage information bound

Cost of one stage. Fix an admissible bias ρ. In a sufficiently large block dimension, Lemma B.1 gives the row threshold

$$r_\rho = \Omega\big(\min\{n, \rho^{-2}\}\big). \tag{16}$$

Before this many distinct rows have been revealed, the probability that any query chosen from the preceding history crosses the gate threshold is at most a fixed constant below $1/2$. Repeated indices and side information independent of the hidden stage do not improve this bound.

Order the distinct indices by their first occurrence. Conditional on the past, exchangeability makes each newly observed row a uniformly chosen remaining row of the fixed table. The mutual information between Θ and the first r such rows is only $O(Dr\rho^2)$, and repeats add nothing. The data processing inequality then bounds the probability of alignment. At $\rho \asymp n^{-1/2}$, this requires $\Omega(n)$ distinct rows.

5.2 Sequential composition across stages

Stage j becomes current after stages $1, \dots, j-1$ have been completed, and it is completed when its alignment first exceeds the gate threshold a. Progress can be faster than this sequential process in two ways: the current stage may finish in unusually few calls, or a query may align with a stage that is not yet current. The one-stage bound controls the first event. The event $\mathcal{E}_{\text{future}}$ covers the second. On $\mathcal{E}^c_{\text{future}}$, unopened future links satisfy $G = G' = 0$, so stages can only be completed in order.

Let ν_N be the number of stages completed within N calls, and let $\mathcal{E}_{\text{future}}$ be the event that a query or the final output crosses the alignment threshold for a stage beyond the current one.

Stage costs add. Assign each call to the stage that is current when the query is made. Suppose every reached stage has conditional probability at most p of finishing in fewer than r assigned calls. If J stages finish within $N \le rJ/2$ calls, at least $J/2$ of them must finish in fewer than r calls. Their expected number is at most pJ, so Markov's inequality gives $\mathbb{P}(\nu_N \ge J) \le 2p$. Together with $\mathbb{P}(\mathcal{E}_{\text{future}}) \le \delta$, Theorem B.5 gives

$$\mathbb{P}(\nu_N < J,\ \mathcal{E}^c_{\text{future}}) \ge 1 - 2p - \delta. \tag{17}$$

On $\mathcal{E}^c_{\text{future}}$, every omitted value and gradient term is zero because its gate satisfies $G = G' = 0$. Appendix B gives the filtration and coupling details.

6 Derivation of the Main Rates

At the optimized bias, one hidden stage costs $\Omega(n)$ distinct rows, and these costs add across the stages, which are revealed in order. The three constructions below specify the number of stages and the error that remains when one of them is unfinished.

6.1 NONCONVEX STATIONARITY

Write the scaled objective as $F(y) = \alpha_\rho R(\beta_\rho y)$. Up to universal constants, individual smoothness, the initial gap, and the gradient left by an unfinished final stage are calibrated by

$$\frac{\alpha_\rho \beta_\rho^2}{\rho} \asymp L_{\max}, \qquad \alpha_\rho T_\rho \asymp \Delta, \qquad \alpha_\rho \beta_\rho \asymp \varepsilon. \tag{18}$$

Solving these relations gives $\alpha_\rho \asymp \varepsilon^2/(L_{\max}\rho)$, $\beta_\rho \asymp L_{\max}\rho/\varepsilon$, and $T_\rho \asymp \Delta L_{\max}\rho/\varepsilon^2$. Appendix C fixes the universal constants so that individual smoothness is at most $L_{\max}$, the initial gap is at most Δ, and the gradient barrier is at least 8ε. At the optimized bias $\rho \asymp n^{-1/2}$, the one-stage row threshold is $r_\rho = \Theta(n)$, and hence $r_\rho T_\rho = \Omega(\sqrt{n}\,\Delta L_{\max}/\varepsilon^2)$. When $T_\rho < 2$, the auxiliary quadratic in Appendix C gives the complementary $\Omega(n)$ term. This proves Theorem 3.1.

6.2 GLOBAL PL: TWO GEOMETRICALLY SCALED CONSTRUCTIONS

Both PL constructions use $\rho \asymp n^{-1/2}$ and hence $r_\rho = \Theta(n)$ rows per stage. Up to universal constants, their scales set Δ, $L_{\max}$, μ, and the remaining objective gap ε as follows:

	Δ	$L_{\max}$	μ	ε
small $\kappa_{\max}$	ζ_1^2/μ	$\zeta_j/(\rho W_j)$	ζ_j/R_j	ζ_{J+1}^2/μ
large $\kappa_{\max}$	$\alpha/(1-r_{\mathrm{w}})$	$\alpha\beta^2/\rho$	$\alpha\beta^2(1-r_{\mathrm{w}})$	Δr_{w}^J

(19)

For small $\kappa_{\max}$, the μ entry is a strong convexity constant. For large $\kappa_{\max}$, it is the PL constant. The last column is the objective gap left by the unrevealed terminal stage or suffix.

Geometric stage scales. Under PL, the gradient at the first unfinished stage must control the objective contribution of every later stage, so the stage scales decrease geometrically. The large $\kappa_{\max}$ construction uses one scale w_j for both the gate and the stage contribution. The first two entries of its row in (19) give $\alpha \asymp \Delta(1-r_{\mathrm{w}})$ and $\beta^2 \asymp L_{\max}\rho/[\Delta(1-r_{\mathrm{w}})]$. Dividing the μ and $L_{\max}$ entries then gives $1 - r_{\mathrm{w}} \asymp 1/(\kappa_{\max}\rho)$. A geometric chain requires $0 < r_{\mathrm{w}} < 1$, so this scaling works only when $\kappa_{\max} \gtrsim \sqrt{n}$. Appendix D makes this concrete by taking $r_{\mathrm{w}} \in [1/2, 1)$. For $3 \le \kappa_{\max} < \sqrt{n}$, the required geometric ratio is no longer admissible. In the small $\kappa_{\max}$ row, the smoothness and curvature entries instead give $R_j/W_j \asymp \kappa_{\max}\rho$. The construction also takes $\vartheta \asymp R_j/W_j$ to control the remaining link curvature. Thus $R_j \ll W_j$ when $\kappa_{\max} \ll \sqrt{n}$. Completion occurs above alignment a. A gate is *fully open* only above b, where $G = 1$ and $G' = 0$.

Small $\kappa_{\max}$: separate gate and contribution scales. The gate scale R_j sets the displacement needed to open the next stage, while the larger contribution scale W_j keeps each component Hessian within the $L_{\max}$ budget. Since $\zeta_{j+1} = \vartheta\zeta_j$, the Δ and ε entries in (19) require $\vartheta^{2J} \asymp \varepsilon/\Delta$. For $3 \le \kappa_{\max} \le \sqrt{n}$, one has $\vartheta \asymp \kappa_{\max}/\sqrt{n}$, and therefore

$$J = \Theta\left(\frac{\log(\Delta/\varepsilon)}{1 + \log(\sqrt{n}/\kappa_{\max})}\right), \qquad r_\rho J = \Theta\left(\frac{n\log(\Delta/\varepsilon)}{1 + \log(\sqrt{n}/\kappa_{\max})}\right). \tag{20}$$

Appendix D verifies that every component is $L_{\max}$-smooth, the average is μ-strongly convex, and an unfinished last stage leaves objective gap at least 8ε. The auxiliary quadratic covers the remaining case, where the target expression is $O(n)$.

Large $\kappa_{\max}$: geometric weights. For a universal constant $c_{\mathrm{w}} > 1$ and $\kappa_{\max} \ge c_{\mathrm{w}}\sqrt{n}$, the weighted construction uses the same scale w_j for the gate and stage contribution. The Δ and ε entries in (19) give $r_{\mathrm{w}}^J \asymp \varepsilon/\Delta$. Together with $1 - r_{\mathrm{w}} \asymp \sqrt{n}/\kappa_{\max}$, this gives

$$J = \Theta\left(\frac{\kappa_{\max}}{\sqrt{n}}\log\frac{\Delta}{\varepsilon}\right), \qquad r_\rho J = \Theta\left(\kappa_{\max}\sqrt{n}\log\frac{\Delta}{\varepsilon}\right). \tag{21}$$

If no stage $k > J$ has been completed, the hidden suffix leaves objective gap at least 8ε. To verify the PL property, let $j_\star$ be the first gate that is not fully open. Appendix D shows that the gradient at this stage controls its scale, while the geometric weights control all remaining stages:

$$\|\nabla R(x)\|^2 \gtrsim w_{j_\star}^2, \qquad \sum_{k \ge j_\star} w_k^2 \lesssim \frac{w_{j_\star}^2}{1 - r_{\mathrm{w}}}. \tag{22}$$

Together with the strong convexity of the preceding fully open blocks, these bounds give the global PL inequality. Global rescaling leaves the condition number unchanged. The two-scale construction covers $\sqrt{n} \le \kappa_{\max} < c_{\mathrm{w}}\sqrt{n}$.

7 Restarted PAGE

What the lower bound suggests. On the small-κ hard instance, revealing a stage costs $\Omega(n)$ calls while the gap shrinks by a factor of $\Theta((\kappa_{\max}/\sqrt{n})^2)$. This leads to PAGE epochs of $\Theta(n)$ calls, each shrinking the gap by $q_{\mathrm{ep}} = \Theta(\kappa_{\mathrm{ms}}/\sqrt{n})$. When $\kappa_{\mathrm{ms}} \ge \sqrt{n}$, longer epochs are needed to obtain a constant contraction factor.

Algorithm 1 implements this schedule. Its expected IFO complexity is

$$\begin{cases} O\left(n + \dfrac{n\log(\Delta/\varepsilon)}{1+\log(\sqrt{n}/\kappa_{\mathrm{ms}})}\right), & 1 \le \kappa_{\mathrm{ms}} \le \sqrt{n}, \\ O\left(n + \kappa_{\mathrm{ms}}\sqrt{n}\log\dfrac{\Delta}{\varepsilon}\right), & \kappa_{\mathrm{ms}} \ge \sqrt{n}. \end{cases} \tag{23}$$

Algorithm 1 Restarted PAGE

Require: $x^{(0)}$, n, a mean-squared smoothness bound L_{ms}, a PL constant μ, a gap bound $F(x^{(0)}) - F^* \le \Delta$, and a target accuracy $\varepsilon \in (0, \Delta)$; $\kappa_{\mathrm{ms}} := L_{\mathrm{ms}}/\mu \ge 1$

$p \leftarrow 1/(n+1)$; $\eta \leftarrow 1/[L_{\mathrm{ms}}(1+\sqrt{n})]$
$A_{\mathrm{ep}} \leftarrow \kappa_{\mathrm{ms}}(1+\sqrt{n})$; $T \leftarrow \lceil 4(n + A_{\mathrm{ep}})\rceil$; $q_{\mathrm{ep}} \leftarrow A_{\mathrm{ep}}/T$
$S_{\mathrm{ep}} \leftarrow \lceil \log(\Delta/\varepsilon)/\log(1/q_{\mathrm{ep}})\rceil$
for $s = 0, \dots, S_{\mathrm{ep}} - 1$ **do**
  $x_0 \leftarrow x^{(s)}$; $g_0 \leftarrow n^{-1}\sum_{i=1}^n \nabla f_i(x_0)$
  for $t = 0, \dots, T-1$ **do**
    $x_{t+1} \leftarrow x_t - \eta g_t$
    if $t < T-1$ **then**
      With probability p, set $g_{t+1} \leftarrow n^{-1}\sum_{i=1}^n \nabla f_i(x_{t+1})$
      Otherwise draw $i_t \sim \mathrm{Unif}[n]$ and set $g_{t+1} \leftarrow g_t + \nabla f_{i_t}(x_{t+1}) - \nabla f_{i_t}(x_t)$
  Draw $U_s \sim \mathrm{Unif}\{0, \dots, T-1\}$ and set $x^{(s+1)} \leftarrow x_{U_s}$
return $x^{(S_{\mathrm{ep}})}$

Lemma E.1 shows that one epoch costs at most $n + 3T$ IFO calls in expectation and that

$$\mathbb{E}\Big[F(x^{(s+1)}) - F^* \mid x^{(s)}\Big] \le q_{\mathrm{ep}}\big(F(x^{(s)}) - F^*\big). \tag{24}$$

When $\kappa_{\mathrm{ms}} \le \sqrt{n}$, each epoch costs $O(n)$, and

$$q_{\mathrm{ep}} = \Theta\left(\frac{\kappa_{\mathrm{ms}}}{\sqrt{n}}\right), \qquad S_{\mathrm{ep}} = O\left(1 + \frac{\log(\Delta/\varepsilon)}{1+\log(\sqrt{n}/\kappa_{\mathrm{ms}})}\right).$$

When $\kappa_{\mathrm{ms}} \ge \sqrt{n}$, each epoch costs $O(\kappa_{\mathrm{ms}}\sqrt{n})$ and shrinks the gap by a constant factor. These two cases give (23), matching the upper bounds in Theorem 3.3. Appendix E proves the contraction and IFO cost; Appendix F converts this expected guarantee to a fixed call budget with any fixed constant success probability.

8 Conclusion and Limitations

We establish tight lower bounds under individual smoothness for nonconvex optimization and for both $\kappa_{\max}$ regimes under PL. Restarted PAGE matches the PL rates, and the small-$\kappa_{\max}$ construction has a globally strongly convex average objective. Open directions include lower bounds in a fixed ambient dimension, noisy finite-sum oracle models, and the range $\kappa < 3$, where the condition number is bounded by a constant. Our lower bounds also assume that n is above a universal threshold.

REFERENCES

Alekh Agarwal and Léon Bottou. A lower bound for the optimization of finite sums. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 78–86, 2015.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51, 2018. URL https://jmlr.org/papers/v18/16-410.html.

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(1–2): 165–214, 2023. doi: 10.1007/s10107-022-01822-7.

Yunyan Bai, Yuxing Liu, and Luo Luo. On the complexity of finite-sum smooth optimization under the Polyak–Łojasiewicz condition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2392–2417, 2024.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1–2):71–120, 2020. doi: 10.1007/s10107-019-01406-y.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points II: First-order methods. *Mathematical Programming*, 185(1–2):315–355, 2021. doi: 10.1007/s10107-019-01431-x.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, volume 27, pp. 1646–1654, 2014.

Nicolas Emmenegger, Rasmus Kyng, and Ahad N. Zehmakan. On the oracle complexity of higher-order smooth non-convex finite-sum optimization. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 10718–10752, 2022. URL https://proceedings.mlr.press/v151/emmenegger22a.html.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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

Jikai Jin. A tight lower bound for smooth nonconvex stochastic optimization with bounded gradient noise. *arXiv preprint arXiv:2608.09004*, 2026. URL https://arxiv.org/abs/2608.09004.

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, 2016.

Ali Kavis, Stratis Skoulakis, Kimon Antonakopoulos, Leello Tadesse Dadi, and Volkan Cevher. Adaptive stochastic variance reduction for non-convex finite-sum minimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23524–23538, 2022.

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical Programming*, 171(1–2):167–215, 2018. doi: 10.1007/s10107-017-1173-0.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6286–6295, 2021.

Zijian Liu, Ta Duy Nguyen, Alina Ene, and Huy L. Nguyen. Adaptive accelerated (extra-)gradient methods with variance reduction. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022.

Nicolas Loizou, Sharan Vaswani, Issam Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for SGD: An adaptive learning rate for fast convergence. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 2021.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Arkadi S. Nemirovski and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2 edition, 2018.

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

Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Blake E. Woodworth and Nathan Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3639–3647, 2016.

Guangzeng Xie, Luo Luo, and Zhihua Zhang. A general analysis framework of lower complexity bounds for finite-sum optimization. *arXiv preprint arXiv:1908.08394*, 2019. URL https://arxiv.org/abs/1908.08394.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer, 1997.

Pengyun Yue, Cong Fang, and Zhouchen Lin. On the lower bound of minimizing Polyak–Łojasiewicz functions. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2948–2968, 2023.

Dongruo Zhou. Sharp first-order lower bounds for higher-order smooth nonconvex optimization. *arXiv preprint arXiv:2606.05438*, 2026. URL https://arxiv.org/abs/2606.05438.

Dongruo Zhou and Quanquan Gu. Lower bounds for smooth nonconvex finite-sum optimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7574–7583, 2019.

AI Use Statement

During the preparation of this manuscript, the authors used large language model assistants to support the development and writing of mathematical proofs and to improve the clarity and readability of the manuscript.

Appendix dependencies. The diagram separates the shared argument for the lower bounds from the nonconvex and PL branches and from the Restarted PAGE analysis. Solid arrows indicate proof dependencies.

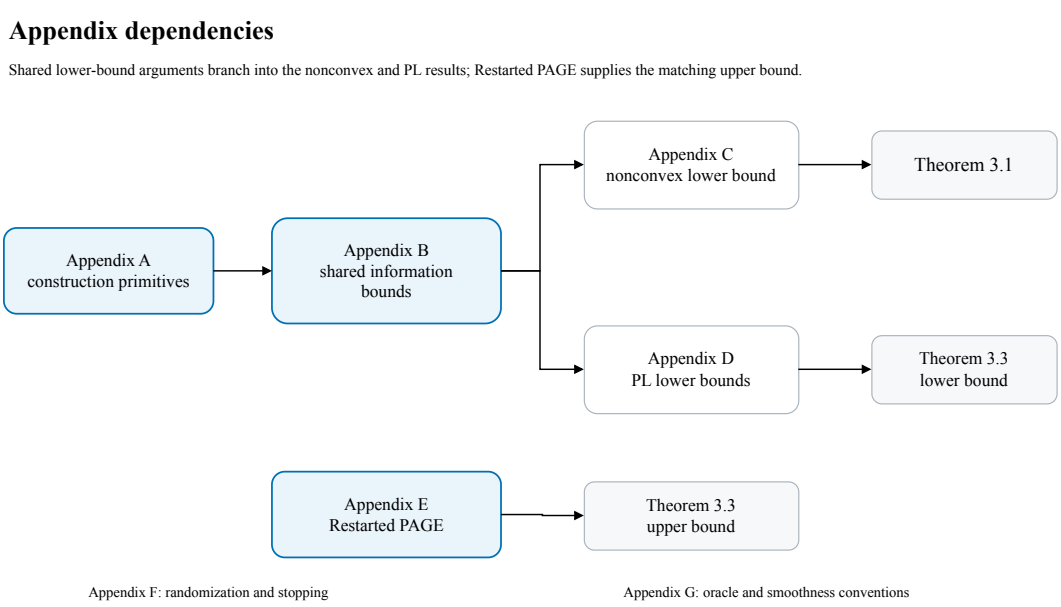

Figure 2: Proof dependencies for the main results. Appendices F and G supply the transfers and conventions used by the branches proving the lower and upper bounds.

A Common Definitions and Calculus

This appendix defines the three objects shared by the lower bounds: a smooth threshold gate, a fixed encoding table, and a bounded radial map. Random objects for stage j carry the superscript (j). The indices $i \in [n]$ and $\ell \in [D]$ refer to a component row and a coordinate.

A.1 Constructions and basic properties

Lemma A.1 (Explicit smooth threshold gate)**.** *Set*

$$a = \frac{5}{32}, \qquad b = \frac{9}{32}. \tag{25}$$

There is an explicit nondecreasing $G \in C^\infty(\mathbb{R})$ with $0 \le G \le 1$ such that

$$G^{(r)}(s) = 0 \quad \text{for } s \le a,\ r \ge 0, \tag{26}$$

$$G(s) = 1, \qquad G^{(r)}(s) = 0 \quad \text{for } s \ge b,\ r \ge 1. \tag{27}$$

Its first two derivatives satisfy the dimension-free bounds

$$\|G'\|_\infty \le 17, \qquad \|G''\|_\infty \le 259. \tag{28}$$

Encoding table and admissible bias grid. Define the admissible bias grid

$$\mathcal{R}_n := \left\{ \frac{m}{n} : m \in \mathbb{Z}_{>0},\ m \equiv n \pmod 2,\ m \le \frac{n}{8} \right\}. \tag{29}$$

Fix any $\rho \in \mathcal{R}_n$; equivalently, choose a positive integer m with the same parity as n and $1 \le m \le n/8$, and set

$$\rho := \frac{m}{n}. \tag{30}$$

The construction and the information bound are uniform over every such choice. Consecutive values of ρ in $\mathcal{R}_n$ are $2/n$ apart, and no relation between m and $\sqrt{n}$ is assumed here.

Before oracle interaction, sample the complete stage pairs independently. For each j, draw $\Theta^{(j)} \sim \mathrm{Unif}\{\pm 1\}^D$ and, conditional on $\Theta^{(j)}$, form the matrix $S^{(j)} = (S_{i\ell}^{(j)})_{i\in[n],\ell\in[D]} \in \{\pm 1\}^{n\times D}$. Its ith row $S_{i:}^{(j)} \in \{\pm 1\}^D$ is stored by component i. For each coordinate ℓ, write its column as $S_\ell^{(j)} = (S_{1\ell}^{(j)}, \dots, S_{n\ell}^{(j)}) \in \{\pm 1\}^n$. Conditional on $\Theta^{(j)}$, draw the columns $S_1^{(j)}, \dots, S_D^{(j)}$ independently and uniformly subject to

$$\mathbf{1}^\top S_\ell^{(j)} = \sum_{i=1}^n S_{i\ell}^{(j)} = m\Theta_\ell^{(j)}. \tag{31}$$

Write the complete stage variable as

$$\mathcal{Z}^{(j)} = (\Theta^{(j)}, S^{(j)}). \tag{32}$$

Lemma A.2 (Properties of the encoding table)**.** *For every stage j:*

(i) *$S^{(j)}$ is sampled once before oracle interaction. Conditional on $\Theta^{(j)}$, its columns are independent, and the pairs $\mathcal{Z}^{(j)} = (\Theta^{(j)}, S^{(j)})$ are independent across stages.*

(ii) *The component rows have the exact average*

$$\frac{1}{n}\sum_{i=1}^n S_{i:}^{(j)} = \rho\Theta^{(j)}. \tag{33}$$

(iii) *After any d distinct rows have been chosen causally, with each new index fixed before its row is observed, the remaining rows are uniformly assigned to the unseen labels subject to the residual column sums.*

Corollary A.3 (Optimized admissible bias)**.** *Let $n \geq 1024$, let m be the least integer not smaller than $\sqrt{n}$ that has the parity of n, and set $\rho = m/n$. Then*

$$\sqrt{n} \leq m < \sqrt{n}+2, \qquad \rho = \frac{m}{n}, \qquad 0 \leq \rho - \frac{1}{\sqrt{n}} < \frac{2}{n}, \qquad \rho < \frac{1}{\sqrt{n}} + \frac{2}{n} \leq \frac{17}{16\sqrt{n}} \leq \frac{17}{512} =: \overline{\rho}. \tag{34}$$

For $g_n(\rho) := \min\{n\rho, \rho^{-1}\}$, this admissible grid point satisfies

$$g_n(\rho) = \rho^{-1} > \frac{\sqrt{n}}{1+2/\sqrt{n}} \geq \frac{16}{17}\sqrt{n}. \tag{35}$$

Bounds for the radial map. For $x \in \mathbb{R}^D$, define

$$\Psi(x) = \frac{x}{\sqrt{1+\|x\|^2}}. \tag{36}$$

For any unit vector u, let

$$q_u(x) = \langle u, \Psi(x)\rangle. \tag{37}$$

Suppressing the stage index, for a hidden direction θ and the rows $S_1, \dots, S_n$ of the preceding encoding table, set

$$u_\theta = \frac{\theta}{\sqrt{D}}, \qquad s_i = \frac{S_i}{\sqrt{D}}, \qquad q_\theta = q_{u_\theta}, \qquad \widetilde{q}_i(x) = \rho^{-1}\langle s_i, \Psi(x)\rangle. \tag{38}$$

Lemma A.4 (Properties of the radial coordinate)**.** *For every unit u, one has $\|\Psi(x)\| < 1$ and, with $r_x = (1+\|x\|^2)^{1/2}$,*

$$\nabla q_u(x) = \frac{u}{r_x} - \frac{\langle u, x\rangle x}{r_x^3}, \qquad \|\nabla q_u(x)\| \leq 1, \tag{39}$$

and

$$\nabla^2 q_u(x) = -\frac{ux^\top + xu^\top + \langle u, x\rangle I}{r_x^3} + \frac{3\langle u, x\rangle xx^\top}{r_x^5}, \qquad \|\nabla^2 q_u(x)\|_{\mathrm{op}} \leq 2. \tag{40}$$

Consequently,

$$|\widetilde{q}_i(x)| \le \frac{1}{\rho}, \qquad \|\nabla \widetilde{q}_i(x)\| \le \frac{1}{\rho}, \qquad \|\nabla^2 \widetilde{q}_i(x)\|_{\mathrm{op}} \le \frac{2}{\rho}. \tag{41}$$

All constants are independent of D. Moreover, the identity for the fixed table gives

$$\frac{1}{n}\sum_{i=1}^{n} \widetilde{q}_i(x) = q_\theta(x) \qquad \forall x. \tag{42}$$

For any normalized stage coordinate z_j used below, $V_j = \Psi(z_j)$ satisfies the unit ball condition in Lemma B.1.

A.2 Proofs of the shared primitives

Proof of Lemma A.1. We use the following explicit construction. Let

$$\delta_{\mathrm{g}} := \frac{1}{512}, \qquad w_{\mathrm{g}} := 1 - 2\delta_{\mathrm{g}} = \frac{255}{256}, \tag{43}$$

and define the triangular density of unit mass

$$p_{\mathrm{g}}(t) := \begin{cases} \dfrac{4(t-\delta_{\mathrm{g}})}{w_{\mathrm{g}}^2}, & \delta_{\mathrm{g}} \le t \le \dfrac{1}{2}, \\ \dfrac{4(1-\delta_{\mathrm{g}}-t)}{w_{\mathrm{g}}^2}, & \dfrac{1}{2} \le t \le 1-\delta_{\mathrm{g}}, \\ 0, & \text{otherwise.} \end{cases} \tag{44}$$

Define the normalized C^∞ bump

$$Z_{\mathrm{g}} := \int_{-\delta_{\mathrm{g}}}^{\delta_{\mathrm{g}}} \exp\!\left(-\frac{1}{\delta_{\mathrm{g}}^2 - u^2}\right) du, \qquad \phi_{\mathrm{g}}(s) := \begin{cases} Z_{\mathrm{g}}^{-1} \exp\!\left(-\dfrac{1}{\delta_{\mathrm{g}}^2 - s^2}\right), & |s| < \delta_{\mathrm{g}}, \\ 0, & |s| \ge \delta_{\mathrm{g}}. \end{cases} \tag{45}$$

With convolution denoted by $*$, set

$$\begin{aligned} g_{\mathrm{g}}(t) &:= (p_{\mathrm{g}} * \phi_{\mathrm{g}})(t), \\ H_0(t) &:= \int_{-\infty}^{t} g_{\mathrm{g}}(u)\, du, \qquad G(s) := H_0\!\left(8s - \frac{5}{4}\right). \end{aligned} \tag{46}$$

Plateaus. The density p_{g} is nonnegative, integrates to one, and satisfies

$$\|p_{\mathrm{g}}\|_\infty = \frac{2}{w_{\mathrm{g}}}, \qquad \mathrm{Lip}(p_{\mathrm{g}}) = \frac{4}{w_{\mathrm{g}}^2}. \tag{47}$$

The function ϕ_{g} is nonnegative, has unit mass, belongs to $C_c^\infty(\mathbb{R})$, and is supported on $[-\delta_{\mathrm{g}}, \delta_{\mathrm{g}}]$. Hence g_{g} is nonnegative, has unit mass, and is supported on $[0,1]$. Thus H_0 is nondecreasing, equals zero on $(-\infty, 0]$, and equals one on $[1, \infty)$.

Derivative bounds. Convolution with a nonnegative kernel of unit mass and $g_{\mathrm{g}}' = p_{\mathrm{g}}' * \phi_{\mathrm{g}}$ give

$$\|H_0'\|_\infty \le \frac{2}{w_{\mathrm{g}}} < \frac{17}{8}, \qquad \|H_0''\|_\infty \le \frac{4}{w_{\mathrm{g}}^2} < \frac{259}{64}. \tag{48}$$

The affine rescaling in (46) gives the two plateaus. It also gives $G'(s) = 8H_0'(8s - 5/4)$ and $G''(s) = 64H_0''(8s - 5/4)$, which yield the stated derivative bounds. On the fully open region, $G = 1$ and only the derivatives of positive order vanish. □

Proof of Lemma A.2. The first claim is part of the sampling rule. Summing each column and using (31) proves the second. For the third, conditioning on revealed positions in a uniformly ordered column with fixed sign counts leaves a uniform ordering of its residual signs; induction applies because each new label is chosen before its row is revealed. □

Proof of Corollary A.3. Consecutive parity-compatible integers differ by two. For $n \geq 1024$, the chosen m also satisfies $m < n/8$. The displayed bounds follow directly from $\sqrt{n} \leq m < \sqrt{n} + 2$ and $\rho \geq n^{-1/2}$. □

Proof of Lemma A.4. Derivative formulas. The formulas follow by differentiation. The Jacobian of Ψ has operator norm at most one. With $s = \|x\|$ and $|\langle u, x\rangle| \leq s$, the Hessian norm is at most

$$3\frac{s}{(1+s^2)^{3/2}} + 3\frac{s^3}{(1+s^2)^{5/2}} < 2. \tag{49}$$

After setting $t = s^2$, the squared inequality is equivalent to

$$4(1+t)^5 - 9t(1+2t)^2 = 4 + 11t + 4t^2 + 4t^3 + 20t^4 + 4t^5 > 0. \tag{50}$$

Since $\|s_i\| = 1$, multiplying the same bounds by $1/\rho$ proves (41); the value bound follows from $\|\Psi(x)\| < 1$. Finally, (33) gives

$$\frac{1}{n}\sum_{i=1}^{n} \widetilde{q}_i(x) = \frac{1}{\rho}\left\langle \frac{1}{n}\sum_{i=1}^{n} \frac{S_i}{\sqrt{D}}, \Psi(x) \right\rangle = \left\langle \frac{\theta}{\sqrt{D}}, \Psi(x) \right\rangle = q_\theta(x),$$

which proves (42). □

A.3 Derivative bounds

The following two lemmas give the local Hessian and radial gradient bounds used in both hard constructions.

Set

$$M_1 := \|G'\|_\infty, \qquad M_2 := \|G''\|_\infty. \tag{51}$$

Let $h : \mathbb{R}^{d_x} \to \mathbb{R}$ and $P : \mathbb{R}^{d_y} \to \mathbb{R}$ be twice continuously differentiable, let $c \in \mathbb{R}$, and suppose

$$|P| \leq B_0, \quad \|\nabla P\| \leq B_1, \quad \|\nabla^2 P\|_{\mathrm{op}} \leq B_2, \qquad \|\nabla h\| \leq g_1, \quad \|\nabla^2 h\|_{\mathrm{op}} \leq g_2. \tag{52}$$

For

$$\mathcal{L}(x, y) = -G(h(x))(P(y) + c), \tag{53}$$

Lemma A.5 (Hessian bounds for a gated link and its block rows)**.** *Under these assumptions, the Hessian blocks satisfy*

$$\|\nabla^2_{yy}\mathcal{L}\|_{\mathrm{op}} \leq B_2, \tag{54}$$

$$\|\nabla^2_{xy}\mathcal{L}\|_{\mathrm{op}} \leq M_1 g_1 B_1, \tag{55}$$

$$\|\nabla^2_{xx}\mathcal{L}\|_{\mathrm{op}} \leq (B_0 + |c|)(M_2 g_1^2 + M_1 g_2). \tag{56}$$

Moreover, for any symmetric block matrix $H = (H_{jk})$,

$$\|H\|_{\mathrm{op}} \leq \max_j \sum_k \|H_{jk}\|_{\mathrm{op}}. \tag{57}$$

Lemma A.6 (Radial gradient lower bound)**.** *Set $\lambda = 33/16$ and*

$$\chi := \frac{(1-b^2)^2 - \lambda b}{\sqrt{1-b^2}} > \frac{5}{18}. \tag{58}$$

For every unit vector u, every $x \in \mathbb{R}^D$, and every $c \geq 1$ such that $q_u(x) \leq b$,

$$\|\lambda x - c\nabla q_u(x)\| \geq \|\lambda x - \nabla q_u(x)\| \geq \chi. \tag{59}$$

The unique stationary alignment for $c = 1$ is a scalar $q_* > b$ satisfying $(1 - q_*^2)^2 - \lambda q_* = 0$.

A.4 Proofs of the derivative bounds

Proof of Lemma A.5. Direct differentiation gives

$$\nabla^2_{yy}\mathcal{L} = -G(h)\nabla^2 P, \qquad \nabla^2_{xy}\mathcal{L} = -G'(h)\nabla h(\nabla P)^\top,$$
$$\nabla^2_{xx}\mathcal{L} = -(P+c)\big[G''(h)\nabla h\nabla h^\top + G'(h)\nabla^2 h\big].$$

The three displayed bounds follow from $0 \le G \le 1$ and the derivative bounds on G.

For the block statement, set $A_{jk} = \|H_{jk}\|_{\mathrm{op}}$. If $z_k = \|v_k\|$, then the vector of block norms of Hv is coordinatewise at most Az, so $\|Hv\| \le \|A\|_2\|v\|$. Symmetry gives

$$\|A\|_2 \le \sqrt{\|A\|_1\|A\|_\infty} = \|A\|_\infty, \tag{60}$$

which is (57). □

Proof of Lemma A.6. Write $x = tu + y$, where $y \perp u$, and set

$$q = q_u(x), \qquad \zeta = (1+\|x\|^2)^{-1}. \tag{61}$$

Lemma A.4 gives

$$\langle x, \nabla q_u(x)\rangle = q\zeta, \qquad \|\nabla q_u(x)\|^2 = \zeta(1-q^2-q^2\zeta), \qquad 0 < \zeta \le 1-q^2. \tag{62}$$

Case 1: $0 \le q \le b$. First expand the dependence on c:

$$\|\lambda x - c\nabla q_u(x)\|^2 = \lambda^2(\zeta^{-1}-1) - 2\lambda c q\zeta + c^2\zeta(1-q^2-q^2\zeta). \tag{63}$$

Its derivative with respect to c is

$$2\zeta\big[-\lambda q + c(1-q^2-q^2\zeta)\big]. \tag{64}$$

Setting this derivative to zero gives the unconstrained minimizer

$$c_0 = \frac{\lambda q}{1-q^2-q^2\zeta} \le \frac{\lambda q}{(1-q^2)^2} < 1. \tag{65}$$

The strict inequality follows because $h(q) := (1-q^2)^2 - \lambda q$ is decreasing on $[0,1]$ and

$$h(b) > \frac{4}{15} > 0. \tag{66}$$

Moreover, $\sqrt{1-b^2} < 24/25$, so the definition of χ gives $\chi > 5/18$.

Thus the minimum over $c \ge 1$ occurs at $c = 1$. For this value,

$$\|\lambda x - \nabla q_u(x)\|^2 = \lambda^2(\zeta^{-1}-1) - 2\lambda q\zeta + \zeta(1-q^2-q^2\zeta). \tag{67}$$

The derivative of the right-hand side is

$$-\frac{\lambda^2}{\zeta^2} - 2\lambda q + (1-q^2) - 2q^2\zeta \le -\lambda^2 + 1 < 0, \tag{68}$$

where $q \ge 0$ and $\zeta \le 1$. Hence the minimum occurs at $\zeta = 1-q^2$, equivalently $y = 0$. There,

$$\|\lambda x - \nabla q_u(x)\| = \frac{(1-q^2)^2 - \lambda q}{\sqrt{1-q^2}}, \tag{69}$$

which decreases on $[0, b]$ and is therefore at least χ.

Case 2: $q \le 0$. Equation (64) is nonnegative for $c \ge 1$, so the squared norm is minimized at $c = 1$. At that value,

$$\begin{aligned}\|\lambda x - \nabla q_u(x)\|^2 &= \lambda^2\|x\|^2 + \|\nabla q_u(x)\|^2 - 2\lambda\langle x, \nabla q_u(x)\rangle \\ &\ge \lambda^2\|x\|^2 + \|\nabla q_u(x)\|^2,\end{aligned} \tag{70}$$

because $\langle x, \nabla q_u(x)\rangle = q\zeta \le 0$.

Moreover, $\nabla q_u(x) = J_\Psi(x)^\top u$, and the smallest singular value of $J_\Psi(x)$ is $(1+\|x\|^2)^{-3/2}$. Since $\|u\| = 1$,

$$\|\nabla q_u(x)\|^2 \ge (1+\|x\|^2)^{-3}. \tag{71}$$

Consequently,

$$\|\lambda x - \nabla q_u(x)\|^2 \ge \lambda^2\|x\|^2 + (1+\|x\|^2)^{-3} \ge 1 > \chi^2. \tag{72}$$

The second inequality uses $1-(1+s)^{-3} \le 3s$ for $s \ge 0$ and $\lambda^2 > 3$. This proves (59). Finally, the stationary equation for $c = 1$ is $h(q_*) = 0$; monotonicity and (66) give the stated uniqueness and $q_* > b$. □

B Dense Weak Hiding: One-Stage Information and Sequential Revelation

The one-stage argument proves that alignment above the gate threshold requires $\Omega(\min\{n, \rho^{-2}\})$ rows, which is $\Omega(n)$ at the optimized bias. The sequential argument then gives an $\Omega(rJ)$ cost for J stages. Lemmas B.1–B.4 prove the row cost and cover the completion query. Lemmas B.7–B.8 and Theorem B.5 give stage independence, coupling, and additivity.

B.1 One-stage information bound

One-stage setting. For this one-stage argument, suppress the stage superscript and write $\Theta = \Theta^{(j)}$ and $S = S^{(j)}$. Write the fixed table as $S = (S_1, \dots, S_D) \in \{\pm 1\}^{n \times D}$, with entries $S_{i\ell}$, column $S_\ell = (S_{1\ell}, \dots, S_{n\ell}) \in \{\pm 1\}^n$, and row $S_{i:} \in \{\pm 1\}^D$. Column S_ℓ encodes the hidden coordinate Θ_ℓ. Throughout the general result, let m be a positive integer with the parity of n and $1 \le m \le n/8$, and set $\rho = m/n$, so $\rho \in \mathcal{R}_n$. The relation $m \asymp \sqrt{n}$ is used only in the optimized result below.

Column sampling under P. The hard distribution is

$$\begin{aligned} P: \quad & \Theta \sim \mathrm{Unif}(\{\pm 1\}^D), \\ & S_\ell \mid \Theta \sim \mathrm{Unif}\big\{s \in \{\pm 1\}^n : \mathbf{1}^\top s = m\Theta_\ell\big\}, \qquad \ell \in [D], \end{aligned} \tag{73}$$

Equivalently, column ℓ contains $(n + m\Theta_\ell)/2$ plus signs and $(n - m\Theta_\ell)/2$ minus signs, uniformly permuted across its rows. The columns are conditionally independent, but entries within a column are not. The table is sampled once before interaction. Querying a row does not resample it.

One-stage protocol and alignment event. Let $W \perp (\Theta, S)$ collect the random tape and all data independent of (Θ, S) that are used in replies before stage completion.

For the information bound, consider the stronger protocol in which every query also reveals its complete encoding row. Let $\mathcal{H}_t$ be the history after response t in this stronger protocol. $\mathcal{H}_0$ includes W and one optional initialization row whose index is fixed from W in advance. At query $t \le B$, $\mathcal{H}_{t-1}$ determines an index $I_t \in [n]$ and an alignment vector V_t with $\|V_t\| \le 1$; repeats reveal no new row. This strengthening is valid because every reply before stage completion is measurable in W, the queries, and the selected rows. The alignment condition is evaluated before the response and is not revealed to the algorithm.

Let K_t count the distinct rows returned before query t, including the initialization row when present. With $a = 5/32$, define

$$E_r = \left\{ \exists t \le B : K_t < r, \quad \frac{1}{\sqrt{D}} \langle \Theta, V_t \rangle > a \right\}. \tag{74}$$

Both one-stage results use the dimension condition

$$D \ge 512 \log(B + 2) + 2048. \tag{75}$$

Lemma B.1 (One-stage row lower bound for a tunable bias)**.** *Let m have the parity of n, let $1 \le m \le n/8$, set $\rho = m/n$, and define*

$$r_\rho := \left\lfloor \frac{1}{4096} \min\{n, \rho^{-2}\} \right\rfloor. \tag{76}$$

If $r_\rho \ge 2$, $B \ge 1$, and D satisfies (75), then

$$P(E_{r_\rho} \mid W) \le \frac{1}{8} \qquad \textit{almost surely}. \tag{77}$$

Lemma B.2 (Optimized one-stage row lower bound)**.** *Choose the least parity-compatible $m \ge \sqrt{n}$. Then $\sqrt{n} \le m < \sqrt{n} + 2$ and $\rho = m/n$. For every $n \ge 1024$, $B \ge 1$, and D satisfying (75), the sharper threshold*

$$r = \left\lfloor \frac{n}{512} \right\rfloor \tag{78}$$

satisfies

$$P(E_r \mid W) \le \frac{1}{8} \qquad \textit{almost surely.} \tag{79}$$

At the optimized bias, this sharpens the tunable threshold to $r = \lfloor n/512 \rfloor$ *while preserving* $P(E_r \mid W) \le 1/8$.

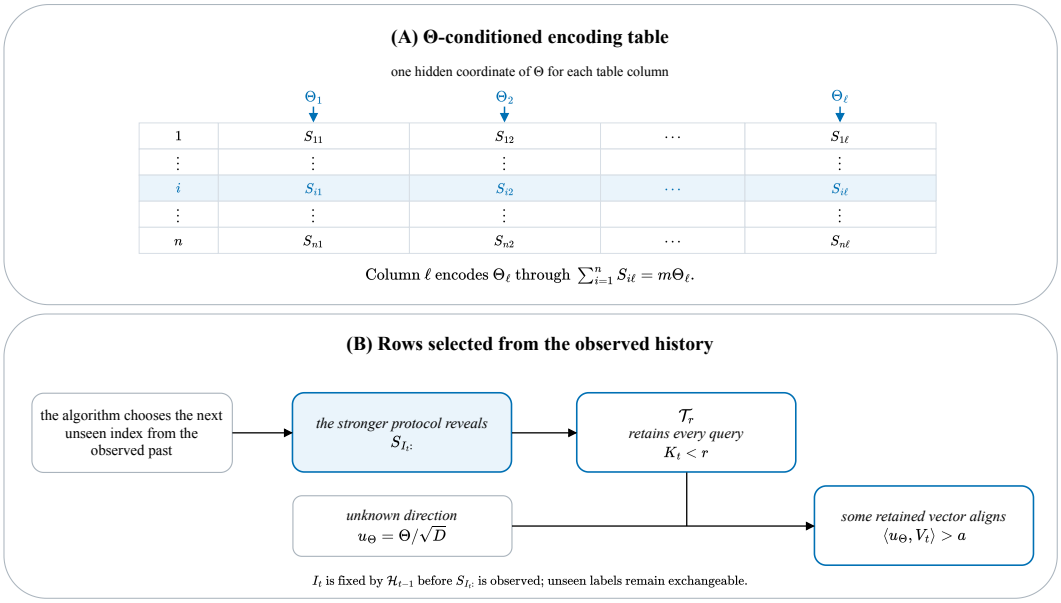

Figure 3: One-stage encoding when row indices are selected from the observed history. Column ℓ of the fixed table encodes Θ_ℓ. In the stronger protocol used for the information bound, a new component index is chosen before its row is observed, so exchangeability determines the conditional law of the next unseen row.

Information under sequential row selection. The one-stage argument has three steps: bound the information in one unseen row, sum over the first r distinct rows, and convert low mutual information into a low probability of early alignment. Throughout the proof of Lemma B.2, fix an admissible value $W = w$. All probabilities and mutual informations in this part are conditional on $W = w$, and we suppress this conditioning. The bounds are uniform in w.

At every query, I_t is fixed by $\mathcal{H}_{t-1}$ before its row is observed. If I_t is new and $K_t = d$, Lemma A.2 therefore makes $S_{I_t:}$ a uniformly chosen remaining row conditional on $(\Theta, \mathcal{H}_{t-1})$. If I_t is repeated, its row is already in the history and contributes no new information.

Let $\mathcal{T}_r$ be this history, including the optional initialization row, stopped immediately after the response that returns the rth distinct row, or after query B if fewer than r rows are returned. Suppose the rth row is returned. The query selecting it is retained and has $K_t = r - 1$, and every later query has $K_t \ge r$. If the rth row is never returned, $\mathcal{T}_r$ contains all B queries. In both cases, $\mathcal{T}_r$ retains exactly every query that can witness E_r, so E_r is measurable with respect to $(\Theta, \mathcal{T}_r)$.

Information in newly revealed rows.

Lemma B.3 (Information in the first r distinct rows)**.** *Let* $1 \le r < n$ *and assume*

$$8m + 9r \le n. \tag{80}$$

For the stopped history $\mathcal{T}_r$*,*

$$\mathrm{I}(\Theta; \mathcal{T}_r) \le \frac{32D}{63} \sum_{d=0}^{r-1} \left(\frac{m}{n-d}\right)^2 \le \frac{32D}{63}\, r \left(\frac{m}{n-r}\right)^2. \tag{81}$$

Proof of Lemma B.3. Information in a newly queried row. Consider a query t retained in $\mathcal{T}_r$ and condition on a realization h of $\mathcal{H}_{t-1}$. Both I_t and K_t are then fixed. If I_t is repeated, its row and the corresponding reply are already determined by h, so the conditional information increment is zero. Suppose instead that I_t is unseen and $K_t = d < r$. For column ℓ, let $\Sigma_{d,\ell}$ be the sum of that coordinate over the d distinct rows already revealed in h. Conditional exchangeability leaves $n - d$ uniformly assigned signs with residual sum $m\Theta_\ell - \Sigma_{d,\ell}$. Hence

$$p^{(d)}_{\theta,\ell} := \mathbb{P}(S_{I_t\ell} = 1 \mid \Theta_\ell = \theta, \mathcal{H}_{t-1} = h) = \frac{n + m\theta - d - \Sigma_{d,\ell}}{2(n-d)}. \tag{82}$$

The optional initialization row obeys the same formula with $d = 0$, because its index is fixed by W before the row is observed.

Expose the coordinates of the new row in the order $\ell = 1, \dots, D$. Conditional independence of the column sequences gives

$$\mathbb{P}(S_{I_t\ell} = 1 \mid \Theta, \mathcal{H}_{t-1} = h, S_{I_t,<\ell}) = p^{(d)}_{\Theta_\ell,\ell}.$$

Thus, under this conditioning, the current sign depends on the hidden direction only through Θ_ℓ. The argument does not require independence among the posterior coordinates of Θ. More explicitly, conditional on $\mathcal{H}_{t-1} = h$ and $S_{I_t,<\ell}$,

$$\Theta \longrightarrow \Theta_\ell \longrightarrow S_{I_t\ell}$$

is a Markov chain. Therefore,

$$\mathrm{I}(\Theta; S_{I_t\ell} \mid \mathcal{H}_{t-1} = h, S_{I_t,<\ell}) = \mathrm{I}(\Theta_\ell; S_{I_t\ell} \mid \mathcal{H}_{t-1} = h, S_{I_t,<\ell}).$$

Condition (80) implies

$$\left| p^{(d)}_{\theta,\ell} - \frac{1}{2} \right| \le \frac{m+d}{2(n-d)} \le \frac{1}{16}.$$

Hence both $p^{(d)}_{+,\ell}$ and $p^{(d)}_{-,\ell}$ lie in $[7/16, 9/16]$. For $Y \sim \mathrm{Bernoulli}(p)$, its Shannon entropy is $H(Y) = h_{\mathrm{b}}(p) := -p\log p - (1-p)\log(1-p)$. Here $\log$ is the natural logarithm, so entropy and mutual information are measured in nats. On the preceding interval,

$$-h''_{\mathrm{b}}(p) = \frac{1}{p(1-p)} \le \frac{256}{63} \qquad (p \in [7/16, 9/16]).$$

Fix also a realization of $S_{I_t,<\ell}$ and write

$$\pi := \mathbb{P}(\Theta_\ell = 1 \mid \mathcal{H}_{t-1} = h, S_{I_t,<\ell}), \qquad \bar{p} := \pi p^{(d)}_{+,\ell} + (1-\pi)p^{(d)}_{-,\ell}.$$

Under this conditioning, $S_{I_t\ell}$ is Bernoulli with parameter $\bar{p}$; after additionally conditioning on $\Theta_\ell = \theta$, its parameter is $p^{(d)}_{\theta,\ell}$. Thus the first conditional entropy below is $h_{\mathrm{b}}(\bar{p})$, while the entropy after conditioning on Θ_ℓ is the posterior average of $h_{\mathrm{b}}(p^{(d)}_{+,\ell})$ and $h_{\mathrm{b}}(p^{(d)}_{-,\ell})$. The identity $\mathrm{I}(X; Y \mid Z) = H(Y \mid Z) - H(Y \mid X, Z)$ gives

$$\begin{aligned}
\mathrm{I}(\Theta; S_{I_t\ell} \mid \mathcal{H}_{t-1} = h, S_{I_t,<\ell}) &= h_{\mathrm{b}}(\bar{p}) - \pi h_{\mathrm{b}}\left(p^{(d)}_{+,\ell}\right) - (1-\pi)h_{\mathrm{b}}\left(p^{(d)}_{-,\ell}\right) \\
&\le \frac{1}{2}\frac{256}{63}\,\pi(1-\pi)\left(p^{(d)}_{+,\ell} - p^{(d)}_{-,\ell}\right)^2 \\
&\le \frac{32}{63}\left(\frac{m}{n-d}\right)^2.
\end{aligned} \tag{83}$$

The first inequality is Taylor's theorem for the concave function h_{b}: its Jensen gap is at most one half of the curvature bound times the variance of the two-point parameter. The last inequality uses

$$\pi(1-\pi) \le \frac{1}{4}, \qquad p^{(d)}_{+,\ell} - p^{(d)}_{-,\ell} = \frac{m}{n-d}.$$

Chain rule along the actual history. Because the coordinate bound is uniform in $S_{I_t,<\ell}$, the chain rule within the new row gives

$$\mathrm{I}(\Theta; S_{I_t:} \mid \mathcal{H}_{t-1} = h) = \sum_{\ell=1}^{D} \mathrm{I}(\Theta; S_{I_t\ell} \mid \mathcal{H}_{t-1} = h, S_{I_t,<\ell}) \leq \frac{32D}{63} \left(\frac{m}{n-d}\right)^2. \tag{84}$$

Pad the stopped history after its stopping time with deterministic null query–response symbols to a total of B rounds. The chain rule then has one increment in conditional mutual information per round. The chosen query is $\mathcal{H}_{t-1}$-measurable and therefore adds no information by itself; a response from a repeated row, or a null response, is already determined by the preceding history and also adds zero. The optional initialization row, when present, is the $d = 0$ increment before the first query. Each response from a new row increases the number of distinct rows, so every $d = 0, \ldots, r-1$ contributes at most once. (Without an initialization row, the first response from a new row corresponds to the $d = 0$ increment.) Applying (84) at these increments gives

$$\mathrm{I}(\Theta; \mathcal{T}_r) \leq \frac{32D}{63} \sum_{d=0}^{r-1} \left(\frac{m}{n-d}\right)^2 \leq \frac{32D}{63}\, r \left(\frac{m}{n-r}\right)^2.$$

Conditioning on $S_{I_t,<\ell}$ in (83) makes posterior independence unnecessary. Finally, $n - d \geq n - r$ for $d \leq r - 1$. This proves (81). □

From row information to early alignment. For any r above, define the product distribution

$$\overline{P}_r := P_\Theta \otimes P_{\mathcal{T}_r}.$$

Under $\overline{P}_r$, every alignment vector belonging to a query retained in $\mathcal{T}_r$ is independent of Θ. Conditional on any retained vector v with $\|v\| \leq 1$, Hoeffding's inequality gives

$$\overline{P}_r\left(\frac{1}{\sqrt{D}}\langle \Theta, v\rangle > a \,\middle|\, v\right) \leq \exp\left(-\frac{a^2 D}{2}\right).$$

There are at most B retained queries, so a union bound gives

$$q_r := \overline{P}_r(E_r) \leq B \exp\left(-\frac{a^2 D}{2}\right) = B \exp\left(-\frac{25D}{2048}\right) < \frac{1}{8}, \tag{85}$$

where the final inequality follows because $D \geq 512 \log(B+2) + 2048$ as required in (75).

We next transfer this bound to the hard distribution. If $q_r = 0$, absolute continuity of $P_{\Theta,\mathcal{T}_r}$ with respect to $P_\Theta \otimes P_{\mathcal{T}_r}$ gives $P(E_r) = 0$. Hence assume below that $q_r > 0$. Since E_r is measurable from $(\Theta, \mathcal{T}_r)$, apply the map $(\Theta, \mathcal{T}_r) \mapsto \mathbf{1}_{E_r}$. The KL divergence between the joint law and the product law equals mutual information, so data processing gives the binary divergence

$$\mathrm{I}(\Theta; \mathcal{T}_r) \geq d_{\mathrm{bin}}(P(E_r) \| q_r).$$

For fixed $q_r < 1/8$, the derivative

$$\frac{\partial}{\partial p} d_{\mathrm{bin}}(p \| q_r) = \log \frac{p(1-q_r)}{q_r(1-p)}$$

is positive for every $p \geq 1/8$. Therefore, if $P(E_r) > 1/8$,

$$\mathrm{I}(\Theta; \mathcal{T}_r) \geq d_{\mathrm{bin}}\left(\frac{1}{8} \middle\| q_r\right) \geq \frac{25}{16384} D - \frac{1}{8} \log B - h_{\mathrm{b}}\left(\frac{1}{8}\right). \tag{86}$$

The last step follows by expanding the binary divergence:

$$d_{\mathrm{bin}}\left(\frac{1}{8} \middle\| q\right) = \frac{1}{8} \log \frac{1}{q} - h_{\mathrm{b}}\left(\frac{1}{8}\right) - \frac{7}{8} \log(1-q) \geq \frac{1}{8} \log \frac{1}{q} - h_{\mathrm{b}}\left(\frac{1}{8}\right),$$

and then substituting (85).

Proof of Lemma B.1. Because $r_\rho \geq 2$ and the coefficient in (76) is $1/4096$,

$$n \geq 8192, \qquad \rho \leq \frac{1}{\sqrt{8192}} < \frac{1}{32}, \qquad r_\rho \leq \frac{n}{4096}.$$

Thus $8m + 9r_\rho < n/4 + 9n/4096 < n$, so Lemma B.3 applies. Moreover, $r_\rho \rho^2 \leq 1/4096$, and hence

$$\mathrm{I}(\Theta; \mathcal{T}_{r_\rho}) \leq \frac{32D}{63 \cdot 4096} \left(\frac{4096}{4095}\right)^2 = \frac{131072}{1056448575} D < \frac{D}{8000}. \tag{87}$$

Suppose, for contradiction, that $P(E_{r_\rho}) > 1/8$. Then (86) gives

$$\mathrm{I}(\Theta; \mathcal{T}_{r_\rho}) \geq \frac{25D}{16384} - \frac{1}{8} \log B - h_{\mathrm{b}}\left(\frac{1}{8}\right).$$

The numerical margins satisfy

$$\frac{25}{16384} - \frac{1}{8000} > \frac{1}{1024}$$

and

$$\frac{D}{1024} \geq \frac{1}{2} \log(B+2) + 2 > \frac{1}{8} \log B + h_{\mathrm{b}}\left(\frac{1}{8}\right),$$

Together these inequalities make the lower bound contradict (87). Therefore $P(E_{r_\rho}) \leq 1/8$. □

Proof of Lemma B.2. Assume $n \geq 1024$ and use the optimized choice $m < \sqrt{n} + 2 \leq (17/16)\sqrt{n}$. With $r = \lfloor n/512 \rfloor$,

$$m \leq \frac{17n}{512}, \qquad 8m + 9r < \frac{145n}{512} < n.$$

Lemma B.3 gives

$$\mathrm{I}(\Theta; \mathcal{T}_r) \leq \frac{32D}{63} \frac{n}{512} \left[\frac{(17/16)\sqrt{n}}{n(511/512)}\right]^2 = \frac{18496}{16450623} D. \tag{88}$$

Suppose, for contradiction, that $P(E_r) > 1/8$. Then

$$\mathrm{I}(\Theta; \mathcal{T}_r) \geq \frac{25D}{16384} - \frac{1}{8} \log B - h_{\mathrm{b}}\left(\frac{1}{8}\right).$$

The coefficient margin satisfies

$$\frac{25}{16384} - \frac{18496}{16450623} > \frac{1}{2500},$$

whereas

$$\frac{D}{2500} \geq \frac{512}{2500} \log(B+2) + \frac{2048}{2500} > \frac{1}{8} \log B + h_{\mathrm{b}}\left(\frac{1}{8}\right).$$

This is a contradiction. Thus $P(E_r) \leq 1/8$, completing the proof of Lemma B.2. □

Lemma B.4 (Coupling up to the first stage completion)**.** *For an independent (Θ, S), let $W \perp (\Theta, S)$ contain the history before initialization, an initialization index fixed in advance, the random tape, and all other data needed for replies. Suppose each IFO reply before completion, including its value and all gradient blocks, has the form*

$$Y_t^{\mathrm{pre}} = \Phi_t(W, (I_s, X_s)_{s \leq t}, (S_{I_s:})_{s \leq t}), \tag{89}$$

where Φ_t does not depend on the alignment condition and reveals neither Θ nor data from the next stage. At query t, suppose completion requires $D^{-1/2}\langle \Theta, V_t \rangle > a$ before the response, with V_t fixed by the preceding history and query. A length-B continuation can then be coupled so that

$$\{\exists t \leq B : \text{ completion at } t,\ K_t < r\} \subseteq E_r \tag{90}$$

and the histories and queries agree through the first completion query. Consequently, whenever $P(E_r \mid W) \leq p$ almost surely,

$$P(\exists t \leq B : \text{ completion at } t,\ K_t < r \mid W) \leq p \qquad \text{almost surely}. \tag{91}$$

Proof. Construct a B-query continuation that remains at the same stage. After alignment, keep using the same maps Φ_t and the same fixed table, and set $V_t = 0$ after the rth distinct row is returned. Couple the two continuations using the same W, stage data, and optional initialization row.

Until stage completion or the return of the rth distinct row, the coupled interactions have the same history before each query and hence choose the same query; every earlier reply that does not complete the stage agrees by (89). If completion occurs with $K_t < r$, that query is made before the rth distinct row is returned and therefore belongs to E_r. The completion event is determined before that response, so the response is not needed for the inclusion. □

B.2 Sequential revelation

Process and notation. At stage j, write

$$\mathrm{align}_j(x_j) := \frac{1}{\sqrt{D}}\langle \Theta^{(j)}, V_j(x_j)\rangle, \qquad \|V_j(x_j)\| \le 1. \tag{92}$$

This is q_j for the constant-scale chain, u_j for the weighted chain, and h_j for the two-scale chain. Stage j completes when $\mathrm{align}_j > a$ first holds, checked before the response. Assign each IFO call to the unique stage that is current when its query is made. The response completing stage $j-1$ may also include one stage-j initialization row, but this costs no additional call and is not assigned again. Let ν_t be the number of stages completed after response t, with $\nu_0 = 0$. The revelation process advances by at most one stage per response; premature alignment at a later stage is included in the event below. For an N-query interaction, define

$$\begin{aligned}\mathcal{E}_{\mathrm{future}} := \{\exists t \le N,\ k > \nu_{t-1} + 1 : \mathrm{align}_k(X_{t,k}) > a\} \\ \cup \{\exists k > \nu_N + 1 : \mathrm{align}_k(\widehat{X}_k) > a\}.\end{aligned} \tag{93}$$

After stage j is completed, we reveal $\Theta^{(j)}$ and one stage-$(j+1)$ initialization row as part of the *augmented interaction*, without charging another IFO call. The *truncated interaction* retains this revealed information but omits links beyond the completed stages and the current stage. Crossing a leaves the region where $G = G' = 0$; full opening occurs only at b.

Theorem B.5 (Sequential revelation: row costs add)**.** *Fix the first J stages, a budget of N IFO calls, and constants $r \ge 2$, $p \in [0, 1/2)$, and $\delta \in [0, 1-2p)$. Assume:*

(S1) conditional on the history immediately before its initialization row is returned, or on the initial history for stage 1, every stage once it becomes current completes after fewer than r calls assigned to it with probability at most p;

(S2) the event $\mathcal{E}_{\mathrm{future}}$ satisfies $\mathbb{P}(\mathcal{E}_{\mathrm{future}}) \le \delta$; and

(S3) on $\mathcal{E}^c_{\mathrm{future}}$, the augmented and truncated interactions have identical histories and final outputs.

If $N \le rJ/2$, then

$$\mathbb{P}(\nu_N \ge J) \le 2p, \qquad \mathbb{P}(\nu_N < J,\ \mathcal{E}^c_{\mathrm{future}}) \ge 1 - 2p - \delta. \tag{94}$$

On the second event, the output of the original IFO interaction satisfies $\mathrm{align}_k(\widehat{X}_k) \le a$ for every $k > J$.

Conditions for applications to radial chains. The optimized corollary below requires:

(R1) The pairs $\mathcal{Z}^{(j)} = (\Theta^{(j)}, S^{(j)})$ are independent samples from P with the least parity-compatible $m \ge \sqrt{n}$ and $a = 5/32$.

(R2) Immediately before an optional initialization row is returned, let W contain its index (fixed in advance), the history, seed, and data from completed stages used in replies. Then $W \perp \mathcal{Z}^{(j)}$, and every full IFO reply that does not complete the stage has the form

$$\Phi_t\Big(W, (I_s, X_s)_{s\le t}, (S^{(j)}_{I_s:})_{s\le t}\Big),$$

including its value and all gradient blocks. The map does not depend on the alignment condition and reveals neither $\Theta^{(j)}$ nor data from later stages.

(R3) Completion uses (92). If the current stage and all stages not yet current have alignment at most a, every omitted link has $G = G' = 0$. The final output is included without an oracle call.

(R4) Assign each call to the stage current when its query is made. A completion response may include the next initialization row without an additional IFO call and is assigned only to the stage it completes. Before the qth call assigned to stage j, at most its initialization row and $q-1$ response rows have been seen.

Corollary B.6 (Sequential revelation for the optimized radial chains)**.** *Consider a radial chain satisfying* (R1)–(R4) *above. Let* $n \geq 1024$, $1 \leq J < M_{\mathrm{tot}}$, *and set*

$$r = \left\lfloor \frac{n}{512} \right\rfloor. \tag{95}$$

If $N \geq 1$, $N \leq rJ/2$, *and*

$$D \geq \max\left\{2048[1+\log(N+3)], \quad \frac{2048}{25}\log\big(128(N+1)M_{\mathrm{tot}}\big)\right\}, \tag{96}$$

then

$$\mathbb{P}(\nu_N \geq J) \leq \frac{1}{4}, \qquad \mathbb{P}(\nu_N < J,\ \mathcal{E}^c_{\mathrm{future}}) \geq \frac{95}{128}. \tag{97}$$

On the second event, the original IFO output satisfies $\mathrm{align}_k(\widehat{X}_k) \leq a$ *for every* $k > J$.

The lemmas below establish stage independence and exact coupling; the counting argument in the next subsection then proves the corollary.

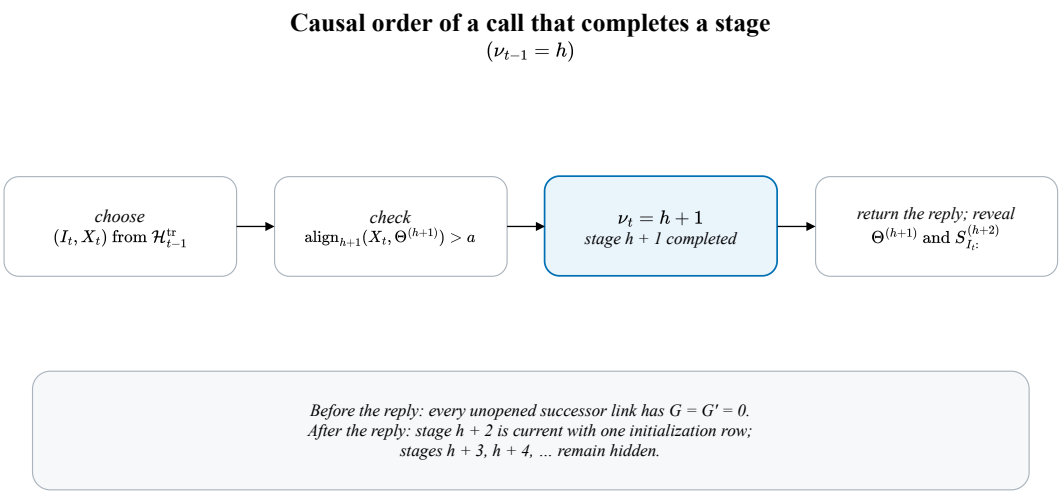

Figure 4: The completion condition is evaluated on a query fixed from the preceding history. Only afterward can the augmented interaction reveal the completed direction and an initialization row for the next stage.

The complete history of the truncated interaction after response t is denoted by $\mathcal{H}^{\mathrm{tr}}_t$ and includes the information returned with completion responses and the private seed ξ. Including the output in $\mathcal{E}_{\mathrm{future}}$ reveals nothing to the algorithm and leaves ν_N unchanged.

Lemma B.7 (Independence before stage revelation and control of future alignment)**.** *Under* (R1)–(R4)*, the history of the truncated interaction before a stage's initialization row is independent of that stage. Thus a stage, once current, completes after fewer than* r *calls assigned to it with conditional probability at most* p *whenever the one-stage bound has parameters* (r, p)*. At query* t *with* $\nu_{t-1} = h$*, every* $\Theta^{(k)}$*,* $k > h+1$*, is independent of* $X_{t,k}$*; the same holds for* $\widehat{X}_k$*. Consequently, for a chain of* M_{tot} *stages,*

$$D \geq \frac{2048}{25}\log\big(128(N+1)M_{\mathrm{tot}}\big) \quad \Longrightarrow \quad \mathbb{P}(\mathcal{E}_{\mathrm{future}}) \leq \frac{1}{128}. \tag{98}$$

This joint probability bound covers every query paired with every future stage, as well as the final output. Replacing 128 *above by* $1/\delta_{\mathrm{f}}$ *gives any target* $\delta_{\mathrm{f}} \in (0,1)$*, with only logarithmic growth in* D.

Lemma B.8 (The augmented and truncated interactions agree exactly)**.** *Under* (R1)–(R4)*, for the same fixed instance and seed, on* $\mathcal{E}^c_{\mathrm{future}}$ *the two interactions have identical component values, all gradient blocks, revealed data, queries, and final outputs, pathwise. Here* $\mathcal{E}_{\mathrm{future}}$ *is evaluated in the truncated interaction.*

Proof of Lemma B.7. Independence before revelation. Maintain the filtration invariant

$$\{\nu_t = g\} \quad \Longrightarrow \quad \sigma(\mathcal{H}_t^{\mathrm{tr}}) \subseteq \sigma(\mathcal{Z}^{(1)}, \dots, \mathcal{Z}^{(g+1)}, \xi). \tag{99}$$

If $\nu_{t-1} = h$, the query uses only $\mathcal{Z}^{(1)}, \dots, \mathcal{Z}^{(h+1)}$ and ξ. A reply that does not complete the current stage reveals nothing from the next stage; a completion reply reveals only one initialization row from $\mathcal{Z}^{(h+2)}$. This proves the invariant and leaves every $\mathcal{Z}^{(k)}$, $k > \nu_t + 1$, independent of the observed history.

Conditional one-stage bound. For $j \geq 2$, condition on the history immediately before its initialization row is returned, including the fixed query, row index, seed, and reply data from completed stages. The invariant gives $W \perp \mathcal{Z}^{(j)}$, while $G = G' = 0$ below the lower threshold puts the full reply in (89) with rows $S^{(j)}_{I_s:}$. Lemma B.4 therefore applies; for $j = 1$, use the initial data and no initialization row.

Because alignment is checked before the response, completion on the qth call assigned to that stage gives

$$K_t \leq q - 1 + \mathbf{1}\{j \geq 2\}, \qquad q < r \Longrightarrow K_t < r.$$

Thus completion after fewer than r assigned calls is contained in E_r and has conditional probability at most $\mathbb{P}(E_r \mid W) \leq p$.

Each direction at a stage not yet current is independent of its query or output block. A Rademacher tail bound over at most $(N+1)M_{\mathrm{tot}}$ pairs (t, k) gives

$$\mathbb{P}(\mathcal{E}_{\mathrm{future}}) \leq (N+1)M_{\mathrm{tot}} e^{-a^2 D/2} \leq \frac{1}{128}.$$

□

Proof of Lemma B.8. We argue by induction, starting from a common history before the query with progress h. If the stage does not complete, $\mathrm{align}_{h+1} \leq a$; after completion, $\mathcal{E}^c_{\mathrm{future}}$ also controls the new successor. Every omitted link therefore has $G = G' = 0$, so its scalar value and every gradient contribution vanish. The replies, revealed data, histories, and next queries agree, and the common seed gives the same output. Completed links remain in both interactions, even if a completed coordinate later falls below b; the output term in $\mathcal{E}^c_{\mathrm{future}}$ is used only to control the final output. □

B.3 Counting calls across stages

Each IFO call is assigned to the stage that is current when its query is made. A completion response may supply the next initialization row without an additional IFO call, but the same call is assigned only to the stage it completes. Hence calls assigned to different stages are disjoint.

Proof of Theorem B.5. Let Z count the stages among $1, \dots, J$ that are reached and finish in fewer than r calls assigned to them. Conditioning when stage j is reached, *(S1)* and the tower property give $\mathbb{E}Z \leq pJ$.

If all first J stages are completed by $N \leq rJ/2$ calls, then necessarily $Z \geq J/2$: otherwise more than $J/2$ stages would each use at least r calls, exceeding the budget. Thus Markov's inequality, without assuming independence across stages, gives

$$\mathbb{P}(\nu_N \geq J) \leq \mathbb{P}\left(Z \geq \frac{J}{2}\right) \leq 2p. \tag{100}$$

The assumed bound on $\mathcal{E}_{\mathrm{future}}$ and a union bound yield

$$\mathbb{P}\big(\nu_N < J,\ \mathcal{E}^c_{\mathrm{future}}\big) \geq 1 - 2p - \delta. \tag{101}$$

On this event, *(S3)* makes the auxiliary histories and outputs identical; the augmented interaction only adds information, so the conclusion transfers to the original IFO.

It remains to check the final output. On the joint event, let $h = \nu_N < J$. Every $k > J$ also satisfies $k > h + 1$; the output term in the definition of $\mathcal{E}_{\text{future}}$ therefore gives $\text{align}_k(\widehat{X}_k) \le a$. □

Proof of Corollary B.6. The first dimension term, Lemma B.2, and Lemma B.4 give $p = 1/8$ for every stage once it becomes current, because completion in $q < r$ calls assigned to that stage uses fewer than r distinct rows, including one optional initialization row. The second dimension requirement and Lemma B.7 give $\delta = 1/128$ jointly for all N queries and the final output. The condition $G = G' = 0$ for future links invokes Lemma B.8; the disjoint call assignments verify the remaining hypothesis of Theorem B.5, which now yields (97) and the claimed bound on the original IFO output. □

B.4 Auxiliary quadratic for the additive n term

This subsection proves the standalone $\Omega(n)$ IFO lower bound that gives the additive n term in Theorems 3.1 and 3.3. The sequential argument above uses the optimized bias. The separate additive n term uses instead the smallest positive parity-compatible bias and the fixed dimension $D_0 = 4$.

Lemma B.9 (Auxiliary quadratic lower bound)**.** *Fix $n \ge n_0 = 1024$ and $0 \le N \le \lfloor n/4 \rfloor$. Set*

$$m_0 := \begin{cases} 1, & n \text{ odd}, \\ 2, & n \text{ even}, \end{cases} \qquad \rho_0 := \frac{m_0}{n}, \qquad D_0 := 4. \tag{102}$$

Draw (Θ, S) from P in (73) with $(m, D) = (m_0, D_0)$, and write $u = \Theta/\sqrt{D_0}$ and $s_i = S_{i:}/\sqrt{D_0}$. For any $\gamma, \tau > 0$, define

$$f_i^{\text{quad}}(x) := \frac{\gamma}{2}\|x\|^2 - \frac{\tau}{\rho_0}\langle s_i, x\rangle, \qquad F_{\text{quad}}(x) := \frac{1}{n}\sum_{i=1}^{n} f_i^{\text{quad}}(x) = \frac{\gamma}{2}\|x\|^2 - \tau\langle u, x\rangle. \tag{103}$$

Every randomized IFO algorithm making at most N exact calls and returning $\widehat{X}$ satisfies

$$\mathbb{P}\big(\|\gamma\widehat{X} - \tau u\| \le \tau/4\big) \le \frac{5}{16}. \tag{104}$$

Proof. Each response reveals the selected row because

$$\nabla f_i^{\text{quad}}(x) = \gamma x - \frac{\tau}{\rho_0} s_i, \qquad s_i = \frac{\rho_0}{\tau}\big(\gamma x - \nabla f_i^{\text{quad}}(x)\big), \tag{105}$$

and, given x, the row determines both the value and gradient.

We next bound the information in all revealed rows. Let $\mathcal{T}_N^{\text{quad}}$ be the complete interaction history, including the private seed ξ, and let $\mathcal{H}_{t-1}^{\text{quad}}$ be the history before query t. The display above shows that every response determines its selected encoding row. Conditional on $\mathcal{H}_{t-1}^{\text{quad}}$, the index I_t is fixed; querying a previously selected component reveals no new encoding row. When a new row is queried with $K_t = d$, let $\Sigma_{d,\ell}$ be the sum in column ℓ over the d distinct rows already revealed. Conditional exchangeability gives

$$p_{\theta,\ell}^{(d)} = \frac{n + m_0\theta - d - \Sigma_{d,\ell}}{2(n-d)}. \tag{106}$$

Since $m_0 \le 2$, $d < N \le n/4$, and $n \ge 1024$,

$$\left| p_{\theta,\ell}^{(d)} - \frac{1}{2} \right| \le \frac{m_0 + d}{2(n-d)} \le \frac{3}{16}. \tag{107}$$

Thus the conditional Bernoulli parameters lie in $[5/16, 11/16]$, where $-h_{\text{b}}'' \le 256/55$. The entropy identity and Taylor bound used in (83) now give, for each coordinate,

$$\text{I}(\Theta; S_{I_t\ell} \mid \mathcal{H}_{t-1}^{\text{quad}}, S_{I_t,<\ell}) \le \frac{1}{2}\frac{256}{55}\frac{1}{4}\left(\frac{m_0}{n-d}\right)^2 = \frac{32}{55}\left(\frac{m_0}{n-d}\right)^2.$$

Summing over the D_0 coordinates yields

$$\mathrm{I}(\Theta; S_{I_t:} \mid \mathcal{H}_{t-1}^{\mathrm{quad}}) \le \frac{32 D_0}{55} \left(\frac{m_0}{n-d} \right)^2 . \tag{108}$$

Since $\xi \perp (\Theta, S)$, queries are predictable from their histories, and each value of d occurs at most once when a new row is queried, the chain rule for mutual information gives

$$\mathrm{I}(\Theta; \mathcal{T}_N^{\mathrm{quad}}) \le \frac{32 D_0 N}{55} \left(\frac{m_0}{n-N} \right)^2 \le \frac{2048}{495 n} \le \frac{2}{495} < \frac{1}{200}. \tag{109}$$

Finally, the 16 vectors $u_\theta = \theta/2$ have pairwise distance at least one. The nearest-codeword decoder

$$\widehat{\Theta}_{\mathrm{dec}} \in \operatorname*{argmin}_{\vartheta \in \{\pm 1\}^{D_0}} \left\| \frac{\gamma \widehat{X}}{\tau} - \frac{\vartheta}{2} \right\| \tag{110}$$

is therefore correct whenever $\|\gamma \widehat{X}/\tau - u_\Theta\| \le 1/4$. Fano's inequality (Yu, 1997) and $\log 2 > 1/2$ give

$$\mathbb{P}(\widehat{\Theta}_{\mathrm{dec}} = \Theta) \le \frac{\mathrm{I}(\Theta; \mathcal{T}_N^{\mathrm{quad}}) + \log 2}{\log 16} < \frac{1}{4} + \frac{1}{400} < \frac{5}{16}, \tag{111}$$

which proves (104). □

C Nonconvex First-Order Lower Bound

Restatement of Theorem 3.1. There are universal constants $c_0, c_1 > 0$ such that the following holds. Fix $n \ge 1024$, $L_{\max}, \Delta > 0$, and $0 < \varepsilon^2 \le c_0 \Delta L_{\max}$. For every randomized IFO algorithm making at most

$$N \le c_1 \left(n + \sqrt{n} \, \frac{\Delta L_{\max}}{\varepsilon^2} \right) \tag{112}$$

calls, there exist a dimension d and a finite sum $F \in \mathcal{F}_{\mathrm{ind}}^{n,d}(\Delta, L_{\max})$ such that

$$\Pr_{\xi} \Big(\|\nabla F(\widehat{X})\| \ge 4\varepsilon \Big) \ge \frac{11}{16}. \tag{113}$$

The algorithm may choose arbitrary queries from its observed history and return any output measurable with respect to that history.

C.1 Construction

Let

$$\mathcal{R}_n := \left\{ \frac{m}{n} : m \in \mathbb{Z}_{>0}, \quad m \equiv n \pmod 2, \quad m \le \frac{n}{8} \right\}. \tag{114}$$

Fix $\rho = m/n \in \mathcal{R}_n$, $T \ge 1$, and write $x = (x_1, \dots, x_T)$ with $x_j \in \mathbb{R}^D$. At stage j, let $\Theta^{(j)} \in \{\pm 1\}^D$ be the hidden direction and let $S^{(j)} \in \{\pm 1\}^{n \times D}$ be its fixed encoding table with bias $\rho = m/n$; row $S_{i:}^{(j)}$ belongs to component i. The table satisfies $n^{-1} \sum_{i=1}^n S_{i:}^{(j)} = \rho \Theta^{(j)}$. For $z \in \mathbb{R}^D$, define

$$\Psi(z) = \frac{z}{\sqrt{1 + \|z\|^2}}, \qquad q_j(z) = \frac{1}{\sqrt{D}} \langle \Theta^{(j)}, \Psi(z) \rangle, \qquad \widetilde{q}_{ij}(z) = \frac{1}{\rho} \left\langle \frac{S_{i:}^{(j)}}{\sqrt{D}}, \Psi(z) \right\rangle. \tag{115}$$

Here q_j is the scalar alignment with the stage-j hidden direction, and $\widetilde{q}_{ij}$ is the corresponding component term determined by the selected row. The displayed table average gives $n^{-1} \sum_{i=1}^n \widetilde{q}_{ij}(z) = q_j(z)$ for every z. Fix

$$C = 1, \qquad \lambda = \frac{33}{16}, \tag{116}$$

and use the gate thresholds

$$a = \frac{5}{32}, \qquad b = \frac{9}{32}. \tag{117}$$

Thus $s \le a$ implies $G(s) = G'(s) = 0$, whereas $s \ge b$ implies $G(s) = 1$ and $G'(s) = 0$. In the link from stage $j-1$ to stage j, q_{j-1} is the gate coordinate and $\widetilde{q}_{ij}$ is component i's stage contribution. Define

$$\begin{aligned} R_i(x) = & -\widetilde{q}_{i1}(x_1) - \sum_{j=2}^{T} G(q_{j-1}(x_{j-1}))\big(\widetilde{q}_{ij}(x_j) + C\big) \\ & + \frac{\lambda}{2}\sum_{j=1}^{T}\|x_j\|^2. \end{aligned} \tag{118}$$

Averaging over components and using $n^{-1}\sum_i \widetilde{q}_{ij} = q_j$ gives

$$\begin{aligned} R(x) := \frac{1}{n}\sum_{i=1}^{n} R_i(x) = & -q_1(x_1) - \sum_{j=2}^{T} G(q_{j-1}(x_{j-1}))(q_j(x_j) + C) \\ & + \frac{\lambda}{2}\sum_{j=1}^{T}\|x_j\|^2. \end{aligned} \tag{119}$$

For progress index $h \in \{0, \dots, T\}$, let $R_i^{[h]}$ be (118) with the sum of chain links truncated at $\min\{h+1, T\}$.

Set

$$M_1 := \|G'\|_\infty, \qquad M_2 := \|G''\|_\infty, \qquad \mathsf{H}(\rho) := \lambda + 2 + (1 + C\rho)(M_2 + 2M_1) + 2M_1. \tag{120}$$

Proposition C.1 (Properties of the nonconvex radial chain)**.** *Let $\alpha, \beta > 0$. Define $f_i(y) := \alpha R_i(\beta y)$ and $F(y) := n^{-1}\sum_i f_i(y) = \alpha R(\beta y)$. Then:*

(i) *Each component satisfies*

$$\mathrm{Lip}(\nabla f_i) \le \frac{\alpha\beta^2 \mathsf{H}(\rho)}{\rho}. \tag{121}$$

(ii) *With $C_0 := C + 1 = 2$,*

$$F(0) - \inf F \le \alpha C_0 T. \tag{122}$$

(iii) *For the constant $\chi > 0$ in Lemma A.6,*

$$\min_j q_j(\beta y_j) < b \quad \Longrightarrow \quad \|\nabla F(y)\| \ge \alpha\beta\chi. \tag{123}$$

(iv) *For $0 \le h < T$, a response that does not complete stage $h+1$ uses only $S_{i:}^{(h+1)}$ from that stage, while the response that completes it may also use $S_{i:}^{(h+2)}$ when $h+2 \le T$. For any $0 \le h \le T$, if every omitted gate input is at most a, then*

$$\big(R_i(x), \nabla R_i(x)\big) = \big(R_i^{[h]}(x), \nabla R_i^{[h]}(x)\big). \tag{124}$$

C.2 Initial gap and individual smoothness

Lemma C.2 (Initial gap)**.** *With $C_0 = C + 1$,*

$$R(0) = 0, \qquad R(0) - \inf_x R(x) \le C_0 T. \tag{125}$$

Proof. At the origin all $q_j = 0$ and $G(0) = 0$, so $R(0) = 0$. Since $|q_j| < 1$, $0 \le G \le 1$, and the quadratic is nonnegative, $R(x) \ge -1 - (T-1)(C+1) \ge -C_0 T$. □

Lemma C.3 (Individual smoothness)**.** *For every $\rho \in \mathcal{R}_n$,*

$$\sup_x \|\nabla^2 R_i(x)\|_{\mathrm{op}} \le \frac{\mathsf{H}(\rho)}{\rho} \qquad \textit{for every } i. \tag{126}$$

For the least parity-compatible $m \geq \sqrt{n}$, one has $\rho = m/n < 17/(16\sqrt{n}) \leq \overline{\rho} := 17/512$, and

$$H_{\mathrm{main}} := 341. \tag{127}$$

Then every component satisfies

$$\sup_x \|\nabla^2 R_i(x)\|_{\mathrm{op}} \leq \frac{H_{\mathrm{main}}}{\rho}. \tag{128}$$

The bound is independent of T, D, and n, except through $1/\rho$.

Proof. One chain link. Consider one chain link

$$A_{ij}(x_{j-1}, x_j) = G(q_{j-1}(x_{j-1}))\big(\widetilde{q}_{ij}(x_j) + C\big). \tag{129}$$

Apply Lemma A.5 to $-A_{ij}$ with

$$(B_0, B_1, B_2) = \frac{1}{\rho}(1, 1, 2), \qquad (g_1, g_2) = (1, 2), \qquad c = C. \tag{130}$$

By Lemma A.1,

$$M_1 \leq 17, \qquad M_2 \leq 259. \tag{131}$$

It gives, respectively, the current diagonal, cross, and previous diagonal block bounds

$$\frac{2}{\rho}, \qquad \frac{M_1}{\rho}, \qquad \frac{(1 + C\rho)(M_2 + 2M_1)}{\rho}. \tag{132}$$

For the last bound we used $1/\rho + C = (1 + C\rho)/\rho$, not the looser factor $(C + 1)/\rho$.

Global block row. The first component term has Hessian norm at most $2/\rho$, and the quadratic contributes λI. A global block row contains at most one current diagonal term, one diagonal term induced by its successor, and two adjacent cross terms. Therefore the block row clause of Lemma A.5 gives

$$\begin{aligned} \max_j \sum_k \|\nabla^2_{x_j x_k} R_i(x)\|_{\mathrm{op}} &\leq \lambda + \frac{2}{\rho} + \frac{(1 + C\rho)(M_2 + 2M_1)}{\rho} + \frac{2M_1}{\rho} \\ &\leq \frac{\mathsf{H}(\rho)}{\rho}. \end{aligned} \tag{133}$$

The second line uses $\rho \leq 1$ and the definition of $\mathsf{H}(\rho)$. This proves (126) with no factor depending on T or D. For the optimized bias, $\rho \leq \overline{\rho}$ and substitution of the displayed derivative bounds gives $\mathsf{H}(\rho) \leq H_{\mathrm{main}}$, proving (128). □

C.3 Gradient lower bound and scaling

Lemma C.4 (Nonconvex gradient lower bound at a constant scale)**.** *Let $\chi > 0$ be the constant in Lemma A.6. Then, for the average objective R in (119),*

$$\min_{1 \leq j \leq T} q_j(x_j) < b \quad \Longrightarrow \quad \|\nabla R(x)\| \geq \chi. \tag{134}$$

Proof. Let j be the first stage with $q_j(x_j) < b$. All previous gates equal one. The jth gradient block is

$$\nabla_{x_j} R(x) = \lambda x_j - \gamma_j \nabla q_j(x_j), \qquad \gamma_j \geq 1. \tag{135}$$

For $j < T$, the successor chain link adds $G'(q_j)(q_{j+1} + C) \geq 0$ to γ_j; for $j = T$, $\gamma_j = 1$. Positivity uses $C = 1$ and $q_{j+1} > -1$. Lemma A.6, applied with $u = u_{\Theta^{(j)}}$ and $c = \gamma_j$, now gives $\|\nabla_{x_j} R(x)\| \geq \chi$, proving the claim. □

Proof of Proposition C.1. Item (i) follows from Lemma C.3 and the chain rule. Items (ii) and (iii) follow from Lemmas C.2 and C.4, respectively, after the same rescaling. For item (iv), inspection of (118) shows that the current stage contribution uses only $S_{i:}^{(h+1)}$ and that a newly active successor term uses only $S_{i:}^{(h+2)}$, when that stage exists. Since $G = G' = 0$ below the lower threshold, the displayed equality holds: every omitted value term contains G, and every omitted gradient term contains G or G'. □

Tunable scaling. For every $\rho \in \mathcal{R}_n$, set

$$H_{\mathrm{tun}} := 368, \tag{136}$$

and, for a sufficiently small universal $c_T > 0$, define

$$\beta_\rho = \frac{L_{\max}\rho\chi}{8H_{\mathrm{tun}}\varepsilon}, \qquad \alpha_\rho = \frac{64H_{\mathrm{tun}}\varepsilon^2}{L_{\max}\rho\chi^2}, \qquad T_\rho = \left\lfloor c_T \frac{\Delta L_{\max}\rho}{\varepsilon^2} \right\rfloor. \tag{137}$$

Lemma C.5 (Analytic bounds for a tunable bias)**.** *With the preceding definitions, H_{tun} is universal and*

$$\sup_x \|\nabla^2 R_i(x)\|_{\mathrm{op}} \le \frac{H_{\mathrm{tun}}}{\rho} \qquad \textit{for every } i. \tag{138}$$

Moreover, define $f_i(y) := \alpha_\rho R_i(\beta_\rho y)$ and $F(y) := n^{-1}\sum_{i=1}^n f_i(y) = \alpha_\rho R(\beta_\rho y)$. Then f_i is $L_{\max}$-smooth for every i, $F(0) - \inf F \le \Delta$, and $\min_j q_j(\beta_\rho y_j) < b$ implies $\|\nabla F(y)\| \ge 8\varepsilon$. In particular, $T_\rho = \Theta(\Delta L_{\max}\rho/\varepsilon^2)$ whenever $T_\rho \ge 2$.

Proof. The proof of Lemma C.3 based on block row sums uses the bias only through $1 + C\rho$. Since $\rho \le 1/8$, the bounds $M_1 \le 17$ and $M_2 \le 259$ give $\mathsf{H}(\rho) \le \mathsf{H}(1/8) < 368 = H_{\mathrm{tun}}$, which proves (138).

Smoothness. For $f_i(y) = \alpha_\rho R_i(\beta_\rho y)$, the chain rule and (138) give

$$\|\nabla^2 f_i(y)\|_{\mathrm{op}} = \alpha_\rho\beta_\rho^2\|\nabla^2 R_i(\beta_\rho y)\|_{\mathrm{op}} \le \frac{\alpha_\rho\beta_\rho^2 H_{\mathrm{tun}}}{\rho} = L_{\max}. \tag{139}$$

Gradient lower bound. If $\min_j q_j(\beta_\rho y_j) < b$, Lemma C.4 gives $\|\nabla R(\beta_\rho y)\| \ge \chi$. Hence

$$\|\nabla F(y)\| = \alpha_\rho\beta_\rho\|\nabla R(\beta_\rho y)\| \ge \alpha_\rho\beta_\rho\chi = 8\varepsilon. \tag{140}$$

Initial gap and stage count. Lemma C.2 and the definition of T_ρ imply

$$F(0) - \inf F \le \alpha_\rho C_0 T_\rho \le \frac{64C_0H_{\mathrm{tun}}c_T}{\chi^2}\,\Delta \le \Delta, \tag{141}$$

after choosing $c_T \le \chi^2/(64C_0H_{\mathrm{tun}})$. Finally, when $T_\rho \ge 2$, the floor in (137) loses at most a factor of two, proving $T_\rho = \Theta(\Delta L_{\max}\rho/\varepsilon^2)$. □

Proposition C.6 (Bias–smoothness tradeoff)**.** *There are universal constants $c > 0$ and $p_\star < 1/2$ such that, for every $n \ge n_0$, $L_{\max}, \Delta, \varepsilon > 0$, and admissible ρ for which the integer row threshold r_ρ and stage count T_ρ are both at least two, every randomized IFO algorithm using at most*

$$c\frac{\Delta L_{\max}}{\varepsilon^2}\min\{n\rho, \rho^{-1}\}$$

calls has, in some dimension d, an instance $F_\rho \in \mathcal{F}_{\mathrm{ind}}^{n,d}(\Delta, L_{\max})$ for which $\Pr_\xi(\|\nabla F_\rho(\widehat{X})\| \le \varepsilon) \le p_\star$.

Proof. We keep ρ arbitrary rather than immediately selecting $\rho \asymp n^{-1/2}$. The corresponding row threshold and number of stages are

$$r_\rho = \left\lfloor \frac{1}{4096}\min\{n, \rho^{-2}\} \right\rfloor, \qquad T_\rho = \left\lfloor c_T \frac{\Delta L_{\max}\rho}{\varepsilon^2} \right\rfloor. \tag{142}$$

Assume $r_\rho, T_\rho \ge 2$ and choose the block dimension large enough for the one-stage bound and the bound on $\mathcal{E}_{\mathrm{future}}$, equivalently

$$D \ge c_D\left[1 + \log\big((N+2)(T_\rho+1)\big)\right] \tag{143}$$

for a universal c_D.

Assembly. Lemma B.1 gives the row cost r_ρ, and Lemma C.5 gives T_ρ individually smooth hidden stages and an 8ε gradient lower bound at the final stage. Apply Theorem B.5 to the first $J = T_\rho - 1$ stages. Its joint event has probability at least $95/128$ whenever $N \le r_\rho(T_\rho - 1)/2$, and on that

event the gradient lower bound at the final stage applies. Since $r_\rho, T_\rho \geq 2$, the floor losses are only universal factors and

$$r_\rho(T_\rho - 1) = \Omega\left(\frac{\Delta L_{\max}}{\varepsilon^2} \min\{n\rho, \rho^{-1}\}\right). \tag{144}$$

Consequently, a sufficiently small universal choice of c makes the call budget in the proposition at most $r_\rho(T_\rho - 1)/2$.

The conversion in Section F.1 now gives the following statement for a fixed instance of the construction at accuracy ε:

$$\Pr_\xi\left(\|\nabla F_\rho(\widehat{X})\| \leq 4\varepsilon\right) \leq \frac{33}{128} < \frac{1}{2}. \tag{145}$$

Since the event with threshold ε is contained in the event with threshold 4ε, this proves the proposition with $p_\star = 33/128$. □

C.4 Optimized bias

When the optimized stage count is at least two, the construction yields the $\Omega(\sqrt{n}\,\Delta L_{\max}/\varepsilon^2)$ lower bound, with the last stage reserved for the algorithm's output.

Here m is the least parity-compatible integer not smaller than $\sqrt{n}$ and

$$\sqrt{n} \leq m < \sqrt{n} + 2, \qquad \rho = \frac{m}{n}, \qquad \frac{1}{\sqrt{n}} \leq \rho < \frac{17}{16\sqrt{n}}. \tag{146}$$

The constants used below are $H_{\text{main}} = 341$ from Lemma C.3, $\chi > 0$ from Lemma A.6, $C_0 = C + 1 = 2$, and the one-stage row threshold $r = \lfloor n/512 \rfloor$.

Set

$$f_i(y) = \alpha R_i(\beta y), \qquad F(y) := \frac{1}{n}\sum_{i=1}^{n} f_i(y) = \alpha R(\beta y), \tag{147}$$

where

$$\beta = \frac{L_{\max}\rho\chi}{8H_{\text{main}}\varepsilon}, \qquad \alpha = \frac{64H_{\text{main}}\varepsilon^2}{L_{\max}\rho\chi^2}. \tag{148}$$

Then

$$\alpha\beta^2\frac{H_{\text{main}}}{\rho} = L_{\max}, \qquad \alpha\beta\chi = 8\varepsilon. \tag{149}$$

Let

$$T = \left\lfloor c_T \frac{\Delta L_{\max}\rho}{\varepsilon^2} \right\rfloor, \tag{150}$$

where c_T is small enough that $\alpha C_0 T \leq \Delta/2$. Proposition C.1 and (149) give

$$\max_i \operatorname{Lip}(\nabla f_i) \leq L_{\max}, \qquad F(0) - F^* \leq \Delta, \qquad \min_j q_j(\beta y_j) < b \Longrightarrow \|\nabla F(y)\| \geq 8\varepsilon. \tag{151}$$

Assume $T \geq 2$; the case $T < 2$ is handled next. Reserve the last block by setting

$$J := T - 1. \tag{152}$$

With $V_j(y_j) = \Psi(\beta y_j)$, Proposition C.1, item (iv), gives *(R2)–(R3)*. Independent stage tables give *(R1)*, and the rule for assigning calls, including the free initialization row convention, gives *(R4)*.

Choose the integer D so that

$$D \geq \max\left\{2048[1 + \log(N+3)], \quad \frac{2048}{25}\log\big(128(N+1)T\big)\right\}. \tag{153}$$

The case $N = 0$ is immediate, so assume $N \geq 1$. The corollary, with $M_{\text{tot}} = T$, now gives

$$\mathbb{P}\big(\nu_N < T - 1,\ \mathcal{E}^c_{\text{future}}\big) \geq \frac{95}{128} \quad \text{whenever} \quad N \leq \frac{r(T-1)}{2}. \tag{154}$$

On this event, Corollary B.6 gives $q_T(\beta\widehat{X}_T) \le a < b$. Therefore (151) yields

$$\|\nabla F(\widehat{X})\| \ge 8\varepsilon. \tag{155}$$

Since $T-1 \ge T/2$ and $n\rho = m = \Theta(\sqrt{n})$, the query threshold in (154) is $\Omega(\sqrt{n}\,\Delta L_{\max}/\varepsilon^2)$. The reserved last stage is what makes the conclusion valid for an output that was never queried.

Finally, $D = O(\log((N+2)(T+2)))$ under the displayed requirement. Since $d = TD$, substituting (150) and (154) gives

$$d = O\left[\left(1 + \frac{\Delta L_{\max}}{\sqrt{n}\,\varepsilon^2}\right)\log\left(2 + n + \sqrt{n}\,\frac{\Delta L_{\max}}{\varepsilon^2}\right)\right]. \tag{156}$$

C.5 The linear sample-size term

When $T < 2$, one auxiliary quadratic gives the full bound.

Because the optimized bias satisfies $\rho = \Theta(n^{-1/2})$, the condition $T \ge 2$ in (150) is equivalent, up to universal constants, to

$$\varepsilon^2 \lesssim \frac{\Delta L_{\max}}{\sqrt{n}} \quad\Longleftrightarrow\quad \sqrt{n}\,\frac{\Delta L_{\max}}{\varepsilon^2} \gtrsim n. \tag{157}$$

Thus the radial chain applies in the regime where its second complexity term is at least a constant multiple of n.

For the case $T < 2$, instantiate the auxiliary quadratic already defined in Lemma B.9 with

$$\tau := 16\varepsilon, \qquad \gamma := \frac{\tau^2}{2\Delta} = \frac{128\varepsilon^2}{\Delta}. \tag{158}$$

Its components f_i^{quad} have Hessian γI, and its average obeys

$$F_{\mathrm{quad}}(0) - F^*_{\mathrm{quad}} = \frac{\tau^2}{2\gamma} = \Delta, \qquad \nabla F_{\mathrm{quad}}(x) = \gamma x - \tau u. \tag{159}$$

Under the standing assumption $\varepsilon^2 \le c_0 \Delta L_{\max}$, choose the theorem's universal constant so that $c_0 \le 1/128$. Then $\gamma \le L_{\max}$, so every component is $L_{\max}$-smooth. The same lemma gives, after reducing the universal budget constant so that $N \le \lfloor n/4 \rfloor$,

$$\mathbb{P}(\|\nabla F_{\mathrm{quad}}(\widehat{X})\| \le 4\varepsilon) \le \frac{5}{16}. \tag{160}$$

In the complementary regime, $\sqrt{n}\,\Delta L_{\max}/\varepsilon^2 = O(n)$. Hence the target budget $n + \sqrt{n}\,\Delta L_{\max}/\varepsilon^2$ is $O(n)$, so (160) gives the full bound after reducing constants. This quadratic uses $D_0 = 4$ in the coarse regime, within the dimension bound in (156). Combining the two regimes proves the fixed budget claim in Theorem 3.1.

D Proofs for the Global PL Bounds

Restatement of the lower bound in Theorem 3.3. Fix $n \ge 1024$, $L_{\max}, \mu, \Delta > 0$, and $\kappa_{\max} = L_{\max}/\mu \ge 3$. There are universal constants $c, c_\varepsilon > 0$ such that, whenever $0 < \varepsilon \le c_\varepsilon \Delta$, every randomized IFO algorithm using at most

$$N \le c \begin{cases} n + \dfrac{n\log(\Delta/\varepsilon)}{1 + \log(\sqrt{n}/\kappa_{\max})}, & 3 \le \kappa_{\max} \le \sqrt{n}, \\ n + \kappa_{\max}\sqrt{n}\log\dfrac{\Delta}{\varepsilon}, & \kappa_{\max} \ge \sqrt{n} \end{cases} \tag{161}$$

calls has a fixed instance $F \in \mathcal{F}^{n,d}_{\mathrm{ind,PL}}(\Delta, L_{\max}, \mu)$, in some dimension d, for which

$$\Pr_{\xi}\Big(F(\widehat{X}) - F^* > \varepsilon\Big) \ge \frac{11}{16}. \tag{162}$$

Because the hard instances are individually smooth, the same lower bound also holds for the mean-squared class. Appendix F gives the conversion to a fixed hard instance and the extension to expected call budgets. The two-scale chain proves the first regime and the bounded transition interval. The geometrically weighted chain proves the large-$\kappa_{\max}$ regime. We analyze the weighted construction first because its block decomposition of the gap isolates the global PL argument. The two-scale construction then treats the range in which the weighted geometric ratio is not admissible.

Motivation for geometric stage scales. Let $j_\star$ be the first unfinished stage. With equal stage scales, the total contribution of stages $j_\star, \dots, M$ grows like $M - j_\star + 1$, while the available lower bound on the squared gradient norm remains constant. Hence the PL constant would shrink as the chain gets longer. Geometric scales instead give

$$\sum_{k=j_\star}^{M} w_k^2 = O(w_{j_\star}^2), \quad \|\nabla F(x)\|^2 = \Omega(w_{j_\star}^2).$$

Thus the first unfinished stage controls the total contribution of all remaining stages without forcing μ to decrease with the chain length. The two constructions differ in how quickly the stages may shrink and still keep every component $L_{\max}$-smooth.

Scale choices at a glance. Let r_{w} be the ratio of consecutive squared weights, ϑ the ratio of consecutive amplitudes, and $\bar{\kappa}_{\max} := \min\{\kappa_{\max}, \sqrt{n}\}$. At the optimized bias, the parameter choices that determine the two rates are

$$\rho \asymp n^{-1/2}, \qquad 1 - r_{\mathrm{w}} \asymp \frac{\sqrt{n}}{\kappa_{\max}}, \qquad \vartheta \asymp \frac{R_j}{W_j} \asymp \frac{\bar{\kappa}_{\max}}{\sqrt{n}}.$$

The first relation makes one hidden stage cost $\Theta(n)$ rows. The second is the admissible geometry using one scale in the large-κ branch; the third is the separated geometry used below the transition. The explicit numerical constants introduced later serve only to verify individual smoothness, the PL inequality, and admissibility. They do not affect these scale relations or the final rates.

Shared encoding. All staged constructions use the smooth threshold gate and independent fixed tables $S^{(j)} \in \{\pm 1\}^{n \times D}$ sampled conditional on $\Theta^{(j)} \in \{\pm 1\}^D$. Fix $\rho = m/n \in \mathcal{R}_n$ as in Appendix A. The row $S_{i:}^{(j)}$ belongs to component i and $n^{-1} \sum_i S_{i:}^{(j)} = \rho \Theta^{(j)}$. We use

$$\Psi(z) := \frac{z}{\sqrt{1 + \|z\|^2}}, \qquad u_{\Theta^{(j)}} := \frac{\Theta^{(j)}}{\sqrt{D}},$$
$$q_{\Theta^{(j)}}(z) := \frac{1}{\sqrt{D}} \langle \Theta^{(j)}, \Psi(z) \rangle, \quad q_j(z) := q_{\Theta^{(j)}}(z),$$
$$\widetilde{q}_{ij}(z) := \frac{1}{\rho} \left\langle \frac{S_{i:}^{(j)}}{\sqrt{D}}, \Psi(z) \right\rangle.$$

Thus q_j is the alignment coordinate with the hidden direction at stage j, $\widetilde{q}_{ij}$ is the component term determined by row i, and $n^{-1} \sum_i \widetilde{q}_{ij} = q_j$ exactly.

Throughout this appendix, $a = 5/32$ and $b = 9/32$ are the gate thresholds: $s \le a$ implies $G(s) = G'(s) = 0$, whereas $s \ge b$ implies $G(s) = 1$ and $G'(s) = 0$.

All three constructions used for the lower bounds share the same fixed encoding table. The nonconvex chain has a constant stage scale, the large-κ PL chain uses one geometric scale w_j per stage, and the small-κ PL chain separates the gate scale R_j from the contribution scale W_j.

Part I: weighted chain for the large-κ branch. For the large-κ branch, we use one geometric scale w_j for both the gate and the stage contribution. We say that a gate is *fully open* when $u_j \ge b$, so $G(u_j) = 1$ and $G'(u_j) = 0$. We reserve *stage completion* for the information threshold $u_j > a$ used in Appendix B.

D.1 Weighted construction

Fix an integer $M \ge 2$ and positive nonincreasing weights

$$1 = w_1 \ge w_2 \ge \cdots \ge w_M > 0. \tag{163}$$

For any stage j, the remaining squared weight is $\sum_{k=j}^{M} w_k^2$. Its largest ratio to the squared weight of the current stage is

$$\mathcal{T}(w) := \max_{1 \le j \le M} \frac{\sum_{k=j}^{M} w_k^2}{w_j^2}. \tag{164}$$

For each stage define

$$u_j(x_j) = q_{\Theta^{(j)}}(x_j/w_j), \qquad Q_j(x_j) = w_j^2 u_j(x_j), \tag{165}$$

and

$$\widetilde{Q}_{ij}(x_j) = \frac{w_j^2}{\rho}\left\langle \frac{S_{i:}^{(j)}}{\sqrt{D}}, \Psi(x_j/w_j)\right\rangle. \tag{166}$$

Then $n^{-1}\sum_i \widetilde{Q}_{ij} = Q_j$ exactly. Here u_j is the gate coordinate, $\widetilde{Q}_{ij}$ is the stage contribution in component i, and Q_j is its component average. In the link from stage $j-1$ to stage j, the factor $G(u_{j-1})$ decides whether the current stage contribution is present. The condition $G = G' = 0$ makes this link and its gradients in both adjacent blocks vanish whenever $u_{j-1} \le a$. Fix

$$C = 1, \qquad \lambda = \frac{33}{16}, \qquad \sigma := \lambda - 2 = \frac{1}{16}, \tag{167}$$

The component hard instance is

$$\begin{aligned} R_i(x) = & -\widetilde{Q}_{i1}(x_1) - \sum_{j=2}^{M} G(u_{j-1}(x_{j-1}))\big(\widetilde{Q}_{ij}(x_j) + Cw_j^2\big) \\ & + \frac{\lambda}{2}\sum_{j=1}^{M}\|x_j\|^2, \end{aligned} \tag{168}$$

and its average is $R := n^{-1}\sum_{i=1}^n R_i$; equivalently, averaging the displayed formula replaces every $\widetilde{Q}_{ij}$ by Q_j.

Proposition D.1 (Properties of the weighted chain)**.** *For $M \ge 2$, $\rho \in \mathcal{R}_n$ with $0 < \rho \le 17/512$, and $1 = w_1 \ge \cdots \ge w_M > 0$, there are universal constants $H_{\mathrm{main}}, C_{\mathrm{PL}}^{\mathrm{gap}} > 0$ such that:*

(i) *Every component is H_{main}/ρ-smooth.*

(ii) *The initial gap satisfies*

$$R(0) - R^* \le 2\mathcal{T}(w). \tag{169}$$

(iii) *The average satisfies*

$$R(x) - R^* \le C_{\mathrm{PL}}^{\mathrm{gap}}\mathcal{T}(w)\|\nabla R(x)\|^2. \tag{170}$$

Equivalently, it is PL with constant $\big(2C_{\mathrm{PL}}^{\mathrm{gap}}\mathcal{T}(w)\big)^{-1}$.

(iv) *If $\mathcal{I} \subseteq \{2,\dots,M\}$ and $G(u_{j-1}(x_{j-1})) = 0$ for every $j \in \mathcal{I}$, then*

$$R(x) - R^* \ge \sum_{j\in\mathcal{I}} w_j^2. \tag{171}$$

The offset Cw_j^2 carries no hidden direction. After averaging, it makes $Q_j + Cw_j^2 = w_j^2(u_j + C) > 0$. In the decomposition below, a link whose predecessor coefficient is zero then leaves a fixed positive residual.

Lemma A.4 gives

$$\begin{aligned} \|\nabla Q_j\| &\le w_j, & \|\nabla^2 Q_j\|_{\mathrm{op}} &\le 2, \\ \|\nabla \widetilde{Q}_{ij}\| &\le \frac{w_j}{\rho}, & \|\nabla^2 \widetilde{Q}_{ij}\|_{\mathrm{op}} &\le \frac{2}{\rho}, \\ \|\nabla u_j\| &\le \frac{1}{w_j}, & \|\nabla^2 u_j\|_{\mathrm{op}} &\le \frac{2}{w_j^2}. \end{aligned} \tag{172}$$

D.2 Decomposition of the gap by blocks

Define the predecessor coefficient

$$c_1 := 1, \qquad c_j := G(u_{j-1}) \quad (2 \le j \le M). \tag{173}$$

The potential for one block,

$$\phi_j(z) := \frac{\lambda}{2}\|z\|^2 - q_{\Theta^{(j)}}(z) \tag{174}$$

is σ-strongly convex by (172). Its minimum does not depend on j or $\Theta^{(j)}$ by rotational symmetry. Write

$$d_* := -\min_z \phi_j(z). \tag{175}$$

The minimizer is $z_j^* = t_* u_{\Theta^{(j)}}$, where $\lambda t_* = (1+t_*^2)^{-3/2}$. Its alignment $q_* = t_*/\sqrt{1+t_*^2}$ satisfies $q_* > b$ by (66). Moreover, $0 < d_* < 1$: the directional derivative of $\phi_j(t u_{\Theta^{(j)}})$ at zero is negative, while at its finite minimizer z_j^* one has $\phi_j(z_j^*) \geq -q_{\Theta^{(j)}}(z_j^*) > -1$.

For $c \in [0,1]$, define the block gap by the exact mixture

$$e_{j,c}(z) := c\big(\phi_j(z) + d_*\big) + (1-c)\left(\frac{\lambda}{2}\|z\|^2 + 1 + d_*\right) \geq 0. \tag{176}$$

Thus $e_{j,c}$ interpolates between two nonnegative terms. Expanding the mixture gives the equivalent formula

$$e_{j,c}(z) = \frac{\lambda}{2}\|z\|^2 - c\big(q_{\Theta^{(j)}}(z) + 1\big) + 1 + d_*. \tag{177}$$

At $c = 1$, $e_{j,1} = \phi_j + d_*$ is the fully open, strongly convex block gap; at $c = 0$, $e_{j,0}(z) = \lambda\|z\|^2/2 + 1 + d_* \geq 1$ is a fixed residual. Since the minimizer of the first gap has alignment $q_* > b$, all block gaps can vanish simultaneously, while every closed gate contributes objective error.

Lemma D.2 (Exact decomposition of the gap by blocks and consequences)**.** *For every x, with $z_j = x_j/w_j$ and c_j from (173),*

$$R(x) - R^* = \sum_{j=1}^{M} w_j^2 e_{j,c_j}(z_j), \tag{178}$$

where

$$R^* = -d_* w_1^2 - (1+d_*)\sum_{j=2}^{M} w_j^2. \tag{179}$$

At the origin,

$$R(0) - R^* = d_* w_1^2 + (1+d_*)\sum_{j=2}^{M} w_j^2 \leq 2\mathcal{T}(w). \tag{180}$$

If $c_j = 0$ for every j in an index set $\mathcal{I}$, then

$$R(x) - R^* \geq \sum_{j\in\mathcal{I}} w_j^2. \tag{181}$$

Proof. The first stage contribution in (168) has no offset, whereas every later stage contribution has offset w_j^2. Consequently,

$$R(x) = w_1^2 + \sum_{j=1}^{M} w_j^2 \left[\frac{\lambda}{2}\|z_j\|^2 - c_j(q_{\Theta^{(j)}}(z_j) + 1)\right]. \tag{182}$$

Substituting (177) into this identity gives (178) with the value in (179). Every summand is nonnegative by (176). Conversely, setting $z_j = z_j^*$ for every j makes every predecessor gate fully open, so $c_j = 1$ and every summand vanishes. This proves both the decomposition and the displayed value of R^*.

At zero, $q_j(0) = 0 \leq a$, so $c_1 = 1$ and $c_j = G(0) = 0$ for $j \geq 2$. Equation (180) follows from $0 < d_* < 1$, $w_1 = 1$, and the definition of $\mathcal{T}(w)$. Finally, if $c_j = 0$, then $e_{j,0}(z_j) = \lambda\|z_j\|^2/2 + 1 + d_* \geq 1$; retaining these terms in (178) proves (181). □

D.3 Weighted gradient lower bound

Write

$$z_j = \frac{x_j}{w_j}, \qquad u_j = q_{\Theta^{(j)}}(z_j), \qquad v_j = \nabla q_{\Theta^{(j)}}(z_j). \tag{183}$$

Define the coefficient

$$\gamma_j := c_j + \mathbf{1}\{j < M\} G'(u_j) \left(\frac{w_{j+1}}{w_j}\right)^2 \left(q_{\Theta^{(j+1)}}(z_{j+1}) + 1\right). \tag{184}$$

Since $q_{\Theta^{(j+1)}}(z_{j+1}) + 1 > 0$, one has $\gamma_j \geq c_j \geq 0$. Direct differentiation gives

$$\nabla_{x_j} R = w_j(\lambda z_j - \gamma_j v_j). \tag{185}$$

Lemma D.3 (Weighted radial gradient lower bound)**.** *If $j_\star$ is the first stage with $u_{j_\star}(x_{j_\star}) < b$, then*

$$\|\nabla R(x)\| \geq \chi w_{j_\star}, \tag{186}$$

where $\chi > 0$ is the constant from Lemma A.6.

Proof. If $j_\star = 1$, then $c_{j_\star} = 1$ by definition. If $j_\star > 1$, every preceding gate is fully open and again $c_{j_\star} = 1$. Hence $\gamma_{j_\star} \geq 1$ in both cases. The block formula (185) and Lemma A.6 imply

$$\|\nabla_{x_{j_\star}} R\| \geq w_{j_\star} \|\lambda z_{j_\star} - v_{j_\star}\| \geq \chi w_{j_\star}. \tag{187}$$

□

By definition of $\mathcal{T}(w)$, every j_0 obeys

$$\sum_{j=j_0}^{M} w_j^2 \leq \mathcal{T}(w) w_{j_0}^2. \tag{188}$$

Figure 5 displays the two estimates used for the global PL inequality: the exact decomposition of the gap by blocks and the gradient bound at the first gate that is not fully open.

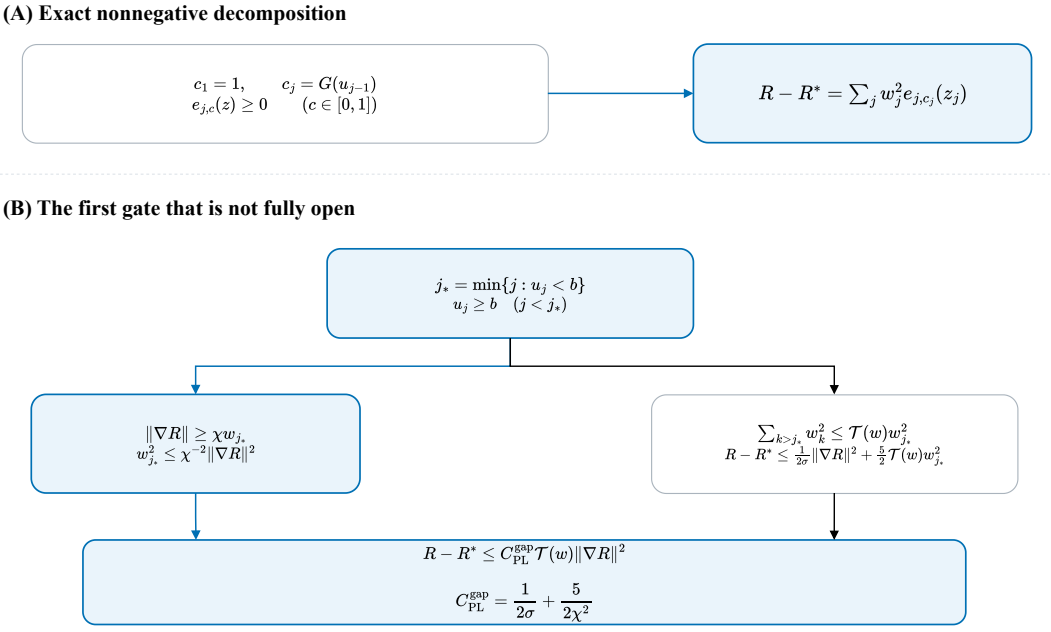

Figure 5: Weighted PL inequality. The decomposition of the gap by blocks gives nonnegative terms. If every gate is fully open, strong convexity of one block directly controls the gap. Otherwise, the first gate that is not fully open gives the gradient barrier, while $\mathcal{T}(w)$ bounds all later contributions. The threshold b for full opening is distinct from the completion threshold a.

D.4 Global PL inequality from the block gaps

Set

$$C_{\mathrm{PL}}^{\mathrm{gap}} := \frac{1}{2\sigma} + \frac{5}{2\chi^2}. \tag{189}$$

We use the standard consequence of differentiable σ-strong convexity

$$f(x) - f^* \le \frac{1}{2\sigma}\|\nabla f(x)\|^2. \tag{190}$$

It follows by minimizing the right-hand side of $f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + (\sigma/2)\|y - x\|^2$ over y.

Lemma D.4 (Gap bound for one block)**.** *For every block j and every x,*

$$w_j^2 e_{j,c_j}(z_j) \le \frac{1}{2\sigma}\|\nabla_{x_j} R(x)\|^2 + \frac{5}{2} w_j^2 \mathbf{1}\{c_j < 1\}. \tag{191}$$

Proof. Case 1: $\gamma_j \ge 1$. If $u_j < b$, the radial comparison (59) and the strong convexity of ϕ_j apply because $e_{j,1}(z_j) = \phi_j(z_j) + d_* = \phi_j(z_j) - \min_z \phi_j(z)$, and give

$$e_{j,1}(z_j) \le \frac{1}{2\sigma}\|\lambda z_j - v_j\|^2 \le \frac{1}{2\sigma}\|\lambda z_j - \gamma_j v_j\|^2. \tag{192}$$

Since $e_{j,c_j} = e_{j,1} + (1 - c_j)(q_{\Theta^{(j)}}(z_j) + 1)$ and $q_{\Theta^{(j)}}(z_j) + 1 \le 2$, this proves (191). If instead $u_j \ge b$, then $G'(u_j) = 0$ and $\gamma_j = c_j \le 1$. Thus $\gamma_j \ge 1$ forces $c_j = \gamma_j = 1$, and the same bound from strong convexity applies without slack.

Case 2: $0 \le \gamma_j < 1$. Here $c_j \le \gamma_j < 1$, so the indicator in (191) equals one. Since $q_{\Theta^{(j)}}(z_j) + 1 \ge 0$ and $d_* < 1$,

$$e_{j,c_j}(z_j) \le \frac{\lambda}{2}\|z_j\|^2 + 2. \tag{193}$$

Write $g_j := \lambda z_j - \gamma_j v_j$. Because $\lambda z_j = g_j + \gamma_j v_j$, $\gamma_j < 1$, and $\|v_j\| \le 1$,

$$\lambda^2\|z_j\|^2 \le 2\|g_j\|^2 + 2. \tag{194}$$

Consequently,

$$e_{j,c_j}(z_j) \le \frac{1}{\lambda}\|g_j\|^2 + 2 + \frac{1}{\lambda} \le \frac{1}{2\sigma}\|g_j\|^2 + \frac{5}{2}. \tag{195}$$

Multiplying by w_j^2 and using $\nabla_{x_j} R = w_j g_j$ proves the claim. □

Proof of Proposition D.1, item (iii). Case 1: every gate is fully open. Then $c_j = 1$ for every j. Summing Lemma D.4 and using (178) gives

$$R - R^* \le \frac{1}{2\sigma}\|\nabla R\|^2, \tag{196}$$

which implies (170) because $\mathcal{T}(w) \ge 1$.

Case 2: a gate is not fully open. Let $j_\star$ be the first index with $u_{j_\star} < b$. Every predecessor gate for a block $j \le j_\star$ is fully open, so $c_j = 1$ for these blocks. Summing Lemma D.4 therefore gives

$$\begin{aligned} R - R^* &\le \frac{1}{2\sigma}\|\nabla R\|^2 + \frac{5}{2}\sum_{j=j_\star+1}^{M} w_j^2 \\ &\le \frac{1}{2\sigma}\|\nabla R\|^2 + \frac{5}{2}\mathcal{T}(w) w_{j_\star}^2. \end{aligned} \tag{197}$$

Lemma D.3 gives $w_{j_\star}^2 \le \chi^{-2}\|\nabla R\|^2$. Therefore

$$\begin{aligned} R - R^* &\le \left(\frac{1}{2\sigma} + \frac{5}{2\chi^2}\mathcal{T}(w)\right)\|\nabla R\|^2 \\ &\le C_{\mathrm{PL}}^{\mathrm{gap}}\mathcal{T}(w)\|\nabla R\|^2, \end{aligned} \tag{198}$$

where the second line uses $\mathcal{T}(w) \ge 1$ and (189). This is (170).

Lemma A.6 gives $\chi > 5/18$ and hence $\chi^2 > 1/13$. Substitution into (189) yields

$$C_{\mathrm{PL}}^{\mathrm{gap}} < 8 + \frac{65}{2} = \frac{81}{2} < 41. \tag{199}$$

□

D.5 Verification of individual smoothness

Lemma D.5 (Explicit weighted Hessian bound)**.** *For every $\rho \in \mathcal{R}_n$ satisfying $0 < \rho \le \overline{\rho} = 17/512$, let $H_{\mathrm{main}} = 341$ be the constant defined in (127). Then every component satisfies*

$$\sup_x \|\nabla^2 R_i(x)\|_{\mathrm{op}} \le \frac{H_{\mathrm{main}}}{\rho}. \tag{200}$$

The bound is independent of $M, D, \mathcal{T}(w)$, and the smallest weight.

Proof. Apply Lemma A.5 to one link $-G(u_{j-1})(\widetilde{Q}_{ij} + Cw_j^2)$ with

$$(B_0, B_1, B_2) = \left(\frac{w_j^2}{\rho}, \frac{w_j}{\rho}, \frac{2}{\rho}\right), \quad (g_1, g_2) = \left(\frac{1}{w_{j-1}}, \frac{2}{w_{j-1}^2}\right), \quad c = Cw_j^2. \tag{201}$$

For this calculation,

$$M_1 := \|G'\|_\infty \le 17, \qquad M_2 := \|G''\|_\infty \le 259. \tag{202}$$

Since $\rho \le \overline{\rho}$ and $w_j/w_{j-1} \le 1$, the current diagonal, cross, and predecessor diagonal bounds are at most

$$\frac{2}{\rho}, \qquad \frac{M_1}{\rho}, \qquad \frac{(1 + C\overline{\rho})(M_2 + 2M_1)}{\rho}. \tag{203}$$

The first stage contribution is at most $2/\rho$, and the quadratic contributes λI. Since $\rho \le 1$, the block row bound is

$$\max_j \sum_k \|\nabla^2_{x_j x_k} R_i(x)\|_{\mathrm{op}} \le \frac{\lambda + 2 + (1 + C\overline{\rho})(M_2 + 2M_1) + 2M_1}{\rho} \le \frac{H_{\mathrm{main}}}{\rho}, \tag{204}$$

which proves (200). □

Proof of the remaining items of Proposition D.1. Item (i) is Lemma D.5. Items (ii) and (iv) follow from Lemma D.2. □

D.6 Scaling, condition number, and information cost

With the optimized bias and $r = \lfloor n/512 \rfloor$, this subsection verifies

$$\rho = \Theta(n^{-1/2}), \qquad 1 - r_{\mathrm{w}} = \Theta\left(\frac{\sqrt{n}}{\kappa_{\max}}\right),$$
$$J = \Theta\left(\frac{\kappa_{\max}}{\sqrt{n}} \log \frac{\Delta}{\varepsilon}\right), \qquad rJ = \Theta\left(\kappa_{\max}\sqrt{n} \log \frac{\Delta}{\varepsilon}\right).$$

D.6.1 Weight geometry and stage count

Geometric weights. Let $\kappa_{\max} = L_{\max}/\mu$ and set

$$\Gamma := 2H_{\mathrm{main}} C_{\mathrm{PL}}^{\mathrm{gap}}. \tag{205}$$

The bounds $H_{\mathrm{main}} = 341$ and $C_{\mathrm{PL}}^{\mathrm{gap}} < 41$ give

$$\Gamma < 28000. \tag{206}$$

Use throughout the weighted branch the optimized choice in (34). Thus

$$\rho = \frac{m}{n}, \qquad \sqrt{n} \le m < \sqrt{n} + 2, \qquad \rho < \frac{17}{16\sqrt{n}} \le \overline{\rho}. \tag{207}$$

Assume

$$\kappa_{\max} \ge \kappa_{\mathrm{w}}(n) := 2\Gamma\sqrt{n}. \tag{208}$$

This explicit threshold comes from the smoothness and PL bounds above. The two-scale construction covers the bounded transition interval, so the theorem still changes regimes at $\kappa_{\max} = \sqrt{n}$. Define the ratio of consecutive squared weights

$$r_{\mathrm{w}} := 1 - \frac{\Gamma}{\kappa_{\max}\rho}. \tag{209}$$

Because $\rho \geq n^{-1/2}$, the threshold above gives

$$\frac{1}{2} \leq r_{\mathrm{w}} < 1, \qquad 0 < 1 - r_{\mathrm{w}} \leq \frac{1}{2}. \tag{210}$$

Accuracy scale and stage count. We may decrease the universal c_ε in the theorem so that

$$\bar{\varepsilon} := \frac{64\varepsilon}{\Delta} \leq \frac{1}{4} \qquad \text{and} \qquad \log\frac{1}{\bar{\varepsilon}} = \Theta\left(\log\frac{\Delta}{\varepsilon}\right). \tag{211}$$

The first inequality follows from $\varepsilon \leq \Delta/256$. Choose

$$J := \left\lfloor \frac{\log(1/\bar{\varepsilon})}{-\log r_{\mathrm{w}}} \right\rfloor, \qquad M := 2J + 1, \qquad w_j := r_{\mathrm{w}}^{(j-1)/2}. \tag{212}$$

The quantity inside the floor is at least two, so $J \geq 2$. The definition also gives

$$\bar{\varepsilon} \leq r_{\mathrm{w}}^J < \frac{\bar{\varepsilon}}{r_{\mathrm{w}}} \leq 2\bar{\varepsilon} \leq \frac{1}{2}. \tag{213}$$

Value of $\mathcal{T}(w)$. For this geometric sequence,

$$\mathcal{T}(w) \leq \sum_{k=0}^{\infty} r_{\mathrm{w}}^k = \frac{1}{1 - r_{\mathrm{w}}} = \frac{\kappa_{\max}\rho}{\Gamma}. \tag{214}$$

D.6.2 Scaling, the gap from an unrevealed suffix, and stage count

Rescale to the target gap and smoothness. A global rescaling $F(y) = \alpha R(\beta y)$ multiplies both the smoothness and PL constants by $\alpha\beta^2$, and therefore preserves their ratio:

$$L_F = \alpha\beta^2 L_R, \qquad \mu_F = \alpha\beta^2\mu_R, \qquad \frac{L_F}{\mu_F} = \frac{L_R}{\mu_R}. \tag{215}$$

Thus the choice of r_{w} fixes the ratio between the PL and smoothness constants. The two global scales below enforce the initial gap and individual smoothness bounds. Define

$$f_i(y) = \alpha R_i(\beta y), \qquad \alpha = \frac{\Delta(1 - r_{\mathrm{w}})}{2}, \qquad \beta^2 = \frac{L_{\max}\rho}{\alpha H_{\mathrm{main}}}. \tag{216}$$

Let

$$F(y) := \frac{1}{n}\sum_{i=1}^{n} f_i(y) = \alpha R(\beta y). \tag{217}$$

Lemma D.5 makes every component $L_{\max}$-smooth, while (180) and (214) give $F(0) - F^* \leq \Delta$.

PL constant after scaling. To track the PL constant under scaling, use $\nabla F(y) = \alpha\beta\nabla R(\beta y)$ and

$$\begin{aligned} F(y) - F^* &= \alpha\big(R(\beta y) - R^*\big) \\ &\leq \frac{\alpha}{2\mu_0}\|\nabla R(\beta y)\|^2 = \frac{1}{2\alpha\beta^2\mu_0}\|\nabla F(y)\|^2. \end{aligned} \tag{218}$$

Proposition D.1, item (iii), and (214) give $\mu_0 \geq (1 - r_{\mathrm{w}})/(2C_{\mathrm{PL}}^{\mathrm{gap}})$. Therefore the scaled PL constant is at least

$$\alpha\beta^2\mu_0 \geq \frac{L_{\max}\rho(1 - r_{\mathrm{w}})}{2H_{\mathrm{main}}C_{\mathrm{PL}}^{\mathrm{gap}}} = \frac{L_{\max}}{\kappa_{\max}} = \mu, \tag{219}$$

where the middle identity uses $1 - r_{\mathrm{w}} = \Gamma/(\kappa_{\max}\rho)$ and $\Gamma = 2H_{\mathrm{main}}C_{\mathrm{PL}}^{\mathrm{gap}}$.

Gap from the unrevealed suffix and stage count. Stage $J+1$ separates the first J stages used in the call count from the hidden suffix $j = J+2, \dots, 2J+1$, whose total weight is

$$\sum_{j=J+2}^{2J+1} w_j^2 = \frac{r_{\mathrm{w}}^{J+1}(1 - r_{\mathrm{w}}^J)}{1 - r_{\mathrm{w}}}. \tag{220}$$

If $u_k \le a$ for every $k > J$, then the predecessor of every block in this suffix is below the lower gate threshold. Thus $c_j = 0$ throughout the hidden suffix. Hence the exact decomposition of the gap by blocks gives $R - R^* \ge \sum_{j=J+2}^{2J+1} w_j^2$. Substituting $\alpha = \Delta(1 - r_{\mathrm{w}})/2$ and $\bar{\varepsilon} \le r_{\mathrm{w}}^J < 2\bar{\varepsilon} \le 1/2$, and using $r_{\mathrm{w}} \ge 1/2$, yields

$$F(y) - F^* \ge \frac{\Delta}{2} r_{\mathrm{w}}^{J+1}(1 - r_{\mathrm{w}}^J) \ge \frac{\Delta}{8}\bar{\varepsilon} = 8\varepsilon. \tag{221}$$

For $0 < 1 - r_{\mathrm{w}} \le 1/2$,

$$1 - r_{\mathrm{w}} \le -\log r_{\mathrm{w}} \le 2(1 - r_{\mathrm{w}}). \tag{222}$$

Since the quantity inside the floor in (212) is at least two, flooring loses at most a factor of two. Consequently,

$$\frac{\kappa_{\max}\rho}{4\Gamma} \log\frac{1}{\bar{\varepsilon}} \le J \le \frac{\kappa_{\max}\rho}{\Gamma} \log\frac{1}{\bar{\varepsilon}}, \qquad J = \Theta\left(\kappa_{\max}\rho \log\frac{\Delta}{\varepsilon}\right). \tag{223}$$

D.6.3 Sequential revelation, query cost, and dimension

Information cost. The one-stage result at the optimized bias now applies directly. Set $r = \lfloor n/512 \rfloor$. Since $n \ge 1024$, one has $r \ge n/1024$; therefore (207) and the lower bound in (223) give

$$rJ \ge \frac{\kappa_{\max} m}{4096\Gamma} \log\frac{1}{\bar{\varepsilon}} \ge \frac{\kappa_{\max}\sqrt{n}}{4096\Gamma} \log\frac{1}{\bar{\varepsilon}}. \tag{224}$$

For $h \in \{0, \dots, M\}$, let $R_i^{[h]}$ keep the first stage contribution and full quadratic sum in (168), but truncate the sum of gated links at $\min\{h+1, M\}$. Thus future blocks retain their deterministic quadratic gradients while omitted gated stage contributions are removed. With $V_j(y_j) = \Psi(\beta y_j / w_j)$, table independence, the dependence of each component on its assigned table row, the condition $G = G' = 0$ for omitted gates, and the call accounting rule give *(R1)–(R4)* of Corollary B.6. Take an integer D satisfying

$$D \ge \max\left\{2048[1 + \log(N+3)], \quad \frac{2048}{25}\log\big(128(N+1)M\big)\right\}. \tag{225}$$

The case $N = 0$ is immediate, so assume $N \ge 1$. The corollary, with $M_{\mathrm{tot}} = M = 2J+1$, gives

$$\mathbb{P}(\nu_N < J,\ \mathcal{E}_{\mathrm{future}}^c) \ge \frac{95}{128} \quad \text{whenever} \quad N \le \frac{rJ}{2}. \tag{226}$$

On this event, Corollary B.6 gives

$$u_j(\beta\widehat{X}_j) \le a \qquad \forall j > J. \tag{227}$$

Equation (221) then gives

$$F(\widehat{X}) - F^* \ge 8\varepsilon. \tag{228}$$

Together with (224), this proves the $\kappa_{\max}\sqrt{n}\log(\Delta/\varepsilon)$ term.

Equations (207) and (223) give $M = O((\kappa_{\max}/\sqrt{n})\log(\Delta/\varepsilon))$. Since $d = MD$, and both $N+3$ and M are bounded by a universal multiple of the quantity inside the logarithm below, the two logarithmic requirements can be absorbed into one:

$$d = O\left[\frac{\kappa_{\max}}{\sqrt{n}} \log\frac{\Delta}{\varepsilon} \log\left(2 + n + \kappa_{\max}\sqrt{n}\log\frac{\Delta}{\varepsilon}\right)\right]. \tag{229}$$

Conclusion for the weighted chain. The preceding calculation establishes, after scaling,

$$\max_i \mathrm{Lip}(\nabla f_i) \le L_{\max}, \qquad F(0) - F^* \le \Delta, \qquad F \text{ is } \mu\text{-PL},$$
$$N \le rJ/2 \implies \mathbb{P}\big(F(\widehat{X}) - F^* \ge 8\varepsilon\big) \ge \frac{95}{128}, \qquad rJ = \Omega\left(\kappa_{\max}\sqrt{n}\log\frac{\Delta}{\varepsilon}\right).$$

D.7 Auxiliary $\Omega(n)$ bound

Lemma D.6 (PL auxiliary quadratic lower bound)**.** *For every $n \ge n_0 = 1024$ and integer $0 \le N \le \lfloor n/4 \rfloor$, let $\mu = L_{\max}/\kappa_{\max}$, where $\kappa_{\max} \ge 1$, and suppose $0 < \varepsilon \le \Delta/128$. There exists a distribution over individually $L_{\max}$-smooth finite sums in dimension 4 whose averages are globally μ-PL and have initial gap Δ such that every algorithm making at most N calls satisfies*

$$\mathbb{P}\big(F(\widehat{X}) - F^* < 8\varepsilon\big) \le \frac{5}{16}. \tag{230}$$

Proof. Instantiate the quadratic f_i^{quad} from Lemma B.9 with $\gamma = \mu$ and $\tau = (2\mu\Delta)^{1/2}$, and write $F := F_{\mathrm{quad}}$. Each component has Hessian $\mu I \preceq L_{\max} I$, while its average satisfies

$$F(0) - F^* = \frac{\tau^2}{2\mu} = \Delta, \qquad \|\nabla F(x)\|^2 = 2\mu(F(x) - F^*). \tag{231}$$

Completing the square gives

$$F(x) - F^* = \frac{1}{2\mu}\|\mu x - \tau u\|^2. \tag{232}$$

Hence the success event implies

$$\|\mu\widehat{X} - \tau u\|^2 < 16\mu\varepsilon \le \frac{\tau^2}{16}, \tag{233}$$

where the last inequality uses $\tau^2 = 2\mu\Delta$ and $\varepsilon \le \Delta/128$. Lemma B.9, with these parameters, bounds this success probability by $5/16$ and proves (230). □

Lower bound from the weighted chain. The weighted chain has failure probability $95/128$, and Lemma D.6 has failure probability $11/16$. The larger of their two query thresholds is at least a universal constant times their sum. Thus, for $\kappa_{\max} \ge \kappa_{\mathrm{w}}(n) := 2\Gamma\sqrt{n}$, where $\Gamma < 28000$, the weighted construction proves

$$\Omega\left(n + \kappa_{\max}\sqrt{n}\log\frac{\Delta}{\varepsilon}\right) \tag{234}$$

using the one-stage bound at the optimized bias for the chain and obtaining the linear term from an auxiliary quadratic at small bias.

Part II: two-scale chain for the small-κ branch. The weighted construction requires a geometric ratio $r_{\mathrm{w}} \in (0,1)$ with $1 - r_{\mathrm{w}} = \Theta(\sqrt{n}/\kappa_{\max})$. This choice is not admissible when $\kappa_{\max} < \sqrt{n}$. In this range, we instead separate the gate scale R_j from the contribution scale W_j. The construction can then use the amplitude ratio $\vartheta = \Theta(\kappa_{\max}/\sqrt{n})$ while preserving sequential hiding and individual smoothness.

D.8 Two-scale chain: construction

Fix an integer $M_{\mathrm{ts}} \ge 2$; the choice depending on accuracy is made only after the analytic properties of this family have been established.

Bias and effective curvature. Use the optimized admissible bias from Corollary A.3: m is the least parity-compatible integer not smaller than $\sqrt{n}$, so

$$\sqrt{n} \le m < \sqrt{n} + 2, \qquad \rho = \frac{m}{n}, \qquad \frac{1}{\sqrt{n}} \le \rho < \frac{17}{16\sqrt{n}}. \tag{235}$$

For $\kappa_{\max} \ge 3$, define

$$\bar{\kappa}_{\max} := \min\{\kappa_{\max}, \sqrt{n}\}, \qquad \bar{\mu} := \frac{L_{\max}}{\bar{\kappa}_{\max}}. \tag{236}$$

Since $\bar{\mu} \ge \mu = L_{\max}/\kappa_{\max}$, every globally $\bar{\mu}$-strongly convex objective is globally μ-PL.

Geometrically decreasing amplitudes and two spatial scales. Define the positive stage amplitudes by

$$\vartheta := \frac{\bar{\kappa}_{\max}\rho}{8192}, \qquad \zeta_1 := \sqrt{2\bar{\mu}\Delta}, \qquad \zeta_j := \zeta_1 \vartheta^{j-1}. \tag{237}$$

For every stage, define the gate scale R_j and contribution scale W_j by

$$R_j := \frac{\zeta_j}{\bar{\mu}}, \qquad W_j := \frac{64\zeta_j}{L_{\max}\rho}. \tag{238}$$

The ratio R_j/W_j is independent of j and satisfies

$$\delta_{\mathrm{rad}} := \frac{R_j}{W_j} = \frac{\bar{\kappa}_{\max}\rho}{64} < \frac{17}{1024}, \tag{239}$$

where we used $\bar{\kappa}_{\max} \le \sqrt{n}$ and the local bias bound in (235).

Gate, component term, and chain. For $x = (x_1, \dots, x_{M_{\mathrm{ts}}})$ with $x_j \in \mathbb{R}^D$, define the gate coordinate

$$h_j(x_j) := q_{\Theta^{(j)}}(x_j/R_j) \tag{240}$$

and the component term

$$P_{ij}(x_j) := \frac{\zeta_j W_j}{\rho} \left\langle \frac{S_{i:}^{(j)}}{\sqrt{D}}, \Psi(x_j/W_j) \right\rangle. \tag{241}$$

By (33), its average is exactly

$$P_j(x_j) := \frac{1}{n}\sum_{i=1}^{n} P_{ij}(x_j) = \zeta_j W_j q_{\Theta^{(j)}}(x_j/W_j). \tag{242}$$

With M_{ts} stages, define

$$\begin{aligned} f_i^{\mathrm{ts}}(x) = \bar{\mu}\sum_{j=1}^{M_{\mathrm{ts}}} \|x_j\|^2 - P_{i1}(x_1) \\ - \sum_{j=2}^{M_{\mathrm{ts}}} G(h_{j-1}(x_{j-1}))P_{ij}(x_j), \end{aligned} \tag{243}$$

and let $F_{\mathrm{ts}} = n^{-1}\sum_i f_i^{\mathrm{ts}}$. The exact table average gives

$$F_{\mathrm{ts}}(x) = \bar{\mu}\sum_{j=1}^{M_{\mathrm{ts}}} \|x_j\|^2 - P_1(x_1) - \sum_{j=2}^{M_{\mathrm{ts}}} G(h_{j-1}(x_{j-1}))P_j(x_j). \tag{244}$$

All components are C^∞.

Proposition D.7 (Properties of the two-scale chain)**.** *For every $M_{\mathrm{ts}} \geq 2$, the construction above satisfies:*

(i) *Every component f_i^{ts} is individually $L_{\max}$-smooth.*

(ii) *The average satisfies*

$$\nabla^2 F_{\mathrm{ts}}(x) \succeq \bar{\mu} I, \tag{245}$$

and hence is globally $\bar{\mu}$-strongly convex and μ-PL. When $3 \leq \kappa_{\max} \leq \sqrt{n}$, one has $\bar{\mu} = \mu$, so F_{ts} is μ-strongly convex.

(iii) *Its unique minimizer is*

$$x_j^* = R_j z_* u_{\Theta^{(j)}}, \qquad 2z_* = (1 + \delta_{\mathrm{rad}}^2 z_*^2)^{-3/2}, \tag{246}$$

where $z_* > 1/3$ is the unique positive solution; every optimal gate satisfies $h_j(x_j^*) > b$.

(iv) *The initial gap and the gap at the final stage satisfy*

$$F_{\mathrm{ts}}(0) - F_{\mathrm{ts}}^* \leq \Delta, \tag{247}$$

and

$$h_{M_{\mathrm{ts}}}(x_{M_{\mathrm{ts}}}) \leq a \quad \Longrightarrow \quad F_{\mathrm{ts}}(x) - F_{\mathrm{ts}}^* \geq \frac{\zeta_{M_{\mathrm{ts}}}^2}{128\bar{\mu}}. \tag{248}$$

For $3 \leq \kappa_{\max} \leq \sqrt{n}$, one has $\bar{\kappa}_{\max} = \kappa_{\max}$ and $\rho \asymp n^{-1/2}$, so $R_j/W_j \asymp \kappa_{\max}/\sqrt{n}$. If the stage contribution used the short gate scale R_j as well, its basic component Hessian estimate would be $2\zeta_j/(\rho R_j) = 2L_{\max}/(\bar{\kappa}_{\max}\rho) > L_{\max}$, already exceeding the smoothness budget for each component. The larger contribution scale W_j reduces this Hessian contribution to the required scale. The choice $R_j = \zeta_j/\bar{\mu}$ sets the displacement needed to activate the next stage. Figure 6 shows the conflict between hiding and individual smoothness, and how the separation $R_j \ll W_j$ resolves it.

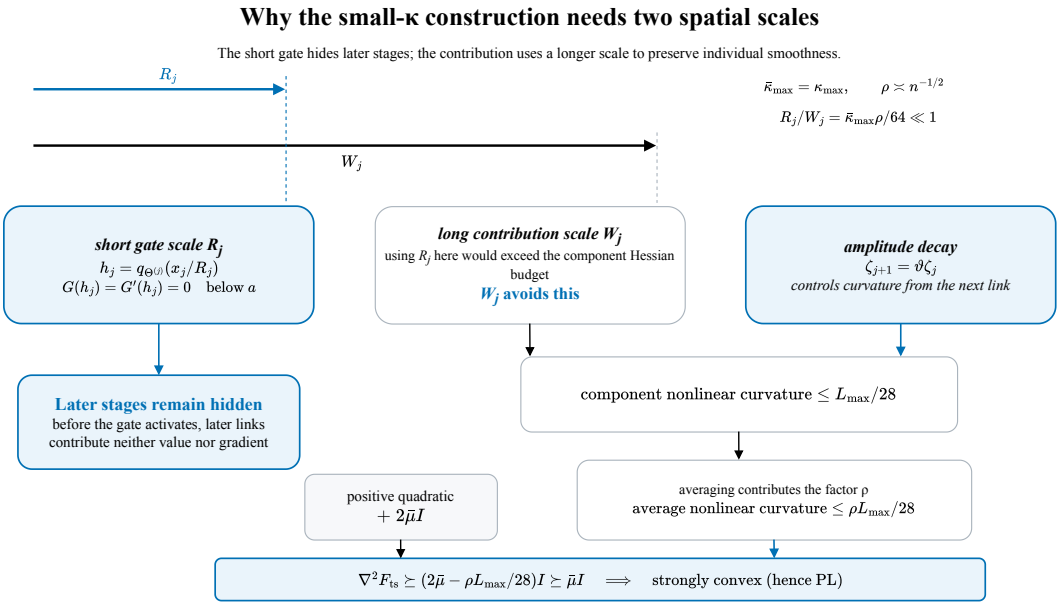

Figure 6: Why two spatial scales are needed in the small-κ construction. The short scale R_j makes the gate locally flat before it opens and thereby hides later stages. Using that same scale for the stage contribution would exceed the $L_{\max}$ Hessian budget of a component. The longer scale W_j, together with the amplitude decay, reduces the nonlinear component curvature to at most $L_{\max}/28$. Averaging contributes the factor ρ, and the positive quadratic then gives $\nabla^2 F_{\mathrm{ts}} \succeq \bar{\mu}I$.

D.9 Two-scale chain: analytic properties

Lemma D.8 (Nonlinear Hessian bound)**.** *Define*

$$B_\star := \frac{1}{28}. \tag{249}$$

Then the nonlinear Hessian of every component in (243) has operator norm at most $B_\star L_{\max}$. After averaging the components, the corresponding bound is $\rho B_\star L_{\max}$.

Proof. Scaled derivatives. Lemma A.4 and the chain rule give

$$\begin{aligned} \|\nabla h_j\| &\le \frac{1}{R_j}, & \|\nabla^2 h_j\|_{\mathrm{op}} &\le \frac{2}{R_j^2}, \\ |P_{ij}| &\le \frac{\zeta_j W_j}{\rho}, & \|\nabla P_{ij}\| &\le \frac{\zeta_j}{\rho}, & \|\nabla^2 P_{ij}\|_{\mathrm{op}} &\le \frac{2\zeta_j}{\rho W_j}. \end{aligned} \tag{250}$$

The corresponding bounds for P_j omit $1/\rho$.

Local link blocks. For the link $\mathcal{L}_{ij} = -G(h_{j-1})P_{ij}$, substitute these bounds into Lemma A.5. The actual scale choices make the current diagonal, both cross blocks together, and the previous diagonal at most

$$\begin{aligned} B_2 &= \frac{2\zeta_j}{\rho W_j} = \frac{L_{\max}}{32}, \\ 2M_1 g_1 B_1 &= \frac{2M_1\zeta_j}{\rho R_{j-1}} = \frac{2M_1 L_{\max}}{8192} \le \frac{L_{\max}}{240}, \\ B_0(M_2 g_1^2 + M_1 g_2) &= \frac{\zeta_j W_j (M_2 + 2M_1)}{\rho R_{j-1}^2} = \frac{64(M_2 + 2M_1)L_{\max}}{8192^2} \le \frac{L_{\max}}{3500}. \end{aligned} \tag{251}$$

Here (28) bounds M_1 and M_2.

With $\delta_{\mathrm{rad}} = R_j/W_j = 128\vartheta$, all three contributions in (251) are bounded by universal fractions of the $L_{\max}$ smoothness budget. Their sum is below $B_\star L_{\max}$ because

$$\frac{1}{32} + \frac{1}{240} + \frac{1}{3500} < \frac{1}{28} = B_\star. \tag{252}$$

Global block row and averaging. Each block row contains at most one current diagonal, one previous diagonal, and two cross blocks. The ungated first stage contribution $-P_{i1}$ obeys the same bound for the current diagonal block. Hence the block row sum is at most $B_\star L_{\max}$, and the block row clause of Lemma A.5 gives the component bound. For every $2 \le j \le M_{\mathrm{ts}}$, the exact average over the fixed table gives

$$\frac{1}{n}\sum_{i=1}^{n} \nabla^2[-G(h_{j-1})P_{ij}] = \nabla^2[-G(h_{j-1})P_j]\,. \tag{253}$$

The ungated term at the first stage obeys the analogous identity. Moreover,

$$|P_j| \le \zeta_j W_j, \qquad \|\nabla P_j\| \le \zeta_j, \qquad \|\nabla^2 P_j\|_{\mathrm{op}} \le \frac{2\zeta_j}{W_j}, \tag{254}$$

so the value, gradient, and Hessian bounds for P_j are each ρ times the corresponding component bounds in (250). Thus every nonlinear block bound is multiplied by ρ. □

Proof of Proposition D.7, items (i)–(ii). Since $\bar\kappa_{\max} \ge 3$,

$$\frac{2\bar\mu + B_\star L_{\max}}{L_{\max}} = \frac{2}{\bar\kappa_{\max}} + B_\star \le \frac{2}{3} + \frac{1}{28} < \frac{3}{4} < 1, \tag{255}$$

which proves individual smoothness. This is the reason for the convenient threshold $\kappa_{\max} \ge 3$: the quadratic and nonlinear terms fit within the component smoothness bound with a uniform margin. This threshold is not optimized. For the average, $\bar\kappa_{\max}\rho < 17/16$ gives

$$\nabla^2 F_{\mathrm{ts}}(x) \succeq \left(2 - \frac{17}{16}\frac{1}{28}\right)\bar\mu I \succeq \bar\mu I, \tag{256}$$

where $2 - 17/(16 \cdot 28) > 1$. Finally, $\bar\mu \ge \mu$ implies the declared μ-PL property. □

Proof of Proposition D.7, item (iii). Scalar stationary point. In the scalar equation in (246), the left side increases strictly from zero and the right side decreases strictly from one. At $z = 1/3$,

$$(1 + \delta_{\mathrm{rad}}^2/9)^{-3/2} > (10/9)^{-3/2} > 2/3, \tag{257}$$

where the final comparison follows by cubing and squaring positive quantities: $1000/729 < 9/4$. Thus the unique crossing has $z_* > 1/3$.

Average minimizer and gate values. At $x_j = R_j z u_{\Theta^{(j)}}$, direct differentiation gives

$$\nabla P_j(x_j) = \zeta_j (1 + \delta_{\mathrm{rad}}^2 z^2)^{-3/2} u_{\Theta^{(j)}}. \tag{258}$$

Because $2\bar{\mu} R_j z = 2\zeta_j z$, (246) is exactly the stationarity condition for $\bar{\mu}\|x_j\|^2 - P_j(x_j)$. Moreover,

$$h_j(x_j^*) = \frac{z_*}{\sqrt{1 + z_*^2}} > \frac{1}{\sqrt{10}} > \frac{9}{32} = b. \tag{259}$$

Every predecessor coefficient therefore equals one, while every derivative of a successor gate vanishes when the gate is fully open. Hence $\nabla F_{\mathrm{ts}}(x^*) = 0$, and uniqueness follows from Proposition D.7, item (ii). □

Proof of Proposition D.7, item (iv). Initial gap. At the origin, $h_j(0) = 0 \le a$, so all successor gates satisfy $G = G' = 0$ and $\nabla F_{\mathrm{ts}}(0) = -\zeta_1 u_{\Theta^{(1)}}$. Strong convexity implies

$$F_{\mathrm{ts}}(0) - F_{\mathrm{ts}}^* \le \frac{\|\nabla F_{\mathrm{ts}}(0)\|^2}{2\bar{\mu}} = \frac{\zeta_1^2}{2\bar{\mu}} = \Delta. \tag{260}$$

Gap at the final stage. By item (iii), $h_{M_{\mathrm{ts}}}(x_{M_{\mathrm{ts}}}^*) > b$. The gate coordinate is $1/R_{M_{\mathrm{ts}}}$-Lipschitz, so

$$\|x_{M_{\mathrm{ts}}} - x_{M_{\mathrm{ts}}}^*\| \ge (b - a) R_{M_{\mathrm{ts}}} = \frac{R_{M_{\mathrm{ts}}}}{8}. \tag{261}$$

Using $\bar{\mu}$-strong convexity and $R_{M_{\mathrm{ts}}} = \zeta_{M_{\mathrm{ts}}}/\bar{\mu}$ proves (248). In particular, $\zeta_{M_{\mathrm{ts}}}^2 \ge 1024\bar{\mu}\varepsilon$ implies a gap of at least 8ε. □

D.10 Two-scale chain: sequential hiding and stage count

The two-scale chain gives $\boldsymbol{\Omega\big(n \log(\Delta/\varepsilon)/[1 + \log(\sqrt{n}/\bar{\kappa}_{\max})]\big)}$ when there are multiple stages; the auxiliary quadratic gives $\boldsymbol{\Omega(n)}$ otherwise.

Truncated construction at each progress level. At progress $g \in \{0, \dots, M_{\mathrm{ts}}\}$, retain links only through the current stage contribution:

$$\begin{aligned}(f_i^{\mathrm{ts}})^{[g]}(x) = \bar{\mu} \sum_{j=1}^{M_{\mathrm{ts}}} \|x_j\|^2 - P_{i1}(x_1) \\ - \sum_{j=2}^{\min\{g+1, M_{\mathrm{ts}}\}} G(h_{j-1}(x_{j-1})) P_{ij}(x_j).\end{aligned} \tag{262}$$

With $V_j(x_j) = \Psi(x_j/R_j)$, table independence, the dependence of each component on its assigned table row, the condition $G = G' = 0$ for omitted gates, and the call accounting rule give *(R1)–(R4)* of Corollary B.6.

Let

$$s_{\mathrm{ts}} := \log \frac{1}{\vartheta} = \log \frac{8192}{\bar{\kappa}_{\max}\rho}, \qquad \ell_\kappa := 1 + \log \frac{\sqrt{n}}{\bar{\kappa}_{\max}}. \tag{263}$$

Write $\ell_\varepsilon := \log(\Delta/\varepsilon)$. For the case with multiple stages, assume

$$\ell_\varepsilon \ge 2 \log 512 + 4 s_{\mathrm{ts}} \tag{264}$$

and set

$$J := \left\lfloor \frac{\ell_\varepsilon - \log 512}{2 s_{\mathrm{ts}}} \right\rfloor, \qquad M_{\mathrm{ts}} := J + 1. \tag{265}$$

Lemma D.9 (Logarithmic scale, stage count, and scale at the final stage)**.** *The scales in* (263) *satisfy*

$$\ell_\kappa < s_{\mathrm{ts}} < 10\ell_\kappa. \tag{266}$$

Under (264)*, moreover,*

$$J \geq \frac{\ell_\varepsilon}{4s_{\mathrm{ts}}} \geq \frac{\ell_\varepsilon}{40\ell_\kappa}, \qquad \zeta_{M_{\mathrm{ts}}}^2 \geq 1024\bar{\mu}\varepsilon. \tag{267}$$

Proof. Compare the logarithmic scales. The upper bound on ρ in (235) gives

$$s_{\mathrm{ts}} > \log\frac{8192\cdot 16}{17} + \log\frac{\sqrt{n}}{\bar{\kappa}_{\max}} > \ell_\kappa. \tag{268}$$

The lower bound $\rho \geq n^{-1/2}$ gives

$$s_{\mathrm{ts}} \leq \log 8192 + \log\frac{\sqrt{n}}{\bar{\kappa}_{\max}} < 10 + (\ell_\kappa - 1) \leq 10\ell_\kappa. \tag{269}$$

Stage count and scale at the final stage. The inequality $\lfloor z \rfloor \geq z - 1$ and (264) give

$$J \geq \frac{\ell_\varepsilon - \log 512 - 2s_{\mathrm{ts}}}{2s_{\mathrm{ts}}} \geq \frac{\ell_\varepsilon}{4s_{\mathrm{ts}}}. \tag{270}$$

The upper comparison $s_{\mathrm{ts}} < 10\ell_\kappa$ gives the second bound on the number of stages. The upper bound on the floor in (265) yields

$$\zeta_{M_{\mathrm{ts}}}^2 = 2\bar{\mu}\Delta e^{-2s_{\mathrm{ts}}J} \geq 2\bar{\mu}\Delta e^{-\ell_\varepsilon + \log 512} = 1024\bar{\mu}\varepsilon. \tag{271}$$

□

Proposition D.10 (Bound for multiple stages of the two-scale construction)**.** *Let* $\ell_\varepsilon = \log(\Delta/\varepsilon)$, $s_{\mathrm{ts}} = \log(8192/(\bar{\kappa}_{\max}\rho))$*, and* $\ell_\kappa = 1 + \log(\sqrt{n}/\bar{\kappa}_{\max})$*. Assume* $\ell_\varepsilon \geq 2\log 512 + 4s_{\mathrm{ts}}$*, and set* $c_{\mathrm{ts}} := 1/163840$*. Then every deterministic algorithm whose IFO call budget* N *satisfies*

$$N \leq c_{\mathrm{ts}}\frac{n\ell_\varepsilon}{\ell_\kappa} \tag{272}$$

fails with probability at least $95/128$ *under the two-scale hard distribution. Specifically,*

$$F_{\mathrm{ts}}(\widehat{X}) - F_{\mathrm{ts}}^* \geq 8\varepsilon. \tag{273}$$

Proof. Dimension and sequential event. The conclusion is immediate for $N = 0$, so assume $N \geq 1$. Take $r = \lfloor n/512 \rfloor$, so $r \geq n/1024$ for $n \geq 1024$. Choose

$$D \geq \max\left\{2048[1 + \log(N+3)], \quad \frac{2048}{25}\log\bigl(128(N+1)M_{\mathrm{ts}}\bigr)\right\}. \tag{274}$$

Corollary B.6, with $\mathrm{align}_j = h_j$, $M_{\mathrm{tot}} = M_{\mathrm{ts}}$, and $J = M_{\mathrm{ts}} - 1$, shows that whenever $N \leq rJ/2$, with probability at least $95/128$ the output of the original IFO interaction satisfies $h_{M_{\mathrm{ts}}}(\widehat{X}_{M_{\mathrm{ts}}}) \leq a$.

Gap at the final stage. On this event, Lemma D.9 and Proposition D.7, item (iv), give

$$F_{\mathrm{ts}}(\widehat{X}) - F_{\mathrm{ts}}^* \geq \frac{\zeta_{M_{\mathrm{ts}}}^2}{128\bar{\mu}} \geq 8\varepsilon. \tag{275}$$

Query cost. Finally, the same lemma on the stage count gives $J \geq \ell_\varepsilon/(40\ell_\kappa)$, and hence

$$\frac{rJ}{2} \geq \frac{n}{2048}\frac{\ell_\varepsilon}{40\ell_\kappa} = \frac{n\ell_\varepsilon}{81920\ell_\kappa}, \tag{276}$$

Since $c_{\mathrm{ts}} = 1/163840$, the budget in (272) therefore ensures $N \leq rJ/2$. □

Corollary D.11 (Two-scale PL lower bound). *For every $n \geq n_0 = 1024$, $\kappa_{\max} \geq 3$, and $0 < \varepsilon \leq c\Delta$, every randomized algorithm with IFO access using fewer than*

$$c' \left[n + \frac{n}{1 + \log(\sqrt{n}/\bar{\kappa}_{\max})} \log \frac{\Delta}{\varepsilon} \right], \qquad \bar{\kappa}_{\max} = \min\{\kappa_{\max}, \sqrt{n}\}, \tag{277}$$

calls fails with probability at least $11/16$ on some individually $L_{\max}$-smooth finite sum with initial gap at most Δ and globally μ-PL average objective.

Proof. Choose $c \leq 1/128$ and $c' \leq (4/5)c_{\mathrm{ts}} = 1/204800$. When the condition for multiple stages holds, we have $\ell_\varepsilon \geq 4s_{\mathrm{ts}} > 4\ell_\kappa$, so

$$c' \left(n + \frac{n\ell_\varepsilon}{\ell_\kappa} \right) < \frac{5}{4} c' \frac{n\ell_\varepsilon}{\ell_\kappa} \leq c_{\mathrm{ts}} \frac{n\ell_\varepsilon}{\ell_\kappa}.$$

Thus Proposition D.10 and the conversion in Section F.1 give a fixed instance with failure probability at least $95/128$.

When this condition fails, suppose $\ell_\varepsilon < 2\log 512 + 4s_{\mathrm{ts}}$. Since $s_{\mathrm{ts}} < 10\ell_\kappa$,

$$\ell_\varepsilon < 2\log 512 + 4s_{\mathrm{ts}} < 54\ell_\kappa, \qquad n + \frac{n\ell_\varepsilon}{\ell_\kappa} < 55n. \tag{278}$$

Our choices also give $c' \leq 1/220$ and $\varepsilon \leq \Delta/128$. Every integer budget satisfying

$$N < c' \left(n + \frac{n\ell_\varepsilon}{\ell_\kappa} \right) < 55c'n \leq \frac{n}{4} \tag{279}$$

falls under Lemma D.6, which gives failure probability at least $11/16$. The two cases prove (277). □

Conclusion for the two-scale chain. In the case with multiple stages, Proposition D.7 and Proposition D.10 establish

$$\max_i \mathrm{Lip}(\nabla f_i^{\mathrm{ts}}) \leq L_{\max}, \qquad \nabla^2 F_{\mathrm{ts}} \succeq \bar{\mu} I, \qquad F_{\mathrm{ts}}(0) - F_{\mathrm{ts}}^* \leq \Delta,$$
$$N \leq \frac{c_{\mathrm{ts}} n \ell_\varepsilon}{\ell_\kappa} \implies \mathbb{P}\big(F_{\mathrm{ts}}(\widehat{X}) - F_{\mathrm{ts}}^* \geq 8\varepsilon\big) \geq \frac{95}{128}.$$

The auxiliary quadratic covers the remaining $O(n)$ case, yielding Corollary D.11.

Assembly across all condition-number regimes

We use the two-scale branch below $\sqrt{n}$, the weighted branch above $2\Gamma\sqrt{n}$, and the two-scale construction again on the bounded transition interval between them.

Each branch with multiple stages costs $\Theta(n)$ calls per stage. Multiplying by its stage count and adding the auxiliary $\Omega(n)$ term gives the bounds below.

Small $\kappa_{\max}$: $3 \leq \kappa_{\max} \leq \sqrt{n}$. Here $\bar{\kappa}_{\max} = \kappa_{\max}$, and Corollary D.11 gives

$$\Omega\left(n + \frac{n \log(\Delta/\varepsilon)}{1 + \log(\sqrt{n}/\kappa_{\max})} \right), \tag{280}$$

which is the small-$\kappa_{\max}$ regime of Theorem 3.3.

Range covered by the weighted chain: $\kappa_{\max} \geq 2\Gamma\sqrt{n}$. The weighted chain proves

$$\Omega\left(n + \kappa_{\max}\sqrt{n} \log \frac{\Delta}{\varepsilon} \right) \tag{281}$$

whenever $\kappa_{\max} \geq \kappa_{\mathrm{w}}(n) = 2\Gamma\sqrt{n}$.

Bounded transition: $\sqrt{n} \leq \kappa_{\max} < 2\Gamma\sqrt{n}$. Here $\bar{\kappa}_{\max} = \sqrt{n}$ and $\bar{\mu} \geq \mu$, so the two-scale instance still belongs to the required μ-PL class. The corollary gives

$$\Omega\left(n + n \log \frac{\Delta}{\varepsilon} \right). \tag{282}$$

Within the displayed transition interval, $\kappa_{\max}\sqrt{n} < 2\Gamma n$, so (282) is at least a universal $1/(2\Gamma)$ multiple of (281). Thus, after shrinking the universal theorem constant, the large-κ bound holds for every $\kappa_{\max} \geq \sqrt{n}$.

Failure probability and dimension. The arguments for multiple stages and for the complementary case have failure probabilities at least $95/128$ and $11/16$, respectively. The dimension choice above permits

$$d = M_{\mathrm{ts}} D = O\left[\left(1 + \frac{\ell_\varepsilon}{1 + \log(\sqrt{n}/\bar{\kappa}_{\max})}\right) \log\left(2 + n + \frac{n\ell_\varepsilon}{1 + \log(\sqrt{n}/\bar{\kappa}_{\max})}\right)\right] \tag{283}$$

for the two-scale construction. For the weighted chain,

$$d = O\left[\frac{\kappa_{\max}}{\sqrt{n}} \log\frac{\Delta}{\varepsilon} \log\left(2 + n + \kappa_{\max}\sqrt{n} \log\frac{\Delta}{\varepsilon}\right)\right], \tag{284}$$

as established in (229). The auxiliary quadratic uses only $D_0 = 4$.

E Restarted PAGE Upper Bound

Restatement of the upper bound in Theorem 3.3. Let $F = n^{-1}\sum_{i=1}^n f_i$ satisfy mean-squared smoothness with constant L_{ms}, the global μ-PL inequality, and $F(x^{(0)}) - F^* \leq \Delta$. Write $\kappa_{\mathrm{ms}} = L_{\mathrm{ms}}/\mu \geq 1$. For $0 < \varepsilon \leq c\Delta$, Algorithm 1 returns $x^{(S_{\mathrm{ep}})}$ satisfying

$$\mathbb{E}\left[F(x^{(S_{\mathrm{ep}})}) - F^*\right] \leq \varepsilon \tag{285}$$

with expected IFO cost bounded, for a universal constant $C > 0$, by

$$\mathbb{E}[\#\text{ IFO calls}] \leq C \begin{cases} n + \dfrac{n\log(\Delta/\varepsilon)}{1 + \log(\sqrt{n}/\kappa_{\mathrm{ms}})}, & 1 \leq \kappa_{\mathrm{ms}} \leq \sqrt{n}, \\ n + \kappa_{\mathrm{ms}}\sqrt{n}\log\dfrac{\Delta}{\varepsilon}, & \kappa_{\mathrm{ms}} \geq \sqrt{n}. \end{cases} \tag{286}$$

For $\kappa_{\mathrm{ms}} \geq 3$, this is the upper bound in Theorem 3.3. The same algorithm and proof also cover $1 \leq \kappa_{\mathrm{ms}} < 3$.

Epoch length and the rate for small condition numbers. At the optimized bias, Appendix B shows that each additional hidden stage costs $\Omega(n)$ IFO calls. Appendix D shows that the gap in the hard instance decreases by $\vartheta^2 = \Theta((\kappa_{\max}/\sqrt{n})^2)$ per stage. Consequently, reducing the gap from Δ to ε requires $\Theta(\log(\Delta/\varepsilon)/[1 + \log(\sqrt{n}/\kappa_{\max})])$ stages.

Lemma E.1 gives $q_T = \Theta(\kappa_{\mathrm{ms}}/\sqrt{n})$ when $T = \Theta(n)$, at $\Theta(n)$ expected cost. The estimator is PAGE (Li et al., 2021), and Proposition E.2 chooses the epoch length in both κ_{ms} regimes. Appendix F gives the conversion to a fixed call budget.

E.1 Contraction and cost over one epoch

One epoch has expected IFO cost at most $n + 3T$ and bounds the expected PL gap by $q_T = \kappa_{\mathrm{ms}}(1 + \sqrt{n})/T$ times the starting gap. The restart choice below makes $q_T \leq 1/4$.

Assume

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L_{\mathrm{ms}}^2\|x - y\|^2 \qquad \forall x, y, \tag{287}$$

and write $\kappa_{\mathrm{ms}} = L_{\mathrm{ms}}/\mu$ when F is μ-PL. Algorithm 1 uses

$$p = \frac{1}{n+1}, \qquad \eta = \frac{1}{L_{\mathrm{ms}}(1 + \sqrt{n})}. \tag{288}$$

Each refresh uses all n components, while each difference update samples one component. The probability p balances an n-component refresh against a one-component difference update, which keeps the expected update cost constant. The step size η satisfies the PAGE stability condition used below. At the epoch start x, compute the exact full gradient, run T PAGE updates, and choose the epoch output uniformly from the iterates used in the standard nonconvex guarantee.

Lemma E.1 (PAGE epoch contraction)**.** *Condition on an arbitrary epoch starting point x, and let $\widehat{x}$ be the random epoch output. Then*

$$\mathbb{E}\big[\|\nabla F(\widehat{x})\|^2 \mid x\big] \le \frac{2(F(x)-F^*)}{\eta T}. \tag{289}$$

If F is globally μ-PL, then

$$\mathbb{E}[F(\widehat{x}) - F^* \mid x] \le q_T(F(x)-F^*), \qquad q_T := \frac{\kappa_{\mathrm{ms}}(1+\sqrt{n})}{T}. \tag{290}$$

The expected number of IFO calls in the epoch is at most $n+3T$.

Proof. Lyapunov descent. For the analysis, extend the estimator recurrence by one auxiliary update to define g_T. Algorithm 1 need not compute it: the epoch output uses only $x_0,\dots,x_{T-1}$, and the next epoch starts with a full gradient. Let $e_t := g_t - \nabla F(x_t)$ and let $\mathbb{E}_t$ condition on the epoch history through the construction of g_t. Objective smoothness and the PAGE estimator update give the two standard recursions

$$\mathbb{E}_t[F(x_{t+1})] \le F(x_t) - \frac{\eta}{2}\|\nabla F(x_t)\|^2 + \frac{\eta}{2}\|e_t\|^2 - \left(\frac{1}{2\eta} - \frac{L_{\mathrm{ms}}}{2}\right)\|x_{t+1}-x_t\|^2, \tag{291}$$

$$\mathbb{E}_t\|e_{t+1}\|^2 \le (1-p)\|e_t\|^2 + (1-p)L_{\mathrm{ms}}^2\|x_{t+1}-x_t\|^2. \tag{292}$$

The second inequality follows by conditioning on the refresh and difference updates and applying mean-squared smoothness; these are the PAGE recursions of Li et al. (2021, proof of Theorem 1).

With an exact full-gradient initialization, define the Lyapunov function

$$\mathcal{L}_t = F(x_t) - F^* + \frac{\eta}{2p}\|g_t - \nabla F(x_t)\|^2. \tag{293}$$

Multiplying (292) by $\eta/(2p)$, adding (291), and taking expectations gives

$$\mathbb{E}[\mathcal{L}_{t+1}] \le \mathbb{E}[\mathcal{L}_t] - \frac{\eta}{2}\mathbb{E}\|\nabla F(x_t)\|^2 - c_\eta \mathbb{E}\|x_{t+1}-x_t\|^2, \qquad c_\eta := \frac{1}{2\eta} - \frac{L_{\mathrm{ms}}}{2} - \frac{(1-p)\eta L_{\mathrm{ms}}^2}{2p}.$$

The condition

$$\eta \le \frac{1}{L_{\mathrm{ms}}\big(1+\sqrt{(1-p)/p}\big)} \tag{294}$$

ensures $c_\eta \ge 0$. Indeed, with $s = \sqrt{(1-p)/p}$ and $t = \eta L_{\mathrm{ms}} \le (1+s)^{-1}$,

$$2\eta c_\eta = 1 - t - s^2t^2 \ge \frac{s}{(1+s)^2} \ge 0.$$

Dropping its nonpositive contribution therefore gives

$$\mathbb{E}[\mathcal{L}_{t+1}] \le \mathbb{E}[\mathcal{L}_t] - \frac{\eta}{2}\mathbb{E}\|\nabla F(x_t)\|^2. \tag{295}$$

Epoch contraction. The parameters in (288) satisfy $(1-p)/p = n$, so the selected stepsize is admissible. Summing (295) for $t = 0,\dots,T-1$ gives

$$\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x_t)\|^2 \le \mathcal{L}_0 - \mathbb{E}[\mathcal{L}_T] \le F(x) - F^*,$$

where the last inequality uses $g_0 = \nabla F(x)$ and $\mathcal{L}_T \ge 0$. Since $\widehat{x}$ is uniform on $x_0,\dots,x_{T-1}$, dividing by $\eta T/2$ proves (289). The pointwise PL inequality gives

$$F(\widehat{x}) - F^* \le \frac{1}{2\mu}\|\nabla F(\widehat{x})\|^2, \tag{296}$$

which proves (290) after substituting $\eta^{-1} = L_{\mathrm{ms}}(1+\sqrt{n})$.

IFO cost. The initial full gradient costs n IFO calls. The algorithm computes $T-1$ estimator updates, omitting the auxiliary g_T. A refresh update costs n calls, while a difference update for one component can be implemented by two calls, one at each point. The expected cost per update is therefore

$$pn + 2(1-p) \le 3, \tag{297}$$

Hence the expected epoch cost is at most $n + 3(T-1) \le n + 3T$, proving the stated bound. □

E.2 Restarts tuned to the condition number

The epoch rule below specializes to $T = \Theta(n)$ for small κ_{ms} and to the standard PAGE schedule with a fixed contraction factor for large κ_{ms} (Li et al., 2021, Corollary 6).

Proposition E.2 (Restarted PAGE)**.** *Let F be globally μ-PL and have mean-squared smoothness constant L_{ms} as in (287), let $\kappa_{\mathrm{ms}} = L_{\mathrm{ms}}/\mu \geq 1$, and suppose $F(x^{(0)}) - F^* \leq \Delta$. Algorithm 1 returns a point with expected gap at most ε using*

$$\begin{cases} O\left(n + \dfrac{n\log(\Delta/\varepsilon)}{1+\log(\sqrt{n}/\kappa_{\mathrm{ms}})}\right), & 1 \leq \kappa_{\mathrm{ms}} \leq \sqrt{n}, \\ O\left(n + \kappa_{\mathrm{ms}}\sqrt{n}\log\dfrac{\Delta}{\varepsilon}\right), & \kappa_{\mathrm{ms}} \geq \sqrt{n} \end{cases} \tag{298}$$

expected IFO calls, provided $0 < \varepsilon \leq c\Delta$ for a universal $c < 1$.

Proof. Use the same epoch choice for every condition number. Set

$$A_{\mathrm{ep}} := \kappa_{\mathrm{ms}}(1+\sqrt{n}), \qquad T := \lceil 4(n + A_{\mathrm{ep}})\rceil, \qquad q_{\mathrm{ep}} := \frac{A_{\mathrm{ep}}}{T} \leq \frac{1}{4}. \tag{299}$$

Repeated conditioning and the tower property give $\mathbb{E}[F(x^{(S_{\mathrm{ep}})}) - F^*] \leq q_{\mathrm{ep}}^{S_{\mathrm{ep}}}\Delta$. Hence Algorithm 1 uses $S_{\mathrm{ep}} = \lceil \log(\Delta/\varepsilon)/\log(1/q_{\mathrm{ep}})\rceil$ epochs, and Lemma E.1 gives the explicit bound on expected cost

$$\mathbb{E}[\#\text{ IFO calls}] \leq (n+3T)\left\lceil \frac{\log(\Delta/\varepsilon)}{\log(1/q_{\mathrm{ep}})}\right\rceil. \tag{300}$$

The ceiling accounts for the case in which one epoch already achieves the target accuracy.

Small condition numbers. Suppose $1 \leq \kappa_{\mathrm{ms}} \leq \sqrt{n}$. Then $A_{\mathrm{ep}} = \Theta(\kappa_{\mathrm{ms}}\sqrt{n})$, and the chosen epoch length satisfies

$$T = \Theta(n), \qquad q_{\mathrm{ep}} = \Theta\left(\frac{\kappa_{\mathrm{ms}}}{\sqrt{n}}\right), \qquad n + 3T = \Theta(n).$$

Thus $\log(1/q_{\mathrm{ep}}) = \Theta(1 + \log(\sqrt{n}/\kappa_{\mathrm{ms}}))$, and (300) gives

$$O\left(n + \frac{n\log(\Delta/\varepsilon)}{1+\log(\sqrt{n}/\kappa_{\mathrm{ms}})}\right). \tag{301}$$

The schedule with a fixed contraction factor underlying Li et al. (2021, Corollary 6) sets $T_{\mathrm{ff}} = \Theta(A_{\mathrm{ep}})$ and keeps only $q_T = \Theta(1)$. In the small-κ_{ms} regime, both epoch choices cost $\Theta(n)$ because each begins with a full gradient. Extending the epoch to $T = \Theta(n)$ leaves its order cost unchanged but strengthens the contraction to $\Theta(\kappa_{\mathrm{ms}}/\sqrt{n})$. Thus this schedule gives $O(n\log(\Delta/\varepsilon))$. Using the sharper $T = \Theta(n)$ contraction produces the denominator in (301); the estimator itself is unchanged.

Large κ_{ms}. If $\kappa_{\mathrm{ms}} \geq \sqrt{n}$, then $A_{\mathrm{ep}} = \Omega(n)$, $T = \Theta(A_{\mathrm{ep}})$, and $q_{\mathrm{ep}} = \Theta(1)$. Equation (300) gives $O(n + \kappa_{\mathrm{ms}}\sqrt{n}\log(\Delta/\varepsilon))$. Hence the schedule reduces to the usual choice with a fixed contraction factor in this regime. These two cases prove (298); the restriction $\varepsilon \leq c\Delta$ absorbs the additive cost of one epoch in the large-κ_{ms} case. □

F Randomization and Expected Complexity

This appendix turns bounds for deterministic algorithms averaged over a hard distribution into bounds for randomized algorithms. It also relates fixed budgets to expected stopping times. The two transfers are, respectively,

$$[\forall A_{\mathrm{det}} :\ Pr_{F\sim\Pi}(\mathsf{Fail}(A_{\mathrm{det}}, F)) \geq p] \Longrightarrow \left[\forall A_\xi\ \exists F \in \operatorname{supp}\Pi :\ \Pr_\xi(\mathsf{Fail}(A_\xi, F)) \geq p\right],$$

and

$$\left[\mathbb{E}\tau \leq Q, \quad \mathbb{E}[\mathrm{err}_F(\widehat{X})] \leq \varepsilon\right] \Longrightarrow \left[\Pr\left(\mathrm{err}_F(\widehat{X}^{(N)}) \leq t\right) \geq 1 - \frac{Q}{N} - \frac{\varepsilon}{t}\right].$$

F.1 From deterministic to randomized algorithms

For each parameter choice, let Π be the distribution over fixed instances sampled before interaction. Consider any deterministic IFO algorithm that makes N calls, chooses its queries from the preceding history, and returns an output measurable with respect to its full IFO history. For every such algorithm, the bounds established in Appendices B–D have the form

$$\Pr_{F\sim\Pi}(\mathsf{Fail}(A,F)) \geq p.$$

Output coverage follows from Lemma B.7 and Lemma B.8 for staged chains and Lemma B.9 for the auxiliary quadratic.

For a randomized algorithm, condition on its private seed $\xi \perp F$ and average the deterministic bound:

$$\mathbb{E}_{F\sim\Pi} \Pr_{\xi}(\mathsf{Fail}(A_\xi, F) \mid F) \geq p. \tag{302}$$

Thus some fixed $F \in \operatorname{supp}\Pi$ has randomized failure probability at least p. Every supported instance satisfies $F(x^{(0)}) - F^* \leq \Delta$ and has individually $L_{\max}$-smooth components; each PL instance additionally satisfies $\|\nabla F(x)\|^2 \geq 2\mu(F(x) - F^*)$ for every x.

F.2 A generic capping lemma

Lemma F.1 (Capping a stopped algorithm)**.** *Fix an instance F and a nonnegative measurable error err_F. Let τ be a stopping time for the algorithm's IFO filtration $(\mathcal{F}_t)_{t\geq 0}$ and let $\widehat{X}$ be $\mathcal{F}_\tau$-measurable. Let $\widehat{X}^{(N)}$ be its output when capped after $N \geq 1$ IFO calls, using any fixed fallback if $\tau > N$. For every $t > 0$,*

$$\Pr\Big(\mathrm{err}_F(\widehat{X}^{(N)}) \leq t\Big) \geq 1 - \frac{\mathbb{E}\tau}{N} - \frac{\mathbb{E}[\mathrm{err}_F(\widehat{X})]}{t}. \tag{303}$$

Proof. On $\{\tau \leq N\} \cap \{\mathrm{err}_F(\widehat{X}) \leq t\}$, the capped output equals $\widehat{X}$. Apply Markov's inequality to the two complement events and take their union; no independence is used. □

F.3 Applications to upper and lower bounds

Upper bound. Fix $\delta \in (0,1)$ and apply Proposition E.2 at target $\delta\varepsilon/2$. If its expected call count is Q, use

$$N = \max\left\{1, \left\lceil \frac{2Q}{\delta} \right\rceil\right\}, \qquad \mathrm{err}_F(x) = F(x) - F^*, \qquad t = \varepsilon.$$

Lemma F.1 gives success probability at least $1-\delta$. For fixed δ, this changes $\log(\Delta/\varepsilon)$ only by an additive constant and proves the upper bound with a fixed budget in Theorem 3.3, together with its extension to $1 \leq \kappa_{\mathrm{ms}} < 3$.

Lower bounds. Suppose that on every fixed instance an algorithm satisfies $\mathbb{E}\tau \leq Q$ and $\mathbb{E}[\mathrm{err}_F(\widehat{X})] \leq \varepsilon$, where

$$\mathrm{err}_F(x) = \begin{cases} \|\nabla F(x)\|, & \text{nonconvex,} \\ F(x) - F^*, & \text{PL.} \end{cases}$$

With $N = \max\{1, \lceil 4Q \rceil\}$ and $t = 4\varepsilon$, Lemma F.1 gives, for every fixed F,

$$\Pr\Big(\mathrm{err}_F(\widehat{X}^{(N)}) \leq 4\varepsilon\Big) \geq \frac{1}{2}. \tag{304}$$

Apply the randomized lower bound with a fixed budget above to this capped algorithm. In each case with multiple stages, $\{\mathrm{err}_F \leq 4\varepsilon\} \subset \{\mathrm{err}_F < 8\varepsilon\}$ and the latter event has probability at most $33/128$; each auxiliary quadratic bounds the former by $5/16$. Both contradict (304).

Absorbing the constant factors from truncation and rescaling the target gives the expected nonconvex lower bound

$$\Omega\left(n + \sqrt{n}\,\frac{\Delta L_{\max}}{\varepsilon^2}\right), \tag{305}$$

and the expected PL lower bound

$$\begin{cases} \Omega\left(n + \dfrac{n\log(\Delta/\varepsilon)}{1+\log(\sqrt{n}/\kappa_{\max})}\right), & 3 \le \kappa_{\max} \le \sqrt{n}, \\ \Omega\left(n + \kappa_{\max}\sqrt{n}\log\dfrac{\Delta}{\varepsilon}\right), & \kappa_{\max} \ge \sqrt{n}. \end{cases} \tag{306}$$

These are the conclusions of Theorems 3.1 and 3.3; the constant target rescaling is absorbed within their stated accuracy regimes.

G Oracle Model and Smoothness Comparisons

This appendix states the formal IFO model, gives the parameter ranges for the cited PL results, and compares individual, mean-squared, and objective smoothness at the same numerical scale. It also verifies that the nonconvex dense chain has comparable individual and mean-squared smoothness constants. Figure 7 displays the two containment relations used throughout the comparison.

G.1 Formal IFO model and theorem conventions

This subsection gives the formal details omitted from Section 3. The function classes used there over arbitrary finite dimensions are

$$\mathcal{F}^n_{\mathrm{ind}}(\Delta, L_{\max}) := \bigcup_{d\ge 1} \mathcal{F}^{n,d}_{\mathrm{ind}}(\Delta, L_{\max}), \qquad \mathcal{F}^n_{\mathrm{ms}}(\Delta, L_{\mathrm{ms}}) := \bigcup_{d\ge 1} \mathcal{F}^{n,d}_{\mathrm{ms}}(\Delta, L_{\mathrm{ms}}), \tag{307}$$

and, for $\mathsf{s} \in \{\mathrm{ind}, \mathrm{ms}\}$,

$$\mathcal{F}^n_{\mathsf{s},\mathrm{PL}}(\Delta, L, \mu) := \bigcup_{d\ge 1} \mathcal{F}^{n,d}_{\mathsf{s},\mathrm{PL}}(\Delta, L, \mu). \tag{308}$$

Fix a dimension d. Let ξ be the algorithm's private random seed and let $Y_t = \mathcal{O}_{I_t}(X_t)$. Define the filtration

$$\mathcal{F}_0 := \sigma(\xi), \qquad \mathcal{F}_t := \sigma\big(\xi, (I_s, X_s, Y_s)_{1\le s\le t}\big), \quad t \ge 1. \tag{309}$$

A randomized IFO algorithm chooses (I_t, X_t) as an $\mathcal{F}_{t-1}$-measurable random variable, stops at a stopping time τ with respect to $(\mathcal{F}_t)_{t\ge 0}$, and returns an $\mathcal{F}_\tau$-measurable output $\widehat{X}$. All spaces are standard Borel. The resulting class is denoted by

$$\mathcal{A}^{\mathrm{rand}}_{\mathrm{IFO}}(d). \tag{310}$$

It permits repeated indices, arbitrary query points, randomized stopping, and unqueried outputs.

For a dimension-free function class, an algorithm is a family indexed by dimension:

$$\mathcal{A}^{\mathrm{rand}}_{\mathrm{IFO}} := \left\{A = (A_d)_{d\ge 1} : A_d \in \mathcal{A}^{\mathrm{rand}}_{\mathrm{IFO}}(d) \text{ for every } d\right\}. \tag{311}$$

Let $\mathcal{A}^{\mathrm{span}}_{\mathrm{IFO}}(d)$ and $\mathcal{A}^{\mathrm{zero}}_{\mathrm{IFO}}(d)$ denote the corresponding linear-span and zero-respecting subclasses, following the usual conventions for lower bounds (Zhou & Gu, 2019; Carmon et al., 2021). They restrict queries to the previous linear span and revealed coordinate support, respectively. With $X_0 = 0$ and $d \ge 2$,

$$\mathcal{A}^{\mathrm{span}}_{\mathrm{IFO}}(d) \subseteq \mathcal{A}^{\mathrm{zero}}_{\mathrm{IFO}}(d) \subseteq \mathcal{A}^{\mathrm{rand}}_{\mathrm{IFO}}(d). \tag{312}$$

For $\mathcal{F} = \bigcup_{d\ge 1} \mathcal{F}^d$ and an error functional $\mathcal{E}_F$, the complexity for a fixed budget used in the main text is

$$\mathrm{Comp}^{\mathrm{IFO}}_{\varepsilon,\delta}(\mathcal{F}; \mathcal{E}) := \inf\left\{ N \in \mathbb{N}_0 : \begin{array}{l} \exists A = (A_d)_{d\ge 1} \in \mathcal{A}^{\mathrm{rand}}_{\mathrm{IFO}}, \\ A_d \text{ makes at most } N \text{ calls for every } d, \text{ and} \\ \sup\limits_{d\ge 1} \sup\limits_{F\in\mathcal{F}^d} \Pr\limits_{\xi}\Big(\mathcal{E}_F(\widehat{X}_{A_d}) > \varepsilon\Big) \le \delta \end{array} \right\}. \tag{313}$$

Once d and F are fixed, $\Pr_\xi$ refers only to the private seed. In the proofs of the lower bounds, $\mathbb{P}$ includes the auxiliary random hard instance as well. Calligraphic $\mathcal{E}$ denotes an error criterion, whereas $\mathbb{E}$ denotes expectation.

Nonconvex theorem parameters. The universal threshold and failure probability in Theorem 3.1 may be taken as $n_0 = 2^{10} = 1024$ and $p_0 = 11/16$, with universal $c_0 > 0$ and a universal implicit constant in (5). Appendix C proves the corresponding claim for a fixed budget and may use $D = \lceil c_D[1 + \log((N+2)(T_{\mathrm{chain}}+1))]\rceil$ coordinates per stage, where c_D is universal. For $n < n_0$, replicating one hard function across all components gives the same asymptotic order after changing universal constants (Carmon et al., 2021; Emmenegger et al., 2022).

The PL lower bound uses the same failure probability $11/16$. Its full quantifiers for a fixed budget are restated at the beginning of Appendix D. Appendix F converts both lower bounds to expected call budgets in the worst case and converts the Restarted PAGE guarantee to complexity with a fixed budget and constant success probability.

Parameter ranges for the PL bounds in Table 1. Yue et al. (2023) count full-gradient calls; their Theorem 6 inherits the requirement $\kappa_F > 2632476000/361$ on the condition number from Theorem 4 and assumes $\varepsilon \le \Delta/16$. The logarithmic Theorem 3.4 of Bai et al. (2024) requires $\kappa_{\mathrm{ms}} \ge 729196\sqrt{n}$ and $\varepsilon < \Delta/200$, whereas their Theorem 3.5, valid for all κ_{ms}, gives only $\Omega(n)$. The new upper entries in Table 1 use κ_{ms}, and the new lower entries use $\kappa_{\max}$. The latter hard instances are individually smooth and hence also lie in the mean-squared class.

G.2 Assumptions on the function classes

Table 2 collects the three regularity assumptions used in the main text and representative analyses for each. Proposition G.1 then gives counterexamples to three inequalities that require separate control of each component when the same numerical smoothness constant is substituted under mean-squared smoothness.

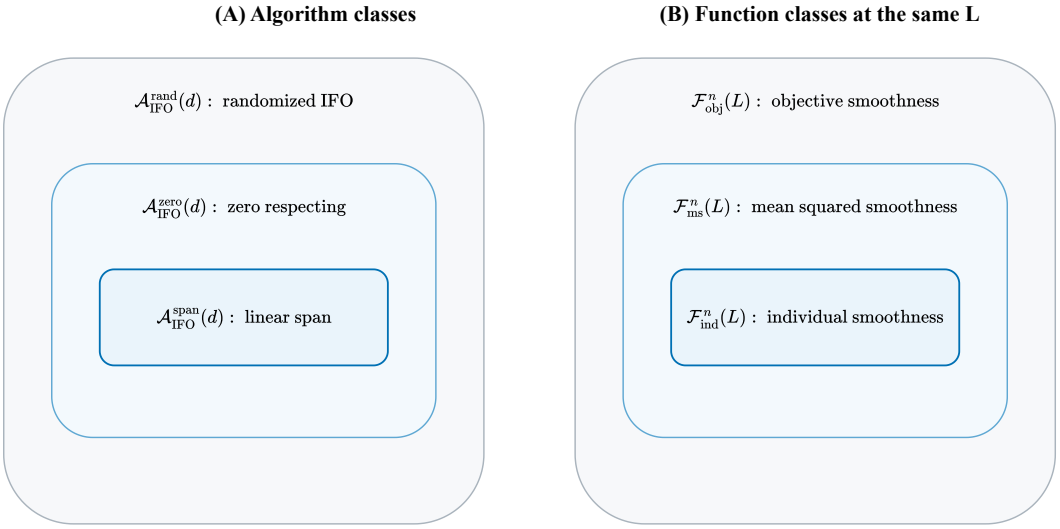

Figure 7: Containment relations for the algorithm and function classes. The lower bounds allow algorithms from the outermost class in panel (A), while the hard instances belong to the innermost class in panel (B). For the function classes, the inclusions use the same numerical smoothness bound L.

Table 2: Definitions and representative analyses for the function classes in Figure 7.

Class	Defining regularity	Representative sources
$\mathcal{F}_{\mathrm{ind}}^{n}$	$\max_i \|\nabla f_i(x) - \nabla f_i(y)\| \le L_{\max}\|x-y\|$	Finite-sum lower bounds and analyses of component gradient updates or stored gradients (Zhou & Gu, 2019; Han et al., 2024; Emmenegger et al., 2022; Defazio et al., 2014; Kavis et al., 2022); this work.
$\mathcal{F}_{\mathrm{ms}}^{n}$	$\left(n^{-1}\sum_i \|\nabla f_i(x) - \nabla f_i(y)\|^2\right)^{1/2} \le L_{\mathrm{ms}}\|x-y\|$	PAGE and SPIDER (Fang et al., 2018; Li et al., 2021); Zhou & Gu (2019, Definition 3.2); PL lower bounds in Bai et al. (2024).
$\mathcal{F}_{\mathrm{obj}}^{n}$	$\|\nabla F(x) - \nabla F(y)\| \le L_F\|x-y\|$	Lower bounds for stationary points of the full objective (Carmon et al., 2020; 2021).

The first inclusion follows by averaging squared bounds; the second follows from Jensen's inequality. A lower bound on a larger aggregate class does not automatically hold on the smaller individual class.

Proposition G.1 (Mean-squared smoothness does not imply componentwise bounds at the same constant)**.** *Fix $n \geq 2$ and $\ell > 0$. For each of the following inequalities, there is a convex quadratic finite sum with mean-squared smoothness constant exactly ℓ for which the inequality fails:*

(i) the same smoothness bound for every component

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq \ell\|x - y\| \qquad \text{for every } i; \tag{314}$$

(ii) the componentwise co-coercivity bound

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(u) - \nabla f_i(x)\|^2 \leq 2\ell\big(F(u) - F(x) - \langle \nabla F(x), u - x\rangle\big); \tag{315}$$

(iii) the bound with a separate reference point for each component

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f_i(\phi_i)\|^2 \leq \ell^2 \frac{1}{n}\sum_{i=1}^{n} \|x - \phi_i\|^2. \tag{316}$$

Proof. Counterexample to a uniform component bound. For (314), consider the following one-dimensional quadratic family with one nonzero component:

$$f_1(x) = \frac{\sqrt{n}\,\ell}{2}x^2, \qquad f_2(x) = \cdots = f_n(x) = 0. \tag{317}$$

For every $x, y \in \mathbb{R}$,

$$\frac{1}{n}\sum_{i=1}^{n} |\nabla f_i(x) - \nabla f_i(y)|^2 = \ell^2|x - y|^2, \tag{318}$$

so the mean-squared smoothness constant is exactly ℓ, whereas the first component has smoothness constant $\sqrt{n}\,\ell$ and violates (314) whenever $x \neq y$.

Thus the gradient of one component can have a Lipschitz constant $\sqrt{n}$ times larger than the mean-squared constant. A proof step can therefore fail if it substitutes the mean-squared constant into a descent or self-bounding inequality for the sampled component.

Counterexample to the componentwise co-coercivity bound. The same family also disproves (315). At $x = 0$ and $u = 1$,

$$\frac{1}{n}\sum_{i=1}^{n} |\nabla f_i(1) - \nabla f_i(0)|^2 = \ell^2, \qquad F(1) - F(0) - \langle \nabla F(0), 1\rangle = \frac{\ell}{2\sqrt{n}}. \tag{319}$$

Thus the right-hand side of (315) is only $\ell^2/\sqrt{n}$. More generally, the coefficient multiplying the term of Bregman type on the right-hand side must be at least $2\sqrt{n}\,\ell$, or equivalently the smoothness parameter in the displayed $2C$ form must satisfy $C \geq \sqrt{n}\,\ell$.

Counterexample with separate reference points for the components. To disprove (316), set the ambient dimension to $d = n$ and define

$$f_i(z) = \frac{\sqrt{n}\,\ell}{2}z_i^2, \qquad i \in [n]. \tag{320}$$

For all $z, w \in \mathbb{R}^n$,

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(z) - \nabla f_i(w)\|^2 = \ell^2\|z - w\|^2, \tag{321}$$

again giving exact mean-squared smoothness constant ℓ. Take $x = 0$ and $\phi_i = e_i$. Then the left- and right-hand sides of (316) are, respectively,

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(0) - \nabla f_i(e_i)\|^2 = n\ell^2, \qquad \ell^2\frac{1}{n}\sum_{i=1}^{n} \|e_i\|^2 = \ell^2. \tag{322}$$

Mean-squared smoothness controls all components at one common pair (z, w). It does not imply the displayed estimate when component i is paired with its own reference point ϕ_i. □

The three counterexamples concern, respectively, descent and self-bounding arguments applied to a sampled component (Vaswani et al., 2019; Loizou et al., 2021; Mishchenko et al., 2020; Kavis et al., 2022), variance-to-Bregman bounds (Liu et al., 2022), and bounds using a separate reference point for each component (Defazio et al., 2014, Lemma 2). These inequalities therefore do not follow from mean-squared smoothness alone at the same numerical constant.

G.3 Mean-squared smoothness of the nonconvex dense chain

The following lemma shows that the nonconvex dense chain has mean-squared smoothness comparable to its individual smoothness. The PL lower bounds do not require this comparison of least constants: individual smoothness with bound L already implies mean-squared smoothness with the same bound L.

Lemma G.2 (The nonconvex dense chain has comparable mean-squared smoothness)**.** *Suppose $T \geq 2$ and scale the components of the dense chain by $f_i(y) = \alpha R_i(\beta y)$, where*

$$\alpha\beta^2 = \frac{L_{\max}\rho}{H_{\main}} \tag{323}$$

as in (148). Here $a = 5/32$ and $b = 9/32$ are the lower and upper gate thresholds, and $H_{\main} = 341$, defined in (127), is the unscaled component Hessian bound. If $L_{\ms}$ denotes their least mean-squared smoothness constant, then

$$\frac{8(1-b^2)^{3/2}}{H_{\main}} L_{\max} \leq L_{\ms} \leq L_{\max}. \tag{324}$$

In particular, $L_{\ms} = \Theta(L_{\max})$ with universal constants, independently of n, D, and T.

Proof. Witness point. Since $G(b) - G(a) = 1$ and $b - a = 1/8$, the mean value theorem gives an $s_0 \in (a, b)$ with $G'(s_0) = 8$. Write $u = u_{\Theta^{(1)}}$ and choose $z_1 = ru$, where $r = s_0/\sqrt{1 - s_0^2}$, so that $q_1(z_1) = s_0$ and

$$\nabla q_1(z_1) = (1 - s_0^2)^{3/2} u. \tag{325}$$

Unscaled Hessian block. Set $z_2 = \cdots = z_T = 0$ and let v be the unit vector supported on the first block with $v_1 = u$. In the second output block, the only Hessian term applied to v is the cross derivative of the second chain link. Hence, for every component i,

$$\left\|(\nabla^2 R_i(z)v)_2\right\| = G'(s_0)(1 - s_0^2)^{3/2} \|\nabla \widetilde{q}_{i2}(0)\| \geq \frac{8(1-b^2)^{3/2}}{\rho}, \tag{326}$$

because $\nabla \widetilde{q}_{i2}(0) = s_{i2}/\rho$ and $\|s_{i2}\| = 1$.

Difference quotient argument. After scaling, $\nabla^2 f_i(z/\beta) = \alpha\beta^2 \nabla^2 R_i(z)$ and $\alpha\beta^2 = L_{\max}\rho/H_{\main}$. Mean-squared smoothness implies, for every unit vector v and every $h \neq 0$,

$$\left(\frac{1}{n}\sum_{i=1}^{n} \left\| \frac{\nabla f_i(x + hv) - \nabla f_i(x)}{h} \right\|^2\right)^{1/2} \leq L_{\ms}. \tag{327}$$

Letting $h \to 0$ at any fixed x gives

$$L_{\ms} \geq \left(\frac{1}{n}\sum_{i=1}^{n} \|\nabla^2 f_i(x)v\|^2\right)^{1/2}. \tag{328}$$

Final scaling and substitution. At $x = z/\beta$ and for the v chosen above, the second block gives $\|\nabla^2 f_i(x)v\| \geq 8\alpha\beta^2(1 - b^2)^{3/2}/\rho$ for every i. Substituting $\alpha\beta^2 = L_{\max}\rho/H_{\main}$ in (328) proves the lower bound in (324); the upper bound follows from individual $L_{\max}$-smoothness. □