OPTIMAL STOCHASTIC OPTIMIZATION WITH BOUNDED GRADIENT NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Lower bounds for stochastic optimization usually rely on oracles that occasionally return very large gradient errors, so it is unclear whether they remain valid when the noise is uniformly bounded, a setting Arjevani et al. left open. We show that bounding the noise does not make optimization easier. For smooth nonconvex, Polyak–Łojasiewicz (PL), strongly convex, and convex objectives, we prove matching upper and lower bounds under almost-surely bounded gradient noise. In particular, finding an ε-stationary point requires $\Theta(\Delta L\sigma^2\varepsilon^{-4})$ stochastic gradients in the noise-dominated regime, and $\Theta(\Delta H\sigma\varepsilon^{-3})$ when the centered noise is H-Lipschitz, so stochastic gradient descent and variance reduction remain optimal. We also find a separation between PL and strongly convex objectives: in the noise-dominated regime, PL objectives require $\kappa\sigma^2/(\mu\eta)$ samples rather than $\sigma^2/(\mu\eta)$. Further results cover additive noise, finite sums, relative noise, higher-order smoothness, and limits on sample reuse. Our lower bounds hold for arbitrary randomized algorithms that may re-evaluate samples adaptively and observe exact function values. They rest on bounded observations drawn from the Poisson kernel on the sphere, together with an information argument that controls all responses from reused samples.

1 INTRODUCTION

Stochastic optimization methods access the objective only through noisy gradient estimates, and the assumptions placed on this noise determine both the guarantees one can prove and the algorithms one should use. Much recent attention has focused on heavy tails: Şimşekli et al. (2019) observe heavy-tailed gradient noise when training several neural-network architectures, and Zhang et al. (2020) relate such noise to the benefits of clipping in attention models. Many problems sit at the opposite extreme, where every gradient error is uniformly bounded. We ask what this stronger assumption buys: which complexity barriers of stochastic optimization persist when large errors cannot occur?

Under uniformly bounded noise, every stochastic gradient lies within a fixed distance σ of the population gradient, at every query point. Binary logistic regression with feature vectors of norm at most B is an example: each sample gradient has norm at most B at every parameter, so its centered noise has norm at most $2B$. Adding a differentiable deterministic regularizer leaves this centered noise unchanged, even when the regularizer's gradient is unbounded.

Because bounded noise is a special case of bounded variance, upper bounds proved under unbiasedness and bounded variance carry over unchanged. Lower bounds do not. The standard hard oracles for the variance-bounded class rely on rare, amplified responses whose errors grow as the target accuracy shrinks, and Arjevani et al. (2023) identify almost-surely bounded noise as an open direction. Whether the classical rates remain optimal under bounded noise is therefore a question about lower bounds. We answer it for convex, strongly convex, Polyak–Łojasiewicz (PL), and smooth nonconvex objectives. For each class, we fix the information revealed by the oracle and measure accuracy by the expected objective gap (convex and PL) or the expected gradient norm (nonconvex), and we prove matching upper and lower bounds.

Our main finding is that bounding the noise does not make optimization easier. For smooth nonconvex objectives, finding an ε-stationary point still requires order $\Delta L\sigma^2\varepsilon^{-4}$ stochastic gradients in the noise-dominated regime, and order $\Delta H\sigma\varepsilon^{-3}$ when the centered noise is H-Lipschitz, so stochastic

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.

gradient descent and variance reduction remain optimal. The convex rates likewise recover the classical benchmarks (Agarwal et al., 2012; Lan, 2012). The bounds also separate PL objectives from strongly convex ones: in the noise-dominated regime, a general PL objective can require $\kappa\sigma^2/(\mu\eta)$ evaluations, whereas a strongly convex objective requires only $\sigma^2/(\mu\eta)$. This comparison stays exact as the condition number κ approaches one.

Noise regularity and sample access refine these rates. Separating the smoothness of the population objective from that of the centered noise lets us quantify when evaluating one sample at two points lowers the statistical cost. Under additive noise, such differences cancel the error exactly, yet estimating a single population gradient still requires order σ^2/ε^2 samples; when each sample can be evaluated only finitely often, we also determine how many distinct samples are needed. In finite sums with n components, exact averaging replaces σ^2/ε^2 by $\min\{n, \sigma^2/\varepsilon^2\}$ in the nonconvex statistical cost. We further treat relative-noise bounds and higher-order population smoothness, with matching rates for the latter in a stated regime. Together, these results separate three resources that coincide when every evaluation uses a fresh sample: gradient evaluations, simultaneous query rounds, and distinct samples.

The main difficulty in the lower bounds is to keep every sample gradient unbiased and within the noise radius while still revealing the hidden instance slowly, even when the algorithm re-evaluates a sample at adaptively chosen points. We encode each hidden direction in a bounded observation drawn from the Poisson kernel on the sphere, and build scalar sample losses around it to realize each objective geometry. An information argument then bounds everything the algorithm can learn from a sample, including repeated gradient evaluations and exact function values at queried points. The matching algorithms are stochastic gradient descent, gradient tracking with paired evaluations, and acceleration, chosen according to the geometry and regularity assumptions.

Section 2 defines the models and criteria, Section 3 states the rates, Section 4 explains the constructions and algorithms, and Section 5 discusses prior work. Complete proofs appear in the appendices.

2 Models and complexity criteria

We consider differentiable objectives $F : \mathbb{R}^d \to \mathbb{R}$ with finite infimum $F_\star$, an initial point x_0, and L-Lipschitz gradient. Unless a convex radius assumption is stated, $F(x_0) - F_\star \le \Delta$. A sample ξ specifies a differentiable scalar loss $f(\cdot;\xi)$ with gradient $g(x;\xi)$, and

$$\mathbb{E}g(x;\xi) = \nabla F(x), \qquad \zeta(x;\xi) := g(x;\xi) - \nabla F(x), \qquad \sup_{x\in\mathbb{R}^d} \|\zeta(x;\xi)\| \le \sigma \quad \text{almost surely.} \tag{1}$$

The exceptional null set in (1) is common to all x. In the unbiasedness condition, we take the expectation over a fresh sample. If a point is chosen adaptively using that sample, we do not assume conditional unbiasedness. The sample losses in our lower bounds are bounded below and, when a sample-smoothness constant is prescribed, are continuously differentiable with Lipschitz gradients.

We use three access models. A *fresh-sample oracle* draws an independent sample for each evaluation. A *retained-sample oracle* provides a handle to a sample, allowing later evaluations at adaptively chosen points without revealing the sample's description. A *K-point oracle* draws one fresh sample after an algorithm chooses at most K points, then returns its evaluations at all of those points. Within a round, all points are chosen before any responses are observed. We count gradient evaluations for retained-sample methods and rounds for K-point methods. An upper-bound algorithm may use fresh samples even when retention is allowed. The lower bounds also allow exact $F(x)$ and $f(x;\xi)$ at each charged evaluation point, but the models do not grant free access to sample descriptions or free function evaluations at additional points.

For finite sums, $F = n^{-1}\sum_{i=1}^n f_i$, the algorithm may select an index, retain it, and choose all later queries adaptively. The centered bound is $\sup_{i,x} \|\nabla f_i(x) - \nabla F(x)\| \le \sigma$. Thus the finite-sum lower bounds apply to active index selection, in addition to random sampling. Throughout, algorithms may be randomized, and their output need not be a point they have queried.

For stationarity the target is $\mathbb{E}\,\|\nabla F(\widehat{x})\| \le \varepsilon$. We write

$$Q := \frac{\Delta L}{\varepsilon^2}, \qquad m := \frac{\sigma^2}{\varepsilon^2}, \qquad r := \min\{n, m\}. \tag{2}$$

Table 1: Optimal online complexities under uniformly bounded gradient noise. The table counts evaluations. The same rates hold for K-point rounds under the restrictions on K stated in the corresponding theorems. The notation is defined in (2)–(3). Centered-noise regularity refers to (4).

Objective	Noise assumptions	Optimal complexity
Nonconvex	Bounded	$Q + Qm$
Nonconvex	Centered regularity H	$Q + m + \min\{Qm, \Delta H\sigma/\varepsilon^3\}$
Nonconvex	Additive	$Q + m$
PL	Bounded	$D_\kappa + \kappa S$
PL	Centered regularity H	$D_\kappa + S + \min\{\kappa S, (H/\mu)\sqrt{S}\}$
PL	Additive	$D_\kappa + S$
Strongly convex	Bounded	$E_\kappa + S$
Convex, radius R	Bounded	$\sqrt{LR^2/\eta} + \sigma^2 R^2/\eta^2$

For objective accuracy the target is $\mathbb{E}[F(\widehat{x}) - F_\star] \le \eta$. A PL objective satisfies $\|\nabla F(x)\|^2 \ge 2\mu(F(x) - F_\star)$ globally. For these objectives and for μ-strongly convex objectives, let

$$\kappa := L/\mu \ge 1, \quad S := \frac{\sigma^2}{\mu\eta}, \quad D_\kappa := 1 + \frac{\log(\Delta/\eta)}{\log(\kappa/(\kappa-1))}, \quad E_\kappa := 1 + \frac{\log(\Delta/\eta)}{\log((\sqrt{\kappa}+1)/(\sqrt{\kappa}-1))}. \tag{3}$$

We set $D_1 = E_1 = 1$. We measure minimax complexity over all finite dimensions. The hard dimension may depend on the evaluation budget and on K, which are fixed before the instance is chosen. Constants in $\Theta(\cdot)$ are universal, including with respect to finite K; constants with a subscript p may depend on a fixed derivative order p. These statements are unchanged by integer rounding.

3 Matching complexity bounds

Table 1 compares minimax rates across objective geometries and noise assumptions. Each row gives the minimax rate for one objective class and noise assumption; the corresponding accuracy criteria are defined in Section 2. The theorems specify the parameter ranges and finite-sum variants; evaluation counts, simultaneous rounds, and distinct-sample counts are stated separately.

3.1 Smooth nonconvex objectives and sample reuse

The regularity relevant to a gradient difference is the variation of its centered noise. We consider

$$\mathbb{E}\,\|\zeta(x;\xi) - \zeta(y;\xi)\|^2 \le H^2\,\|x - y\|^2\,. \tag{4}$$

For lower bounds we impose the stronger samplewise version of (4). Population smoothness and H are separate assumptions: a large deterministic component can make the sample gradient vary rapidly without increasing the variation of the noise.

Theorem 1 (Nonconvex objectives)**.** *There is a universal $c > 0$ such that, for $0 < \varepsilon \le c\sqrt{\Delta L}$, the optimal retained-sample evaluation complexity under (1) is $\Theta(Q + Qm)$ (5). Under the additional condition (4), it is*

$$\Theta\left(Q + m + \min\left\{Qm, \frac{\Delta H\sigma}{\varepsilon^3}\right\}\right). \tag{6}$$

The same rates hold for K-point round complexity for every fixed finite $K \ge 1$ in (5), and every fixed finite $K \ge 2$ in (6). The upper bounds use fresh samples in (5) and at most two evaluations of each sample in (6).

The lower-bound samples in (5) can be chosen smooth, without prescribing a common numerical smoothness constant. If every sample instead has an L_s-Lipschitz gradient, set $L_* := \min\{L, L_s\}$ and $Q_* := \Delta L_*/\varepsilon^2$. The corresponding rate is

$$\Theta\left(Q_* + m + \min\left\{Q_* m, \frac{\Delta L_s \sigma}{\varepsilon^3}\right\}\right), \qquad \varepsilon \le c\sqrt{\Delta L_*}. \tag{7}$$

Indeed F is L_*-smooth and its centered noise is at most (L_s+L_*)-Lipschitz. For the lower bound, we allocate smoothness budgets $L_*/2$ to the population gradient and $L_s/2$ to the centered noise, so each sample gradient is L_s-Lipschitz. When $L_s = L$, the expression is equivalent to $\Theta(Q + m + Q\sqrt{m})$. When $H = 0$, sample gradients differ from ∇F by constant vectors on a common probability-one set (Lemma 38), and (6) reduces to $\Theta(Q + m)$.

The crossover between the two stochastic terms has a simple form. In the regime $m \geq 1$, tracking is favorable when $H \lesssim L\sigma/\varepsilon$. When $H \asymp L$ and $m \geq 1$, the interaction term is of order $Q\sqrt{m}$; the full rate also includes $Q + m$. At $H = 0$, taking a difference removes the noise entirely, but an initial estimate of one population gradient is still necessary. These transitions explain why a support-radius assumption and a sample-smoothness assumption need separate parameters.

The term m accounts for estimating an unknown mean even when gradient differences are exact. The minimum in (6) compares the cost of independent noisy steps with the cost of tracking the gradient through paired evaluations. For every fixed finite number K of simultaneous query points, the stated minimax round rates remain unchanged when the worst case ranges over all finite dimensions. This conclusion depends on choosing the hard dimension after K is fixed.

Theorem 2 (Finite sums). *Under the same stationarity and accuracy assumptions, active access to n components has optimal evaluation complexity*

$$\Theta\big(Q(1+r)\big), \qquad r = \min\{n, m\}. \tag{8}$$

If the centered component gradients are H-Lipschitz, the rate is

$$\Theta\left(Q + r + \min\left\{Qr, \frac{\Delta H\sqrt{r}}{\varepsilon^2}\right\}\right). \tag{9}$$

With individual component smoothness L_s, set $L_* := \min\{L, L_s\}$ and require $0 < \varepsilon \leq c\sqrt{\Delta L_*}$. Replace Q by $Q_* := \Delta L_*/\varepsilon^2$ and H by L_s in (9). In particular, common smoothness $L_s = L$ gives $\Theta(Q + r + Q\sqrt{r})$.

The replacement of m by r reflects two ways to estimate an anchor gradient: an average of independent component samples, each with centered second moment at most σ^2, or the exact average of all n components. For the lower bound, the finite collection must average to exactly F, a property that independent draws from the online construction do not provide. We obtain it with a balanced table of continuous observations and control the joint distribution of all indices selected by the algorithm.

Theorem 3 (Relative noise). *Suppose in addition that $\mathbb{E}\,\|\zeta(x;\xi)\|^2 \leq \rho\,\|\nabla F(x)\|^2$ for all x. The optimal complexities are*

$$\Theta(Q[1 + \min\{m, \rho\}]) \quad \textit{and} \quad \Theta(Q[1 + \min\{n, m, \rho\}]) \tag{10}$$

for the online and finite-sum models, respectively. The online round statement holds for every fixed finite $K \geq 1$. The lower bounds satisfy the stronger pointwise relative bound $\|\zeta(x;\xi)\|^2 \leq \rho\,\|\nabla F(x)\|^2$.

The relative bound is imposed on the gradient error; it does not require a shared sample minimizer. Vaswani et al. (2019) establish the strong-growth upper bound; our result includes the absolute-noise and finite-population caps. The lower construction places noise only where the population gradient is bounded away from zero.

3.2 PL geometry, strong convexity, and convexity

The distinction between PL geometry and strong convexity appears in both the deterministic and stochastic terms.

Theorem 4 (Global PL objectives). *For $0 < \eta \leq \Delta/128$ and $\kappa \geq 1$, the optimal retained-sample evaluation complexity under (1) is $\Theta(D_\kappa + \kappa S)$ (11). Under (4), the optimal retained-sample evaluation complexity is*

$$\Theta\left(D_\kappa + S + \min\left\{\kappa S, \frac{H}{\mu}\sqrt{S}\right\}\right). \tag{12}$$

For sample smoothness $L_s \geq \mu$, set $\kappa_* := \min\{L, L_s\}/\mu$. The optimal rate is

$$\Theta\left(D_{\kappa_*} + S + \min\left\{\kappa_* S, \frac{L_s}{\mu}\sqrt{S}\right\}\right). \tag{13}$$

The round bounds hold for every fixed finite $K \geq 1$ in (11) and $K \geq 2$ in (12)–(13). The same evaluation rates hold for active finite sums when $n \geq S$.

For large κ, D_κ has order $\kappa \log(\Delta/\eta)$. Its exact form also captures the improvement as $\kappa \downarrow 1$. When $L_s = L$, the complete rate in (13) is equivalently $\Theta(D_\kappa + S + \kappa\sqrt{S})$. Additive noise again leaves the estimation cost S. For $1 \leq \kappa \leq 3$, fresh samples attain the general and centered-noise rate $\Theta(D_\kappa + S)$. For the sample-smooth rate (13), this conclusion uses κ_* in place of κ. The finite-sum statement has a further restriction: $n \geq S$ is the regime in which the online statistical obstruction fits within a balanced finite table.

Theorem 5 (Strongly convex objectives). *For μ-strongly convex, L-smooth objectives, $0 < \eta \leq \Delta/128$, and $\kappa \geq 1$, the optimal evaluation complexity and the optimal K-point round complexity for every fixed finite $K \geq 1$ are $\Theta(E_\kappa + S)$ (14). The upper bound uses fresh samples. The lower bound holds when every sample is itself μ-strongly convex and L-smooth, with additive noise. Consequently, imposing any centered-noise smoothness bound $H \geq 0$ leaves (14) unchanged. A separate sample-smoothness bound $L_s \geq \mu$ replaces L by $\min\{L, L_s\}$, so $\kappa_* := \min\{L, L_s\}/\mu \geq 1$. Active finite sums have the same evaluation rate when $n \geq S$.*

The two geometry classes already differ at fixed accuracy: in the noise-dominated regime, a general PL objective can require κS evaluations, whereas a strongly convex objective requires S. When κ is bounded away from one, strong convexity also reduces the deterministic factor from order κ to order $\sqrt{\kappa}$, with the same logarithmic dependence on accuracy. The lower bound in Theorem 5 is uniform throughout the stated accuracy range; it does not require accuracy to shrink with the condition number.

Theorem 6 (Smooth convex objectives). *Let F be convex and L-smooth, with a minimizer $x_\star$ satisfying $\|x_0 - x_\star\| \leq R$. For $0 < \eta \leq cLR^2$, the optimal retained-sample evaluation complexity and the optimal K-point round complexity for every fixed finite $K \geq 1$ are*

$$\Theta\left(\sqrt{\frac{LR^2}{\eta}} + \frac{\sigma^2 R^2}{\eta^2}\right). \tag{15}$$

The upper bound uses fresh samples. The lower-bound sample losses are convex and smooth, and the statistical lower bound uses additive noise.

The convex expressions are classical in the moment-bounded oracle model. We establish these rates when each response is the gradient of a scalar sample loss and its centered error has bounded support. The lower bounds also allow sample reuse and exact sample and population values at counted query points. The deterministic convex and strongly convex constructions are proved directly in Appendix E and Appendix C.

3.3 Higher-order population smoothness and distinct samples

Higher derivatives of F can improve stationarity complexity even if no higher derivative of a sample loss is available. Let $p \geq 2$ be fixed and assume $D^p F$ is L_p-Lipschitz in multilinear operator norm. Define

$$A_p := \frac{\Delta L_p^{1/p}}{\varepsilon^{1+1/p}}, \qquad C_p := \min\{Q, A_p\}. \tag{16}$$

Theorem 7 (Higher-order smooth objectives). *For $0 < \varepsilon \leq c\sqrt{\Delta L}$, the optimal complexity is*

$$\Theta_p\big((1+m)(1+C_p)\big) \quad \textit{whenever} \quad Q \leq (1+m)(1+C_p). \tag{17}$$

The lower bound allows retained samples and charged exact values. A fresh-sample algorithm attains the rate. For fixed positive Δ, L, L_p, σ and sufficiently small ε, it becomes

$$\Theta_p\left(\Delta L_p^{1/p} \sigma^2 \varepsilon^{-(3+1/p)}\right). \tag{18}$$

In particular, Lipschitz Hessians give exponent $7/2$.

The proof gives an upper bound $O_p(Q + m + mC_p)$ and a lower bound $\Omega_p((1+m)(1+C_p))$. These bounds match under the condition in (17). This restriction reflects the deterministic cost Q

that remains in the upper bound. The smoothness assumption here controls only derivatives of the population objective.

A separate resource question arises for additive noise. Here an evaluation has the form $g(x;\xi) = \nabla F(x) + Z_\xi$ with $\|Z_\xi\| \le \sigma$. Fix a finite integer $J \ge 2$ and allow at most J adaptive evaluations of each retained sample. Unlike a simultaneous K-point round, these evaluations may depend on earlier responses from that sample.

Corollary 8 (Distinct samples under additive noise)**.** *Under the smooth nonconvex assumptions of Theorem 1, for* $0 < \varepsilon \le c\sqrt{\Delta L}$ *the optimal number of distinct samples needed for stationarity is* $\Theta(1 + m + Q/J)$ (19)*, and this number can be achieved with* $\Theta(Q + m)$ *total gradient evaluations.*

To attain the rate, estimate $\nabla F(x_0)$ once using $O(1+m)$ independent samples. Each later handle is evaluated at x_0 and then at up to $J-1$ adaptively chosen iterates. Subtracting its response at x_0 cancels that handle's noise exactly. The resulting descent method has one fixed estimation error, which is controlled in the usual smooth descent inequality. The lower bound combines mean estimation, the total deterministic number of evaluations, and the cap J.

4 Proof Ideas

Every lower bound in Section 3 needs hard instances with three properties at once: sample gradients are unbiased, their errors stay within radius σ, and an algorithm that reuses samples adaptively still learns the instance slowly. Standard constructions obtain slow learning from rare, large responses, which the second property rules out. We instead hide the instance in bounded observations that are individually uninformative (Section 4.1), build scalar sample losses around them for each objective geometry (Sections 4.2 and 4.3), and close with the matching algorithms (Section 4.4).

4.1 A continuous observation of a hidden direction

Each construction consists of T *stages*, each with its own hidden unit vector in a separate block of coordinates of dimension D, and each sample carries one observation per stage. Fix one stage with hidden direction θ, uniform on $\mathbb{S}^{D-1}$. Conditional on θ, the sample's observation $S \in \mathbb{S}^{D-1}$ is drawn from the Poisson kernel, with density relative to uniform surface measure

$$p_{\beta\theta}(s) = \frac{1-\beta^2}{\|s-\beta\theta\|^D}, \qquad \mathbb{E}[S \mid \theta] = \beta\theta, \qquad 0 < \beta \le \tfrac{1}{2}. \tag{20}$$

The Poisson integral formula gives the normalization and the mean. Thus S/β is an unbiased estimate of θ whose error has norm at most $2/\beta$. In the nonconvex construction, for example, θ enters the population gradient with a coefficient of order ε, and a sample replaces θ by S/β; the resulting gradient error has norm $O(\varepsilon/\beta)$, which is $O(\sigma)$ once we set $\beta \asymp \varepsilon/\sigma$. Each observation is also weakly informative: if U is uniform on the sphere,

$$\mathrm{KL}(\mathcal{L}(S \mid \theta) \,\|\, \mathcal{L}(U)) \le 2D\beta^2. \tag{21}$$

This bound concerns the whole vector S, so it holds no matter how many times the algorithm evaluates the same sample, even if it eventually reconstructs S exactly.

Continuity of the observation is what makes sample reuse harmless. With a threshold-based observation, two queries of one sample could reveal whether the threshold lies between their projections, an event whose law can depend on θ. Here, every response of a fixed sample at a fixed stage is a deterministic function of the stored vector S, the directions of earlier stages, and the exact projections at the queried points (defined below). A sample's information is therefore counted once, when S is first used, and (21) bounds all later responses, including those of gradient-difference algorithms.

Queries reveal θ in a second way. Each construction depends on a query x only through scalar *projections* $\langle \theta, w(x)\rangle$, where $w(x)$ is a known encoding of x with $\|w(x)\| \le 1$. At every query counted in the budget, we give the algorithm these projections exactly; the exact values $F(x)$ and $f(x;\xi)$ can be computed from them and from S. We say that a query *crosses* the stage if $|\langle \theta, w(x)\rangle| > a$ for a fixed threshold a. Let W collect everything the algorithm can use that is independent of θ, such as its internal randomness and the other stages' directions. If M encodings,

including one for the final output, are each chosen as functions of W and earlier projections, then the total squared projection of θ onto their span is stochastically dominated by a $\mathrm{Beta}\big(\frac{M}{2}, \frac{D-M}{2}\big)$ (22) variable, whose mean is M/D. Consequently, for $D \geq 16M/a^2$, some query or the output crosses the stage with probability at most $2\exp(-a^2 D/16)$. The output must be included because an optimization method need not return a point it has queried.

We now bound how many samples the algorithm must use before it crosses a given *target* stage. Compare the true interaction with a reference interaction in which the first b samples used at the target stage carry independent uniform vectors instead of Poisson observations. The rest of the transcript is generated by the same rules from the algorithm's randomness, the observations, and the exact projections, so its relative entropy is at most that of the b replaced observations, namely $2bD\beta^2$. Let E be the event that the target stage is crossed before more than b of its samples are used, and let $\mathbb{P}$ and $\mathbb{P}_0$ be probabilities under the true and reference interactions. The data-processing inequality gives

$$\mathbb{P}(E) \log \frac{1}{\mathbb{P}_0(E)} \leq \mathrm{KL}(\mathbb{P} \| \mathbb{P}_0) + \log 2. \tag{23}$$

In the reference interaction the algorithm's information depends on the target direction only through projections, so the projection bound gives $\log(1/\mathbb{P}_0(E)) \gtrsim a^2 D$, and therefore $\mathbb{P}(E) \lesssim (bD\beta^2 + 1)/(a^2 D)$. This is a small constant when b is a small multiple of β^{-2} and D is large. Hence each stage consumes order β^{-2} distinct samples, and summing over stages shows that a T-stage chain consumes order $T\beta^{-2}$ with constant probability. Lemma 13 extends this to skipped stages, random stopping times, and K-point rounds, with D chosen after the budget and K are fixed.

For finite sums the observations need one more property: since F is the exact average of the n components, the n observations at a stage must average to exactly $\beta\theta$. We therefore use a table with one row per component, $S_i = \beta\theta + U_i$, where $U_i = Z_i - \overline{Z}$ for independent Gaussian vectors Z_i with covariance I/D, and we condition the whole table on $\max_i \|U_i\| \leq 2$ so that the noise stays bounded. The rows are then dependent, so we compare joint laws instead of multiplying one-row bounds. Shift the table by $v_i = \beta\theta$ on the b rows the algorithm has observed and by $v_i = -b\beta\theta/(n-b)$ on the others. The shift sums to zero, so the exact mean is preserved, and its Gaussian relative-entropy cost is $\frac{D}{2}\sum_i \|v_i\|^2 = Db\beta^2/(2(1-b/n))$, which is $O(Db\beta^2)$ for $b \leq n/8$; a slightly larger truncation radius in the reference table absorbs the shifted support. Because the rows are exchangeable, a row can be assigned to a component when the algorithm first selects it, which covers adaptive index choices. A stage thus forces order $\min\{n, \beta^{-2}\}$ component evaluations, replacing m by r.

To convert constant failure probability into the expectation criteria, each construction makes the gradient norm, or the gap, at every *unfinished* point, one at which some stage has not been crossed, a large constant multiple of the target accuracy. Since M includes the output, this covers outputs that were never queried.

4.2 Scalar sample losses and the regularity tradeoff

The nonconvex construction follows the chain approach of Carmon et al. (2020). Each stage j contributes a scalar coordinate $q_j(x)$, the projection of block j onto θ_j after a bounded radial map, and the objective is, roughly, $A\, h_T(q(x/\delta))$ for a fixed scalar chain h_T, amplitude A, and spatial scale δ. The chain enforces sequential progress: smooth *gates* switch on the dependence on q_{j+1} only after $|q_j|$ is large, and near zero F is affine in the first coordinate that is still small. Consequently, at a point whose last crossed stage is p, the population gradient depends only on $\theta_1, \ldots, \theta_{p+1}$. A radial quadratic *guard* ensures that every point with small gradient lies where the chain is active, and the bounded radial map limits what queries far from the origin can learn. Up to constants, with the last bound at every unfinished point x,

$$F(x_0) - F_\star \lesssim AT, \qquad \mathrm{Lip}(\nabla F) \lesssim A/\delta^2, \qquad \|\nabla F(x)\| \gtrsim A/\delta. \tag{24}$$

The noise acts at the first unfinished stage. On the affine interval, F depends on the next direction θ_{p+1} only through a term linear in θ_{p+1}. A sample replaces θ_{p+1} in this term by its observation S_{p+1}/β; the difference is a scalar function $N(x;\xi)$ with $\mathbb{E}N(x;\xi) = 0$, and the sample loss is $f = F + N$. Sample responses therefore depend on θ_{p+1} only through S_{p+1}, as the information argument requires. Because N is multiplied by gates and cutoffs, its gradient also contains their derivatives; we keep these terms, so every sample gradient is the exact gradient of a scalar loss. At

each point at most one stage's noise term is active, so the bounds do not grow with T:

$$\sup_x \|\nabla N(x;\xi)\| \lesssim \frac{A}{\beta\delta}, \qquad \mathrm{Lip}(\nabla N(\cdot;\xi)) \lesssim \frac{A}{\beta\delta^2}, \qquad \mathbb{E}N(x;\xi) = 0. \tag{25}$$

Since N is bounded, each sample loss is bounded below. Sample and population values are also functions of the observations and projections, so Section 4.1 covers them.

Now choose the scales. Setting $A/\delta \asymp \varepsilon$ makes the gradient at every unfinished point of order ε, and $\beta \asymp \varepsilon/\sigma$ makes the noise bound $A/(\beta\delta)$ of order σ; each stage then needs $\beta^{-2} \asymp m$ samples. Without a sample-smoothness constraint, set $A/\delta^2 \asymp L$. Then $A \asymp \varepsilon^2/L$, and the gap bound $AT \lesssim \Delta$ allows $T \asymp Q$ stages, for Qm samples in total. If the centered noise must be H-Lipschitz, the second bound in (25) requires $A/(\beta\delta^2) \lesssim H$, so the population smoothness A/δ^2 is at most of order $\min\{L, \beta H\}$. The same computation gives

$$T\beta^{-2} \asymp \min\left\{\frac{\Delta L\sigma^2}{\varepsilon^4}, \frac{\Delta H\sigma}{\varepsilon^3}\right\}. \tag{26}$$

The terms Q and m come from two simpler instances, a noiseless chain and a mean-estimation problem, and the maximum of the three lower bounds is within a constant factor of their sum. For finite sums, the balanced table replaces m by r. For relative noise, the perturbation lives where the chain gradient is bounded below, with $\beta^{-2} \asymp \min\{m, \rho\}$.

4.3 PL and strongly convex constructions

For PL objectives the chain must satisfy $\|\nabla F\|^2 \geq 2\mu(F - F_\star)$ everywhere, not only have a large gradient at unfinished points. We therefore replace the gates by transition functions for which the remaining gap is controlled by the squared gradient, and use the radial guard to extend this inequality to the whole space. The resulting construction has smoothness $L \asymp A/\delta^2$ and PL constant $\mu \asymp A/(T\delta^2)$, so a condition number κ allows $T \asymp \kappa$ stages. For the statistical term we take the total gap of order η and $\beta^{-2} \asymp \sigma^2/(\mu\eta)$, so the chain needs order $\kappa\sigma^2/(\mu\eta)$ samples. Under (4), the smoothness limit of the nonconvex case gives $\min\{\kappa S, (H/\mu)\sqrt{S}\}$. The deterministic term uses separate constructions: for large κ, a chain with geometrically decreasing stage weights forces order $\kappa\log(\Delta/\eta)$ queries, and flat-link chains of the kind described next give the sharper D_κ for bounded κ.

For strong convexity we start from the classical quadratic chain, whose Hessian is tridiagonal and whose minimizer decays geometrically along the chain with a ratio $u \in (0,1)$ set by the condition number (for the pure quadratic chain, $u = (\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$). Each coupling is replaced by a convex function that is flat near zero, so a small projection onto a direction not yet found contributes nothing to the gradient. Restricting the quadratic chain to its first t coordinates leaves an optimality gap that we compute exactly. Solving the tridiagonal system, this gap, in units of the gap of the infinite chain at the origin, is

$$\frac{(1+u)u^{2t}}{1+u^{2t+1}}, \qquad 0 < u < 1. \tag{27}$$

Lemma 31 writes the gap of a length-T chain restricted to its first t coordinates as a scaled difference of these quantities. Since the gap decays like u^{2t}, reaching accuracy η requires $t \gtrsim \log(\Delta/\eta)/\log(1/u)$, which is of order E_κ. Because the formula is exact, the lower bound holds over the whole stated accuracy range, without requiring η to shrink with κ.

4.4 Matching algorithms

Without extra regularity, minibatch stochastic gradient descent attains the bounded-noise rates. Under centered-noise regularity (4) we use a recursive gradient estimator in the style of Fang et al. (2018); Li et al. (2021). Every q steps it resets its estimate v_t to an *anchor*, the average of B fresh stochastic gradients; in between, it sets $v_t = v_{t-1} + \ell_t^{-1}\sum_{k=1}^{\ell_t}[g(x_t;\xi_{t,k}) - g(x_{t-1};\xi_{t,k})]$, using ℓ_t fresh samples each evaluated at two consecutive iterates. Its error $e_t = v_t - \nabla F(x_t)$ satisfies

$$\mathbb{E}\|e_t\|^2 \leq \mathbb{E}\|e_{t-1}\|^2 + \frac{H^2}{\ell_t}\mathbb{E}\|x_t - x_{t-1}\|^2. \tag{28}$$

An anchor with $B \asymp 1 + m$ has squared error of order ε^2. With steps of size of order $1/L$, taking $\ell_t \asymp 1 + qH^2/L^2$ keeps the error accumulated over q steps at the same order, so it can be absorbed into the smooth descent inequality. The budget for $O(Q)$ steps is

$$O\left(B + Q\left[1 + \frac{qH^2}{L^2} + \frac{B}{q}\right]\right). \tag{29}$$

Balancing the last two terms gives $q \asymp L\sqrt{B}/H$, clipped to lie between 1 and the number of steps; taking the better of this method and stochastic gradient descent gives (6). For finite sums, the anchor can instead be the exact average of all n components, so its cost is $O(1 + r)$, which gives (9). Each sample is drawn after its evaluation points are chosen, as the unbiasedness assumption requires.

For PL objectives we restart this method with target gaps $\Delta_j = 2^{-j}\Delta$, $j = 0, \dots, j_\eta$, where $\Delta_{j_\eta} \asymp \eta$. By the PL inequality, a point with $\|\nabla F\|^2 \le \mu\Delta_j$ has gap at most $\Delta_j/2$, so stage j is a stationarity problem with tolerance $\sqrt{\mu\Delta_j}$, and it takes order κ steps. The statistical costs form geometric series dominated by the last stage, so restarting adds no logarithm:

$$\sum_{j=0}^{j_\eta} \frac{\sigma^2}{\mu\Delta_j} = O(S), \qquad \sum_{j=0}^{j_\eta} \frac{H\sigma}{\mu^{3/2}\sqrt{\Delta_j}} = O\left(\frac{H}{\mu}\sqrt{S}\right). \tag{30}$$

The deterministic stage costs sum to order $\kappa \log(\Delta/\eta)$; a direct minibatch schedule attains the sharper D_κ as κ approaches one. For convex objectives, an accelerated method gives expected error $O(LR^2/N^2 + \sigma R/\sqrt{N})$ after N iterations, and restarting it gives the strongly convex rate.

For higher-order smoothness we use normalized momentum, $x_{t+1} = x_t - h\, v_t/\|v_t\|$ with $v_t = (1-\alpha)v_{t-1} + \alpha\widehat{g}_t$ (Cutkosky & Mehta, 2020; He, 2025). Momentum is biased by the change of ∇F along the path; $\widehat{g}_t$ removes this bias by combining fresh gradients at p points extrapolated along the last step, with weights of absolute sum $O_p(1)$ that cancel the Taylor terms of degree below p. The remaining bias is $O_p(L_p h^p/\alpha^p)$; keeping it and the noise $\sigma\sqrt{\alpha}$ below ε gives $\alpha \asymp 1/m$, $h \asymp \alpha(\varepsilon/L_p)^{1/p}$, and $O_p(mA_p)$ evaluations. Taking the better of this method and the smoothness-based one, and adding the descent and initial estimation costs, gives $O_p(Q + m + mC_p)$.

5 Relation to previous lower bounds

Classical complexity bounds (Agarwal et al., 2012; Lan, 2012; Ghadimi & Lan, 2012; 2013b) are the benchmarks for our convex and strongly convex results. For nonconvex objectives, Arjevani et al. (2023) establish the ε^{-4} rate under bounded variance and the ε^{-3} rate under mean-square sample smoothness, matching stochastic descent and variance reduction (Ghadimi & Lan, 2013a; Fang et al., 2018; Li et al., 2021), and identify almost-surely bounded error as an open direction; our constructions address it, building on the deterministic chains of Carmon et al. (2020; 2021). Fradin et al. (2026) obtain sharp centered-noise interpolation under moment assumptions for zero-respecting algorithms; (6) has the same statistical scaling with bounded noise and unrestricted randomized queries. Clipping (Zhang et al., 2020) responds to large observations but introduces bias, so it does not produce the unbiased bounded-noise oracle studied here.

For finite sums, PAGE uses anchor size $\min\{n, \sigma^2/\varepsilon^2\}$ (Li et al., 2021); lower bounds include Woodworth & Srebro (2016) and, under individual smoothness, the nonconvex and PL results of Peng et al. (2026). Our table also enforces the noise radius and the exact component average. Lower bounds under PL and gradient dominance (Masiha et al., 2026; Ou & Wang, 2026) provide context for the term κS. For higher-order smoothness, normalized methods (Cutkosky & Mehta, 2020), extrapolation (He, 2025), and second-order lower bounds (Arjevani et al., 2020) give the exponent $7/2$ under moment assumptions; Theorem 7 adds the bounded-noise lower bound and its matching regime.

6 Discussion

Bounded noise preserves the familiar statistical barriers, while extra regularity and sample access can reduce their cost. Open cases include fresh-sample complexity under sample or centered-noise smoothness (nonconvex and large-κ PL), PL and strongly convex finite sums with $n < S$, and higher-order smoothness when $Q > (1+m)(1+C_p)$.

STATEMENT ON THE USE OF AI

AI tools were used to develop mathematical claims and proof constructions, check derivations, survey the literature, and prepare and edit the manuscript. The mathematical development included the construction of lower-bound instances, the analysis of oracle transcripts, and the derivation of matching upper bounds.

REPRODUCIBILITY STATEMENT

The appendices specify the oracle models, constructions, parameter choices, and algorithms used in every theorem. They include the information inequalities for adaptive queries and the full analytic estimates for the sample losses. The results are theoretical and do not rely on numerical experiments or external datasets.

REFERENCES

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012. doi: 10.1109/TIT.2011.2182178. URL https://arxiv.org/abs/1009.0571v3.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Ayush Sekhari, and Karthik Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Proceedings of the Thirty-Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 242–299, 2020. URL https://proceedings.mlr.press/v125/arjevani20a.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. URL https://arxiv.org/abs/1912.02365v2.

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. URL https://link.springer.com/article/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. URL https://arxiv.org/abs/1711.00841v1.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2260–2268, 2020. URL https://proceedings.mlr.press/v119/cutkosky20b.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. URL https://proceedings.neurips.cc/paper/2018/hash/1543843a4723ed2ab08e18053ae6dc5b-Abstract.html.

Adrien Fradin, Abdurakhmon Sadiev, Laurent Condat, and Peter Richtárik. Tight lower bounds and optimal algorithms for stochastic nonconvex optimization with heavy-tailed noise, 2026. URL https://arxiv.org/abs/2512.18713v2. arXiv:2512.18713, version 2.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012. doi: 10.1137/110848864. URL https://epubs.siam.org/doi/10.1137/110848864.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013a. doi: 10.1137/120880811. URL https://epubs.siam.org/doi/10.1137/120880811.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: Shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013b. doi: 10.1137/110848876. URL https://epubs.siam.org/doi/10.1137/110848876.

Chuan He. A stochastic first-order method with multi-extrapolated momentum for highly smooth unconstrained optimization, 2025. URL https://arxiv.org/abs/2412.14488v4. arXiv:2412.14488, version 4. First posted in 2024.

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1–2):365–397, 2012. doi: 10.1007/s10107-010-0434-y. URL https://link.springer.com/article/10.1007/s10107-010-0434-y.

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. URL https://proceedings.mlr.press/v139/li21a.html.

Saeed Masiha, Negar Kiyavash, and Patrick Thiran. Sharp first-order lower bounds under α-Polyak–Lojasiewicz conditions, 2026. URL https://arxiv.org/abs/2606.28278v3. arXiv:2606.28278, version 3.

Weiming Ou and Xiao Wang. Heavy-tailed first-order optimization for Polyak–Łojasiewicz condition: High-dimensional minimax bounds, high-probability guarantee, and fixed-dimensional improvements, 2026. URL https://arxiv.org/abs/2609.03990v1. arXiv:2609.03990, version 1.

Yuxing Peng, Zhiqing Tang, and Weijia Jia. Dense weak hiding: Closing complexity gaps in nonconvex and PL finite-sum optimization under individual smoothness, 2026. URL https://arxiv.org/abs/2609.00045v2. arXiv:2609.00045, version 2.

Umut Şimşekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5827–5837, 2019. URL https://proceedings.mlr.press/v97/simsekli19a.html.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1195–1204, 2019. URL https://proceedings.mlr.press/v89/vaswani19a.html.

Blake Woodworth and Nathan Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/645098b086d2f9e1e0e939c27f9f2d6f-Abstract.html.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems*, volume 33, pp. 15383–15393, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/b05b57f6add810d3b7490866d74c0053-Abstract.html.

A INFORMATION BOUNDS FOR BOUNDED NOISE

This appendix proves the information estimates used by the constructions. We take all random variables and decision rules below to be measurable on standard Borel spaces. The query points themselves may have arbitrarily large norm; the bounds in the projection estimates apply to their encoded vectors. We also allow the algorithm to output a point that it has never queried.

A.1 ADAPTIVE EXACT MEASUREMENTS

Lemma 9 (Adaptive spherical projections)**.** *Let Θ be uniform on $\mathbb{S}^{D-1}$, and let W be independent of Θ. For $t = 1, \dots, M$, let $v_t \in \mathbb{R}^D$, $\|v_t\| \le 1$, be a measurable function of W and the preceding exact observations $y_s = \langle \Theta, v_s \rangle$, $s < t$. Assume $1 \le M < D$. There is, on an extension of this probability space, a random variable*

$$B \sim \mathrm{Beta}\big(M/2, (D-M)/2\big) \quad \textit{such that} \quad \max_{t \le M} |\langle \Theta, v_t \rangle|^2 \le B.$$

Consequently, if $0 < a \le 1$ and $D \ge 16M/a^2$, then

$$\mathbb{P}\left\{ \max_{t \le M} |\langle \Theta, v_t \rangle| > a \right\} \le 2\exp(-a^2 D/16). \tag{31}$$

The assertion permits an adaptively stopped list, padded by zero vectors, and an arbitrary transcript-measurable output appended as one further vector.

Proof. Condition first on W. Apply Gram–Schmidt to the selected vectors, retaining only nonzero residuals. Suppose $w_1, \dots, w_j$ are the orthonormal vectors retained so far. Their scalar observations can be recovered from the original observations by triangular substitution. A query in their span reveals no additional information. Conditional on the observations so far, write

$$\Theta = \sum_{h=1}^{j} \xi_h w_h + U_j, \qquad \xi_h = \langle \Theta, w_h \rangle .$$

The vector U_j is uniform on the sphere of radius $(1 - \sum_{h=1}^{j} \xi_h^2)^{1/2}$ in the orthogonal complement of this span. For $j = 0$ this is the defining rotational invariance of uniform surface measure. For the inductive step, the next residual direction is fixed by the already observed information. Rotations of its orthogonal complement leave both that information and uniform surface measure invariant. Conditioning on the next scalar coordinate therefore leaves a uniform residual sphere in one lower dimension. This proves the assertion at every step.

If fewer than M independent queries occur, complete the orthonormal list by applying Gram–Schmidt to the standard basis in its fixed order. Reveal the additional scalar coordinates for purposes of this proof. This completion is measurable, and each newly chosen direction is determined by previously revealed information. The preceding conditional-distribution calculation shows that $(\xi_1, \dots, \xi_M)$ has the same successive kernels as the first M coordinates of a uniform sphere point in a fixed basis. Thus $B = \sum_{h=1}^{M} \xi_h^2$ has the asserted beta distribution. Every query belongs to the span formed by the time it is processed, so Cauchy–Schwarz bounds its squared observation by B. The conditional law of B does not depend on W, which removes the initial conditioning.

For the tail estimate, represent $B = Y/(Y+Z)$ using independent $Y \sim \chi^2_M$ and $Z \sim \chi^2_{D-M}$. If $B > a^2$, then either $Y + Z < D/2$ or $Y > a^2 D/2$. The Gaussian moment-generating formula gives

$$\mathbb{P}(Y + Z < D/2) \le e^{D/4} 2^{-D/2} \le e^{-D/16}, \qquad \mathbb{P}(Y > a^2 D/2) \le 2^{M/2} e^{-a^2 D/8}.$$

Since $D \ge 16M/a^2$, we have $(M/2)\log 2 \le a^2 D/16$. Each displayed probability is therefore at most $e^{-a^2 D/16}$. A stopping decision based on already available observations is an admissible decision rule for selecting zero vectors afterward. A normalized or otherwise bounded encoding of the final output is also such a decision rule. Including that vector increases M by one and proves the last assertion. □

For independent directions $\Theta_1, \ldots, \Theta_T$, the lemma can be applied to one direction after including all other directions in W. Every additional variable included in W must be independent of the target direction; in particular, W cannot include an observation whose distribution depends on that direction. We retain exact measurements received before entering a stage in the measurement history, without assuming a uniform posterior at the random time of entry.

A.2 A SPHERICAL CODE WITH BOUNDED SUPPORT

Lemma 10 (Poisson code)**.** *Let $D \geq 3$, let μ_D be uniform probability measure on $\mathbb{S}^{D-1}$, and let $a \in \mathbb{R}^D$ satisfy $\|a\| \leq 1/2$. The formula*

$$p_a(s) = \frac{1 - \|a\|^2}{\|s - a\|^D} \tag{32}$$

defines a probability density with respect to μ_D. Its law P_a satisfies

$$\mathbb{E}_{P_a} S = a, \qquad \|S - a\| \leq 1 + \|a\|, \qquad \mathrm{KL}(P_a \| \mu_D) \leq 2D \|a\|^2. \tag{33}$$

These laws form a Borel probability kernel and admit a realization with one fixed sample law whose values lie on the sphere for every parameter and sample. In particular, $P_{\beta\theta}$ has mean $\beta\theta$ and relative entropy at most $2D\beta^2$ for $\theta \in \mathbb{S}^{D-1}$ and $0 \leq \beta \leq 1/2$.

Proof. The case $a = 0$ is immediate. Put $r = \|a\| > 0$, let ω_D denote the surface area of the unit sphere, and consider

$$G_a(x) = \|x - a\|^{2-D} - r^{2-D} \left\|x - a/r^2\right\|^{2-D}.$$

The second pole is outside the closed unit ball. On its boundary, $\left\|s - a/r^2\right\| = \|s - a\| / r$, so $G_a(s) = 0$. Differentiating in the outward normal direction gives

$$\partial_\nu G_a(s) = (2 - D)\frac{1 - \langle s, a\rangle}{\|s - a\|^D} - (2 - D)\frac{r^2 - \langle s, a\rangle}{\|s - a\|^D} = (2 - D)p_a(s).$$

Let $h(x) = 1$ or $h(x) = x_k$. Both h and G_a are harmonic on the unit ball with a ball of radius ρ about a removed. Green's second identity on this punctured domain gives

$$0 = \int_{\partial\Omega_\rho} (h\partial_\nu G_a - G_a \partial_\nu h)\, dS.$$

On the inner boundary its outward normal points toward a. The singular term consequently contributes $(D - 2)\omega_D h(a) + o(1)$ to the first integral. The term involving $G_a \partial_\nu h$ is $O(\rho)$, and all terms from the nonsingular pole tend to zero there. On the outer boundary, $G_a = 0$. Letting ρ decrease to zero therefore yields

$$(2 - D)\omega_D h(a) = \int_{\mathbb{S}^{D-1}} h(s)\partial_\nu G_a(s)\, dS(s).$$

Division by $(2 - D)\omega_D$ proves $h(a) = \int h(s)p_a(s)\, d\mu_D(s)$. Taking $h = 1$ proves normalization; taking each coordinate function proves the mean identity. The support bound follows from the triangle inequality.

For the entropy bound put $t = \langle a, S\rangle$ and $r = \|a\|$. Since

$$\|S - a\|^2 = 1 - 2t + r^2 \geq (1 - t)^2,$$

and $\log(1 - r^2) \leq 0$, we have $\log p_a(S) \leq -D \log(1 - t)$. For $|t| \leq 1/2$, $-\log(1 - t) \leq t + t^2$. The derivative of $t + t^2 + \log(1 - t)$ is $t(1 - 2t)/(1 - t)$, which is nonpositive on $[-1/2, 0]$ and nonnegative on $[0, 1/2]$. Thus $t + t^2 + \log(1 - t)$ is minimized at zero. Under P_a the mean identity gives $\mathbb{E}t = r^2$, whereas $t^2 \leq r^2$ pointwise. Taking expectations proves $\mathrm{KL}(P_a \| \mu_D) \leq 2Dr^2$.

The density is continuous jointly in (a, s) on the parameter half-ball and the sphere, hence is a Borel kernel. An explicit fixed-law realization uses independent uniform sphere proposals V_m and independent uniforms $U_m \in [0, 1]$. Since $p_a(s) \leq 2^D$, return the first V_m for which $U_m \leq 2^{-D} p_a(V_m)$. Each trial has acceptance probability 2^{-D}, so failure forever has probability zero for each fixed a. Assign the fixed vector e_1 if it fails forever. The resulting map is jointly Borel and sphere-valued for every $(a, (V_m, U_m)_{m \geq 1})$. Its distribution for each a is P_a, and its pointwise support bound consequently holds on one common sample space for all parameters. □

A.3 A FINITE POPULATION WITH AN EXACT PRESCRIBED MEAN

Lemma 11 (Balanced continuous tables)**.** *Fix $n \geq 2$, $D \geq 1$, $\theta \in \mathbb{S}^{D-1}$, and $0 < \beta \leq 1/4$. There is an exchangeable random table $(S_1, \ldots, S_n)$ with*

$$\frac{1}{n}\sum_{a=1}^{n} S_a = \beta\theta, \qquad \|S_a/\beta - \theta\| \leq 2/\beta \quad (a \in [n]) \tag{34}$$

for every table in its support. For any integer $1 \leq b \leq n/8$, a θ-independent law Q_b on b rows satisfies

$$\mathrm{KL}\big(\mathcal{L}(S_1, \ldots, S_b \mid \theta) \| Q_b\big) \leq Db\beta^2, \qquad D \geq 8\log(4n/\beta). \tag{35}$$

The same comparison applies to the first b distinct component indices chosen adaptively from earlier revealed rows, exact measurements of θ, and independent auxiliary data. Repeated indices return their cached rows.

Proof. Let

$$\mathcal{H} = \{(u_1, \ldots, u_n) \in (\mathbb{R}^D)^n : \sum_a u_a = 0\}.$$

On this subspace let γ have density proportional to

$$\exp\left[-\frac{D}{2}\sum_a \|u_a\|^2\right].$$

Equivalently, take independent $Z_a \sim N(0, I_D/D)$ and put $U_a = Z_a - n^{-1}\sum_c Z_c$. Write

$$E_R = \{\max_a \|U_a\| \leq R\}, \qquad Z_R = \gamma(E_R).$$

Sample U from γ conditional on E_2, and set $S_a = \beta\theta + U_a$. The conditioning is well defined because its event has positive probability. The resulting law is invariant under row permutations and under $U \mapsto -U$, so $\mathbb{E}U = 0$. The identity $\sum_a U_a = 0$ and the conditioning give (34) pointwise.

Let Q_b be the first b rows of an independent vector having law γ conditional on $E_{2+\beta}$. Define the balanced shift

$$h_a = \begin{cases} \beta\theta, & a \leq b, \\ -\dfrac{b\beta}{n-b}\theta, & a > b. \end{cases}$$

For $U \in E_2$ the shifted vector $V = U + h$ belongs to $E_{2+\beta}$, since $b/(n-b) \leq 1$. The law of V is thus absolutely continuous with respect to $\gamma(\cdot \mid E_{2+\beta})$. At $V = U + h$ its log density ratio is exactly

$$D\,\langle U, h\rangle + \frac{D}{2}\|h\|^2 + \log\frac{Z_{2+\beta}}{Z_2}.$$

The linear term has mean zero. Moreover,

$$\|h\|^2 = b\beta^2 + (n-b)\frac{b^2\beta^2}{(n-b)^2} = \beta^2\frac{bn}{n-b}.$$

Applying data processing to the first b rows, which are precisely $(S_1, \ldots, S_b)$, proves

$$\mathrm{KL}\big(\mathcal{L}(S_1, \ldots, S_b \mid \theta) \| Q_b\big) \leq \frac{D\beta^2}{2}\frac{bn}{n-b} + \log\frac{Z_{2+\beta}}{Z_2}. \tag{36}$$

For the normalization error, each unconditioned U_a has covariance $(1 - 1/n)I_D/D$. Chernoff's inequality for a chi-squared variable yields

$$\gamma(\|U_a\| > 2) \leq \mathbb{P}(\chi^2_D > 4D) \leq \exp[-D(3 - \log 4)/2] \leq e^{-D/2}.$$

Hence $Z_2 \geq 1 - ne^{-D/2}$. The stated dimension bound implies $ne^{-D/2} \leq \beta^2/16$, so

$$0 \leq \log(Z_{2+\beta}/Z_2) \leq -\log Z_2 \leq \beta^2/8.$$

Since $b \le n/8$, the first coefficient in (36) is at most $(4/7)Db\beta^2$. The normalization term is at most $Db\beta^2/8$, which proves (35).

We finally justify active selection of indices. Conditional on θ, auxiliary data independent of the table, and any already exposed rows, permuting the component labels not yet exposed preserves their conditional joint law. A choice of the next unexposed label depends only on these conditioned data and consequently cannot distinguish its row from any other unexposed row. One can therefore simulate the interaction by drawing the ordered list $(S_1, \dots, S_b)$ first, assigning its next unused row to each newly requested label, and storing that assignment. Induction on the number of distinct requested labels proves equality with the original interaction through its bth distinct row. Exact measurements of θ are deterministic when conditioning on θ, so they do not affect this exchangeability argument. Repeated labels use their stored rows, so they reveal no additional rows; if fewer than b labels are selected, the rest of the ordered list remains unobserved.

For the relative-entropy comparison, include θ and this b-row list in the underlying joint law, and replace only the list by Q_b in the reference. Its joint relative entropy is bounded by (35), averaged over any chosen prior on θ. The adaptive assignment, cached repetitions, and exact measurements are the same measurable maps in both experiments, so data processing gives the same bound for the resulting transcript. Since the comparison concerns the joint list, the argument does not assert independence of successive rows. □

Lemma 12 (Testing with reused codes and exact values)**.** *Let Θ be uniform on $\mathbb{S}^{D-1}$ and let W be independent auxiliary data. Conditional on $\Theta = \theta$, a code pool C is independent of W and has law R_θ. Suppose a θ-independent reference R_0 satisfies $\mathrm{KL}(R_\theta \| R_0) \le H_0$ uniformly in θ. An interaction may use this pool repeatedly and take at most M adaptive scalar measurements $\langle \Theta, v_t \rangle$, with $\|v_t\| \le 1$, including a normalized encoding of its arbitrary output. Assume its other responses and decisions are measurable functions of W, C, and measurements already available at the respective decision. If $D \ge 16M/a^2$, then*

$$\mathbb{P}\left\{ \max_{t \le M} |\langle \Theta, v_t \rangle| > a \right\} \le \frac{H_0 + \log 2}{a^2 D/16 - \log 2}. \tag{37}$$

For b independent Poisson codes one can take $H_0 = 2Db\beta^2$; for the b distinct rows in Lemma 11, one can take $H_0 = Db\beta^2$. The number b counts distinct codes, irrespective of how many adaptive evaluations reuse them. In particular, with at most J evaluations per sampled handle, $M \le bJ + 1$ is valid for a budget of b handles.

Proof. We compare the actual transcript with a reference transcript. Let P denote the joint law of (Θ, W, C) and the transcript, and let Q replace R_Θ by R_0 while leaving the other generating maps unchanged. The initial relative entropy is at most H_0, and subsequent choices and exact measurements are deterministic functions of the same data, so the transcript has no larger divergence. Under Q, (W, C) is independent of Θ, and the projection lemma bounds the crossing event E. Applying binary data processing to $p = P(E)$ and $q = Q(E)$ gives

$$H_0 \ge p \log(p/q) + (1-p)\log((1-p)/(1-q)) \ge p \log(1/q) - \log 2,$$

which proves the claim. For independent codes, the product chain rule gives the stated H_0; for finite tables, the joint-list comparison gives the same bound. Repeated evaluations reuse stored codes and therefore do not enlarge the list. There are at most bJ queried points under the last budget convention, and appending the output gives $M \le bJ + 1$. □

A.4 A COMPARISON STARTED BEFORE THE FIRST QUERY

The sequential comparison must handle two sources of additional information: a sample may be reused, and one fresh sample may be queried at several points. We therefore assign one hidden direction Θ_i to each block and encode every query through a bounded vector $w_i(x)$. The scalar projection $z_i(x)$ records how much the query has learned about that direction. All subsequent sample and value responses are expressed through these projections. The prefix maps below agree with the true responses until the corresponding coordinate is crossed. For each block let Θ_i be uniform on $\mathbb{S}^{D-1}$, independently across $i \in [T]$, and define a bounded encoded query vector $w_i(x) \in \mathbb{R}^D$. Put

$$z_i(x) = \langle \Theta_i, w_i(x) \rangle, \qquad \|w_i(x)\| \le 1, \qquad p(x) = \max\big(\{i : |z_i(x)| > a\} \cup \{0\}\big),$$

where $0 < a \le 1$ is fixed. All w_i are known measurable maps. A scalar coordinate such as $q_i(x) = 2z_i(x)$ or $4z_i(x)$ is therefore covered after rescaling its crossing threshold. We grant every $z_i(x)$ at every charged query, even when the original oracle supplies only the population value. Any value determined by these scalars can then be returned as well.

For each handle h and stage i, let $S_{h,i}$ be a code. Conditional on Θ, the codes are independent over (h, i), with law P_{Θ_i}, and suppose a θ-independent law P_0 satisfies

$$\mathrm{KL}(P_\theta \| P_0) \le \Lambda D \beta^2. \tag{38}$$

Throughout the interaction, each handle refers to the same collection of codes. In the alternative finite-population model, each stage has an exchangeable table of n rows, independent of the tables at other stages. Its first b rows have conditional law $P_{\theta,b}$ and a θ-independent reference satisfying

$$\mathrm{KL}(P_{\theta,b} \| Q_b) \le \Lambda D b \beta^2. \tag{39}$$

The table in Lemma 11 has $\Lambda = 1$ for $b \le n/8$.

Assume that globally defined maps Γ_p, $0 \le p < T$, have the factorization

$$\Gamma_p(x; \Theta_{\le p}, S_{h,\le p+1}) \tag{40}$$

and equal the actual sample response at every point whose last crossed coordinate is p. A response can contain both a sample value and its full gradient. We allow Γ_T to be arbitrary and equal to the actual response on the terminal region. A hidden-independent branch, for example an exterior response, is allowed by taking every Γ_p equal to that branch there. For this branch as well, the query encoding must obey the stated norm bound.

Lemma 13 (Prefix information bound)**.** *Under the preceding assumptions, consider either N adaptive evaluations using arbitrarily retained handles, N active finite-population evaluations, or R fresh-sample rounds with at most K nonadaptive points per round. In the last model the points in a round are all selected before any response from that round, and all use its one new handle. Set*

$$(M, B) = \begin{cases} (N+1, N), & \textit{retained or active evaluations}, \\ (RK+1, 2R), & \textit{fresh simultaneous rounds}. \end{cases}$$

Fix an integer $b \ge 1$ for which (38) or (39) applies, and assume $D \ge 16M/a^2$. There is a capped comparison interaction, with frontier ν increasing by at most one per evaluation or round and by at most one at the arbitrary final output, with the following properties. Writing

$$\delta_D = 2Te^{-a^2 D/16}, \qquad \pi_b = \frac{\Lambda D b \beta^2 + \log 2}{a^2 D/16 - \log 2}, \tag{41}$$

there is an event J of probability at most δ_D outside which the comparison and actual interactions agree. For each i there is a nonnegative integer L_i counting its relevant codes or rounds, with $\sum_i L_i \le B$. If A_i is the event that the capped frontier first crosses stage i with fewer than b prior relevant codes or rounds, then

$$\mathbb{P}(A_i \cap J^c) \le \pi_b. \tag{42}$$

For every integer $1 \le m \le T$ with $m > B/b$,

$$\mathbb{P}\{p(\widehat{x}) \ge m\} \le \delta_D + \frac{m\pi_b}{m - B/b}. \tag{43}$$

In particular this probability is at most $\delta_D + 2\pi_b$ when $B \le mb/2$.

Proof. We index interaction steps by t. Each step is a single evaluation in the retained-sample or active finite-population model, or a complete simultaneous round in the fresh-sample model. Write $x_{t,k}$ for its selected points, with one point in the former models and at most K in the latter. All points in a step are fixed by the preceding ordinary transcript and the private algorithmic seed. Start at $\nu_0 = 0$, and use the common before-step frontier to define

$$\rho_{t,k} = \max\big(\{j \le \min(T, \nu_{t-1} + 1) : |z_j(x_{t,k})| > a\} \cup \{0\}\big). \tag{44}$$

Return $\Gamma_{\rho_{t,k}}$ with the indicated handle and all exact projection responses. Set $\nu_t = \max(\nu_{t-1}, \max_k \rho_{t,k})$. Thus $\nu_t \le \nu_{t-1} + 1$. At the final output use the same capped test once more, without a sample response, and call the resulting frontier ν_{out}. Define J to be the event that some selected point, including the output, crosses an index strictly larger than its before-step frontier plus one.

Skipped coordinates. Fix a target i. Stop the comparison immediately before the first step containing a response index $\rho_{t,k} \geq i-1$, retaining all of that step's preselected points and their exact measurements. Stop before returning any sample response in that step. If this never occurs, retain the output as well. Before the stopping step every response index is at most $i-2$ and every frontier is at most $i-2$. The retained responses therefore depend only on $\Theta_{<i}$ and codes from stages below i, by (40). These data, every other hidden direction, and the private seed are independent of Θ_i. The stop can be implemented from these data and the available measurements: before it occurs the test in (44) examines only indices at most $i-1$. Conditional on the independent data, the retained query vectors access Θ_i only through its preceding exact linear measurements. Pad the retained list with zero vectors to length M. Lemma 9 bounds the probability of a target crossing in this list by $2e^{-a^2D/16}$.

Any skipped crossing of i occurs before the first response of index at least $i-1$. Otherwise the nondecreasing frontier would already be at least $i-1$, contradicting the definition of a skip. Its point therefore belongs to the retained list, including when the point is in the entry round or is the arbitrary output. Taking a union over i proves $\mathbb{P}(J) \leq \delta_D$. To compare with the actual oracle, use the same hidden directions, sample data, and private seed. On J^c every actual last crossed index equals its cap at every selected point. The prefix identity makes both responses equal, and induction gives identical ordinary transcripts and output. Under this coupling, the transcripts and outputs agree exactly on J^c.

Relevant counts. In the retained or active model, let L_i count distinct handles or component labels used at an index-$i-1$ response in steps with $\nu_{t-1} < i$. Let $c_i(t)$ count those distinct labels strictly before step t. In the fresh-sample model let L_i count steps t with $\nu_{t-1} < i$ that contain at least one index-$i-1$ response; $c_i(t)$ is the corresponding count strictly before t. At the output it is the count over all response steps. These definitions may count a response in the very step that crosses i, but such a response is excluded from the before-step count.

If the before-step frontier is m, a counted index must have $i > m$ and $\rho = i-1 \leq m+1$. Thus only $i = m+1$ and $i = m+2$ can be counted. A single evaluation has just one response index and adds at most one new counted pair. A simultaneous round adds at most one count to each of these two stages, regardless of K. Summing proves $\sum_i L_i \leq N$ or $2R$, respectively. Define A_i by the first advancement of the capped frontier from $i-1$ to i, including a possible advancement at output, with $c_i(t) < b$. The frontier cannot skip an integer, so this first advancement is well defined whenever stage i is reached.

Continuation from the initial query. For a fixed i, run a separate experiment from time zero. Preserve all exact measurements throughout. Replace the response index at every point by

$$\lambda_i(x) = \max\big(\{j < i : |z_j(x)| > a\} \cup \{0\}\big). \tag{45}$$

Maintain the same handle caches and a counter c of distinct target-stage codes used at index $i-1$. A fresh round uses a new code once if any of its points has that index. When $c < b$, return the indicated prefix responses, then increase c for a new relevant handle, or for a relevant fresh round. In the evaluation models this increases c by at most one per step; in the round model it also increases c by at most one. Once $c = b$, return fixed zero sample responses at every subsequent step, while continuing to return all exact projections. All these rules are defined without testing a crossing of the target direction. Let E_i be the event that a point, or the output, crosses the target at a step whose before-step counter is less than b.

Let P be this continuation with actual codes. For independent codes, construct Q by replacing each newly used stage-i code by an independent P_0 vector and caching the replacement. Include in an augmented transcript the hidden directions, private seed, and every code vector when it is first read. This augmentation is for the analysis: the algorithm still uses only its ordinary transcript and is not given the additional code vectors. Conditional on this augmented past and the hidden directions, a previously read code has a Dirac kernel at its cached value. A new non-target code has exactly the same kernel P_{Θ_j} in both laws. A new target code has kernels P_{Θ_i} and P_0, whose relative entropy is at most $\Lambda D\beta^2$. The selected queries, indices, exact measurements, and prefix responses are identical deterministic functions of these variables in the two experiments. In particular, even an exact real measurement has zero conditional relative-entropy contribution: the conditioning includes the hidden directions, and both laws use the same Dirac kernel. There are at most b target-code revelations. The

chain rule gives

$$\mathrm{KL}(P\|Q) \le \Lambda D b \beta^2. \tag{46}$$

We augment the transcript only with realized codes, since using the raw random seeds that generate biased codes would produce different deterministic kernels and would not justify this inequality.

For active finite populations use a lazy list of b target rows, assigning its next row to each newly used component label. The exchangeability proof of Lemma 11 shows that this exactly simulates the actual continuation. Pad the list by unused rows if needed. Conditional on all hidden directions, the target list has law $P_{\Theta_i,b}$ in P and Q_b in Q; all other-stage tables and the private seed have the same law. The entire continuation is a common measurable function of these data. Data processing applied to the joint list gives (46) from (39). This reasoning permits the rows to be correlated and does not use conditional one-row entropy bounds.

Under Q, the target reference codes or reference list, all non-target code data and directions, and the private seed are independent of Θ_i. They can be included in W for the purpose of bounding the reference event. With these variables fixed, (45) and every sample response depend on Θ_i only through its prior exact measurements; the cap forbids direct revelation of that direction. Lemma 9, with the output appended, proves

$$Q(E_i) \le 2e^{-a^2 D/16}. \tag{47}$$

For this argument, preassigning b target reference vectors suffices; the entropy comparison includes no unused infinite array of biased codes.

Couple P to the original capped comparison using the actual directions, code data, and private seed. On J^c, before the first crossing of i, no coordinate at or beyond i is crossed. All the prior capped indices therefore agree with (45). The interactions and counters agree until the first target crossing or the exhaustion of b target codes. If a simultaneous tuple first crosses i, the complete tuple was already chosen before any response from it; its crossing is therefore tested with the shared before-round count. Consequently the coupling gives $A_i \cap J^c \subseteq E_i$. No independence statement is made after conditioning on J^c.

For completeness, if $p = P(E_i)$ and $q = Q(E_i)$, binary data processing and the entropy bound for a Bernoulli variable give

$$\mathrm{KL}(P\|Q) \ge p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \ge p \log(1/q) - \log 2.$$

If $q = 0$, finite relative entropy implies $P \ll Q$, and hence $p = 0$. Combining this inequality with (46) and (47) proves $P(E_i) \le \pi_b$, hence (42).

Counting completed stages. If $\nu_{\mathrm{out}} \ge m$, each of the first m stages was crossed. Among them at most B/b can have $L_i \ge b$, since $\sum_i L_i \le B$. Every other one has before-crossing count at most $L_i < b$ and thus belongs to A_i. Pointwise,

$$(m - B/b)\, \mathbf{1}\{\nu_{\mathrm{out}} \ge m, J^c\} \le \sum_{i=1}^{m} \mathbf{1}\{A_i \cap J^c\}.$$

Take expectations and use (42). On J^c, the actual output agrees with the comparison output and $p(\widehat{x}) \le \nu_{\mathrm{out}}$. Adding $\mathbb{P}(J)$ proves (43). □

For the threshold $a = 1/16$ used in the radial constructions, it is enough to take

$$D \ge \max\{4096M,\, 2^{22},\, 8192 \log(256T)\}, \qquad \Lambda b \beta^2 \le 2^{-20}.$$

Then $\delta_D \le 1/128$ and $\pi_b < 1/128$. Finite tables also require the dimension and row-count conditions in Lemma 11. Thus a constant probability of reaching m stages needs order mb retained or active evaluations, or order mb fresh simultaneous rounds. The latter conclusion counts rounds directly; it is not obtained by dividing an evaluation bound by K.

A.5 Deterministic prefix depth

Corollary 14 (Exact responses and simultaneous depth)**.** *Use the independent hidden directions, bounded encodings, and crossing threshold above, but suppose the exact response has a global prefix extension $\Gamma_p(x;\Theta_{\le p+1})$, agreeing with the actual response whenever $p(x)=p<T$. The response may include an exact gradient; all scalar projections are supplied as additional information. For an algorithm using R simultaneous rounds of at most K points each, followed by an arbitrary output, take $M=RK+1$ and $D\ge 16M/a^2$. Then*

$$\mathbb{P}\{p(\widehat{x})>R+1\}\le 2Te^{-a^2D/16}. \tag{48}$$

The assertion holds for adaptive single-point evaluations by taking $K=1$. It also permits hidden-independent response branches on arbitrary query regions.

Proof. Define the capped indices by (44), using a common before-round frontier for all points in that round. Return the corresponding exact prefix responses and every scalar projection. The update obeys $\nu_t\le\nu_{t-1}+1$ identically, so $\nu_R\le R$ and the output update obeys $\nu_{\mathrm{out}}\le R+1$. For a fixed target i, stop before the first response index at least $i-1$, retaining every preselected point of that round. All earlier responses have index at most $i-2$, hence depend directly only on $\Theta_{\le i-1}$. Conditional on the private seed and every other hidden direction, the target has supplied only its adaptive exact measurements. The stopping test before the first such response examines only indices at most $i-1$. Pad the list with zero vectors if it stops early, and include the output if it does not. Lemma 9 bounds the probability of a target crossing in this list by $2e^{-a^2D/16}$. A crossing skipped by the capped frontier must be in this list, since any preceding response index at least $i-1$ would already have advanced the frontier to at least $i-1$. A union over i bounds the probability of any skipped coordinate by $2Te^{-a^2D/16}$. Outside that event, every cap equals the actual last crossed index. The prefix identities and induction then identify the exact and comparison interactions, including the output. Its last crossed index is at most $\nu_{\mathrm{out}}\le R+1$, proving the claim. On a hidden-independent response region, the response can be represented by every prefix map, so the preceding dependence and coupling statements remain valid there. □

B Nonconvex objectives and conservative sampled losses

This appendix proves the lower bounds for the nonconvex, finite-sum, and relative-noise results. Continuous hidden directions make the construction valid even when a charged query returns both the sampled loss and the exact population value; the one-point fresh-sample result follows as a special case. All constants below are absolute. We use

$$C_1=10^4,\qquad B_0=2^{16},\qquad R_0=2^{60},\qquad c_0=2^{-20}. \tag{49}$$

The accuracy assumption $\varepsilon\le c_0\sqrt{\Delta L}$ implies $Q=\Delta L/\varepsilon^2\ge 2^{40}$. The dimension may depend on the proposed finite query budget, but it is fixed before the algorithm makes any random choices.

B.1 An affine-plateau scalar chain

To make each chain link locally affine while keeping later links inactive, define a quintic step s, a gate Ψ that vanishes on a half-line, and a primitive Φ with an affine interval:

$$s(t)=\begin{cases}0, & t\le 0,\\ 10t^3-15t^4+6t^5, & 0<t<1,\\ 1, & t\ge 1,\end{cases}\qquad \Psi(t)=\begin{cases}0, & t\le \frac12,\\ \exp\bigl(1-(2t-1)^{-2}\bigr), & t>\frac12,\end{cases} \tag{50}$$

$$\phi(t)=1-s(|t|-1),\qquad \Phi(t)=2+\int_0^t\phi(u)\,du. \tag{51}$$

The absolute value causes no loss of differentiability at zero because the functions involving it are constant in a neighborhood of zero. For $q\in\mathbb{R}^T$ set

$$h_T(q)=-\Phi(2q_1)+\sum_{j=2}^{T}\bigl[\Psi(-2q_{j-1})\Phi(-2q_j)-\Psi(2q_{j-1})\Phi(2q_j)\bigr]. \tag{52}$$

Lemma 15 (Scalar chain)**.** *For every $T \geq 1$,*

$$h_T(0) - \inf h_T \leq 20T, \qquad \|\nabla h_T(q)\|_\infty \leq 64, \qquad \left\|\nabla^2 h_T(q)\right\|_{\mathrm{op}} \leq 2400. \tag{53}$$

Every coordinate derivative is nonpositive. If j is the first index with $|q_j| \leq 1/2$, then $\partial_j h_T(q) \leq -2$. If $|q_k| \leq 1/8$ for all $k > p$, then

$$h_T(q) = h_T(q^{\leq p}) - 2a_{p+1}(q^{\leq p})q_{p+1}, \qquad p < T, \tag{54}$$

where $q^{\leq p} = (q_1, \ldots, q_p, 0, \ldots, 0)$, $a_1 = 1$, and $a_j(q) = \Psi(-2q_{j-1}) + \Psi(2q_{j-1})$ for $j \geq 2$.

Proof. On $(0,1)$, $s'(t) = 30t^2(1-t)^2$ and $s''(t) = 60t(1-t)(1-2t)$. Thus $0 \leq s' \leq 15/8$, $|s''| \leq 6$, and the step and its first two derivatives match their constant extensions at both endpoints. Consequently

$$0 \leq \Phi \leq 4, \qquad 0 \leq \Phi' = \phi \leq 1, \qquad |\Phi''| \leq 2, \qquad \Phi(t) = 2 + t \quad (|t| \leq 1). \tag{55}$$

Here the integral of ϕ on each half-line is at most two. For $v = (2t-1)^{-2}$, direct differentiation gives

$$\Psi'(t) = 4ee^{-v}v^{3/2}, \qquad \Psi''(t) = 8ee^{-v}v^2(2v-3).$$

All derivatives extend by zero at $t = 1/2$, since a polynomial times e^{-v} tends to zero as $v \to \infty$. The identity $\sup_{v \geq 0} v^k e^{-v} = k^k e^{-k}$ gives

$$0 \leq \Psi < 3, \qquad 0 \leq \Psi' < 3\sqrt{6/e} < 5, \qquad |\Psi''| \leq 432/e^2 + 96/e < 100. \tag{56}$$

At most one of $\Psi^{(k)}(t)$ and $\Psi^{(k)}(-t)$ is nonzero for $k = 0, 1, 2$.

Write $\bar{h}_T(z) = h_T(z/2)$. Differentiating every possible incoming and outgoing term gives

$$\begin{aligned}\partial_j \bar{h}_T(z) = &-\mathbf{1}_{\{j=1\}}\Phi'(z_1) \\ &-\mathbf{1}_{\{j\geq 2\}}[\Psi(-z_{j-1})\Phi'(-z_j) + \Psi(z_{j-1})\Phi'(z_j)] \\ &-\mathbf{1}_{\{j<T\}}[\Psi'(-z_j)\Phi(-z_{j+1}) + \Psi'(z_j)\Phi(z_{j+1})].\end{aligned} \tag{57}$$

Every term is nonpositive. The root, incoming, and outgoing contributions have magnitudes at most $1, 3, 20$, respectively; the root and incoming contributions are mutually exclusive. Hence $\|\nabla h_T\|_\infty \leq 2(3+20) = 46 < 64$. The diagonal Hessian entries of $\bar{h}_T$ are

$$\begin{aligned}\partial^2_{jj} \bar{h}_T(z) = &-\mathbf{1}_{\{j=1\}}\Phi''(z_1) \\ &+\mathbf{1}_{\{j\geq 2\}}[\Psi(-z_{j-1})\Phi''(-z_j) - \Psi(z_{j-1})\Phi''(z_j)] \\ &+\mathbf{1}_{\{j<T\}}[\Psi''(-z_j)\Phi(-z_{j+1}) - \Psi''(z_j)\Phi(z_{j+1})],\end{aligned}$$

and its adjacent entries are

$$\partial^2_{j,j+1} \bar{h}_T(z) = \Psi'(-z_j)\Phi'(-z_{j+1}) - \Psi'(z_j)\Phi'(z_{j+1}).$$

All remaining entries vanish. The diagonal bound is 406, and each adjacent entry has magnitude at most five. Thus the maximum absolute row sum is at most 416, and $\left\|\nabla^2 h_T\right\|_{\mathrm{op}} \leq 4 \cdot 416 < 2400$. Also $h_T(q) \geq -4 - 12(T-1)$ and $h_T(0) = -2$, proving the gap bound.

For the first unfinished index, the derivative of the incoming primitive is one: $\Phi'(\pm 2q_j) = 1$. If $j > 1$, its active incoming gate is at least $\Psi(1) = 1$, since $|q_{j-1}| > 1/2$. If $j = 1$, the root term replaces that gate. The factor two and the nonpositive outgoing contribution give $\partial_j h_T \leq -2$. Finally, every outgoing gate with a hidden predecessor vanishes because $|2q_k| \leq 1/4 < 1/2$, while $\Phi(\pm 2q_{p+1}) = 2 \pm 2q_{p+1}$. Expanding the only surviving link proves (54), including its derivatives on the closed plateau. □

B.2 Radial encoding and a global stationarity certificate

Let $\theta_1, \ldots, \theta_T$ be independent uniform directions on S^{D-1}. For $x = (x_1, \ldots, x_T) \in (\mathbb{R}^D)^T$, write

$$u_j = x_j/\delta, \quad r_j = \sqrt{1 + \|u_j\|^2}, \quad r_v(x_j) = \frac{2\langle v, u_j\rangle}{r_j}, \quad q_j(x) = r_{\theta_j}(x_j), \tag{58}$$

and define

$$F_\theta(x) = Ah_T(q(x)) + B(x), \qquad B(x) = 4A\sum_{j=1}^{T} \|x_j/\delta\|^2. \tag{59}$$

The notation r_v denotes a scalar encoding, whereas r_j denotes the positive radial factor in block j.

Lemma 16 (Radial bounds and guard)**.** *For every v,*

$$|r_v| \le 2\|v\|, \quad \|\nabla r_v\| \le 2\|v\|/\delta, \quad \left\|\nabla^2 r_v\right\|_{\mathrm{op}} \le 6\|v\|/\delta^2. \tag{60}$$

The objective in (59) is bounded below and satisfies

$$F_\theta(0) - \inf F_\theta \le 20AT, \qquad \mathrm{Lip}(\nabla F_\theta) \le C_1 A/\delta^2. \tag{61}$$

At every point in the entire ambient space,

$$\min_{j\le T} |q_j(x)| \le \tfrac{1}{2} \quad \Longrightarrow \quad \|\nabla F_\theta(x)\| \ge A/\delta. \tag{62}$$

Proof. In one block put $u = x/\delta$, $r = \sqrt{1+\|u\|^2}$, and $c = \langle v, u\rangle$. Differentiation gives

$$\nabla r_v = \frac{2}{\delta}(r^{-1}v - r^{-3}cu),$$
$$\nabla^2 r_v = \frac{1}{\delta^2}\left[-2r^{-3}(vu^\top + uv^\top + cI) + 6cr^{-5}uu^\top\right].$$

The matrix $r^{-1}I - r^{-3}uu^\top$ has eigenvalues r^{-1} and r^{-3}, proving the gradient estimate. If $t = \|u\|$, the Hessian norm is at most $6\|v\|\,(t/(1+t^2)^{3/2} + t^3/(1+t^2)^{5/2})/\delta^2$. Each fraction is at most $1/2$, giving (60). The coordinate encodings have disjoint blocks, so $\|Dq\| \le 2/\delta$ and

$$\nabla^2 F_\theta = A(Dq)^\top \nabla^2 h_T(q) Dq + A\sum_j \partial_j h_T(q)\nabla^2 q_j + (8A/\delta^2)I,$$
$$\left\|\nabla^2 F_\theta\right\|_{\mathrm{op}} \le (4\cdot 2400 + 6\cdot 64 + 8)A/\delta^2 < C_1 A/\delta^2.$$

The guard is nonnegative and vanishes at zero, proving the gap estimate.

For (62), choose the first unfinished block and set $b = \partial_j h_T(q) \le -2$. Its gradient, divided by A/δ, is

$$h = \frac{2b}{r}\theta + \left(8 - \frac{bq}{r^2}\right)u, \qquad q = 2\langle\theta, u\rangle/r. \tag{63}$$

Write $c = \langle\theta, u\rangle$ and $v = u - c\theta$. If $c \le 0$ and $\|u\| \ge 1/8$, then $\langle h, u\rangle = 2bc/r^3 + 8\|u\|^2 \ge 8\|u\|^2$, so $\|h\| \ge 1$. If $c \le 0$ and $\|u\| < 1/8$, the chain term has norm at least $2|b|/r^3$, whereas the guard has norm $8\|u\| < 1$. Hence $\|h\| \ge 4/(1+1/64)^{3/2} - 1 > 2$.

If $c \ge 0$, the coefficient of u in (63) is at least eight. Thus $\|v\| \ge 1/8$ again gives $\|h\| \ge 1$. Otherwise $\|v\| < 1/8$, and $|q| \le 1/2$ yields

$$c^2 \le (1+\|v\|^2)/15 < 65/960, \qquad r^2 \le 65/60.$$

Consequently

$$\langle h, \theta\rangle = \frac{2b}{r}(1 - q^2/4) + 8c \le -\frac{15}{4}\sqrt{60/65} + \sqrt{65/15} < -1.$$

These cases exhaust the ambient space and prove the certificate. □

B.3 A SPARSE CONSERVATIVE NOISE POTENTIAL

Define the tail gates

$$\gamma(t) = s(8|t| - 1), \qquad \chi(u) = 1 - s((4u-1)/3), \qquad \Theta_j(q) = \chi\left(\sum_{k=j}^{T}\gamma(q_k)^2\right). \tag{64}$$

Thus $\gamma = 0$ on $[-1/8, 1/8]$, $\gamma = 1$ when $|t| \ge 1/4$, $\chi = 1$ on $(-\infty, 1/4]$, and $\chi = 0$ on $[1, \infty)$. The first two derivatives vanish on these constant regions, including their endpoints.

Lemma 17 (Cutoff bounds and sparsity)**.** *For the incoming gates a_j from Lemma 15,*

$$|a_j| \le 3, \quad \|\nabla_q a_j\| \le 10, \quad \left\|\nabla_q^2 a_j\right\|_{\mathrm{op}} \le 400, \quad \|\nabla_q \Theta_j\| \le 75, \quad \left\|\nabla_q^2 \Theta_j\right\|_{\mathrm{op}} \le 13000. \tag{65}$$

At each q, at most one product $a_j(q)\Theta_j(q)$ has nonzero value, gradient, or Hessian.

Proof. The first three inequalities follow from (56), including $a_1 = 1$. Direct differentiation gives $|\gamma'| \le 15$, $|\gamma''| \le 384$, $|\chi'| \le 5/2$, and $|\chi''| \le 11$. Put $u = \sum_{k \ge j} \gamma(q_k)^2$. Wherever a derivative of χ is nonzero, $u \le 1$, and

$$\|\nabla_q u\| \le 2\,\|\gamma'\|_\infty \sqrt{u} \le 30, \qquad \left\|\nabla_q^2 u\right\|_{\mathrm{op}} \le 2(15^2 + 384) = 1218.$$

The chain rule yields $\|\nabla_q \Theta_j\| \le 75$ and $\left\|\nabla_q^2 \Theta_j\right\|_{\mathrm{op}} \le 11 \cdot 900 + (5/2)1218 = 12945 < 13000$. Where the derivatives of χ vanish, these estimates hold trivially.

Let $p = \max(\{k : |q_k| > 1/4\} \cup \{0\})$. For $j \le p$, the tail contains $\gamma(q_p)^2 = 1$, so Θ_j and its first two derivatives vanish. For $j \ge p + 2$, the incoming gate a_j and its first two derivatives vanish. Only $j = p + 1$ can survive; if $p = T$, no term survives. At equality $|q_k| = 1/4$, all endpoint derivatives vanish, so there is no boundary exception. □

Let $0 < \beta \le 1/4$. One online sample is an array $S = (S_1, \dots, S_T)$, independent across stages conditional on θ, with $S_j \in S^{D-1}$ and law

$$\frac{dP_{\beta\theta_j}}{d\mu}(s) = \frac{1 - \beta^2}{\|s - \beta\theta_j\|^D}. \tag{66}$$

Lemma 10 proves both $\mathbb{E}[S_j \mid \theta] = \beta\theta_j$ and $\mathrm{KL}(P_{\beta\theta_j} \| \mu) \le 5D\beta^2$. Each handle retains this entire array for all later evaluations of the sample. Define

$$\begin{aligned} e_j(x, S_j) &= r_{S_j/\beta - \theta_j}(x_j), \\ N_\theta(x, S) &= -2A \sum_{j=1}^{T} a_j(q(x))\Theta_j(q(x))e_j(x, S_j), \\ f_\theta(x, S) &= F_\theta(x) + N_\theta(x, S), \qquad g_\theta(x, S) = \nabla_x f_\theta(x, S). \end{aligned} \tag{67}$$

To cover finite sums as well, the next lemma allows arbitrary, possibly dependent codes satisfying $\|S_j/\beta - \theta_j\| \le 2/\beta$.

Lemma 18 (Conservative uniformly bounded noise)**.** *The sample losses in (67) are C^2, bounded below, and have globally Lipschitz gradients. For every sample array in the common support and every point,*

$$|N_\theta(x, S)| \le 24A/\beta, \qquad \|\nabla N_\theta(x, S)\| \le \frac{4000A}{\beta\delta}, \qquad \left\|\nabla^2 N_\theta(x, S)\right\|_{\mathrm{op}} \le \frac{1.5 \cdot 10^6 A}{\beta\delta^2}. \tag{68}$$

Whenever $\mathbb{E}S_j = \beta\theta_j$, one has $\mathbb{E}f_\theta = F_\theta$ and $\mathbb{E}g_\theta = \nabla F_\theta$. Moreover,

$$\nabla N_\theta(x, S) \ne 0 \quad \Longrightarrow \quad \min_j |q_j(x)| \le 1/4. \tag{69}$$

Proof. The centered-code bound and (60) give

$$|e_j| \le 4/\beta, \qquad \|\nabla e_j\| \le 4/(\beta\delta), \qquad \left\|\nabla^2 e_j\right\|_{\mathrm{op}} \le 12/(\beta\delta^2).$$

Composing (65) with the encodings gives

$$\begin{aligned} |a_j| \le 3, \qquad \|\nabla_x a_j\| &\le 20/\delta, \qquad \left\|\nabla_x^2 a_j\right\|_{\mathrm{op}} \le (4 \cdot 400 + 6 \cdot 10)/\delta^2 = 1660/\delta^2, \\ \|\nabla_x \Theta_j\| &\le 150/\delta, \qquad \left\|\nabla_x^2 \Theta_j\right\|_{\mathrm{op}} \le (4 \cdot 13000 + 6 \cdot 75)/\delta^2 < 53000/\delta^2. \end{aligned}$$

In the second Hessian term the matrices $\nabla^2 q_k$ have disjoint blocks, so its norm is bounded by $6\,\|\nabla_q \Theta_j\|_\infty / \delta^2$; there is no sum growing with T. For the gradient of $a_j\Theta_j e_j$, the three product terms give

$$\|\nabla(a_j\Theta_j e_j)\| \le \frac{20 \cdot 4 + 3 \cdot 150 \cdot 4 + 3 \cdot 4}{\beta\delta} < \frac{2000}{\beta\delta}.$$

For completeness, the Hessian product rule reads

$$\begin{aligned} \nabla^2(a\Theta e) = {} & a\Theta\nabla^2 e + ae\nabla^2\Theta + \Theta e\nabla^2 a \\ & + a(\nabla\Theta\nabla e^\top + \nabla e\nabla\Theta^\top) + \Theta(\nabla a\nabla e^\top + \nabla e\nabla a^\top) \\ & + e(\nabla a\nabla\Theta^\top + \nabla\Theta\nabla a^\top). \end{aligned}$$

After multiplication by $\beta\delta^2$, the six norm bounds are

$$36 + 636000 + 6640 + 3600 + 160 + 24000 = 670436 < 750000.$$

Lemma 17 allows at most one nonzero summand, also after composition and differentiation. Multiplying by $2A$ proves the two derivative bounds; the value bound is $2A \cdot 3 \cdot 4/\beta$. These uniform bounds justify differentiation under expectation, and the linearity of r_v in v proves both mean identities. The value bound and the lower boundedness of F_θ prove lower boundedness of every sampled loss. Finally, if $p = T$ in the sparsity proof, the noise and its gradient vanish. Otherwise $|q_{p+1}| \le 1/4$, proving (69). □

B.4 Exact prefix identities, including both value channels

Lemma 19 (All-point prefix potential). *For $p < T$ define on the entire ambient space*

$$\begin{aligned} V_p(x; \theta_{\le p}, S_{\le p+1}) = B(x) + Ah_T(q^{\le p}(x)) \\ - 2A \sum_{j=1}^{p} a_j(q^{\le p}(x))\Theta_j(q^{\le p}(x))e_j(x, S_j) \\ - \frac{2A}{\beta} a_{p+1}(q^{\le p}(x))r_{S_{p+1}}(x_{p+1}). \end{aligned} \tag{70}$$

Set $V_T = f_\theta$. If $|q_k(x)| \le 1/8$ for all $k > p$, then

$$(f_\theta(x, S), \nabla f_\theta(x, S)) = (V_p(x), \nabla V_p(x)). \tag{71}$$

Furthermore, granting all scalars

$$\langle \theta_j, w_j(x) \rangle, \qquad w_j(x) = \frac{x_j/\delta}{\sqrt{1 + \|x_j/\delta\|^2}}, \qquad j \le T, \tag{72}$$

provides enough information to recover the exact population value at that point, using one linear measurement per hidden block with a measurement vector of norm at most one.

Proof. Equation (54) writes the hidden population part as $-2Aa_{p+1}q_{p+1}$. All hidden γ values and their first derivatives vanish. Thus replacing hidden coordinates by zero leaves each old cutoff and its gradient unchanged; $\Theta_{p+1} = 1$ with zero gradient; and all future incoming gates and their gradients vanish. The frontier population and noise terms sum to

$$-2Aa_{p+1}\big(q_{p+1} + e_{p+1}\big) = -\frac{2A}{\beta} a_{p+1} r_{S_{p+1}}.$$

This proves the value identity and, using the displayed derivative properties, its gradient identity. The endpoints also follow by continuity. Formula (70) is defined even when the plateau event fails; hence it supplies genuine response maps for the comparison experiments in Lemma 13. The measurements in (72) determine every q_j and therefore $F_\theta(x)$, and $\|w_j(x)\| \le 1$. □

B.5 Quantitative information accounting for this construction

We check how the construction satisfies the assumptions of the common information lemmas. A crossing is $|q_i(x)| > 1/8$, equivalently $|\langle \theta_i, w_i(x) \rangle| > 1/16$. Let M include all charged points and the arbitrary final output. In a retained-sample or active-component interaction, $M = N + 1$; in R nonadaptive batches of at most K points, $M = RK + 1$. Choose

$$D \ge \max\{2^{24},\ 4096M,\ 8192 \log(256T)\}. \tag{73}$$

For a finite table, also require $D \ge 8 \log(4n/\beta)$. Let

$$b = \lfloor 2^{-24}\beta^{-2} \rfloor \ge 2. \tag{74}$$

Online spherical codes have joint information at most $5Db\beta^2$ per target stage; Lemma 11 gives at most $Db\beta^2$ for the first $b \le n/8$ distinct finite-table rows. Lemma 13 applies to the exact maps of Lemma 19. Its skipped-coordinate event J has probability at most

$$\mathbb{P}(J) \le 2Te^{-D/4096} \le 1/128. \tag{75}$$

Writing $A_{i,b}$ for a first crossing of stage i with fewer than b prior relevant code revelations, the same lemma gives

$$\mathbb{P}(A_{i,b} \cap J^c) \le \frac{5Db\beta^2 + \log 2}{D/4096 - \log 2} \le \frac{6 \cdot 2^{-24}}{2^{-13}} < \frac{1}{64}. \tag{76}$$

The continuation starts at time zero, so this is an unconditional bound on the intersection event; it does not assume that the hidden direction remains uniform after earlier exact population values have been observed. Throughout the reference procedure, every projection channel is retained, while stochastic replies are erased once the target-code budget is exhausted.

Let L_i count first revelations of a handle–stage pair, or of a component–stage pair, before crossing stage i. One evaluation can charge at most one such pair, so $\sum_i L_i \le N$. In simultaneous fresh batches, one batch is charged to at most two stages, giving $\sum_i L_i \le 2R$. Consequently, if

$$N \le Tb/2 \quad \text{or, respectively,} \quad R \le Tb/4, \tag{77}$$

completion forces at least $T/2$ events $A_{i,b}$. Markov's inequality and (76) bound completion on J^c by $1/32$. If $A/\delta = 4\varepsilon$, Lemma 16 then implies

$$\mathbb{P}\{\|\nabla F_\Theta(\widehat{x})\| > 2\varepsilon\} \ge 1 - 1/32 - 1/128 = 123/128. \tag{78}$$

The corresponding expected norm exceeds ε. Since all directions, tables, and sampled-function laws are chosen before interaction, we can average this inequality over their prior to fix an instance on which the given randomized algorithm fails. The dimension in (73) depends on the budget and K, not on the algorithm's random realization.

B.6 Deterministic and statistical lower bounds

Lemma 20 (Deterministic depth). *For $\varepsilon \le c_0\sqrt{\Delta L}$, an exact first-order oracle requires $\Omega(Q)$ evaluations. It also requires $\Omega(Q)$ rounds with nonadaptive batches of any fixed finite size K, even with exact values and an arbitrary unqueried output.*

Proof. Use (59) with

$$A = 16C_1\varepsilon^2/L, \quad \delta = 4C_1\varepsilon/L, \quad T = \lfloor Q/(320C_1)\rfloor \ge Q/(640C_1). \tag{79}$$

The analytic bounds give admissibility and $A/\delta = 4\varepsilon$. The exact all-point deterministic prefix potential is

$$W_p(x) = B(x) + Ah_T(q^{\le p}(x)) - 2Aa_{p+1}(q^{\le p}(x))q_{p+1}(x), \qquad p < T.$$

It depends only on $\theta_{\le p+1}$ and agrees in value and gradient with F_θ whenever all coordinates beyond p lie in the hidden plateau. Set $W_T = F_\theta$.

For each simultaneous batch, cap every response index using the same before-batch frontier ν: at a point use the largest crossed index at most $\nu + 1$. After the batch, update ν to the largest of these indices and its previous value. Thus $\nu_t \le \nu_{t-1} + 1$, whence $\nu_R \le R$ and the arbitrary output has capped frontier at most $R + 1$. For a target i, retain every preselected point through the first batch containing a response of prefix index at least $i - 1$, and stop before returning any response from that batch. All of its points are chosen from the preceding transcript, before a gradient can disclose the target direction. Earlier prefix responses depend directly only on $\theta_{<i}$; the target direction has supplied only the exact scalar measurements already granted. This is precisely the stopped list in Corollary 14, including the complete entry batch and the output when there is no entry. Its projection bound and the union bound over i, with (73), show that a query or output skips the permitted next coordinate with probability at most $1/128$. On its complement the capped and exact transcripts agree. If $R + 1 < T$, the output has an unfinished coordinate and therefore gradient norm at least 4ε. For instance $R \le T/2$ suffices here because $T \ge 4$. This proves the round assertion; take $K = 1$ for evaluations. Taking every component identical gives the same bound in every finite-sum class. □

Lemma 21 (Statistical term with exact values). *Under the same accuracy condition, the online class with additive noise requires $\Omega(m)$ evaluations or simultaneous fresh-sample rounds, where $m = \sigma^2/\varepsilon^2$. The finite-population version requires $\Omega(r)$ evaluations, where $r = \min\{n, m\}$. These lower bounds allow both value channels and arbitrary reuse.*

Proof. It suffices first to consider $m \geq R_0$ online or $r \geq R_0$ in the finite case. Set $a = 8\varepsilon$ and

$$\beta = B_0/\sqrt{m} \quad \text{online}, \qquad \beta = \max\{n^{-1/2}, B_0/\sqrt{m}\} \quad \text{in the finite case.}$$

Take a uniform $\theta \in S^{D-1}$. Online, use the spherical code in (66); in the finite case use the fixed table $S_i = \beta\theta + U_i$ of Lemma 11, which satisfies $\sum_i U_i = 0$ and $\|U_i\| \leq 2$ identically. Define

$$F_\theta(x) = \frac{L}{2}\|x\|^2 - a\langle\theta, x\rangle, \qquad f(x, S) = \frac{L}{2}\|x\|^2 - \frac{a}{\beta}\langle S, x\rangle. \tag{80}$$

In the finite case assign the ith row to f_i. Exact averaging gives $F_\theta = n^{-1}\sum_i f_i$. The gap is $a^2/(2L) = 32\varepsilon^2/L \leq \Delta$. Every sampled or component loss is an L-smooth, lower-bounded convex quadratic. Its centered gradient noise is constant in x, so it has Lipschitz constant zero, and has norm at most $2a/\beta \leq (16/B_0)\sigma \leq \sigma$. For a single sample or component, the entire transcript is a function of its one code, while an exact population value supplies only the scalar $\langle\theta, x\rangle$.

If $\|\nabla F_\theta(\widehat{x})\| \leq 2\varepsilon$, then $\|L\widehat{x}/a - \theta\| \leq 1/4$. Therefore $\widehat{x} \neq 0$ and

$$\langle\theta, \widehat{x}/\|\widehat{x}\|\rangle \geq \frac{1 - 1/4}{1 + 1/4} = 3/5. \tag{81}$$

Append that normalized output as one additional exact measurement. Under the reference spherical codes, or the reference joint row pool, the codes are independent of θ. Lemma 9, even at the weaker threshold $1/16$, bounds the reference probability of success by $2e^{-D/4096}$. With b as in (74) and D as in (73) with $T = 1$, binary data processing bounds the actual success probability by the right side of (76). Adaptive row choices are allowed because Lemma 11 controls the whole first-b row pool; cached repetitions reveal no additional row. The same statement applies to retained online handles, and a whole simultaneous batch uses only one code. Hence fewer than b evaluations, or b online fresh-sample rounds, fail the expected-norm criterion.

For online samples $\beta^{-2} = m/B_0^2$. In the finite case

$$r/B_0^2 \leq \beta^{-2} = \min\{n, m/B_0^2\} \leq r. \tag{82}$$

In either high-parameter regime $\beta^{-2} \geq 2^{28}$, so $b \geq 2^{-25}\beta^{-2}$; also $b \leq n/8$ in the finite case. This gives the asserted statistical orders. If $m < R_0$ or $r < R_0$, the zero-noise lower bound of Lemma 20, together with $Q \geq 1$, already dominates that statistical term up to the universal factor R_0. Thus every parameter case is covered. □

B.7 Online noise regularity: complete parameter choices

To prove the regularity-sensitive lower bound, fix $L, \Delta > 0$ and $\sigma, H \geq 0$. When $m \geq R_0$ and $H > 0$, choose

$$\begin{aligned} &\beta = B_0\varepsilon/\sigma, && \lambda = \min\{L, \beta H/256\}, \\ &A = 16C_1\varepsilon^2/\lambda, && \delta = 4C_1\varepsilon/\lambda, \\ &T = \left\lfloor \frac{\Delta\lambda}{320C_1\varepsilon^2} \right\rfloor. \end{aligned} \tag{83}$$

These choices give $\beta \leq 2^{-14}$, $A/\delta = 4\varepsilon$, and

$$\begin{aligned} F_\theta(0) - \inf F_\theta \leq \Delta, \qquad \mathrm{Lip}(\nabla F_\theta) \leq \lambda \leq L, \\ \sup_{x,S} \|g_\theta(x, S) - \nabla F_\theta(x)\| \leq 16000\varepsilon/\beta < \sigma/4, \\ \mathrm{Lip}(\nabla N_\theta(\cdot, S)) \leq 150\lambda/\beta \leq (150/256)H < H. \end{aligned} \tag{84}$$

These hard instances satisfy the stronger samplewise centered-noise bound and therefore the prescribed mean-square bound. Set $X = \Delta\lambda/(320C_1\varepsilon^2)$. If $X \geq 2$, then $T \geq X/2$ and $b \geq m/2^{57}$. Substituting these quantities into the resource and failure bounds gives a lower bound proportional to

$$\frac{\Delta\lambda m}{\varepsilon^2} = \min\left\{Qm, 256\frac{\Delta H\sigma}{\varepsilon^3}\right\} \asymp \min\left\{Qm, \frac{\Delta H\sigma}{\varepsilon^3}\right\}. \tag{85}$$

The constants are the same for every finite K in the nonadaptive-batch round statement, although dimension increases with RK.

If $X < 2$, the desired interaction term is at most $640C_1m$ by (85), so Lemma 21 supplies it. If $H = 0$, the interaction term is zero and the deterministic and additive-noise quadratic constructions suffice. Taking the maximum over these separate instances is at least one third the sum of the three lower bounds. Thus the full lower rate is

$$\Omega\left(Q + m + \min\left\{Qm, \frac{\Delta H\sigma}{\varepsilon^3}\right\}\right). \tag{86}$$

If $m < R_0$, the entire displayed expression is at most $(1 + 2R_0)Q$, so Lemma 20 proves the same assertion. This covers $\sigma = 0$, all integer-budget endpoints, and every stated accuracy/noise ratio.

With no prescribed sample smoothness, set $\lambda = L$ in (83). The assumption $Q \geq 2^{40}$ ensures $T \geq Q/(640C_1) \geq 1$, so the same chain gives $\Omega(Qm)$ when $m \geq R_0$. Its sampled losses have a finite common smoothness constant at most $L(1 + 150/\beta)$, which this class permits. For $m < R_0$ the deterministic lower bound supplies the full expression. Hence

$$\Omega(Q + Qm) \tag{87}$$

holds for arbitrary retained samples in evaluation cost, and for every finite nonadaptive common-sample batch size in round cost. The construction is conservative, so it also establishes the weaker gradient-oracle theorem with one query per fresh sample.

If ordinary sample gradients are mean-square L_s-Lipschitz, Jensen gives population smoothness $L_* := \min\{L, L_s\}$, and the variance identity gives

$$\begin{aligned}\mathbb{E}\left\|[g(x,Z) - \nabla F(x)] - [g(y,Z) - \nabla F(y)]\right\|^2 \\ = \mathbb{E}\left\|g(x,Z) - g(y,Z)\right\|^2 - \left\|\nabla F(x) - \nabla F(y)\right\|^2 \leq L_s^2\left\|x - y\right\|^2.\end{aligned} \tag{88}$$

For the lower bound, apply the proved construction with population budget $L_*/2$ and centered-noise budget $L_s/2$; each sample gradient is then samplewise L_s-Lipschitz. After reducing c_0 by the universal factor $\sqrt{2}$ for this rescaling, its lower rate is

$$\Omega\left(Q_* + m + \min\left\{Q_*m, \frac{\Delta L_s\sigma}{\varepsilon^3}\right\}\right), \qquad Q_* = \Delta L_*/\varepsilon^2. \tag{89}$$

The deterministic and statistical auxiliary instances also have sample smoothness at most L_s. When $L_s = L$, the interaction is $\min\{Qm, Q\sqrt{m}\}$. These lower bounds hold for fresh-only access as well, but the matching regularity-sensitive upper bounds use common-sample pairs. Setting $H = 0$ yields the additive-noise rate $\Omega(Q + m)$.

Additive noise with a finite evaluation cap per sample. Suppose each retained sample may be queried at most $J \geq 2$ times and the resource is the number S of distinct samples. The optimal lower bound is

$$S = \Omega(1 + m + Q/J), \tag{90}$$

including exact population values at the charged points. Since S samples permit at most JS evaluations, Lemma 20 gives $JS = \Omega(Q)$. For the statistical term, when $m \geq R_0$, use the online quadratic from Lemma 21, with $\beta = B_0/\sqrt{m}$ and $b = \lfloor 2^{-24}\beta^{-2}\rfloor$. If $S \leq b$, the interaction exposes at most b code vectors and $Jb + 1$ exact measurements, including the output. Apply Lemma 12 with the joint-code divergence at most $5Db\beta^2$ and $D \geq \max\{2^{24}, 4096(Jb + 1)\}$. Its bound is precisely (76); success at gradient threshold 2ε requires (81), so $S = \Omega(m)$. An algorithm that makes no queries cannot guarantee the target accuracy, because its output is independent of the hidden direction in (80). Thus at least one sample is necessary. The same projection bound gives a failing admissible quadratic with zero noise. For $m < R_0$, this one-sample lower bound is also $\Omega(1 + m)$. Taking the maximum proves (90). The upper budget in (205) is $O(1 + m + Q/(J - 1))$, which is within a factor two of the lower rate for every integer $J \geq 2$.

Without an exact population-value channel, unrestricted retention has optimal distinct-sample cost $\Theta(1 + m)$: the entire quadratic transcript, irrespective of the number of queries, is a function of its sample codes; Lemma 12 then needs only the normalized output measurement. The retained-sample algorithm in Section F.3 matches this lower bound. With exact population values, this statistical argument also requires a bound on the number of evaluations. Counting only distinct samples while allowing unlimited reuse would permit arbitrarily many exact population-value queries through one retained sample; our lower bound does not cover that model.

B.8 Fixed finite populations with arbitrary component indices

At every stage j, take an independent balanced table from Lemma 11:

$$S_a^j = \beta\theta_j + U_a^j, \qquad \sum_{a=1}^n U_a^j = 0, \qquad \|U_a^j\| \le 2, \qquad a \in [n]. \tag{91}$$

We fix the entire table before interaction, drawing it from a law that is exchangeable in its rows. Define actual component functions by

$$f_a(x) = F_\theta(x) - 2A\sum_{j=1}^T a_j(q(x))\Theta_j(q(x))r_{S_a^j/\beta-\theta_j}(x_j). \tag{92}$$

Here a indexes components, while a_j denotes the incoming gate at stage j. The exact row-balance identity gives, pointwise,

$$\frac{1}{n}\sum_{a=1}^n f_a(x) = F_\theta(x), \qquad \frac{1}{n}\sum_{a=1}^n \nabla f_a(x) = \nabla F_\theta(x). \tag{93}$$

Since $S_a^j/\beta - \theta_j = U_a^j/\beta$ has norm at most $2/\beta$, Lemmas 18 and 19 apply to every component, without independence between rows or a unit-norm assumption on S_a^j. In particular every component is lower bounded, C^2, and globally smooth; its centered-noise bounds are exactly (68).

For clarity, the bounded table is a centered Gaussian on $\{(u_1,\dots,u_n) : \sum_a u_a = 0\}$ with density proportional to $\exp[-(D/2)\sum_a \|u_a\|^2]$, conditioned on $\max_a \|u_a\| \le 2$. Lemma 11 proves that its first $b \le n/8$ shifted rows have a θ_j-independent joint reference with KL at most $Db\beta^2$ when $D \ge 8\log(4n/\beta)$. Assigning the next row of this pool to the next previously unqueried component index, and caching that assignment, simulates every adaptive index policy. Index decisions based on earlier component values, exact population values, and other-stage tables do not inspect unseen target-stage rows. Thus the joint-row KL bound applies to the entire adaptive transcript. The argument relies on exchangeability and data processing, without positing a conditional one-row information bound. A component index may already have been used at other stages; those stages have independent tables. The explicit prefix formula also shows that both returned component values and gradients depend on the next hidden direction only through its one selected table row. Exact $F(x)$ values are simulated by (72). These observations verify every hypothesis of the information accounting in (73)–(78) for active component queries.

Assume first $r = \min\{n,m\} \ge R_0$. Choose

$$\beta = \max\{n^{-1/2}, B_0\varepsilon/\sigma\}, \qquad k = \beta^{-2} = \min\{n, m/B_0^2\}, \qquad b = \lfloor 2^{-24}k \rfloor. \tag{94}$$

Then $\beta \le 2^{-14} < 1/4$, $r/B_0^2 \le k \le r$, $b \ge 2^{-25}k \ge r/2^{57}$, and $b \le n/8$. With no prescribed component smoothness, use (79). The noise radius is $16000\varepsilon/\beta < \sigma/4$, the component smoothness is finite and at most $L(1+150/\beta)$, and the lower cost Tb is $\Omega(Qr)$. The deterministic bound supplies $\Omega(Q)$. If $r < R_0$, the same deterministic instance gives the entire rate, since $Q(1+r) \le (1+R_0)Q$. This proves the lower bound with no prescribed component-smoothness constant

$$\Omega\big(Q(1+\min\{n,m\})\big), \tag{95}$$

including $n = 1$ and $\sigma = 0$.

Now prescribe a centered-component Lipschitz constant $H > 0$ and take $\lambda = \min\{L, \beta H/256\}$, with A, δ, T as in (83). Equation (84) remains valid for every component. If the unfloored chain length is at least two, its cost is proportional to

$$\frac{\Delta\lambda k}{\varepsilon^2} = \min\left\{Qk, \frac{\Delta H\sqrt{k}}{256\varepsilon^2}\right\}. \tag{96}$$

By $r/B_0^2 \le k \le r$ and $256B_0 \le B_0^2$, this quantity lies between $B_0^{-2}I_r$ and I_r, where

$$I_r = \min\left\{Qr, \frac{\Delta H\sqrt{r}}{\varepsilon^2}\right\}. \tag{97}$$

If the unfloored chain length is less than two, then $\Delta\lambda/\varepsilon^2 < 640C_1$ and $I_r \le B_0^2\Delta\lambda k/\varepsilon^2 < 640C_1B_0^2 r$. Lemma 21 covers this case. For $H = 0$ the interaction is zero. For $r < R_0$, both r and I_r are bounded by universal multiples of Q. Taking the maximum with the deterministic and finite-statistical instances proves the complete lower half of Theorem 2:

$$\Omega\left(Q + r + \min\left\{Qr, \frac{\Delta H\sqrt{r}}{\varepsilon^2}\right\}\right). \tag{98}$$

All statements concern component evaluations, and impose no restriction on index order or reuse.

If every component gradient is L_s-Lipschitz, the finite-average version of (88) gives centered mean-square regularity L_s, and the effective population constant is $L_* := \min\{L, L_s\}$. For the lower bound prescribe population constant $L_*/2$ and centered constant $L_s/2$, so the sum of the two bounds is at most L_s for every component. The rate becomes

$$\Omega\left(Q_* + r + \min\left\{Q_* r, \frac{\Delta L_s\sqrt{r}}{\varepsilon^2}\right\}\right), \qquad Q_* = \Delta L_*/\varepsilon^2, \tag{99}$$

under $\varepsilon \le (c_0/\sqrt{2})\sqrt{\Delta L_*}$. For the usual one-constant class, $L = L_s$ and $r \ge 1$ give $\min\{Qr, Q\sqrt{r}\} = Q\sqrt{r}$; if $r < 1$ this interaction and the extra $Q\sqrt{r}$ are both absorbed by Q. Hence the equivalent one-constant expression is $\Omega(Q + r + Q\sqrt{r})$.

B.9 Relative variance and strong growth

Proposition 22 (Relative-noise lower bounds)**.** *Suppose the noise has radius σ and satisfies*

$$\mathbb{E}\left\|g(x,Z) - \nabla F(x)\right\|^2 \le \rho\left\|\nabla F(x)\right\|^2. \tag{100}$$

Without a prescribed sample-smoothness bound, the online lower complexity is $\Omega(Q[1+\min\{m,\rho\}])$. For a population of n components, it is $\Omega(Q[1+\min\{n,m,\rho\}])$. The online result also holds in nonadaptive shared-sample rounds, uniformly in finite K. The hard instances satisfy the stronger pointwise condition

$$\|g(x,z) - \nabla F(x)\| \le \sqrt{\rho}\,\|\nabla F(x)\| \quad \textit{for every } x, z. \tag{101}$$

Both value channels and arbitrary retained samples or adaptive component indices remain allowed.

Proof. Define

$$\sigma_{\text{eff}} = \min\{\sigma, \varepsilon\sqrt{\rho}\}, \qquad m_{\text{eff}} = \min\{m, \rho\}. \tag{102}$$

When the effective online or finite information parameter is at least R_0, use the construction without a prescribed sample-smoothness constant, with noise radius σ_{eff} and population scale $\lambda = L$. By (69) and Lemma 16, whenever its noise is nonzero, $\|\nabla F_\theta(x)\| \ge 4\varepsilon$. Consequently

$$\|g_\theta(x,z) - \nabla F_\theta(x)\| \le \sigma_{\text{eff}} \le \varepsilon\sqrt{\rho} \le \frac{\sqrt{\rho}}{4}\|\nabla F_\theta(x)\|.$$

At points with zero noise, (101) is immediate. Thus every sample satisfies the pointwise condition and therefore (100); in the finite case this holds for every component, so also for its uniform mean relative variance.

When $m_{\text{eff}} \ge R_0$ online, the chain lower bound is $\Omega(Qm_{\text{eff}})$. In the finite case the same argument applies when $\min\{n, m_{\text{eff}}\} \ge R_0$, and gives $\Omega(Q\min\{n, m_{\text{eff}}\})$. Below the corresponding fixed threshold the exact-gradient construction, which obeys every relative bound, supplies the entire claimed expression up to the universal factor $1 + R_0$. This includes $\rho = 0$, $\sigma = 0$, and $n = 1$. No additive statistical instance is used for this argument: such an instance need not satisfy strong growth. □

By unbiasedness, (100) is equivalent to $\mathbb{E}\left\|g(x,Z)\right\|^2 \le (1+\rho)\left\|\nabla F(x)\right\|^2$. Thus $1+\rho$, rather than ρ, is the uncentered strong-growth coefficient. Every population stationary point in the constructed instances is stationary for every sampled loss, by (101). No common-minimizer assumption is needed.

B.10 SMOOTHNESS AND ACCESS REFINEMENTS

Infinitely differentiable sampled losses. The C^2 sample restriction can be strengthened to C^∞ in all the preceding lower bounds. To verify this without changing the plateaus, let ρ_η be a nonnegative C^∞ bump of integral one, supported on $[-\eta, \eta]$, with $\eta = 1/100$, and replace s everywhere by

$$s_\infty(t) = (s * \rho_\eta)((1+2\eta)t - \eta).$$

This is zero for $t \le 0$, one for $t \ge 1$, and monotone. Differentiating the convolution gives $|s'_\infty| < 2$ and $|s''_\infty| < 7$. The primitive retains the affine plateau, $0 \le \Phi \le 4$, and $|\Phi''| \le 2$, so the scalar and radial objective bounds remain valid. For the changed cutoffs,

$$|\gamma'| \le 16, \quad |\gamma''| \le 448, \quad |\chi'| \le 3, \quad |\chi''| \le 13.$$

On the derivative support of the tail cutoff, $\|\nabla u\| \le 32$ and $\left\|\nabla^2 u\right\|_{\mathrm{op}} \le 1408$. Hence the q-gradient and Hessian of Θ_j are at most 96 and 18000, and their composed bounds are $192/\delta$ and $73000/\delta^2$. The gradient product bound is less than $2500/(\beta\delta)$. The six Hessian product bounds are now

$$36 + 876000 + 6640 + 4608 + 160 + 30720 = 918164 < 10^6.$$

The resulting common bounds are

$$\|\nabla N\| \le 5000A/(\beta\delta), \qquad \left\|\nabla^2 N\right\|_{\mathrm{op}} \le 2 \cdot 10^6 A/(\beta\delta^2). \tag{103}$$

The previous scale choices still respect the prescribed parameters: $5000A/(\beta\delta) = 20000\varepsilon/\beta < \sigma$ and $2 \cdot 10^6 A/(\beta\delta^2) = 200\lambda/\beta \le (200/256)H$. Sparsity, the perturbation-support property, and exact prefix identities are unchanged. This refinement gives derivatives of every order without prescribing numerical bounds on them.

A globally bounded population gradient. If desired, replace the quadratic guard in (59) by $\widetilde{B}(x) = 8A\sum_j(\sqrt{1 + \|x_j/\delta\|^2} - 1)$. Its gradient and Hessian in a block are

$$\nabla\widetilde{B} = \frac{8A}{\delta}\frac{u}{r}, \qquad \nabla^2\widetilde{B} = \frac{8A}{\delta^2}(r^{-1}I - r^{-3}uu^\top),$$

so it is nonnegative, vanishes at zero, and has Hessian norm at most $8A/\delta^2$. The previous gap and smoothness bounds are unchanged. For its stationarity certificate set $w = u/r$, $t = \langle\theta, w\rangle = q/2$, $v = w - t\theta$, and $b \le -2$. The scaled block gradient becomes

$$h = \frac{2b}{r}(\theta - tw) + 8w.$$

If $t \le 0$, then $\langle h, w\rangle = 2bt/r^3 + 8\|w\|^2 \ge 8\|w\|^2$. Thus $\|w\| \ge 1/8$ gives $\|h\| \ge 1$. If $\|w\| < 1/8$, the chain term has norm at least $4(63/64)^{3/2}$ and the guard term has norm less than one, also giving $\|h\| > 1$. If $t \ge 0$, the coefficient of v is $8 - 2bt/r \ge 8$, so $\|v\| \ge 1/8$ suffices. Otherwise $|t| \le 1/4$ and $\|v\| < 1/8$ imply $\|w\|^2 < 5/64$, and

$$\langle h, \theta\rangle \le -\frac{15}{4}\sqrt{59/64} + 2 < -1.$$

Thus (62) still holds. Since the new guard is known, every prefix and information proof is unchanged. The additional global bound, under (79) or (83), is

$$\|\nabla F_\theta(x)\| \le \frac{A}{\delta}(2 \cdot 64 + 8)\sqrt{T} = 544\varepsilon\sqrt{T} \le \frac{544}{\sqrt{320C_1}}\sqrt{\Delta\lambda} < 0.305\sqrt{\Delta L}.$$

These modified examples also satisfy $\|g(x,z)\| \le \sigma + 0.305\sqrt{\Delta L}$ for every x, z.

Sequential use of one sample within a round in the high-noise regime. There is also a stronger round statement when $m \ge R_0$: each round may make up to a finite K adaptively chosen queries to its fresh sample, provided the sample is discarded at the end of the round. Keep the per-evaluation capped frontier and the retained-sample information proof. If a sample's consecutive block starts with frontier a and ends with frontier c, the stages charged to that sample lie in $\{a+1, \dots, c+1\} \cap [T]$. Indeed an eligible target stage is one more than a response prefix, is beyond the before-query frontier,

and is at most one beyond the after-query frontier. Therefore the sample is charged at most $c - a + 1$ times, and summing the consecutive blocks telescopes to

$$\sum_i L_i \le R + \nu_{RK} \le R + T.$$

Here $b \ge 16$ by the chosen high-noise regime. If $R \le Tb/4$, then $R + T \le Tb/2$, so the same argument counting stages crossed before b observations proves an $\Omega(Qm)$ round lower bound for the general class. We choose the dimension using $M = RK + 1$. Fresh-sample SGD gives the matching upper bound with only one query per round. This additional statement is restricted to the displayed high-noise regime: it is not a round bound for arbitrary sample retention or for the low-noise endpoint. In particular, bounded-error descent with step $1/L$ on a single retained sample satisfies

$$\frac{1}{N}\sum_{t<N} \|\nabla F(x_t)\|^2 \le \frac{2\Delta L}{N} + \sigma^2,$$

which follows by summing

$$F(x_{t+1}) \le F(x_t) - \frac{\|\nabla F(x_t)\|^2}{2L} + \frac{\|g(x_t, z) - \nabla F(x_t)\|^2}{2L}.$$

Thus, in the low-noise regime, adaptive evaluations of one sample can support multiple descent steps within a round.

C Gradient-dominated and strongly convex objectives

We prove the lower bounds in Theorems 4 and 5. Throughout this section, $0 < \eta \le \Delta/128$, $\kappa = L/\mu \ge 1$, and $S = \sigma^2/(\mu\eta)$. We use zero as the starting point and take all constants below to be numerical and independent of dimension. The hard-instance dimension may nevertheless depend on the announced query budget and finite batch width. The common information lemmas in Appendix A apply to arbitrary randomized algorithms and include their final, possibly unqueried output.

C.1 A scalar conservative extension

The following lemma hides directions outside a prescribed ball while preserving conservativity and curvature; we also use it in the convex constructions.

Lemma 23 (Scalar cutoff extension)**.** *Let $h : \mathbb{R}^d \to \mathbb{R}$ be globally convex with $h(0) = 0$, and suppose its gradient is L_0-Lipschitz on a neighborhood of the closed ball $B(0, 2R)$. Suppose $\nabla h(x^*) = 0$ for some $\|x^*\| \le R$. Define*

$$s(t) = \begin{cases} 0 & t \le 0, \\ 10t^3 - 15t^4 + 6t^5 & 0 < t < 1, \\ 1 & t \ge 1, \end{cases} \qquad C(x) = 1 - s((\|x\| - R)/R).$$

For $\|x\| < 2R$, set

$$\widetilde{h}(x) = C(x)h(x) + 1000L_0(\|x\| - R)_+^2,$$

and use the guard term alone for $\|x\| \ge 2R$. Then $\widetilde{h}$ is globally convex and $3000L_0$-smooth, equals h on $B(0, R)$, and has the same minimum and initial gap. If h is globally μ-strongly convex, so is $\widetilde{h}$. Outside $B(0, 2R)$ its value and gradient are independent of h. Every query inside that ball uses only $(h(x), \nabla h(x))$. Every output x can be converted using at most one additional such query to a point $y \in B(0, 2R)$ satisfying

$$h(y) - h(x^*) \le \widetilde{h}(x) - \widetilde{h}(x^*).$$

Proof. Smoothness and stationarity imply $\|\nabla h(0)\| \le L_0 R$. Thus for $r = \|x\| \le 2R$,

$$\|\nabla h(x)\| \le 3L_0 R, \qquad |h(x)| \le L_0 Rr + L_0 r^2/2 \le 4L_0 R^2.$$

On $R < r < 2R$, write $t = (r - R)/R$ and $e = x/r$. For $c(t) = 1 - s(t)$, direct differentiation gives

$$|c'| \le 30t^2, \quad |c''| \le 60t, \quad 1 - c \le 10t^3, \quad \nabla C = \frac{c'}{R}e, \quad \nabla^2 C = \frac{c''}{R^2}ee^\top + \frac{c'}{Rr}(I - ee^\top).$$

At points of twice differentiability,

$$\nabla^2(Ch) = C\nabla^2 h + D, \quad D = h\nabla^2 C + \nabla C \nabla h^\top + \nabla h \nabla C^\top.$$

The preceding estimates yield

$$\|D\|_{\mathrm{op}} \le 4L_0(60t + 30t^2) + 180L_0 t^2 \le 540 L_0 t.$$

The guard Hessian has radial eigenvalue $2000L_0$ and tangential eigenvalues

$$2000L_0 \frac{r-R}{r} = 2000L_0 \frac{t}{1+t} \ge 1000 L_0 t.$$

Consequently $\nabla^2 \widetilde{h} \succeq C\nabla^2 h + 460 L_0 t I \succeq 0$ and

$$\left\|\nabla^2 \widetilde{h}\right\|_{\mathrm{op}} \le L_0 + 540 L_0 + 2000 L_0 < 3000 L_0.$$

If $\nabla^2 h \succeq \mu I$, then

$$C\mu + 460 L_0 t \ge \mu,$$

because $\mu(1-C) \le 10\mu t^3 \le 10 L_0 t$. Inside the inner ball the claimed bounds are inherited from h. Outside the outer ball the guard Hessian eigenvalues lie between $1000L_0$ and $2000L_0$. The cutoff and its first two derivatives vanish at the outer join, and the guard and its first derivative vanish at the inner join. The resulting function is $C^{1,1}$. Integrating these almost-everywhere Hessian inequalities along line segments proves the global smoothness and curvature inequalities; boundary-tangent segments follow by continuity.

Since $h^* = h(x^*) \le h(0) = 0$, one has $Ch \ge h^*$: for $h \le 0$, multiplication by $C \in [0,1]$ increases the value, and for $h > 0$ the product is nonnegative. The nonnegative guard vanishes at x^*, so the minimum and initial gap are preserved. For $r < 2R$, choose $y = x$ if $h(x) \le 0$ and $y = 0$ otherwise. Then $h(y) = \min\{h(x), 0\} \le \widetilde{h}(x)$. For $r \ge 2R$, choose $y = 0$. Subtracting the common minimum proves the output assertion. The value and product-gradient formulas establish the query simulation directly. □

C.2 Affine hidden intervals and global PL geometry

Let $h_3(t) = 3t^2 - 2t^3$ on $[0,1]$, and define

$$\Theta(t) = \begin{cases} 0 & t \le 1/2, \\ h_3(2t-1) & 1/2 < t < 1, \\ 1 & t \ge 1, \end{cases} \quad B(t) = \begin{cases} 2 + (t+1/2)^2/2 & t \le -1/2, \\ 2 & -1/2 < t \le 1/2, \\ 2(1-\Theta(t)) & 1/2 < t < 1, \\ (t-1)^2/2 & t \ge 1. \end{cases}$$

Thus $\|\Theta'\|_\infty \le 3$, $\mathrm{Lip}(\Theta') \le 24$, and $\mathrm{Lip}(B') \le 48$. The gates are $C^{1,1}$; twice differentiability at their joins is unnecessary. Starting from $\phi(t) = 1 - e^{-(t-1)^2/2}$, put

$$\xi(t) = \begin{cases} 1 & |t| \le 1/4, \\ 1 - h_3(4|t| - 1) & 1/4 < |t| < 1/2, \\ 0 & |t| \ge 1/2, \end{cases}$$

and define the affine modification by

$$\phi_a'(t) = (1 - \xi(t))\phi'(t) - \xi(t)/2, \qquad \phi_a(t) = \phi(1/2) + \int_{1/2}^{t} \phi_a'(v)\, dv. \tag{104}$$

On the support of ξ, both ϕ' and $-1/2$ are negative with magnitude at most one. Outside that support the derivative is unchanged. Therefore ϕ_a is nonnegative, has its unique zero at one, and its derivative has the sign of $t - 1$. Since the derivative perturbation is supported on an interval of length one and bounded by $1/2$, while its integral vanishes on $[1/2, \infty)$,

$$0 \le \phi_a \le 2, \quad |\phi_a'| \le 1, \quad |\phi_a''| \le |\phi''| + |\xi'|\,|\phi' + 1/2| \le 1 + 6/2 = 4. \tag{105}$$

The matching derivatives of ξ at the joins make ϕ_a a C^2 function. On $[-1/4, 1/4]$, it equals $\phi_a(0) - t/2$. Moreover $\phi_a(0) \le \phi(1/2) + 1/2 < 1$.

The same geometric estimate applies to the stochastic chain and the longer deterministic chain used for the logarithmic lower bound. In this construction, Q is a local weight parameter, distinct from the stationarity parameter defined in Section 2. Choose either $w_i = 1$ and $Q = T$, or

$$w_i = r^{i-1}, \qquad r = 1 - 1/Q, \qquad Q \geq 2.$$

In both cases write $r = w_{i+1}/w_i \leq 1$, and note $\sum_{i=j}^{T} w_i \leq Q w_j$. Set

$$a_1 = 1, \qquad a_i(z) = \Theta(z_{i-1}), \qquad f_w(z) = \sum_{i=1}^{T} w_i \{a_i(z)\phi_a(z_i) + B(z_i)\}.$$

Its normalized partial derivatives are

$$d_i = \frac{\partial_i f_w}{w_i} = a_i \phi_a'(z_i) + \mathbf{1}_{i<T} r \Theta'(z_i)\phi_a(z_{i+1}) + B'(z_i). \tag{106}$$

Lemma 24 (Weighted scalar and radial PL estimates)**.** *On* $[-4,4]^T$, $|d_i| \leq 13$ *and*

$$f_w(z) \leq 50000 Q \sum_i w_i d_i^2. \tag{107}$$

For all z, $d_i \leq 0$ *if* $z_i \leq 1$ *and* $d_i \geq 0$ *if* $z_i \geq 1$. *For arbitrary unit vectors* $u_i \in \mathbb{R}^D$, *define*

$$\rho(y) = \frac{y}{\sqrt{1 + \|y\|^2/16}}, \quad P(v) = 8(v-2)_+^2, \quad y_i = \frac{x_i}{\delta\sqrt{w_i}}, \quad z_i = \langle u_i, \rho(y_i)\rangle,$$

$$F_u(x) = A\left[f_w(z(x)) + \sum_i w_i P(\|y_i\|)\right]. \tag{108}$$

Then $F_u^* = 0$, $F_u(0) \leq 3AQ$, *and*

$$\mathrm{Lip}(\nabla F_u) \leq 256A/\delta^2, \qquad \|\nabla F_u(x)\|^2 \geq \frac{2A}{10^6 Q\delta^2} F_u(x). \tag{109}$$

Proof. For $z_i \in (1/2, 1)$ the outgoing term and barrier derivative combine as

$$\Theta'(z_i)\{\mathbf{1}_{i<T} r\phi_a(z_{i+1}) - 2\} \leq 0.$$

The incoming derivative is also nonpositive. On the other regions the sign assertion follows immediately from (106). In particular $d_i z_i < 0$ implies $0 < z_i < 1$. On the cube, $B \leq 9$, $|B'| \leq 6$, and (105) gives $|d_i| \leq 1 + 3 \cdot 2 + 6 = 13$, with each scalar loss at most eleven.

Let j be the first index outside $[7/8, 9/8]$, taking $j = T + 1$ if there is none. For $i < j$ the incoming gate is at least $\Theta(7/8) = 27/32$. If $z_i < 1$, the derivative sign calculation gives

$$|d_i| \geq \frac{27}{32} e^{-1/128} |z_i - 1| \geq \frac{1}{2}|z_i - 1|.$$

If $z_i \geq 1$, $B'(z_i) = z_i - 1$ gives the stronger bound. In that interval $\phi_a(z_i) = \phi(z_i) \leq (z_i - 1)^2/2$. For $h = 1 - z_i \in [0, 1/8]$, direct substitution into the cubic gives $B(z_i) = 24h^2 - 32h^3 \leq 24h^2$; on the other side $B(z_i) = (z_i - 1)^2/2$. Hence

$$a_i\phi_a(z_i) + B(z_i) \leq 25(z_i - 1)^2 \leq 100 d_i^2 \qquad (i < j).$$

If $j \leq T$, then $z_j > 9/8$ gives $d_j \geq 1/8$, and $z_j < -3/4$ gives $d_j \leq -1/4$. In the remaining interval $[-3/4, 7/8)$, $\phi'(z_j) = (z_j - 1)e^{-(z_j-1)^2/2} \leq -1/16$; its convex mixture with $-1/2$ has the same upper bound. Since the incoming gate is at least $27/32$ and all other terms are nonpositive, $d_j \leq -1/64$. Therefore

$$\sum_{i=j}^{T} w_i\{a_i\phi_a(z_i) + B(z_i)\} \leq 11 Q w_j \leq 45056 Q w_j d_j^2.$$

Combining prefix and tail proves (107). Nonnegativity is termwise, and all scalar terms vanish at $z_i = 1$.

To transfer the inequality to the whole space, we use the following radial estimates. Put $v = \|y\|$, $\alpha = (1 + v^2/16)^{-1/2}$, and, for $v > 0$, $e = y/v$. Then

$$J\rho(y) = \alpha(I - ee^\top) + \alpha^3 ee^\top, \quad \|\rho(y)\| < 4, \quad \|J\rho(y)\|_{\mathrm{op}} \le 1.$$

For a unit vector u, differentiating $q(y) = \alpha \langle u, y\rangle$ gives

$$\nabla^2 q = -\frac{\alpha^3}{16}(uy^\top + yu^\top + \langle u, y\rangle I) + \frac{3\alpha^5}{256}\langle u, y\rangle\ yy^\top.$$

Writing $t = v/4$, its norm is at most

$$\frac{3}{4}\frac{t}{(1+t^2)^{3/2}} + \frac{3}{4}\frac{t^3}{(1+t^2)^{5/2}} \le \frac{3}{2} < 2.$$

With $c = \langle u_i, e\rangle$, the normalized radial gradient block is

$$v_i = d_i\{\alpha(u_i - ce) + \alpha^3 ce\} + P'(v)e.$$

If $v \le 2$, the smallest Jacobian eigenvalue is at least $1/3$, and $P' = 0$, so $\|v_i\| \ge |d_i|/3$. If $2 < v \le 4$ and $d_i c \ge 0$, radial components align and the same estimate holds. If $2 < v \le 4$ and $d_i c < 0$, the sign assertion above implies $0 < z_i < 1$. Since $\alpha v \ge 4/\sqrt{5}$, one has $0 < c < \sqrt{5}/4$. The tangential component therefore has norm

$$|d_i|\alpha\sqrt{1 - c^2} \ge |d_i|/3.$$

For $v > 4$, $|d_i| \le 13$ because $z_i \in (-4, 4)$, and

$$\|v_i\| \ge 16(v - 2) - 13 \ge 8(v - 2) > 16 \ge |d_i|/3.$$

Thus $|d_i| \le 3\|v_i\|$ in every case. If the chain and guard radial components align, $P' \le \|v_i\|$. In the opposite-sign case with $2 < v \le 4$,

$$\frac{|d_i|\alpha^3 c}{|d_i|\alpha\sqrt{1 - c^2}} = \frac{\alpha^2 c}{\sqrt{1 - c^2}} \le 1,$$

so $P' \le 2\|v_i\|$ and $P = (P')^2/32 \le \|v_i\|^2/8$. The estimate for $v > 4$ gives the same conclusion, and $P = 0$ for $v \le 2$. In particular $P \le \|v_i\|^2/2$ globally. Consequently

$$F_u/A \le (450000Q + 1/2)\sum_i w_i \|v_i\|^2 \le 500000Q \sum_i w_i \|v_i\|^2.$$

The physical gradient blocks equal $(A/\delta)\sqrt{w_i}v_i$, which proves the PL inequality.

For smoothness, let $W = \mathrm{diag}(w_i)$. The diagonal entries of $W^{-1/2}\nabla^2 f_w W^{-1/2}$ have magnitude at most $4 + 24 \cdot 2 + 48 = 100$, and its adjacent off-diagonals equal $\sqrt{r}\Theta'(z_i)\phi_a'(z_{i+1})$, of magnitude at most three. The row-sum bound therefore gives operator norm at most $106 < 128$. In the Hessian of (108), this term contributes at most $128A/\delta^2$, the scalar-composition term at most $26A/\delta^2$, and the radial guard at most $16A/\delta^2$. Their sum is below $256A/\delta^2$. The weights cancel because the block scale is $\delta\sqrt{w_i}$. The piecewise Hessian bounds integrate globally as in Lemma 23. Finally $y_i = 4u_i/\sqrt{15}$ gives $z_i = 1$ and zero guard, while $F_u(0) = A(2\sum_i w_i + \phi_a(0)) \le 3AQ$. □

C.3 Conservative PL noise and its exact prefix

For the stochastic construction use $w_i = 1$, $Q = T$, and write $q_i = \langle u_i, \rho(x_i/\delta)\rangle$. Keep the quintic s of Lemma 23, and set

$$\gamma(t) = s(4|t| - 1), \qquad \chi(v) = 1 - s((4v - 1)/3), \qquad \tau_j(q) = \chi\left(\sum_{k=j}^{T}\gamma(q_k)^2\right).$$

Thus γ is zero on $[-1/4, 1/4]$ and one outside $(-1/2, 1/2)$; χ is one below $1/4$ and zero above one. For each sample choose codes S_j with $\mathbb{E}S_j = \beta u_j$ and $\|S_j/\beta - u_j\| \le 2/\beta$, and put

$$\begin{aligned} e_j &= \langle S_j/\beta - u_j, \rho(x_j/\delta)\rangle, \\ N(x, S) &= -\frac{A}{2}\sum_{j=1}^{T} a_j(q)\tau_j(q)e_j, \qquad f(x, S) = F_u(x) + N(x, S). \end{aligned} \tag{110}$$

All evaluations of a retained handle use this same fixed code array.

Lemma 25 (Uniform conservative noise bounds)**.** *The sample functions in* (110) *are bounded below and $C^{1,1}$, have mean F_u, and satisfy*

$$|N| \le \frac{4A}{\beta}, \qquad \|\nabla N\| \le \frac{163A}{\beta\delta}, \qquad \mathrm{Lip}(\nabla N) \le \frac{15000A}{\beta\delta^2}. \tag{111}$$

For every prefix length $p < T$ there is a globally defined sample potential depending only on $u_{\le p}$ and $S_{\le p+1}$ that agrees with the sample value and gradient whenever $|q_k| \le 1/4$ for every $k > p$.

Proof. At a point let $p = \max(\{k : |q_k| > 1/2\} \cup \{0\})$. For $j \le p$ the tail sum is at least one, so τ_j and its derivatives vanish. For $j \ge p+2$, a_j and its derivatives vanish. Thus only $j = p+1$ can contribute to the value, gradient, or almost-everywhere Hessian. There is no sum over T in the estimates below.

The radial derivative bounds in the preceding proof give

$$|e_j| \le 8/\beta, \quad \|\nabla e_j\| \le 2/(\beta\delta), \quad \left\|\nabla^2 e_j\right\|_{\mathrm{op}} \le 4/(\beta\delta^2),$$

and

$$|a_j| \le 1, \quad \|\nabla a_j\| \le 3/\delta, \quad \left\|\nabla^2 a_j\right\|_{\mathrm{op}} \le (24 + 3\cdot 2)/\delta^2 = 30/\delta^2.$$

Direct differentiation of s gives $|s'| \le 15/8$ and $|s''| \le 6$, hence $|\gamma'| \le 15/2$, $|\gamma''| \le 96$, $|\chi'| \le 5/2$, and $|\chi''| \le 11$. For $v(q) = \sum_{k\ge j} \gamma(q_k)^2$, on the derivative support of χ one has $v \le 1$ and therefore

$$\|\nabla v\| \le 15\sqrt{v} \le 15, \quad \left\|\nabla^2 v\right\|_{\mathrm{op}} \le 2((15/2)^2 + 96) = 609/2.$$

It follows that

$$\|\nabla_q \tau_j\| \le 75/2, \qquad \left\|\nabla_q^2 \tau_j\right\|_{\mathrm{op}} \le 11\cdot 225 + (5/2)(609/2) < 3237.$$

After composing with the block radial maps,

$$\|\nabla_x \tau_j\| \le 75/(2\delta), \qquad \left\|\nabla_x^2 \tau_j\right\|_{\mathrm{op}} \le 3400/\delta^2.$$

For the last inequality, the extra diagonal composition term is bounded by $2\left\|\nabla_q \tau_j\right\|_\infty/\delta^2 \le 75/\delta^2$. Applying the product rule to the sole active term gives the gradient coefficient

$$\tfrac{1}{2}(3\cdot 8 + (75/2)\cdot 8 + 2) = 163,$$

and the Hessian coefficient

$$\tfrac{1}{2}(4 + 3400\cdot 8 + 30\cdot 8 + 2(75/2)\cdot 2 + 2\cdot 3\cdot 2 + 2\cdot 3(75/2)\cdot 8) = 14703 < 15000.$$

These prove (111). Products and compositions are locally $C^{1,1}$, and integrating the uniform almost-everywhere Hessian bound proves that the gradient is globally Lipschitz. The potential perturbation is bounded, while the radial guard is coercive, so every sample is bounded below. Since $\mathbb{E}e_j = 0$, the sample values are unbiased; the uniform gradient bound permits differentiation under expectation and gives $\mathbb{E}\nabla N = 0$.

For the exact prefix write $\bar{q} = (q_1, \dots, q_p, 0, \dots, 0)$, $c = -1/2$, and let $f_T = f_w$ with unit weights. Define, for every x,

$$\begin{aligned} V_p(x) = A\sum_j P(\|x_j/\delta\|) + A f_T(\bar{q}) + Ac\sum_{j=1}^{p} a_j(\bar{q})\tau_j(\bar{q})e_j(x) \\ + \frac{Ac}{\beta} a_{p+1}(\bar{q})\left\langle S_{p+1}, \rho(x_{p+1}/\delta)\right\rangle. \end{aligned} \tag{112}$$

For $p = T$ omit the last line and use the full first line. On the stated hidden region, every hidden barrier equals two, every hidden outgoing gate vanishes, and the only additional population loss beyond $f_T(\bar{q})$ is $ca_{p+1}q_{p+1}$. The earlier tail cutoffs equal their truncated versions, with zero hidden derivatives; $\tau_{p+1} = 1$ and later incoming gates vanish. The identity

$$ca_{p+1}q_{p+1} + ca_{p+1}e_{p+1} = \frac{c}{\beta} a_{p+1}\left\langle S_{p+1}, \rho(x_{p+1}/\delta)\right\rangle$$

proves $f = V_p$ and $\nabla f = \nabla V_p$, including boundaries by continuity. The comparison formula is defined globally, without conditioning its definition on the hidden-region event. □

C.4 Deterministic PL depth at every condition number

We first obtain logarithmic depth for large condition numbers from the weighted geometry. Put $C_{\mathrm{PL}} = 256 \cdot 10^6$. When $\kappa \geq 2C_{\mathrm{PL}}$ choose

$$Q = \kappa/C_{\mathrm{PL}}, \quad r = 1 - 1/Q, \quad A = \Delta/(3Q), \quad \delta^2 = 256A/L,$$

and use (108) with

$$M = \left\lfloor \frac{Q}{2} \log \frac{\Delta}{24\eta} \right\rfloor, \qquad T = M + \lceil Q \rceil.$$

Lemma 24 makes this objective L-smooth and μ-PL with initial gap at most Δ. If its last crossed coordinate

$$p(x) = \max(\{i : |z_i(x)| > 1/4\} \cup \{0\})$$

is at most M, every later barrier equals two. There are at least $\lceil Q \rceil$ later coordinates, and

$$\sum_{i=p+1}^{p+\lceil Q \rceil} w_i \geq Q r^{p+Q} \geq \frac{Q}{4} r^p.$$

The largest exponent is $p + \lceil Q \rceil - 1 \leq p + Q$, so each of the $\lceil Q \rceil \geq Q$ summands is at least r^{p+Q}; the second inequality follows from $r^Q = (1 - 1/Q)^Q \geq 1/4$ for $Q \geq 2$. Also $-\log r \leq 2/Q$, so

$$F_u(x) \geq 2A \sum_{i>p} w_i \geq \frac{AQ}{2} r^p \geq \frac{\Delta}{6} \exp(-2M/Q) \geq 4\eta. \tag{113}$$

The unfloored definition of M is at least $\log(128/24) > 1$, and $\log(\Delta/(24\eta))/\log(\Delta/\eta)$ is bounded below by a positive numerical constant. Thus $M = \Omega(\kappa \log(\Delta/\eta))$.

The exact gradient on the region $|z_i| \leq 1/4$ for $i > p$ depends only on $u_{\leq p+1}$. A global prefix potential is obtained by keeping the first $p + 1$ coordinates in the scalar potential and setting the later coordinates to zero; the radial guard remains unchanged. The stronger exact-value channel consists of all $z_i(x)$, which are four times projections along

$$v_i(x) = \tfrac{1}{4}\rho(x_i/(\delta\sqrt{w_i})), \qquad \|v_i(x)\| \leq 1.$$

A crossing implies $|\langle u_i, v_i(x)\rangle| > 1/16$, independent of the weights. Lemma 14, with $D \geq 4096(r_0 K + 1)$ and $D \geq 8192 \log(256T)$, implies that after r_0 nonadaptive rounds the output frontier is at most $r_0 + 1$, except with probability at most $1/128$. Consequently $r_0 \leq M - 1$ and (113) imply expected gap at least $4\eta(127/128) > \eta$. Every sample may equal F_u, so this is also a lower bound with arbitrary retained handles.

We next prove the depth bound for bounded condition numbers and for condition numbers near one; a one-query bound alone does not cover these regimes. Define the flat convex edge

$$\varphi_c(v) = \begin{cases} 0 & |v| \leq c, \\ 2(|v| - c)^2 & c < |v| \leq 2c, \\ v^2 - 2c^2 & |v| > 2c. \end{cases} \tag{114}$$

Its derivative is continuous, nondecreasing, and four-Lipschitz, and direct evaluation on each of the three regions gives

$$0 \leq v^2 - \varphi_c(v) \leq 2c^2. \tag{115}$$

For example in the middle region the difference equals $-v^2 + 4c|v| - 2c^2$, increasing from c^2 to $2c^2$.

Lemma 26 (A flat-link depth family)**.** *Let $m > 0$, $0 < d \leq m$, and $B_0 = 65536$. Put*

$$a = d/B_0, \quad b = \sqrt{2m\Delta}, \quad R = B_0 b/d, \quad t = \frac{a}{m + 2a}.$$

There are $C^{1,1}$ objectives with curvature in $[m - d/16, m + d/16]$, initial gap at most Δ, and deterministic nonadaptive finite-width round complexity

$$\Omega\left(1 + \frac{\log(\Delta/\eta)}{\log(1/t)}\right).$$

The statement permits exact values, unrestricted query norms, and arbitrary outputs.

Proof. Let $q_i = \langle u_i, x_i \rangle$ in separate blocks, $q_{T+1} = 0$, and define

$$H_u(x) = -bq_1 + \frac{a}{2}\sum_{i=1}^{T} \varphi_c(q_i - q_{i+1}),$$
$$C(x) = 1 - s((\|x\| - R)/R), \qquad F_u(x) = \frac{m}{2}\|x\|^2 + C(x)H_u(x).$$

The block-difference map has norm at most two. Hence $\mathrm{Lip}(\nabla H_u) \le 8a$ and $\|\nabla H_u(0)\| = b$. On $\|x\| \le 2R$,

$$\|\nabla H_u(x)\| \le b + 16aR, \qquad |H_u(x)| \le 2bR + 16aR^2.$$

The loose cutoff estimates $\|\nabla C\| \le 30/R$ and $\left\|\nabla^2 C\right\|_{\mathrm{op}} \le 90/R^2$ give, on its annulus,

$$\begin{aligned}\left\|\nabla^2(CH_u)\right\|_{\mathrm{op}} &\le 8a + 60(b/R + 16a) + 90(2b/R + 16a)\\ &= 2408a + 240b/R = 2648d/B_0 < d/16.\end{aligned}$$

Inside the inner ball the perturbation is convex with Hessian norm at most $8a < d/16$; outside the outer ball it vanishes. The joins have matching first derivatives. Integration proves the announced global curvature bounds.

The uncut objective $m\|x\|^2/2 + H_u(x)$ is m-strongly convex. Its minimizer x^c satisfies $\|x^c\| \le b/m < R$ by strong monotonicity applied to its gradient at zero. Thus x^c is stationary for F_u and is its unique minimizer. The nonnegative edge terms give $F_u(0) - F_u(x^c) \le b^2/(2m) = \Delta$. Orthogonal block components only add quadratic energy, so write its minimizer as $x_i^c = q_i^c u_i$.

Compare its scalar objective with

$$J(q) = \frac{m}{2}\|q\|^2 - bq_1 + \frac{a}{2}\sum_{i=1}^{T}(q_i - q_{i+1})^2.$$

If A_{path} is the path adjacency matrix, the Hessian is $(m+2a)I - a(A_{\mathrm{path}} + e_1 e_1^\top)$, including $T = 1$. The matrix $t(A_{\mathrm{path}} + e_1 e_1^\top)$ is nonnegative with row sums at most $2t < 1$, so

$$q^* = \frac{b}{m+2a}\sum_{k=0}^{\infty} t^k (A_{\mathrm{path}} + e_1 e_1^\top)^k e_1.$$

The length $j-1$ path from one to j contributes one to the corresponding matrix entry, whence $q_j^* \ge bt^{j-1}/(m+2a)$. By (115), the flat scalar objective lies between $J - aTc^2$ and J. Its optimality therefore gives

$$J(q^c) \le h(q^c) + aTc^2 \le h(q^*) + aTc^2 \le J(q^*) + aTc^2,$$

and strong convexity gives $\|q^c - q^*\| \le c\sqrt{2aT/m}$.

Choose

$$T = 1 + \left\lfloor \frac{\log(\Delta/(64\eta))}{2\log(1/t)} \right\rfloor, \quad z_* = \frac{b}{m+2a}t^{T-1}, \quad c = \frac{z_*}{100(1+\sqrt{2aT/m})}.$$

Then $q_T^c \ge 99z_*/100$ and, since $(1 + 2/B_0)^2 < 2$,

$$z_*^2 \ge \frac{b^2}{(m+2a)^2}\frac{64\eta}{\Delta} \ge \frac{64\eta}{m}.$$

Every point with $\|x\| \le R$ and $|q_T(x)| \le c/4$ is at distance at least $z_*/2$ from x^c. Using the weaker curvature $m/2$ gives gap at least $mz_*^2/16 \ge 4\eta$. Every point outside radius R has gap at least

$$\frac{m}{4}(R - b/m)^2 = \frac{\Delta}{2}(B_0 m/d - 1)^2 > 4\eta.$$

For the information argument use the global prefix perturbation

$$H_p(x) = -bq_1 + \frac{a}{2}\sum_{i=1}^{p}\varphi_c(q_i - q_{i+1}), \qquad \Gamma_p(x) = mx + C\nabla H_p + H_p\nabla C.$$

It depends only on $u_{\le p+1}$, and equals the true gradient when all later $|q_i| \le c/4$, since every omitted edge argument then has magnitude at most $c/2$. Outside radius $2R$, every prefix response equals the known gradient mx. The same identities hold for values. Grant all projections

$$\langle u_i, x_i / \max\{2R, \|x\|\} \rangle ,$$

which determine the value inside the outer ball, while the exterior value is known. A crossing inside the outer ball implies a normalized projection exceeding $\alpha = c/(8R)$. For r_0 rounds choose

$$D \ge 16(r_0 K + 1)/\alpha^2, \qquad D \ge 32 \log(256T)/\alpha^2.$$

Lemma 14 then bounds skipped directions by probability $1/128$. On its complement, an output inside the outer ball has frontier at most $r_0 + 1$. If $r_0 \le T - 2$, either it lies outside radius R or its final coordinate has $|q_T| \le c/4$. The preceding gap certificates prove expected gap greater than η.

When the displayed floor gives a depth of at least two, the resulting $T - 1$ is a constant multiple of $\log(\Delta/\eta)/\log(1/t)$ once that ratio exceeds a fixed numerical constant; the subtracted $\log 64$ is absorbed because $\eta \le \Delta/128$. For bounded ratio, a one-query lower bound suffices. Indeed the two quadratics $m(x \mp \sqrt{2\Delta/m})^2/2$ have average query-free expected gap at least Δ, for any randomized output. They can be embedded into the same block dimension by adding known m-quadratics in the remaining coordinates. This proves the assertion in all regimes. □

Proposition 27 (All-condition deterministic PL lower bound)**.** *Define*

$$D_\kappa = \begin{cases} 1 + \dfrac{\log(\Delta/\eta)}{\log(\kappa/(\kappa-1))} & \kappa > 1, \\ 1 & \kappa = 1. \end{cases}$$

For the class of globally L-smooth, μ-PL objectives satisfying the stated initial-gap bound, the minimax complexity is $\Omega(D_\kappa)$ rounds of nonadaptive exact-oracle queries of any fixed finite width, with constants independent of that width. The same lower bound holds for individual retained-handle evaluations.

Proof. The deterministic depth requires separate constructions in three condition-number regimes. For large κ, the weighted chain gives the logarithmic depth directly. For bounded $\kappa \ge 3$, the flat-depth construction gives the same order with fixed curvature. When $1 < \kappa < 3$, choose the curvature parameters in Lemma 26 so that the chain lies inside $[\mu, L]$; the resulting logarithmic comparison is

$$\kappa - 1 \le \frac{1}{\log(\kappa/(\kappa-1))} \le \kappa,$$

so this is the required order.

For $1 < \kappa < 3$, take $m = (L + \mu)/2$ and $d = (L - \mu)/2$ in Lemma 26; its curvature lies inside $[\mu, L]$. Write $\Lambda = \log(\kappa/(\kappa-1))$ and $u = \kappa/(\kappa-1)$. Then

$$1/t = B_0(2u - 1) + 2, \qquad \Lambda \le \log(1/t) \le \Lambda + \log(2B_0 + 2) \le 32\Lambda,$$

where $\Lambda \ge \log(3/2)$. More explicitly, if $X = \log(\Delta/\eta)/\Lambda$, the unfloored depth in that lemma is at least $X/64 - 1/6$. For $X \ge 256$, $T - 1 \ge X/128$; for $X < 256$, the two-quadratic one-query obstruction is already $\Omega(1 + X)$. At $\kappa = 1$ the same obstruction proves $\Omega(1)$. A strongly convex differentiable function is PL because

$$F(y) \ge F(x) + \langle \nabla F(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \ge F(x) - \frac{\|\nabla F(x)\|^2}{2\mu};$$

taking the infimum over y proves the PL inequality. Thus all the bounded-condition examples belong to the claimed PL class. □

C.5 Statistical cost, interpolation, and finite sums

Lemma 28 (The additive quadratic obstruction)**.** *For $S \ge 256$, a μ-smooth, μ-strongly convex quadratic with bounded additive conservative noise requires $\Omega(S)$ retained-handle evaluations and $\Omega(S)$ fresh shared-sample nonadaptive rounds of any finite width. Sample values and exact population values may be included. For a finite sum with $n \ge S \ge 1024$, the same $\Omega(S)$ bound holds for active component evaluations, with every component μ-smooth and μ-strongly convex.*

Proof. Let u be uniform on the unit sphere, $b = 4\sqrt{\mu\eta}$, $r = b/\mu$, and $\beta = 2b/\sigma = 8/\sqrt{S}$. With a Poisson code V from Lemma 10, define

$$F_u(x) = \frac{\mu}{2}\|x - ru\|^2, \qquad f_u(x, V) = \frac{\mu}{2}\|x\|^2 - \frac{\sigma}{2}\langle V, x\rangle + \frac{b^2}{2\mu}.$$

Its expectation is F_u, its initial gap is $8\eta \le \Delta$, and its centered gradient has norm at most $\sigma/2 + b \le 3\sigma/4$. Its centered field is constant in x. One revealed code determines the sample value and gradient at every point. Thus even a stronger oracle revealing each whole new code has at most N such observations for N evaluations, or r_0 for r_0 shared-sample rounds. Replace these codes by independent uniform sphere points. Conditional on u, the exact value replies are the same deterministic kernels, so for B new codes the joint divergence is at most

$$2DB\beta^2 = 128DB/S.$$

Under the reference law, the other observations are independent of u, and every exact population value gives only the scalar $\langle u, x\rangle$. Include all query vectors and the arbitrary output, normalized to length at most one. With $M = N + 1$ or $r_0K + 1$ and $D \ge 256M$, Lemma 9 bounds the probability of any normalized correlation exceeding $1/4$ by $2e^{-D/256}$. If $F_u(\widehat{x}) \le 2\eta$, then $\|\widehat{x} - ru\| \le r/2$, so

$$\langle u, \widehat{x}\rangle \ge r/2, \qquad \|\widehat{x}\| \le 3r/2,$$

and the output correlation is at least $1/3$. For $B \le S/2^{20}$ and $D \ge 8192$, binary data processing gives

$$\mathbb{P}(F_u(\widehat{x}) \le 2\eta) \le \frac{D/8192 + \log 2}{D/256 - \log 2} < \frac{1}{8}.$$

The expected gap exceeds η. Averaging the direction prior fixes one hard objective for any randomized algorithm.

For the finite sum take the balanced table in Lemma 11, $V_a = \beta u + U_a$, with $\sum_a U_a = 0$ and $\|U_a\| \le 2$, and use

$$f_a(x) = \frac{\mu}{2}\|x\|^2 - \frac{b}{\beta}\langle V_a, x\rangle + \frac{b^2}{2\mu}.$$

The average is exactly F_u, and its centered gradient has norm at most $2b/\beta = \sigma$. An evaluation can reveal the entire row, while a repeat reveals no new row. For $N \le S/2^{24} \le n/8$ and

$$D \ge \max\{2^{20}, 256(N+1), 8\log(4n/\beta)\},$$

the joint table comparison gives divergence at most $DN\beta^2 = 64DN/S \le D/2^{18}$. Under its reference law the revealed rows are independent of u. The same projection and binary-event calculation gives success probability below $1/8$. Exchangeability and lazy assignment in Lemma 11 allow adaptive labels and cached repeats. Averaging the table and direction prior fixes a deterministic finite sum. □

For the PL chain use independent uniform hidden directions and independent Poisson codes of mean βu_j in (110). Their norm is one and their per-code divergence to the uniform sphere is at most $2D\beta^2 \le 5D\beta^2$. The code can be realized on a parameter-independent probability space as in Lemma 10; the base seed is not an additional observation. Grant all $q_j(x)$, a stronger oracle than exact population values. Since $\|\rho\| < 4$, these are normalized projections with crossing threshold $1/16$. By (112), a sample response at an unresolved frontier is determined by the old prefix and a single next-stage code. This verifies all hypotheses of Lemma 13, for values as well as gradients.

For clarity, the numerical substitution into that lemma is

$$D \ge \max\{2^{24}, 4096M, 8192\log(512T)\}, \qquad n_0 = \left\lfloor \frac{1}{2^{24}\beta^2} \right\rfloor, \tag{116}$$

where $M = N + 1$ for retained evaluations and $M = r_0K + 1$ for rounds. Its skipped-stage event J has probability at most $1/256$. If A_i denotes a crossing before n_0 distinct frontier codes have been charged, the capped comparison, run from the initial query, gives

$$\mathbb{P}(A_i \cap J^c) \le \frac{5Dn_0\beta^2 + \log 2}{D/4096 - \log 2} \le 40960n_0\beta^2 + 8192\log 2/D < 1/256. \tag{117}$$

The comparison retains all earlier codes and exact projections, without asserting a uniform conditional direction when a stage is entered. If L_i counts distinct unresolved-frontier codes, its charge budgets are $\sum_i L_i \le N$ for evaluations and $\sum_i L_i \le 2r_0$ for rounds. If $N \le Tn_0/8$, or $r_0 \le Tn_0/16$, crossing half the chain requires at least $T/4$ cheap crossings. Markov's inequality and (117) bound this event outside J by $1/64$. Thus with probability at least $251/256$ the output frontier is below $T/2$. Every coordinate beyond that frontier has barrier two, and hence $F_u(\widehat{x}) > AT$.

Choose an effective curvature $\lambda \le L$, $\lambda \ge 2C_{\mathrm{PL}}\mu$, and set

$$T = \left\lfloor \frac{\lambda}{C_{\mathrm{PL}}\mu} \right\rfloor, \quad A = \frac{4\eta}{T}, \quad \delta^2 = \frac{256A}{\lambda}, \quad \beta = \frac{256A}{\delta\sigma}. \tag{118}$$

The objective is λ-smooth and μ-PL, has gap at most $12\eta \le \Delta$, and its noise radius is at most $163\sigma/256$. Direct substitution gives

$$\beta^2 = \frac{1024\lambda}{T\mu S}, \qquad \frac{1024C_{\mathrm{PL}}}{S} \le \beta^2 \le \frac{2048C_{\mathrm{PL}}}{S}. \tag{119}$$

For $S \ge S_0 = 2^{40}C_{\mathrm{PL}}$, $\beta \le 1/4$, $n_0 \ge 16$, and the preceding information calculation proves

$$\Omega(T/\beta^2) = \Omega((\lambda/\mu)S) \tag{120}$$

evaluations and fresh nonadaptive rounds. Since $AT = 4\eta$, its failure probability implies expected gap greater than η.

Proposition 29 (PL lower-rate assembly)**.** *With D_κ as in Proposition 27, the lower bounds are*

$$\Omega(D_\kappa + \kappa S)$$

for general uniformly bounded noise, and

$$\Omega\left(D_\kappa + S + \min\{\kappa S, (H/\mu)\sqrt{S}\}\right) \tag{121}$$

under any prescribed samplewise or mean-square centered-noise Lipschitz bound $H \ge 0$. With full sample-gradient smoothness L_s instead, put $L_* = \min\{L, L_s\}$ and $\kappa_* = L_*/\mu \ge 1$; the lower bound is

$$\Omega\left(D_{\kappa_*} + S + \min\{\kappa_* S, (L_s/\mu)\sqrt{S}\}\right). \tag{122}$$

These bounds hold for retained-handle evaluations and for fresh nonadaptive rounds of every finite width, including sample values and charged exact population values.

Proof. The deterministic and additive constructions already provide the D_κ and S terms. For general noise, the chain construction gives the interaction term when S and κ are large; the additive or deterministic construction absorbs it in the remaining regimes. For centered regularity, choose $\lambda = \min\{L, H/(64\sqrt{S})\}$. In the regime requiring the chain construction, this choice gives the required samplewise Lipschitz bound and the interaction term $\min\{\kappa S, (H/\mu)\sqrt{S}\}$. If the chain parameters fall below their threshold, the additive or deterministic construction supplies the same order. The cases $H = 0$ and $\sigma = 0$ reduce to those endpoint constructions.

For centered regularity, the preceding choice of λ gives the required samplewise Lipschitz bound when $H, S > 0$ and $S \ge S_0$, and (119) implies $\beta \ge 1/\sqrt{S}$ whenever $\lambda \ge 2C_{\mathrm{PL}}\mu$. In that regime, (111) gives

$$\mathrm{Lip}(\nabla N) \le \frac{15000\lambda}{256\beta} \le \frac{15000}{16384}H < H.$$

The lower bound (120) equals, up to a factor 64, $J = \min\{\kappa S, (H/\mu)\sqrt{S}\}$. If $\lambda < 2C_{\mathrm{PL}}\mu$, then $J \le 64(\lambda/\mu)S < 128C_{\mathrm{PL}}S$, absorbed by the additive example. If $S < S_0$ and $\kappa \ge 3$, $J \le \kappa S_0$ is absorbed by the deterministic example. If $H = 0$ or $\sigma = 0$, the interaction is zero. This covers every regime.

For ordinary sample smoothness use the effective population parameter L_* and centered target $H = L_s/2$. In the only regime needing the chain,

$$\lambda \le L_s/(128\sqrt{S}), \qquad \mathrm{Lip}(\nabla N) \le L_s/2,$$

and $S \geq S_0 > 1$, so each full sample gradient is Lipschitz with constant at most $\lambda + L_s/2 < L_s$. The deterministic auxiliary functions have smoothness at most L_* and the quadratic samples have smoothness μ. These samplewise lower instances also satisfy the corresponding mean-square conditions. On the upper-bound side, Jensen gives population smoothness L_* and

$$\begin{aligned}
&\mathbb{E}\,\|[g(x,Z) - \nabla F(x)] - [g(y,Z) - \nabla F(y)]\|^2 \\
&\quad = \mathbb{E}\,\|g(x,Z) - g(y,Z)\|^2 - \|\nabla F(x) - \nabla F(y)\|^2 \\
&\quad \leq L_s^2\,\|x - y\|^2\,,
\end{aligned}$$

which explains the same effective parameter in the matching algorithms. The maximum of the at most three hard-subclass bounds is at least one third of their sum; no simultaneous direct-product construction is required. □

When $L = L_s$, expression (122) is equivalent to $D_\kappa + S + \kappa\sqrt{S}$. For $S \geq 1$ this is the displayed minimum. For $S < 1$ and $\kappa \geq 3$, both possible extra terms are absorbed by $D_\kappa = \Omega(\kappa)$; for $\kappa < 3$ they are bounded by a constant multiple of $1 + S$. Although the lower round bounds hold even at width one, the matching prescribed-regularity algorithms use paired samples at unbounded condition numbers. The round classification over all condition numbers therefore concerns $K \geq 2$; it does not assert a fresh-only prescribed-smoothness classification at unbounded κ.

Corollary 30 (Active finite-sum PL lower bounds)**.** *If $F = n^{-1}\sum_{a=1}^n f_a$ and $n \geq S$, the individual-evaluation lower bounds in Proposition 29 hold with arbitrary adaptive component labels, cached repeats, component values, and charged exact population values.*

Proof. Replicate each deterministic hard objective across all n components. The finite quadratic in Lemma 28 supplies the additive term. For each chain stage independently use a balanced table $S_a^j = \beta u_j + U_a^j$ from Lemma 11, and use its row in (110) to define the fixed component f_a. The row mean is exactly βu_j, so the component mean is exactly F_u. The analytic estimates need only $\|U_a^j\| \leq 2$, and therefore (111) and the prefix identity hold unchanged.

In the scaling (118), for $S \geq S_0$,

$$n_0 = \left\lfloor \frac{1}{2^{24}\beta^2} \right\rfloor \leq \frac{S}{2^{34} C_{\mathrm{PL}}} \leq n/8.$$

Enlarge (116) by $D \geq 8\log(4n/\beta)$. In the target-stage capped comparison, condition on all other-stage directions and complete tables. Replace the first n_0 distinct target-stage rows by the reference joint rows of Lemma 11. Their joint divergence is at most $Dn_0\beta^2$, smaller than the $5Dn_0\beta^2$ used in (117). Adaptive label choices are handled by exchangeable lazy assignment; repeated labels return the cached row. Each component evaluation reveals at most one previously unseen row at a stage that has not yet been crossed, so $\sum_i L_i \leq N$. The same half-chain counting proves (120).

The remaining parameter cases in Proposition 29 use these components or the replicated deterministic and additive quadratic examples. They cover short chains, bounded condition numbers, small or zero noise, and prescribed sample smoothness. The matching online algorithms can draw independent uniform component indices: conditional on the past, each new index is uniform for the fixed finite sum, even if its label has occurred before. Using the same newly drawn index for each paired difference lets the online fixed budgets apply without change. This transfer concerns individual evaluations and makes no claim for $n < S$. □

C.6 Strong convexity: exact chain energy at every accuracy

An estimate based on the distance to the tail can lose a condition-number factor inside the accuracy logarithm. We avoid this loss by calculating the restricted quadratic minimum exactly.

Lemma 31 (Exact energy of the rooted chain)**.** *Let $\mu, a, b > 0$, and let $\rho \in (0,1)$ satisfy $\mu = a(1-\rho)^2/\rho$. For $t \geq 1$ define*

$$J_t(z) = \frac{\mu}{2}\sum_{i=1}^t z_i^2 - bz_1 + \frac{a}{2}\sum_{i=1}^t (z_i - z_{i+1})^2, \qquad z_{t+1} = 0.$$

Writing

$$\Delta_\infty = \frac{b^2\rho}{2a(1-\rho)}, \quad g_t = \frac{1-\rho^{2t}}{1+\rho^{2t+1}}, \quad e_t = 1-g_t = \frac{(1+\rho)\rho^{2t}}{1+\rho^{2t+1}},$$

one has $\min J_t = -\Delta_\infty g_t$ *and* $\rho^{2t} \le e_t \le 2\rho^{2t}$. *Set* $g_0 = 0$, $e_0 = 1$. *For a length-*T *chain restricted to its first* p *coordinates, its restricted optimality gap is exactly*

$$\min_{z_{p+1:T}=0} J_T(z) - \min J_T = \Delta_\infty(e_p - e_T). \tag{123}$$

Proof. The first Hessian diagonal is $\mu + a$, all others are $\mu + 2a$, and adjacent entries are $-a$. The relation $\mu/a = \rho + \rho^{-1} - 2$ implies that ρ^i and ρ^{-i} solve the interior recurrence. The proposed solution is

$$z_i^{(t)} = \frac{b\rho}{a(1-\rho)} \frac{\rho^{i-1} - \rho^{2t-i+1}}{1+\rho^{2t+1}}. \tag{124}$$

It vanishes at $i = t+1$. To verify its first equation, put $K = b\rho/[a(1-\rho)(1+\rho^{2t+1})]$. Since $(\mu+a)/a = \rho^{-1} + \rho - 1$,

$$\begin{aligned}\frac{(\mu+a)z_1^{(t)} - az_2^{(t)}}{aK} &= (\rho^{-1}+\rho-1)(1-\rho^{2t}) - (\rho - \rho^{2t-1}) \\ &= (1-\rho)(\rho^{-1}+\rho^{2t}) = \frac{1-\rho}{\rho}(1+\rho^{2t+1}).\end{aligned}$$

Thus the first equation equals b. For $t = 1$, the same computation uses the zero boundary $z_2^{(1)} = 0$ and remains valid. Positive definiteness makes (124) the unique minimizer. At a quadratic minimizer $Hz = be_1$, the value is $-bz_1/2$, giving the formula for g_t. Because $1 \le 1+\rho^{2t+1} \le 1+\rho \le 2$, the bounds for e_t follow. Setting the last $T-p$ coordinates to zero leaves exactly J_p, proving (123). □

Let $q_i = \langle u_i, x_i\rangle$ in separate blocks, and use the flat edge (114) to form

$$h_u(x) = \frac{\mu}{2}\|x\|^2 - bq_1 + \frac{a}{2}\sum_{i=1}^{T}\varphi_c(q_i - q_{i+1}), \qquad q_{T+1} = 0. \tag{125}$$

It is globally μ-strongly convex and $(\mu+8a)$-smooth. Its minimizer lies in the block-direction span and has norm at most b/μ, since its gradient at zero has norm b. Let J_T^u denote the same function with quadratic edges. Equation (115) gives

$$J_T^u - aTc^2 \le h_u \le J_T^u. \tag{126}$$

For $1 \le p < T$, embed the minimizer $z^{(p)}$ of J_p into the full block space with its last coordinates zero. Its only nonzero full gradient block is $-az_p^{(p)}u_{p+1}$, and $\|z^{(p)}\| \le b/\mu$. The supporting-hyperplane inequality for the convex quadratic gives

$$J_T^u(x) \ge J_T^u(z^{(p)}) - az_p^{(p)}q_{p+1}(x).$$

Therefore, whenever $|q_{p+1}(x)| \le c/4$, subtracting the full minimum and using (123) and (126) yields the global certificate

$$h_u(x) - \min h_u \ge \Delta_\infty(e_p - e_T) - \frac{abc}{4\mu} - aTc^2. \tag{127}$$

Indeed $h_u(x) \ge J_T^u(x) - aTc^2$ and $\min h_u \le \min J_T^u$. Orthogonal block components add nonnegative quadratic energy, so this argument places no restriction on query norms or spans.

Suppose now $\kappa \ge 6000$, and choose

$$L_0 = L/3000, \quad a = (L_0-\mu)/8, \quad \Delta_\infty = \Delta/2, \quad b^2 = a(1-\rho)\Delta/\rho, \quad R = b/\mu. \tag{128}$$

Put $\theta = \log(1/\rho)$ and

$$M = \left\lfloor \frac{\log(\Delta/(32\eta))}{2\theta} \right\rfloor, \qquad T = M + \left\lceil \frac{\log 4}{2\theta} \right\rceil, \qquad c = \min\left\{ \frac{\mu\eta}{ab}, \sqrt{\frac{\eta}{aT}}, \frac{R}{100} \right\}.$$

The nonconstant-depth argument uses $M \geq 4$. By construction,

$$\rho^{2M} \geq 32\eta/\Delta, \qquad e_T \leq 2\rho^{2T} \leq \tfrac{1}{2}\rho^{2M}, \qquad g_T \geq 1/2.$$

Since $h_u(0) = 0$, its gap is at least $\Delta_\infty g_T \geq \Delta/4$ and at most $\Delta_\infty g_T + aTc^2 \leq \Delta/2 + \eta \leq \Delta$. Moreover $aTc^2 \leq \eta$ and $abc/(4\mu) \leq \eta/4$, so (127) at $p = M$ gives

$$|q_{M+1}(x)| \leq c/4 \quad \Longrightarrow \quad h_u(x) - \min h_u \geq \frac{\Delta}{4}\rho^{2M} - \frac{5}{4}\eta \geq \frac{27}{4}\eta > 4\eta. \tag{129}$$

The chain in (125) is globally μ-strongly convex and has a stationary minimizer in $B(0, R)$. Apply Lemma 23 with this radius and smoothness $L_0 = \mu + 8a$. It gives

$$F_u(x) = C(x)h_u(x) + 1000L_0(\|x\| - R)_+^2 \quad (\|x\| < 2R),$$

with the guard alone outside. The minimum and initial gap are preserved, and F_u is globally μ-strongly convex and L-smooth. If $\|x\| \geq 2R$, its value is nonnegative and its gap is at least $\Delta/4 \geq 32\eta$. The same holds if $\|x\| < 2R$ and $h_u(x) > 0$. Otherwise $C(x)h_u(x) \geq h_u(x)$, so (129) transfers directly to F_u. Consequently an output with gap below 4η must lie inside the outer ball and have $|q_{M+1}| > c/4$.

The required prefix potential is explicit:

$$h_p(x) = \frac{\mu}{2}\|x\|^2 - bq_1 + \frac{a}{2}\sum_{i=1}^{p}\varphi_c(q_i - q_{i+1}).$$

Extend it with the same cutoff and guard to define a global prefix value and gradient. If all $|q_i| \leq c/4$ for $i > p$, the omitted flat edges and their derivatives vanish. Hence both responses agree with the true ones, including the term $h_p\nabla C$. The prefix depends only on $u_{\leq p+1}$, while every exterior response is independent of the directions. Granting

$$\langle u_i, x_i/\max\{2R, \|x\|\}\rangle$$

at each charged point determines all needed exact values. A crossing inside the ball has normalized magnitude greater than $\alpha = c/(8R)$. For a proposed r_0-round budget, take

$$D \geq 16(r_0K + 1)/\alpha^2, \qquad D \geq 32\log(256T)/\alpha^2.$$

Lemma 14 makes the true and capped interactions agree except with probability $1/128$, and gives output frontier at most $r_0 + 1$. For $r_0 \leq M - 2$ it cannot cross coordinate $M + 1$ on that event, so the expected gap exceeds η. The argument includes exact sample values because every sample can be set equal to F_u.

It remains to express this lower bound uniformly in the declared condition number and accuracy. The relation between μ/a and ρ is

$$\mu/a = 2(\cosh\theta - 1) \geq \theta^2, \qquad a/\mu = (\kappa/3000 - 1)/8 \geq \kappa/48000.$$

Writing $y = \log(\Delta/\eta) \geq \log 128$, one has $y - \log 32 \geq 2y/7$. Therefore

$$M \geq \frac{y}{7\theta} - 1 \geq \frac{\sqrt{\kappa}\,y}{7\sqrt{48000}} - 1. \tag{130}$$

If $M \geq 4$, the round bound $\Omega(M)$ is $\Omega(\sqrt{\kappa}\log(\Delta/\eta))$ by (130). If $M < 4$, that target expression is bounded by a universal constant, so the two-quadratic one-query obstruction suffices.

Proposition 32 (All-condition strongly convex lower bounds)**.** *Define*

$$E_\kappa = \begin{cases} 1 + \dfrac{\log(\Delta/\eta)}{\log((\sqrt{\kappa}+1)/(\sqrt{\kappa}-1))} & \kappa > 1, \\ 1 & \kappa = 1. \end{cases}$$

The lower bound in the strongly convex class is $\Omega(E_\kappa + S)$ for retained-handle evaluations and fresh nonadaptive rounds of every finite width. It holds with individually μ-strongly convex, L-smooth, bounded-below sample losses, and with centered noise constant in x in the statistical subclass. Exact sample and charged population values may be supplied.

Proof. For $\kappa \geq 3$, put $v = 1/\sqrt{\kappa} \leq 1/\sqrt{3}$. Using the identity

$$\log \frac{1+v}{1-v} = 2 \int_0^v \frac{dt}{1-t^2}$$

we obtain $2v \leq \log((1+v)/(1-v)) \leq 2v/(1-v^2) \leq 3v$. Thus E_κ is comparable to $1 + \sqrt{\kappa}\log(\Delta/\eta)$. For $\kappa \geq 6000$ the exact-energy construction proves this lower bound. For $3 \leq \kappa < 6000$ apply Lemma 26 with $m = d = 2\mu$; its $\Omega(\log(\Delta/\eta))$ depth is $\Omega(E_\kappa)$ on this bounded interval. For $1 < \kappa < 3$ the identity

$$\log \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} = \log \frac{\kappa}{\kappa-1} + 2\log(1 + 1/\sqrt{\kappa})$$

places this denominator between $\Lambda = \log(\kappa/(\kappa-1))$ and 5Λ, because $\Lambda \geq \log(3/2)$. Hence E_κ is comparable to D_κ, and the strongly convex near-one construction in Proposition 27 applies. The two-quadratic obstruction handles $\kappa = 1$.

All deterministic samples can equal the hard population, so they meet the individual curvature conditions and every centered-noise regularity cap. Lemma 28 supplies $\Omega(S)$ with sample Hessian μI and additive centered noise. For $S < 256$, $E_\kappa \geq 1$ absorbs S. Taking the maximum of the deterministic and statistical lower bounds proves $\Omega(E_\kappa + S)$. □

Since the two lower subclasses have centered-noise Lipschitz constant zero, imposing any prescribed samplewise or mean-square cap $H \geq 0$ leaves this order unchanged. With a full sample-gradient cap $L_s \geq \mu$, use $L_* = \min\{L, L_s\}$ and replace E_κ by $E_{L_*/\mu}$. The same conclusion holds with individual sample strong convexity. The matching algorithms need only fresh gradients, with no centered-regularity assumption; the upper-bound appendix gives their proofs and deterministic budgets.

Corollary 33 (Active finite-sum strongly convex bounds)**.** *If $n \geq S$, the dimension-free minimax number of individual active component evaluations in the strongly convex class is $\Theta(E_\kappa + S)$. The rate remains valid with adaptive component reuse and exact component and population values at counted query points. It also holds under any centered-noise regularity bound $H \geq 0$ and when every component is strongly convex. A full component-gradient cap $L_s \geq \mu$ replaces E_κ by E_{κ_*}, where $\kappa_* := \min\{L, L_s\}/\mu \geq 1$.*

Proof. Replicate the deterministic objectives of Proposition 32 across all indices. For $S \geq 1024$, use the balanced-table quadratic in Lemma 28, whose exact component mean is the desired population and whose centered gradients are constant with norm at most σ. Its information argument counts distinct adaptively chosen labels and includes exact population values. For $S < 1024$, the deterministic lower bound absorbs S. Conversely, independently drawn uniform component indices give unbiased gradients whose conditional centered second moment is at most σ^2 for the fixed finite sum, even when an index has previously appeared. The strongly convex upper algorithms therefore apply with the same deterministic budget. No round classification or conclusion for $n < S$ is inferred from this finite-sum corollary. □

D EVERY FIXED ORDER OF POPULATION SMOOTHNESS

Throughout this appendix, $p \geq 2$ is a fixed integer. The population is L-smooth, $F(0) - \inf F \leq \Delta$, and $D^p F$ is L_p-Lipschitz in the full multilinear operator norm. These smoothness assumptions concern the population objective: the oracle still returns only first derivatives, and no prescribed higher derivative bound is required of its sample functions. Write

$$Q = \frac{\Delta L}{\varepsilon^2}, \qquad m = \frac{\sigma^2}{\varepsilon^2}, \qquad A_p = \frac{\Delta L_p^{1/p}}{\varepsilon^{1+1/p}}, \qquad C = \min\{Q, A_p\}. \tag{131}$$

The scaling A used below is distinct from A_p; Proposition 34 proves the higher-order result.

Proposition 34. *For $0 < \varepsilon \leq 2^{-20}\sqrt{\Delta L}$, the minimax complexity is at least $c_p(1+m)(1+C)$ and at most $C'_p(Q + m + mC)$. Both statements hold for total evaluations with arbitrary retained*

opaque samples and for fresh, simultaneous common-sample rounds of any finite width $K \geq 1$. The constants depend only on p. Consequently, if

$$Q \leq (1+m)(1+C), \tag{132}$$

the exact order is $\Theta_p((1+m)(1+C))$. The lower bounds allow exact sample and population values at every charged point. Their population functions are C^∞, and their conservative sample functions are C^2 and bounded below.

D.1 A SMOOTH AFFINE PLATEAU AND DIMENSION-FREE CHAIN DERIVATIVES

Let $a_0 = 1/8$ and define the normalized smooth density

$$\kappa(t) = Z^{-1} \exp\left(-\frac{1}{1-(t/a_0)^2}\right) \mathbf{1}_{\{|t|<a_0\}}, \qquad Z = \int_{-a_0}^{a_0} \exp\left(-\frac{1}{1-(t/a_0)^2}\right) dt.$$

Set

$$h(t) = \frac{4}{3}\int_{a_0}^{1-a_0} \kappa(t-v)\, dv, \qquad s(t) = \int_{-\infty}^{t} h(v)\, dv, \qquad \Phi(t) = 2 + \int_0^t [1 - s(|u|-1)]\, du. \tag{133}$$

The support of h is $[0,1]$, its integral is $(4/3)(1-2a_0) = 1$, and $h(1-t) = h(t)$. Also $0 \leq h \leq 4/3$. Thus s is a smooth step, equal to zero on $(-\infty, 0]$ and to one on $[1, \infty)$. The expression involving $|u|$ is constant near zero, so it introduces no loss of differentiability. It follows directly that

$$0 \leq \Phi \leq 4, \quad 0 \leq \Phi' \leq 1, \quad |\Phi''| \leq 4/3 < 2, \qquad \Phi(t) = 2 + t \quad (|t| \leq 1). \tag{134}$$

For example, the integral of Φ' over $[0, \infty)$ is at most two, which proves the stated range using $\Phi(0) = 2$ and the evenness of Φ'. All derivatives of positive order are bounded: Φ' is bounded and supported in $[-2, 2]$, and its higher derivatives are smooth and compactly supported. Put $B_r = \sup_t |\Phi^{(r)}(t)|$ for $r \geq 0$.

Use the gate

$$\Psi(t) = \begin{cases} 0, & t \leq 1/2, \\ \exp(1 - (2t-1)^{-2}), & t > 1/2. \end{cases}$$

Every derivative tends to zero at the join: after differentiation it is a polynomial in $(2t-1)^{-1}$ times $\exp(-(2t-1)^{-2})$. At infinity these derivatives are bounded as well. Hence $P_r = \sup_t |\Psi^{(r)}(t)| < \infty$. We retain $B_0 \leq 4$, $B_1 \leq 1$, $B_2 \leq 2$ and $P_0 \leq 3$, $P_1 \leq 5$, $P_2 \leq 100$. Define

$$\bar{f}_T(z) = -\Phi(z_1) + \sum_{j=1}^{T-1} [\Psi(-z_j)\Phi(-z_{j+1}) - \Psi(z_j)\Phi(z_{j+1})], \qquad f_T(q) = \bar{f}_T(2q). \tag{135}$$

The proof of Lemma 15 uses only the displayed low-order bounds, nonnegativity of the gates and their first derivatives, and the exact affine plateau. It therefore gives, for this choice too,

$$f_T(0) - \inf f_T \leq 20T, \qquad \|\nabla f_T\|_\infty \leq G := 64, \qquad \|D^2 f_T\|_{\mathrm{op}} \leq \ell := 2400. \tag{136}$$

In particular, at the first coordinate satisfying $|q_j| \leq 1/2$, the corresponding partial derivative is at most -2.

Lemma 35. *For $r \geq 2$, uniformly in T and q,*

$$\|D^r f_T(q)\|_{\mathrm{op}} \leq M_r, \qquad M_r = 2^r \left[B_r + \sum_{s=0}^{r} \binom{r}{s} P_s B_{r-s} \right]. \tag{137}$$

For $r = 2$ one may instead take $M_2 = 2400$.

Proof. Fix arbitrary unit vectors $v^1, \ldots, v^r \in \mathbb{R}^T$. In differentiating one link, let $I \subseteq [r]$ be the derivatives falling on its left coordinate. The coefficient is at most $P_{|I|}B_{r-|I|}$. The positive and negative gate branches, including all their derivatives, have disjoint supports. Their combined coefficient has the same bound, without a factor of two. Generalized Hölder gives

$$\sum_{j=1}^{T-1} \prod_{k \in I} |v_j^k| \prod_{k \notin I} |v_{j+1}^k| \leq \prod_{k=1}^{r} \|v^k\|_r \leq \prod_{k=1}^{r} \|v^k\|_2 = 1.$$

The first inequality applies Hölder with all r exponents equal to r; restricting or shifting a coordinate range can only reduce its norm. The second uses $r \geq 2$. Sum over I, add the root contribution B_r, and multiply by 2^r for the argument scaling. This bounds arbitrary multilinear directions, as required by the operator norm. □

D.2 All derivatives of the radial blocks and their composition

For a unit vector θ in one block, let

$$q_\theta(x) = \frac{2\langle \theta, u\rangle}{\sqrt{1+\|u\|^2}}, \qquad u = x/\delta. \tag{138}$$

There are dimension-free constants R_r satisfying

$$\|D^r q_\theta(x)\|_{\mathrm{op}} \leq R_r \delta^{-r}. \tag{139}$$

To verify this bound at every order, write $w(u) = 1 + \|u\|^2$, $r_u = \sqrt{w(u)}$, and $R(u) = w(u)^{-1/2}$. Only singleton and pair blocks occur in the partition formula for $D^n R$: $D^k w = 0$ for $k \geq 3$. A partition with a singletons and b pairs has $n = a+2b$ and $k = a+b$ blocks. Its scalar outer derivative has absolute value $(2k-1)!!\, r_u^{-1-2k}/2^k$, while the product of its inner derivatives is bounded by $2^{a+b}\|u\|^a \prod_{i=1}^n \|v^i\|$. Since $\|u\| \leq r_u$, their product is at most $(2k-1)!!\, r_u^{-1-n} \prod_i \|v^i\|$. There are $n!/((n-2b)!b!2^b)$ such partitions. Consequently, with $H_0 = 1$,

$$H_n = \sum_{b=0}^{\lfloor n/2 \rfloor} \frac{n!}{(n-2b)!b!2^b}(2(n-b)-1)!!, \qquad \|D^n R(u)\|_{\mathrm{op}} \leq H_n r_u^{-1-n}. \tag{140}$$

Differentiate $2\langle \theta, u\rangle R(u)$. Either every derivative falls on R, or exactly one falls on the linear factor; the latter has n choices. Therefore

$$\|D^n q_\theta(u)\|_{\mathrm{op}} \leq 2(H_n + nH_{n-1}) r_u^{-n} \leq 2(H_n + nH_{n-1}).$$

This proves (139) after scaling. We use the sharper $R_1 = 2$, $R_2 = 6$ from Lemma 16, and the displayed formula for $r \geq 3$. For example, $H_1 = 1$, $H_2 = 4$, $H_3 = 24$, and $R_3 = 72$.

For disjoint blocks $q = (q_1, \ldots, q_T)$, the same constants give

$$\left\|D^s q(x)[v^1, \ldots, v^s]\right\|_2 \leq R_s \delta^{-s} \prod_{k=1}^s \|v^k\|_2. \tag{141}$$

Indeed, the square of the left side is at most $R_s^2 \delta^{-2s} \sum_j \prod_k \|v_j^k\|^2$; the sum is bounded by $\prod_k \sum_j \|v_j^k\|^2$, because its summands are among the nonnegative terms obtained by expanding that product.

Take independent hidden block directions and define

$$F_\theta(x) = A f_T(q(x)) + 4A \sum_{j=1}^T \|x_j/\delta\|^2. \tag{142}$$

For $r \geq 3$ the guard has zero rth derivative. The multivariate Faà di Bruno formula, indexed by the partitions π of $[r]$, is

$$D^r(f_T \circ q)[v^1, \ldots, v^r] = \sum_\pi D^{|\pi|} f_T(q)\big[D^{|B|} q[v^k : k \in B] : B \in \pi\big]. \tag{143}$$

For $|\pi| \geq 2$, Lemma 35 and (141) bound the corresponding term by

$$M_{|\pi|} \prod_{B \in \pi} R_{|B|} \delta^{-r} \prod_k \|v^k\|.$$

The single-block partition requires a separate estimate, since a Euclidean bound on ∇f_T would introduce $\sqrt{T}$. Instead,

$$\begin{aligned}\left|\sum_j (\partial_j f_T) D^r q_j[v_j^1, \ldots, v_j^r]\right| &\leq G R_r \delta^{-r} \sum_j \prod_k \|v_j^k\| \\ &\leq G R_r \delta^{-r} \prod_k \|v^k\|.\end{aligned}$$

The last step is Hölder with exponent $r \geq 2$, followed by monotonicity of the sequence norms, exactly as in Lemma 35. Thus the finite constant

$$C_r^{\mathrm{der}} = GR_r + \sum_{\substack{\pi \text{ a partition of } [r] \\ |\pi| \geq 2}} M_{|\pi|} \prod_{B \in \pi} R_{|B|} \tag{144}$$

satisfies

$$\|D^r F_\theta(x)\|_{\mathrm{op}} \leq C_r^{\mathrm{der}} A \delta^{-r} \qquad (r \geq 3). \tag{145}$$

Integrating $D^{p+1} F_\theta$ on the segment joining any two points gives $\mathrm{Lip}(D^p F_\theta) \leq C_{p+1}^{\mathrm{der}} A/\delta^{p+1}$. All constants are independent of both the dimension and the number of blocks.

D.3 Lower-bound scaling, information, and parameter endpoints

Define

$$C_{1,p} = \max\{10^4, (4^{1-p} C_{p+1}^{\mathrm{der}})^{1/p}\}.$$

Assume $L_p > 0$ and choose

$$\lambda = \min\{L, (L_p \varepsilon^{p-1})^{1/p}\}, \quad A = \frac{16 C_{1,p} \varepsilon^2}{\lambda}, \quad \delta = \frac{4 C_{1,p} \varepsilon}{\lambda}, \quad T = \left\lfloor \frac{C}{320 C_{1,p}} \right\rfloor. \tag{146}$$

These choices give $A/\delta = 4\varepsilon$ and $\Delta\lambda/\varepsilon^2 = C$. The nonnegative guard then keeps the initial gap below Δ, while Lemma 16 gives

$$\mathrm{Lip}(\nabla F_\theta) \leq 10^4 A/\delta^2 = (10^4/C_{1,p})\lambda \leq L.$$

The higher derivative bound is

$$\frac{C_{p+1}^{\mathrm{der}} A}{\delta^{p+1}} = \frac{4^{1-p} C_{p+1}^{\mathrm{der}}}{C_{1,p}^p} \frac{\lambda^p}{\varepsilon^{p-1}} \leq L_p. \tag{147}$$

Finally, any unfinished coordinate with $|q_j| \leq 1/2$ has gradient norm at least $A/\delta = 4\varepsilon$, by the stationarity certificate in Lemma 16.

Use the globally defined conservative sample potential of Lemma 18, with independent spherical codes having mean $\beta\theta_j$. Its construction and prefix identity (Lemma 19) remain valid: they use precisely (134) on the noise-supported region. In particular,

$$\sup_{x,z} \|\nabla f_\theta(x,z) - \nabla F_\theta(x)\| \leq \frac{4000A}{\beta\delta} = \frac{16000\varepsilon}{\beta}. \tag{148}$$

The noise potential has zero expectation and is bounded in absolute value for each sample, so adding it to the bounded-below population preserves a finite lower bound. Its tail cutoffs are C^2, which is all the sample regularity needed here.

For $m \geq M_0 := 2^{60}$, take

$$\beta = 2^{16}/\sqrt{m}, \qquad b = \lfloor 2^{-18} \beta^{-2} \rfloor = \lfloor m/2^{50} \rfloor. \tag{149}$$

Thus $\beta \leq 2^{-14}$, $b \geq 1024$, and (148) is less than $\sigma/4$. Grant all radial scalar projections at charged points and the output. They determine the exact population values, and their normalized measurement vectors have norm at most one. Lemmas 9 and 13 therefore apply to the conservative prefix identity. For clarity, the constants are compatible with the quantitative information bound: its code entropy coefficient is $\Lambda = 2$, the crossing threshold is $1/16$, and $2b\beta^2 \leq 2^{-17}$. Hence, by increasing the block dimension, the cheap-crossing probability

$$\frac{2Db\beta^2 + \log 2}{D/4096 - \log 2}$$

is at most $1/16$, and the skip probability can be made at most $1/16$. The prefix-counting lemma then prevents completion of all T stages with probability at least $3/4$ when there are fewer than $Tb/2$ retained sample evaluations, or fewer than $Tb/4$ simultaneous fresh rounds. On noncompletion the stationarity certificate gives a gradient norm at least 4ε, so the expected norm is larger than ε.

For $C \geq 640C_{1,p}$, $T \geq C/(640C_{1,p})$ and $b \geq m/2^{51}$; this proves $\Omega_p(mC)$. Every finite K only changes the finite dimension needed to include all charged points in the projection event.

For the deterministic term, Lemma 14 limits progress through the same chain to one stage per round, giving $\Omega_p(C)$ when $C \geq 640C_{1,p}$. If C is smaller, consider the two one-dimensional quadratics $Lx^2/2 \mp 4\varepsilon x$. Their initial gaps are $8\varepsilon^2/L \leq \Delta$ and all derivatives above order two are zero. For any query-free randomized output X,

$$\tfrac{1}{2}\mathbb{E}\big[|LX - 4\varepsilon| + |LX + 4\varepsilon|\big] \geq 4\varepsilon.$$

One of the signs therefore requires a call. These arguments together give $\Omega_p(1 + C)$, including $L_p = 0$.

Second, for $m \geq M_0$, use the spherical quadratic of Lemma 21:

$$F_\theta(x) = \tfrac{L}{2}\|x\|^2 - 8\varepsilon\langle\theta, x\rangle, \qquad f_\theta(x, S) = \tfrac{L}{2}\|x\|^2 - \frac{8\varepsilon}{\beta_s}\langle S, x\rangle, \qquad \beta_s = 16\varepsilon/\sigma.$$

The quadratic displayed above satisfies the required initial-gap, higher-order smoothness, and noise-radius bounds. Exact values reveal only scalar projections, so the spherical testing argument gives $\Omega(m)$ even with retained samples and exact values. This covers $m(1 + C)$ when C is bounded. For larger C, the chain gives $\Omega_p(mC)$; when m is small, the deterministic bound supplies the remaining factor.

D.4 A FRESH-GRADIENT ALGORITHM USING LAGRANGE INTERPOLATION

We prove the upper bound with a fixed number of calls, assuming only conditional unbiasedness and a conditional centered second-moment bound of σ^2. The algorithm requires neither sample smoothness nor sample reuse. Define the Lagrange basis at $0, \ldots, p - 1$ by

$$\ell_j(s) = \prod_{\substack{0 \leq k \leq p-1 \\ k \neq j}} \frac{s - k}{j - k}, \qquad 0 \leq j \leq p - 1.$$

For $0 < \alpha \leq 1$ write $\vartheta = 1 - \alpha$ and put

$$c_j(\alpha) = \frac{\ell_j(0) - \vartheta\ell_j(-\alpha)}{\alpha} = \ell_j(0) + \vartheta\frac{\ell_j(0) - \ell_j(-\alpha)}{\alpha}. \tag{150}$$

The finite constants

$$K_j = |\ell_j(0)| + \max_{s \in [-1,0]} |\ell_j'(s)|, \qquad V_p = \max\{1, \sum_j K_j^2\}, \qquad B_p = \frac{1 + \sum_j K_j j^p}{p!}$$

depend only on p. The mean value theorem gives $|c_j(\alpha)| \leq K_j$. For every vector polynomial P of degree at most $p - 1$, interpolation gives the exact cancellation identity

$$\vartheta P(-\alpha) + \alpha \sum_j c_j(\alpha) P(j) = P(0). \tag{151}$$

Indeed substitute $P(s) = \sum_j \ell_j(s)P(j)$ at $s = 0, -\alpha$ and use (150). Taking P constant also gives $\sum_j c_j = 1$. The coefficients need not be positive.

Average B independent fresh gradients at $x_0 = 0$ to form v_0. Define $\mathrm{dir}(v) = v/\|v\|$ for $v \neq 0$, and zero for $v = 0$, and set

$$x_{t+1} = x_t - h\,\mathrm{dir}(v_t).$$

At each $t \geq 1$, let $d_t = x_t - x_{t-1}$, query fresh independent samples at

$$y_{t,j} = x_t + \frac{j}{\alpha} d_t \quad (0 \leq j \leq p - 1), \qquad v_t = \vartheta v_{t-1} + \alpha \sum_j c_j(\alpha) g(y_{t,j}, Z_{t,j}). \tag{152}$$

The past fixes these query points before any samples are drawn, so each call is fresh and single-point.

The operator-norm assumption on D^pF implies the vector Taylor bound

$$\left\| \nabla F(x+u) - \sum_{k=0}^{p-1} \frac{1}{k!} D^k(\nabla F)(x)[u^k] \right\| \le \frac{L_p}{p!} \|u\|^p. \tag{153}$$

To see the constant explicitly, apply the integral Taylor formula through degree $p-2$ to ∇F on $x+tu$, and subtract the degree $p-1$ term. The remaining integral is

$$\frac{1}{(p-2)!} \int_0^1 (1-t)^{p-2} [D^{p-1}(\nabla F)(x+tu) - D^{p-1}(\nabla F)(x)][u^{p-1}]\, dt.$$

Its norm is at most $L_p\|u\|^p \int_0^1 t(1-t)^{p-2}dt/(p-2)! = L_p\|u\|^p/p!$. The vector norm of the derivative difference equals the supremum over an additional unit vector, and is therefore controlled by the assumed full tensor norm.

Apply (153) along the line $x_t + sd_t/\alpha$. The point x_{t-1} has $s = -\alpha$, so (151) cancels the complete Taylor polynomial. Since $\|d_t\| \le h$, the predictable transport error

$$R_t = \vartheta \nabla F(x_{t-1}) + \alpha \sum_j c_j \nabla F(y_{t,j}) - \nabla F(x_t)$$

satisfies

$$\|R_t\| \le \frac{L_p h^p}{p!} + \frac{L_p h^p \alpha^{1-p}}{p!} \sum_j K_j j^p \le B_p L_p h^p \alpha^{1-p}. \tag{154}$$

Let $e_t = v_t - \nabla F(x_t)$ and $\xi_t = \sum_j c_j[g(y_{t,j}, Z_{t,j}) - \nabla F(y_{t,j})]$. Conditional independence of the new samples and their zero means imply

$$\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}[\|\xi_t\|^2 \mid \mathcal{F}_{t-1}] \le \sigma^2 \sum_j c_j^2 \le V_p \sigma^2.$$

Here $\mathcal{F}_{t-1}$ includes the points already determined for step t. The error recursion and its explicit solution are

$$\begin{aligned} e_t &= \vartheta e_{t-1} + \alpha \xi_t + R_t, \\ e_t &= \vartheta^t e_0 + \alpha \sum_{s=1}^{t} \vartheta^{t-s} \xi_s + \sum_{s=1}^{t} \vartheta^{t-s} R_s. \end{aligned} \tag{155}$$

Distinct martingale differences are orthogonal in expectation: for $s < r$, conditioning on $\mathcal{F}_{r-1}$ gives $\mathbb{E}\langle \xi_s, \xi_r \rangle = 0$. Thus the squared norm of the middle sum has expectation at most

$$\alpha^2 V_p \sigma^2 \sum_{j=0}^{t-1} \vartheta^{2j} \le V_p \sigma^2 \frac{\alpha}{2-\alpha} \le V_p \sigma^2 \alpha.$$

The initial average has $\mathbb{E}\|e_0\| \le \sigma/\sqrt{B}$. Apply the triangle inequality to the predictable remainders, without assuming any independence for them, and sum their geometric weights. We obtain

$$\mathbb{E}\|e_t\| \le \vartheta^t \frac{\sigma}{\sqrt{B}} + \sigma \sqrt{V_p \alpha} + \frac{B_p L_p h^p}{\alpha^p}. \tag{156}$$

For any a, e, the elementary direction inequality is

$$\langle a, \mathrm{dir}(a+e) \rangle \ge \|a\| - 2\|e\|.$$

If $a + e \ne 0$, its left side is at least $\|a+e\| - \|e\| \ge \|a\| - 2\|e\|$; if $a + e = 0$, the right side is $-\|a\| \le 0$. Smoothness and (156) therefore give, after summing $F(x_{t+1}) \le F(x_t) - h\|\nabla F(x_t)\| + 2h\|e_t\| + Lh^2/2$,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(x_t)\| \le \frac{\Delta}{hT} + \frac{2\sigma}{\sqrt{B}\alpha T} + 2\sigma\sqrt{V_p \alpha} + \frac{2B_p L_p h^p}{\alpha^p} + \frac{Lh}{2}. \tag{157}$$

In using the geometric sum here, $\sum_{t=0}^{T-1} \vartheta^t \le 1/\alpha$ also covers $\alpha = 1$.

Choose the deterministic parameters

$$\alpha = \min\{1, \varepsilon^2/(64V_p\sigma^2)\}, \qquad B = \max\{1, \lceil 64m \rceil\},$$
$$h = \min\left\{\frac{\varepsilon}{4L}, \alpha\left(\frac{\varepsilon}{16B_pL_p}\right)^{1/p}\right\}, \qquad T = \left\lceil \max\left\{\frac{4\Delta}{h\varepsilon}, \frac{2}{\alpha}\right\}\right\rceil. \tag{158}$$

Set $\alpha = 1$ when $\sigma = 0$, and omit the second restriction on h when $L_p = 0$. The five terms in (157) are bounded respectively by $\varepsilon/4, \varepsilon/8, \varepsilon/4, \varepsilon/8, \varepsilon/8$. For the second term use $\sigma/\sqrt{B} \le \varepsilon/8$ and $\alpha T \ge 2$; for the third use the definition of α. Their sum is $7\varepsilon/8$. An independent uniform choice from $x_0, \dots, x_{T-1}$ consequently meets the required expected gradient-norm criterion. The exact number of calls is $B + p(T-1)$; the unused final move to x_T requires no oracle call. Since $\alpha^{-1} = O_p(1+m)$, its size is

$$O_p\big(1 + m + Q + A_p(1+m)\big). \tag{159}$$

For comparison, a fixed SGD schedule costs $O(Q(1+m))$ whenever $Q \ge 1$. Use $x_{t+1} = x_t - \gamma g_t$ with $\gamma = \min\{1/L, \varepsilon^2/(2L\sigma^2)\}$, omitting the second restriction when $\sigma = 0$. Conditional smooth descent gives

$$\frac{1}{N}\sum_{t<N} \mathbb{E}\|\nabla F(x_t)\|^2 \le \frac{2\Delta}{\gamma N} + L\gamma\sigma^2.$$

Choose $N \ge 4\Delta/(\gamma\varepsilon^2)$ and use an independent uniform output and Jensen's inequality. This proves the stated cost. When $m < 1$, this schedule costs $O(Q)$, and $Q \le Q + m + mC \le 3Q$. When $m \ge 1$, the extrapolation schedule costs $O_p(Q + m + A_p m)$, while SGD costs $O(Q + Qm)$. Choosing in advance the algorithm with the smaller evaluation budget gives

$$O_p\big(Q + m + \min\{Qm, A_p m\}\big) = O_p(Q + m + mC).$$

Indeed, if $A_p \le Q$ choose extrapolation, and if $A_p > Q$ choose SGD. These comparisons also cover $L_p = 0$ and $\sigma = 0$.

Finally, under (132), $Q + m + mC \le 2(1+m)(1+C)$, which proves Proposition 34. For fixed positive Δ, L, L_p, σ, the active higher-order branch eventually has $C = A_p$ and $mA_p \ge Q$, so its leading complexity is $\Theta_p(\Delta L_p^{1/p}\sigma^2\varepsilon^{-3-1/p})$. Outside (132), both bounds still hold; they do not give a complete minimax classification there.

Corollary 36 (Lipschitz population Hessian)**.** *For $p = 2$, put $A_2 = \Delta\sqrt{L_2}/\varepsilon^{3/2}$ and $C = \min\{Q, A_2\}$. Under the accuracy condition of Proposition 34 and $Q \le (1+m)(1+C)$, the exact complexity is $\Theta((1+m)(1+C))$. For fixed positive Δ, L, L_2, σ and sufficiently small ε, this becomes $\Theta(\Delta\sqrt{L_2}\,\sigma^2\varepsilon^{-7/2})$. Only the population Hessian is required to be Lipschitz.*

Proof. Specialize Proposition 34 to $p = 2$. For the fixed-parameter assertion, $A_2/Q = (\sqrt{L_2}/L)\varepsilon^{1/2} \to 0$ and $mA_2/Q = (\sigma^2\sqrt{L_2}/L)\varepsilon^{-3/2} \to \infty$. Thus the higher-order branch and the matching condition eventually both hold, while $m, A_2 \to \infty$. □

E SMOOTH CONVEX OBJECTIVES WITH UNRESTRICTED QUERIES

We give a direct proof of the deterministic and statistical lower bounds in Theorem 6 for unrestricted queries. For the matching upper bound, we use the fresh-gradient accelerated scheme with budget (224). All objectives in this appendix are globally differentiable, convex, and have Lipschitz gradients; the deterministic flat-link examples are $C^{1,1}$. In addition, the conservative sample functions in every lower-bound example are individually convex and L-smooth.

Proposition 37. *There are universal constants $c, c_0 > 0$ such that, for $L, R > 0$, $\sigma \ge 0$, and $0 < \eta \le c_0 LR^2$, attaining expected function gap at most η requires at least*

$$c\left(\sqrt{\frac{LR^2}{\eta}} + \frac{\sigma^2R^2}{\eta^2}\right) \tag{160}$$

evaluations over convex L-smooth functions with a minimizer of norm at most R. The same lower bound holds in simultaneous fresh common-sample rounds of any finite width K, with constants independent of K. Queries and outputs may be unrestricted; exact population and sample values are allowed at charged points, as is arbitrary retention and reuse of opaque samples in the evaluation model.

E.1 Exact energy and a robust flat convex chain

For $T \geq 2$, $a, \gamma > 0$, and $q_{T+1} = 0$, consider

$$J_T(q) = \frac{\gamma}{2}\left[q_1^2 - 2aq_1 + \sum_{i=1}^{T}(q_i - q_{i+1})^2\right]. \tag{161}$$

Its Hessian is the positive-definite tridiagonal matrix with diagonal 2γ and neighboring off-diagonal entries $-\gamma$. Positivity also follows directly from $q_1^2 + \sum_i (q_i - q_{i+1})^2 > 0$ for nonzero q. The stationary equations are $2q_1 - q_2 = a$ and $2q_i - q_{i-1} - q_{i+1} = 0$ for $i \geq 2$. Their unique solution is the linear profile

$$q_i^{(T)} = a\frac{T+1-i}{T+1}, \qquad \min J_T = -\frac{\gamma a^2 T}{2(T+1)}. \tag{162}$$

The profile satisfies each displayed equation and the boundary value $q_{T+1} = 0$. To compute its value, write $J_T(q) = \frac{1}{2}q^T Hq - \gamma a q_1$ and use $Hq^{(T)} = \gamma a e_1$, which gives $J_T(q^{(T)}) = -\gamma a q_1^{(T)}/2$. This calculation also holds for a chain of length one.

Restricting (161) to $q_{p+1:T} = 0$ gives precisely J_p. For $1 \leq p \leq T/2$, its restricted minimum therefore satisfies

$$\begin{aligned}\min_{q_{p+1:T}=0} J_T - \min J_T &= \frac{\gamma a^2}{2}\left(\frac{1}{p+1} - \frac{1}{T+1}\right) \\ &\geq \frac{\gamma a^2}{4(T+1)}.\end{aligned} \tag{163}$$

For the inequality, $p + 1 \leq (T + 2)/2$ and hence the parenthesis is at least $T/((T + 2)(T + 1)) \geq 1/(2(T + 1))$.

For a flatness width $c > 0$, use the even function

$$\varphi_c(s) = \begin{cases} 0, & |s| \leq c, \\ 2(|s| - c)^2, & c < |s| \leq 2c, \\ s^2 - 2c^2, & |s| > 2c. \end{cases} \tag{164}$$

On the nonnegative half-line its derivative is respectively $0, 4(s - c), 2s$, with matching values at c and $2c$. The odd extension of this derivative is continuous and nondecreasing, so φ_c is convex; its derivative is 4-Lipschitz. Direct subtraction in the three regions gives

$$0 \leq s^2 - \varphi_c(s) \leq 2c^2. \tag{165}$$

In the middle region the difference is $-s^2 + 4cs - 2c^2$, increasing from c^2 to $2c^2$; the other two regions are immediate.

Let $\theta_1, \ldots, \theta_T$ be independent unit vectors in disjoint D-dimensional blocks. For $x = (x_1, \ldots, x_T)$ put $q_i = \langle \theta_i, x_i \rangle$ and define

$$h_\theta(x) = \frac{\gamma}{2}\left[q_1^2 - 2aq_1 + \sum_{i=1}^{T}\varphi_c(q_i - q_{i+1})\right]. \tag{166}$$

The function h_θ is convex and 9γ-smooth, with $h_\theta(0) = 0$. For the smoothness bound, the root quadratic contributes a gradient Lipschitz constant of γ. The difference map $Bq = (q_i - q_{i+1})_{i=1}^T$ obeys

$$\|Bq\|^2 \leq 2\sum_i (q_i^2 + q_{i+1}^2) \leq 4\|q\|^2,$$

so its operator norm is at most two. The edge term has Lipschitz gradient at most $(\gamma/2)\cdot 4\cdot\|B\|^2 \le 8\gamma$. The block projection $x \mapsto q$ has norm at most one, proving the claimed bound in the ambient space as well.

The chain has a stationary global minimizer of controlled norm. Indeed, let P be scalar projection onto $[0, a]$. This map satisfies

$$|P(u) - P(v)| \le |u - v|, \qquad |P(u)| \le |u|, \qquad |P(u) - a| \le |u - a|.$$

The root term is $(q_1 - a)^2 - a^2$, and φ_c is nondecreasing in absolute value. Applying P to every coordinate, including the fixed terminal coordinate zero, therefore cannot increase the objective. The continuous scalar objective attains its minimum on the compact cube $[0, a]^T$; the clipping inequality makes that minimum global. Embedding its minimizer as $x_i^* = q_i^*\theta_i$ yields

$$\|x^*\|^2 = \sum_i (q_i^*)^2 \le Ta^2. \tag{167}$$

Differentiability and unconstrained global minimality imply $\nabla h_\theta(x^*) = 0$.

Let $J_T^\theta(x) = J_T(q(x))$. Summing (165) gives

$$J_T^\theta(x) - \gamma T c^2 \le h_\theta(x) \le J_T^\theta(x). \tag{168}$$

We need a gap certificate at points whose hidden tail is small, rather than exactly zero. Embed the length-p quadratic minimizer as $z_i = a(p+1-i)/(p+1)$ for $i \le p$ and zero thereafter. The gradient of J_T at this point vanishes in all coordinates except $p+1$, where it equals $-\gamma z_p = -\gamma a/(p+1)$. The supporting-hyperplane inequality for the convex quadratic gives

$$J_T(q) \ge J_p(z) - \frac{\gamma a}{p+1} q_{p+1}.$$

Consequently, if $|q_{p+1}| \le c/4$ and $1 \le p \le T/2$, then

$$h_\theta(x) - \min h_\theta \ge \frac{\gamma a^2}{4(T+1)} - \frac{\gamma ac}{4(p+1)} - \gamma T c^2. \tag{169}$$

Here (168) gives the lower bound on the first term $h_\theta(x)$, while $\min h_\theta \le \min J_T$ gives the needed upper bound on its subtracted minimum.

Choose

$$\gamma = \frac{L}{27000}, \qquad a = \frac{R}{2\sqrt{T}}, \qquad c = \frac{a}{64T}, \qquad L_0 = 9\gamma. \tag{170}$$

The minimizer in (167) then has norm at most $R/2$. Since $T + 1 \le 2T$,

$$\frac{\gamma ac}{4(p+1)} \le \frac{\gamma a^2}{128(T+1)}, \qquad \gamma T c^2 \le \frac{\gamma a^2}{2048(T+1)}.$$

Subtracting these two bounds from (169) leaves more than $\gamma a^2/(8(T+1))$. As $T + 1 \le 2T$ and $a^2 = R^2/(4T)$, we have proved the useful implication

$$|q_{p+1}(x)| \le c/4, \quad 1 \le p \le T/2 \quad \Longrightarrow \quad h_\theta(x) - \min h_\theta \ge \frac{\gamma R^2}{64T^2}. \tag{171}$$

This implication places no restriction on the other coordinates of x.

E.2 A SCALAR EXTENSION AND THE UNRESTRICTED DETERMINISTIC LOWER BOUND

The chain h_θ is globally convex and has a stationary minimizer in $B(0, R)$. Apply Lemma 23 with smoothness L_0 and radius R. To specify the resulting function, let

$$s(t) = \begin{cases} 1, & t \le 0, \\ 1 - 10t^3 + 15t^4 - 6t^5, & 0 < t < 1, \\ 0, & t \ge 1, \end{cases} \qquad C(x) = s((\|x\| - R)/R).$$

Then

$$F_\theta(x) = C(x)h_\theta(x) + 1000L_0(\|x\| - R)_+^2 \quad (\|x\| < 2R), \tag{172}$$

and only the guard term is used for $\|x\| \geq 2R$. The cited lemma proves global convexity, smoothness at most $3000L_0 = L$, and equality of the minimum values of F_θ and h_θ. The minimizer from (167) is preserved. Outside the radius-$2R$ ball both its value and its gradient are independent of every hidden direction. Inside that ball its gradient is exactly

$$\nabla F_\theta = C\nabla h_\theta + h_\theta \nabla C + \nabla[1000L_0(\|x\| - R)_+^2]. \tag{173}$$

Thus the extended oracle returns the gradient of the scalar function F_θ, together with its value.

We record a direct output certificate, so no output-radius assumption is needed in the information argument. By (162) and (168),

$$-\min h_\theta \geq \frac{\gamma a^2 T}{2(T+1)}. \tag{174}$$

If $\|x\| \geq 2R$, then $F_\theta(x) \geq 0$. The same is true for an interior point with $h_\theta(x) > 0$. In both cases its gap is at least the right side of (174), and hence exceeds $\gamma R^2/(64T^2)$. If $\|x\| < 2R$ and $h_\theta(x) \leq 0$, then $C(x)h_\theta(x) \geq h_\theta(x)$ and the guard is nonnegative. Thus (171) transfers to the extended objective whenever its coordinate hypothesis holds. In particular, an output with gap less than $\gamma R^2/(64T^2)$ must be inside the outer ball and cross the relevant hidden coordinate.

To apply the information bound to arbitrary randomized algorithms, we define responses that depend only on a prefix of the hidden directions. For $0 \leq p < T$, define the scalar prefix

$$h_{\theta,p}(x) = \frac{\gamma}{2}\left[q_1^2 - 2aq_1 + \sum_{i=1}^{p} \varphi_c(q_i - q_{i+1})\right]. \tag{175}$$

It depends only on $\theta_1, \ldots, \theta_{p+1}$. If all $|q_i| \leq c/4$ for $i > p$, every omitted difference has magnitude at most $c/2$, or $c/4$ at the terminal edge. Each corresponding edge term $\varphi_c(q_i - q_{i+1})$ and its derivative therefore vanish, giving $h_\theta = h_{\theta,p}$ and $\nabla h_\theta = \nabla h_{\theta,p}$ at that point. Substituting these equalities in (172) and (173) proves equality of both extended value and gradient responses with the extended prefix. The term $h_{\theta,p}\nabla C$ is included in this identity.

Grant at every charged query and at the final output all scalar measurements

$$y_i(x) = \left\langle \theta_i, \frac{x_i}{\max\{2R, \|x\|\}} \right\rangle. \tag{176}$$

Their measurement vectors have norm at most one. They determine all hidden coordinates at any charged point and thus its exact population value. Inside the outer ball, $|q_i| > c/4$ implies $|y_i| > \alpha$, where $\alpha = c/(8R)$. For a proposed budget of $N \geq 1$ fresh rounds of finite width K, put

$$T = 2(N+2), \qquad p = N+1, \qquad D \geq \max\left\{\frac{16(NK+1)}{\alpha^2}, \frac{16\log(256T)}{\alpha^2}\right\}. \tag{177}$$

Round the dimension upward to an integer.

Lemma 14 now applies to the explicit prefix (175). To make its use transparent, until a direction is entered, all responses are functions of previously entered directions, independent auxiliary randomness, and its scalar projections (176). Conditional on the other directions and the algorithm's random seed, its unexplored residual is uniform on a sphere. Lemma 9, with at most $NK+1$ measurements, therefore bounds a premature crossing of that direction by $2\exp(-\alpha^2 D/16)$. A union bound over T directions and (177) bounds all skipped entries by $1/128$. The capped interaction stops on a skip; before that event its prefix responses agree exactly with the true responses just verified. All K points in a round are chosen before any of their replies, so the cap can reveal at most one additional direction in that round. The additional measurement for the arbitrary output is included in the same bound. Consequently, except with probability $1/128$, an interior output has $|q_{p+1}| \leq c/4$ with $p = N+1 \leq T/2$.

The output certificate proves that the average expected gap under this fixed spherical prior is at least

$$\frac{127}{128}\frac{\gamma R^2}{64T^2} \geq \frac{LR^2}{13824000\,(N+2)^2}. \tag{178}$$

Indeed $T^2 = 4(N+2)^2$, $\gamma = L/27000$, and replacing $127/128$ by $1/2$ gives the displayed denominator. The prior and dimension are fixed before interaction, independently of the algorithm's random

seed; averaging then gives an individual hard instance. Set $f(\cdot, z) = F_\theta$ for every sample: the noise is zero and sample reuse adds no information. Taking $K = 1$ proves the evaluation lower bound; the preceding argument itself proves the round lower bound for every finite K.

For a zero-call algorithm, use $F_\pm(x) = L\|x \mp Re_1\|^2/2$. For every random output X independent of the sign, its average risk over the two signs is $L\mathbb{E}\|X\|^2/2 + LR^2/2 \geq LR^2/2$. Together with (178), this yields a universal $\Omega(\sqrt{LR^2/\eta})$ complexity lower bound for $\eta \leq c_0 LR^2$, after choosing $c_0 > 0$ sufficiently small to absorb the additive constant in $N + 2$.

E.3 A BOUNDED STATISTICAL QUADRATIC WITH EXACT POPULATION VALUES

Fix a proposed budget $N \geq 1$, and suppose $\sigma > 0$. Define

$$b = \min\{LR, \sigma/(128\sqrt{N})\}, \qquad \nu = b/R, \qquad \beta = 2b/\sigma. \tag{179}$$

Let U be uniform on S^{D-1}. Given U, let V have the Poisson-kernel law in Lemma 10, so that $\|V\| = 1$ and $\mathbb{E}[V \mid U] = \beta U$. Define globally

$$F_U(x) = \frac{\nu}{2}\|x - RU\|^2, \qquad f_U(x, V) = \frac{\nu}{2}\|x\|^2 - \frac{\sigma}{2}\langle V, x\rangle + \frac{bR}{2}. \tag{180}$$

Since $\nu R = b = \sigma\beta/2$, expanding the square shows exactly that $\mathbb{E}[f_U(x, V) \mid U] = F_U(x)$. The minimizer is RU, of norm R. Every sample is a bounded-below convex ν-smooth quadratic, with $\nu \leq L$. Its centered gradient noise is the constant vector $bU - \sigma V/2$, and hence simultaneously at every point and every sample

$$\|\nabla f_U(x, V) - \nabla F_U(x)\| \leq b + \sigma/2 \leq (1/128 + 1/2)\sigma < \sigma. \tag{181}$$

Strengthen the oracle by revealing the complete vector V at its first use. All subsequent values and gradients from that retained sample are deterministic functions of V and the query; reuse yields no additional random observations. A common-sample round likewise contains exactly one such vector, regardless of its finite width K. In either model, the comparison therefore involves at most N new vectors. In the reference experiment, we replace each by an independent uniform sphere point and keep the prior on U unchanged. Conditionally on U and the preceding transcript, the entropy bound in Lemma 10 is $2D\beta^2$ for a new vector and zero for a cached one. Exact population values are the same deterministic functions of U and the query in both experiments; in the joint laws including U their conditional KL contribution is zero. The chain rule therefore gives

$$\mathrm{KL}(P\|P_0) \leq 2ND\beta^2 \leq D/2048. \tag{182}$$

The last inequality uses $\beta \leq 1/(64\sqrt{N})$.

Each exact population value in (180) reveals only the scalar product $\langle U, x\rangle$: its other terms are known. Grant these projections directly, normalize every nonzero query direction to length one, and add the output as one further measurement. For N evaluations let $M = N + 1$; for N rounds of width K let $M = NK + 1$. Choose

$$D \geq \max\{8192, 256M\}. \tag{183}$$

Under the reference, U is independent of the code vectors and the private randomness. At threshold $1/4$, Lemma 9 bounds the probability that any measured direction has absolute correlation larger than $1/4$ by $2e^{-D/256}$. This includes the output even if it has never been queried.

If $F_U(\widehat{x}) \leq bR/8$, then $\|\widehat{x} - RU\| \leq R/2$. Such an output is nonzero and obeys

$$\left\langle U, \frac{\widehat{x}}{\|\widehat{x}\|}\right\rangle \geq \frac{R - \|\widehat{x} - RU\|}{R + \|\widehat{x} - RU\|} \geq \frac{1}{3}.$$

Its success event E is consequently contained in the preceding projection-crossing event under the reference. Binary data processing and the inequality

$$\mathrm{kl}(P(E), P_0(E)) \geq P(E)\log(1/P_0(E)) - \log 2$$

imply

$$P(E) \leq \frac{D/2048 + \log 2}{D/256 - \log 2} \leq \frac{1}{4}. \tag{184}$$

For the final numerical inequality, multiply by the positive denominator: it suffices that $5 \log 2 \leq D/512$, which follows from $D \geq 8192$. The population gap is nonnegative, so failure of E and (184) give

$$\mathbb{E}[F_U(\widehat{x}) - \min F_U] \geq \frac{3bR}{32} = \frac{3}{32} \min\left\{LR^2, \frac{\sigma R}{128\sqrt{N}}\right\}. \tag{185}$$

Again average over the spherical prior to obtain an individual fixed hard instance. This proof simultaneously covers unrestricted exact-valued queries, arbitrary retained handles counted by evaluations, and fresh common-sample rounds with any finite K.

If $\eta \leq c_0 LR^2$ with $c_0 < 3/32$, a guarantee of gap at most η is incompatible with the first branch in (185), and the second branch forces $N \geq (3/4096)^2 \sigma^2 R^2/\eta^2$. When this numerical bound is below one, the deterministic lower bound already supplies the required constant. If $\sigma = 0$, the statistical term is zero and this construction is unnecessary. Taking the maximum of the independent deterministic and statistical hard subclasses gives at least half the sum of their lower bounds. This proves Proposition 37. Together with the explicit fresh-gradient upper bound, it yields Theorem 6 for both resource measures.

F Algorithms and matching upper bounds

The upper bounds use fresh stochastic gradients, paired evaluations, exact finite-sum anchors, and acceleration according to the objective class. We choose all horizons, batch sizes, and refresh times from the announced parameters, so the evaluation budgets are worst-case bounds even though the guarantees are in expectation. The algorithms do not use extra sample values or population values.

F.1 Fresh-sample descent and relative variance

For fresh-sample descent, write $g(x, Z) = \nabla F(x) + \zeta(x, Z)$. Independence of each fresh sample from the preceding transcript, unbiasedness, and the bound $\mathbb{E}\,\|\zeta(x, Z)\|^2 \leq \sigma^2$ give the conditional descent inequality below. Uniformly bounded noise implies this centered second-moment bound. We allow the starting point X to depend on earlier answers, but draw every sample in the current run independently of X and of earlier samples. For $x_0 = X$ and $x_{t+1} = x_t - hg(x_t, Z_t)$, conditional expectation in the smoothness inequality gives

$$\mathbb{E}[F(x_{t+1}) \mid \mathcal{F}_t] \leq F(x_t) - h(1 - Lh/2)\,\|\nabla F(x_t)\|^2 + \frac{Lh^2\sigma^2}{2}. \tag{186}$$

Here $\mathcal{F}_t$ contains all information available before Z_t is drawn. Consequently, for $h \leq 1/L$ and an independent index R uniform on $\{0, \ldots, T-1\}$,

$$\mathbb{E}\,\|\nabla F(x_R)\|^2 \leq \frac{2\Delta}{hT} + Lh\sigma^2. \tag{187}$$

Set $Q = \Delta L/\varepsilon^2$ and $m = \sigma^2/\varepsilon^2$. The fixed choices

$$h = \frac{1}{L(1+2m)}, \qquad T = \lceil 4Q(1+2m) \rceil \tag{188}$$

make the two terms in (187) at most $\varepsilon^2/2$ each. Jensen's inequality therefore gives $\mathbb{E}\,\|\nabla F(x_R)\| \leq \varepsilon$. The budget is $O(1+Q+Qm)$, and when $Q \geq 1$ it is at most $9Q(1+m)$. This proof includes $m = 0$ and, in particular, costs $O(Q)$ when $m < 1$. If $\varepsilon^2 \geq 2L\Delta$, no call is needed: smoothness and the lower bound on F imply $\|\nabla F(X)\|^2 \leq 2L(F(X) - F^*)$ pointwise. All stationarity comparisons below are in the nontrivial regime $Q \geq 1$.

If also $\mathbb{E}\,\|\zeta(x, Z)\|^2 \leq \rho\,\|\nabla F(x)\|^2$, unbiasedness gives $\mathbb{E}\,\|g(x, Z)\|^2 \leq (1+\rho)\,\|\nabla F(x)\|^2$. The same calculation, with

$$h = \frac{1}{L(1+\rho)}, \qquad T = \lceil 2Q(1+\rho) \rceil, \tag{189}$$

yields

$$\mathbb{E}[F(x_{t+1}) \mid \mathcal{F}_t] \leq F(x_t) - \frac{h}{2}\,\|\nabla F(x_t)\|^2, \qquad \mathbb{E}\,\|\nabla F(x_R)\|^2 \leq \frac{2\Delta}{hT} \leq \varepsilon^2.$$

Choosing in advance the smaller of the budgets in (188) and (189) proves $O(Q[1+\min\{m,\rho\}])$. For an active finite sum, exact gradient descent is a third option: compute all n component gradients at each of $\lceil 2Q \rceil$ iterations and step by $1/L$. Its cost is $n\lceil 2Q \rceil$. Choosing the least of these three deterministic budgets proves $O(Q[1+\min\{n,m,\rho\}])$. Omitting the relative-variance option gives the general finite-sum cap $O(Q[1+\min\{n,m\}])$.

F.2 A RECURSIVE ESTIMATOR CONTROLLED BY CENTERED NOISE

Suppose that, for every deterministic pair x, y,

$$\mathbb{E}\,\|\zeta(x,Z)-\zeta(y,Z)\|^2 \le H^2\,\|x-y\|^2\,. \tag{190}$$

Each pair of query points is chosen before drawing its sample, so the mean-square assumption for every fixed pair suffices. Requiring the Lipschitz bound to hold simultaneously for all points on one probability-one sample set is stronger. Set

$$h=\frac{1}{2L},\quad T=\lceil 8Q\rceil,\quad b=\max\left\{1,\left\lceil\frac{H^2 q}{2L^2}\right\rceil\right\},\qquad 1\le q\le T. \tag{191}$$

At each deterministic anchor index $s\in\{0,q,2q,\dots\}\cap[0,T-1]$, form an estimate v_s with conditional mean-square error at most $\varepsilon^2/2$. For online sampling it will be an average of B fresh gradients. At every nonanchor index t, draw b independent fresh samples and set

$$v_t=v_{t-1}+\frac{1}{b}\sum_{i=1}^{b}\big(g(x_t,Z_{t,i})-g(x_{t-1},Z_{t,i})\big),\qquad x_{t+1}=x_t-hv_t. \tag{192}$$

We use the same step formula at anchor indices. In (192), both query points are chosen before the sample is drawn, so each sample needs only a nonadaptive pair of queries.

To bound the error of the recursive estimator, put $e_t=v_t-\nabla F(x_t)$. Within an epoch starting at s, the increment

$$e_t-e_{t-1}=\frac{1}{b}\sum_{i=1}^{b}\big(\zeta(x_t,Z_{t,i})-\zeta(x_{t-1},Z_{t,i})\big)$$

has conditional mean zero, and its conditional second moment is at most $H^2\,\|x_t-x_{t-1}\|^2/b$ by independence across the batch. The mixed inner product with e_{t-1} therefore has expectation zero. Induction gives

$$\mathbb{E}\,\|e_t\|^2\le\frac{\varepsilon^2}{2}+\frac{H^2h^2}{b}\sum_{j=s}^{t-1}\mathbb{E}\,\|v_j\|^2\,. \tag{193}$$

The deterministic-pair assumption applies at random pairs by conditioning on the preceding transcript. Summing over each epoch, including the final shorter epoch, gives

$$\sum_{t<T}\mathbb{E}\,\|e_t\|^2\le\frac{T\varepsilon^2}{2}+\frac{H^2h^2q}{b}\sum_{t<T}\mathbb{E}\,\|v_t\|^2\,. \tag{194}$$

Smoothness and the identity $2\langle\nabla F,v\rangle=\|\nabla F\|^2+\|v\|^2-\|v-\nabla F\|^2$ give the pathwise inequality

$$F(x_{t+1})\le F(x_t)-\frac{h}{2}\,\|\nabla F(x_t)\|^2-\frac{h}{4}\,\|v_t\|^2+\frac{h}{2}\,\|e_t\|^2\,. \tag{195}$$

Since $H^2h^2q/b\le 1/2$, the positive term proportional to $\sum_{t<T}\mathbb{E}\,\|v_t\|^2$ contributed by (194) is canceled by the negative term from (195). After telescoping and using the expected initial gap,

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\,\|\nabla F(x_t)\|^2\le\frac{2\Delta}{hT}+\frac{\varepsilon^2}{2}\le\varepsilon^2. \tag{196}$$

Choosing an independent uniform output index gives the required expected norm guarantee, without a pathwise bound on the random starting gap.

Online anchors and the exact count. If $m < 1$, the SGD schedule already has the desired cost. Otherwise use an anchor batch of size $B = \lceil 2m \rceil$; its error variance is at most $\sigma^2/B \le \varepsilon^2/2$. When $H = 0$, take one epoch. When $0 < H \le L\sqrt{B}$, choose

$$a = \frac{L\sqrt{B}}{H}, \qquad q = \min\{T, \max\{1, \lfloor a \rfloor\}\}. \tag{197}$$

For $H > L\sqrt{B}$, return to SGD. The recursive schedule has exactly $J = \lceil T/q \rceil$ anchors; consequently the numbers of evaluations and fresh-sample rounds are, respectively,

$$N_{\mathrm{rec}} = BJ + 2b(T - J), \qquad R_{\mathrm{rec}} = BJ + b(T - J). \tag{198}$$

Each round queries one fresh sample at an anchor point or at the two points of an increment, so any finite batch width $K \ge 2$ suffices. This counts rounds directly, without dividing an evaluation bound by K. Using $J \le 1 + T/q$ and $b \le 1 + H^2 q/(2L^2)$ gives

$$N_{\mathrm{rec}} \le B + BT/q + 2T + TH^2 q/L^2. \tag{199}$$

If $q = T$, then $T \le a$ and $TH/L \le \sqrt{B}$; if $q < T$, then $a/2 \le q \le a$. In both cases

$$N_{\mathrm{rec}} \le 2B + 2T + 3T(H/L)\sqrt{B} = O\left(Q + m + \frac{\Delta H \sigma}{\varepsilon^3}\right).$$

For $H = 0$ the exact count is $B + 2(T - 1) = O(Q + m)$. Choose the cheaper valid recursive or SGD schedule before interaction. When $H > L\sqrt{B}$, the term $(H/L)Q\sqrt{m}$ is larger than Qm, so SGD is already the relevant branch. We obtain

$$O\left(Q + m + \min\left\{Qm, \frac{\Delta H \sigma}{\varepsilon^3}\right\}\right). \tag{200}$$

For $m < 1$, the displayed expression is between Q and $3Q$, and SGD has this order. Thus zero noise and small noise require no exceptional statistical schedule.

Finite sums and exact anchors. For finite sums, replace the centered-noise expectation by uniform sampling of components and set $r = \min\{n, m\}$. If $r < 1$, SGD suffices; if $n = 1$, use exact gradient descent. Otherwise choose

$$B = \min\{n, \lceil 2m \rceil\}, \qquad r \le B \le 3r. \tag{201}$$

When $B = n$, the anchor is exact; when $B < n$, averaging B sampled components gives the required error. For each recursive update, draw fresh component indices and evaluate each selected component at both prescribed points. The preceding error and cost bounds then apply unchanged. The choices (191), (197), and (201) yield $O(Q + r + (\Delta H/\varepsilon^2)\sqrt{r})$ whenever $H \le L\sqrt{B}$. The cheaper of SGD and full-gradient descent independently costs $O(Q + Qr)$. For $H > L\sqrt{B}$, the interaction term is at least Qr. Taking the cheaper available deterministic schedule therefore proves

$$O\left(Q + r + \min\left\{Qr, \frac{\Delta H}{\varepsilon^2}\sqrt{r}\right\}\right). \tag{202}$$

The $r < 1$ and $n = 1$ branches agree with this expression because $Q \ge 1$.

Full sample-gradient smoothness. If $\mathbb{E}\,\|g(x, Z) - g(y, Z)\|^2 \le L_s^2 \|x - y\|^2$, Jensen gives population smoothness at most L_s. The exact variance decomposition is

$$\mathbb{E}\,\|\zeta(x, Z) - \zeta(y, Z)\|^2 = \mathbb{E}\,\|g(x, Z) - g(y, Z)\|^2 - \|\nabla F(x) - \nabla F(y)\|^2 \le L_s^2 \|x - y\|^2. \tag{203}$$

Thus all preceding algorithms apply with population constant $L_* = \min\{L, L_s\}$ and centered constant $H = L_s$. The finite-sum identity is identical, with expectation replaced by $n^{-1}\sum_i$. A samplewise L_s-Lipschitz condition implies the mean-square condition. In particular, with one common smoothness constant the finite-sum rate is $O(Q + r + Q\sqrt{r})$; when $r < 1$, the extra terms are absorbed by Q.

F.3 Additive noise and retained-sample budgets

Assume the common-set identity $g(x,z) = \nabla F(x) + \xi(z)$, with $\mathbb{E}\xi = 0$ and $\|\xi(z)\| \le \sigma$, holds for every x. Use

$$B = \max\{1, \lceil 2m \rceil\}, \qquad T = \lceil 4Q \rceil, \qquad v_0 = B^{-1} \sum_{i=1}^{B} g(x_0, Z_i).$$

Write $e = v_0 - \nabla F(x_0)$, so $\mathbb{E}\,\|e\|^2 \le \varepsilon^2/2$. Take $x_{t+1} = x_t - v_t/L$, and, when another direction is needed, use a fresh sample at the already chosen pair (x_t, x_{t+1}) and set $v_{t+1} = v_t + g(x_{t+1}, Z) - g(x_t, Z)$. Exact cancellation proves $v_t = \nabla F(x_t) + e$ by induction. Smoothness then gives the following inequality for every realization:

$$F(x_{t+1}) \le F(x_t) - \frac{1}{2L}\,\|\nabla F(x_t)\|^2 + \frac{1}{2L}\,\|e\|^2 . \tag{204}$$

The squared-gradient bound for a uniform output is $2L\Delta/T + \mathbb{E}\,\|e\|^2 \le \varepsilon^2$. The number of evaluations is $B + 2(T-1)$ and the number of distinct samples is $B + T - 1$. This proof never assumes that $F + \langle e, \cdot \rangle$ is bounded below.

Under unlimited adaptive retention, retain Z_1 from the anchor and compute every later direction as

$$v_t = v_0 + g(x_t, Z_1) - g(x_0, Z_1).$$

The budget is $B + T - 1$ evaluations and exactly $B = O(1+m)$ distinct samples. If each retained sample may be evaluated at most $J \ge 2$ times, divide the $T-1$ later gradient estimates into blocks of at most $J-1$. At the start of each block, evaluate a new sample at an iterate whose estimate v_t is already known. Each subsequent evaluation of that sample yields another estimate by differencing against its stored anchor response. Therefore

$$N \le B + (T-1) + \left\lceil \frac{T-1}{J-1} \right\rceil, \qquad S_{\text{tokens}} \le B + \left\lceil \frac{T-1}{J-1} \right\rceil . \tag{205}$$

These are $O(1+Q+m)$ evaluations and $O(1+m+Q/J)$ samples uniformly over integer $J \ge 2$. This budget concerns adaptive retention and does not assert a parallel-round bound. In a finite sum, an exact or subsampled anchor from (201) followed by differences of one component gives $O(Q+r)$ evaluations.

Lemma 38 (Additive representation at zero centered regularity)**.** *Under the bounded scalar-loss model, mean-square centered regularity with $H = 0$ implies that, on one probability-one sample set,*

$$g(x;\xi) = \nabla F(x) + Z_\xi \quad \textit{for every } x \in \mathbb{R}^d, \qquad \mathbb{E}Z_\xi = 0, \qquad \|Z_\xi\| \le \sigma.$$

Consequently the retained-sample implementations above apply at this endpoint without any additional continuity assumption on sample gradients.

Proof. For each fixed x, the sample value $f(x;\xi)$ is measurable in ξ as an oracle response. Its continuity in x gives joint measurability: if $q_k(x) = 2^{-k}\lfloor 2^k x \rfloor$ coordinatewise, each $f(q_k(x);\xi)$ is jointly measurable by its countable partition into rational grid cells, and $f(q_k(x);\xi) \to f(x;\xi)$. Moreover

$$\partial_i f(x;\xi) = \lim_{k\to\infty} 2^k \big[f(x + 2^{-k}e_i;\xi) - f(x;\xi) \big]$$

is jointly measurable. This justifies the product integrals below.

Put $n_\xi(x) = f(x;\xi) - F(x)$ and $Z_\xi = \nabla n_\xi(0)$. On the common bounded-noise event, the mean-value theorem applied to $t \mapsto n_\xi(y + t(x-y))$ yields

$$|n_\xi(x) - n_\xi(y)| \le \sup_{0\le t\le 1} \|\nabla n_\xi(y + t(x-y))\|\,\|x-y\| \le \sigma\,\|x-y\| .$$

The condition $H = 0$ gives $\mathbb{E}\,\|\nabla n_\xi(x) - Z_\xi\|^2 = 0$ for every fixed x. For every positive integer k, Tonelli's theorem therefore gives

$$\mathbb{E}\int_{B(0,k)} \|\nabla n_\xi(x) - Z_\xi\|^2\,dx = \int_{B(0,k)} \mathbb{E}\,\|\nabla n_\xi(x) - Z_\xi\|^2\,dx = 0.$$

The integrand is nonnegative. Intersecting the resulting probability-one sets over k proves that $\nabla n_\xi(x) = Z_\xi$ for Lebesgue-almost-every x, simultaneously for almost every ξ.

Fix a sample in this set and write $v(x) = n_\xi(x) - \langle Z_\xi, x\rangle$. This function is 2σ-Lipschitz and has gradient zero almost everywhere. Its restriction to every coordinate line is Lipschitz and hence absolutely continuous. For a smooth compactly supported test function ψ, integration by parts on those lines and Fubini give

$$\int_{\mathbb{R}^d} v(x)\partial_i\psi(x)\,dx = -\int_{\mathbb{R}^d} \partial_i v(x)\psi(x)\,dx = 0.$$

Thus its weak gradient is zero. Let φ_δ be a nonnegative smooth approximate identity supported on $B(0,\delta)$, with integral one, and put $v_\delta = v * \varphi_\delta$. The last identity implies $\nabla v_\delta = 0$, and therefore

$$v_\delta(x) - v_\delta(0) = \int_0^1 \langle \nabla v_\delta(tx), x\rangle\ dt = 0.$$

The Lipschitz bound also gives

$$|v_\delta(x) - v(x)| \le \int \varphi_\delta(y)|v(x-y) - v(x)|\,dy \le 2\sigma\delta.$$

Letting $\delta \downarrow 0$ proves $v(x) = v(0)$ for every x. Differentiating this identity yields the claimed common-set additive representation. Finally, unbiasedness at zero gives $\mathbb{E}Z_\xi = 0$, and the common support bound gives $\|Z_\xi\| \le \sigma$. □

F.4 PL RESTARTS AND THE CONDITION-NUMBER ENDPOINT

For PL objectives, reduce objective-gap accuracy to stationarity by running the preceding algorithms at target squared norm $e^2 = \mu\delta$ whenever the current expected gap is at most δ. Assume globally $\|\nabla F(x)\|^2 \ge 2\mu(F(x) - F^*)$, let $\kappa = L/\mu \ge 1$, and seek $\mathbb{E}[F(\widehat{x}) - F^*] \le \eta$. Their parameters satisfy $Q_\delta = \delta L/e^2 = \kappa$ and $m_\delta = \sigma^2/(\mu\delta)$. Equations (187) and (196) were proved for an expected initial gap, so their uniform output Y satisfies

$$\mathbb{E}\,\|\nabla F(Y)\|^2 \le \mu\delta, \qquad \mathbb{E}[F(Y) - F^*] \le \delta/2. \tag{206}$$

PL then halves the expected gap after each stage. The stage costs are

$$O(\kappa + \kappa m_\delta), \qquad O\left(\kappa + m_\delta + \min\{\kappa m_\delta, (H/\mu)\sqrt{m_\delta}\}\right)$$

for general and centered-smooth noise, respectively. All samples of a new stage are independent of previous stages. Run J stages with target gaps δ_j, where

$$J = \lceil \log_2(\Delta/\eta)\rceil, \qquad \delta_j = 2^{-j}\Delta \quad (0 \le j < J).$$

Induction using (206) then gives a final expected gap of at most η. Put $S = \sigma^2/(\mu\eta)$. Since $\eta < \delta_{J-1} \le 2\eta$,

$$\sum_{j<J} m_{\delta_j} \le 2S, \qquad \sum_{j<J} \sqrt{m_{\delta_j}} \le \frac{\sqrt{S}}{1 - 2^{-1/2}}. \tag{207}$$

Also $\sum_j \min\{a_j, b_j\} \le \min\{\sum_j a_j, \sum_j b_j\}$. The resulting deterministic budgets are

$$O\big(\kappa \log(2\Delta/\eta) + \kappa S\big), \tag{208}$$

$$O\left(\kappa \log(2\Delta/\eta) + S + \min\{\kappa S, (H/\mu)\sqrt{S}\}\right). \tag{209}$$

For $\kappa \ge 3$, these have the claimed deterministic term because $\kappa^{-1} \le \log(\kappa/(\kappa-1)) \le (\kappa-1)^{-1}$.

An optimized minibatch schedule. The restart bound does not capture the faster contraction as $\kappa \downarrow 1$, so we use a separate minibatch schedule in this regime. For $\kappa > 1$, define

$$\tau = 1 - 1/\kappa, \quad N = \left\lceil \frac{\log(2\Delta/\eta)}{\log(1/\tau)} \right\rceil, \quad W_N = \sum_{j=0}^{N-1} \tau^{j/2}, \quad A = \frac{\sigma^2}{L\eta}. \tag{210}$$

At iteration t, average B_t fresh samples at x_t and step by $1/L$, using the predetermined sizes

$$B_t = \max\left\{1, \left\lceil AW_N \tau^{(N-1-t)/2} \right\rceil\right\}. \tag{211}$$

Writing the averaged gradient as $\nabla F(x_t) + e_t$, the same cancellation as in (204) holds regardless of the dependence of e_t on previous errors. With $d_t = \mathbb{E}[F(x_t) - F^*]$, conditional batch variance and PL give

$$d_{t+1} \le \tau d_t + \frac{\sigma^2}{2LB_t}, \qquad d_N \le \tau^N \Delta + \frac{\sigma^2}{2L} \sum_{t=0}^{N-1} \frac{\tau^{N-1-t}}{B_t}. \tag{212}$$

The first term is at most $\eta/2$. When $\sigma > 0$,

$$\sum_{t=0}^{N-1} \frac{\tau^{N-1-t}}{B_t} \le \frac{1}{AW_N} \sum_{t=0}^{N-1} \tau^{(N-1-t)/2} = \frac{1}{A},$$

so the second term is also at most $\eta/2$. When $\sigma = 0$, all batches have size one and the second term vanishes. Since $\max\{1, \lceil u \rceil\} \le 1 + u$ for $u \ge 0$,

$$\sum_{t<N} B_t \le N + AW_N^2 \le N + \kappa S(1 + \sqrt{\tau})^2 \le N + 4\kappa S. \tag{213}$$

Here $W_N \le (1 - \sqrt{\tau})^{-1} = \kappa(1 + \sqrt{\tau})$ and $A = S/\kappa$. If batch sizes may take arbitrary positive real values, this allocation minimizes the total batch size subject to the same variance constraint: Cauchy–Schwarz gives $(\sum_t B_t)(\sum_t \tau^{N-1-t}/B_t) \ge W_N^2$, with equality when B_t is proportional to $\tau^{(N-1-t)/2}$. Integer rounding adds at most N calls, as (213) shows.

For $0 < \eta \le \Delta/128$, define

$$D_1 = 1, \qquad D_\kappa = 1 + \frac{\log(\Delta/\eta)}{\log(\kappa/(\kappa-1))} \quad (\kappa > 1). \tag{214}$$

Because $\log(\Delta/\eta) \ge 7 \log 2$, equation (210) implies $N \le (8/7) D_\kappa$. At $\kappa = 1$, use the single step

$$B = \max\{1, \lceil \sigma^2/(2L\eta) \rceil\}, \qquad \widehat{x} = x_0 - \frac{1}{LB} \sum_{i=1}^{B} g(x_0, Z_i). \tag{215}$$

The contraction coefficient in (212) is zero, so its expected gap is at most $\sigma^2/(2LB) \le \eta$, and $B \le 1 + S/2$. Thus general noise has the upper budget $O(D_\kappa + \kappa S)$ for every $\kappa \ge 1$. For $1 \le \kappa < 3$, the same budget is $O(D_\kappa + S)$, and the centered-noise interaction is at most $3S$. Combining this algorithm with (209) for $\kappa \ge 3$ proves the full bound

$$O\left(D_\kappa + S + \min\{\kappa S, (H/\mu)\sqrt{S}\}\right). \tag{216}$$

At $H = 0$, the budget is $O(D_\kappa + S)$. For large condition numbers, the algorithm uses fresh samples at predetermined pairs of points, so its analysis follows directly from the mean-square condition.

The general-noise method uses one fresh sample per call; the recursive method uses at most two predetermined points per fresh sample. Their budgets therefore also bound fresh-sample rounds, uniformly over finite $K \ge 1$ in the former case and $K \ge 2$ in the latter. For $1 \le \kappa < 3$, the latter upper also uses $K = 1$. Applying (203) gives the formula for smooth sample gradients with $L_* = \min\{L, L_s\}$, $\kappa_* = L_*/\mu$, and $H = L_s$. In the one-constant case it is equivalent to $O(D_\kappa + S + \kappa\sqrt{S})$: for $S < 1$ the extra square-root term is absorbed by $D_\kappa + S$, while for $S \ge 1$ it is the smaller interaction. Every online algorithm in this subsection is also a legal finite-sum algorithm after replacing each fresh sample by a newly drawn uniform component index; the conditional-mean argument after (201) applies verbatim. In particular it proves all the matching finite-sum upper bounds asserted in the regime $n \ge S$. No round statement for active finite sums is inferred from this transfer.

F.5 A SELF-CONTAINED ACCELERATED STOCHASTIC CONVEX ALGORITHM

For the convex and strongly convex upper bounds, we use the following accelerated stochastic algorithm and analyze it through a potential function. Let $\mathcal{X}$ be a closed convex set with available Euclidean projection, and let $x^* \in \mathcal{X}$ minimize a convex L-smooth function on $\mathcal{X}$. The set may be $\mathbb{R}^d$. Start from $y_0 = z_0 = X \in \mathcal{X}$, where $\mathbb{E}\|X - x^*\|^2 \le R_0^2$, and assume each fresh gradient has conditional mean ∇F and conditional centered second moment at most σ^2. For a fixed horizon $T \ge 1$, use a constant $\beta \ge 4L$. Define $a_t = t$, $A_0 = 0$, and $A_t = \sum_{i=1}^t a_i = t(t+1)/2$, and perform, for $t = 1, \dots, T$,

$$x_t = \frac{A_{t-1}y_{t-1} + a_t z_{t-1}}{A_t}, \qquad G_t = g(x_t, Z_t), \tag{217}$$

$$z_t = \mathrm{proj}_{\mathcal{X}}\left(z_{t-1} - \frac{a_t}{\beta} G_t\right), \qquad y_t = \frac{A_{t-1}y_{t-1} + a_t z_t}{A_t}. \tag{218}$$

Return y_T. The sole oracle call in iteration t is at x_t, before z_t and y_t are formed. All iterates lie in $\mathcal{X}$.

Lemma 39 (Accelerated stochastic potential)**.** *For the preceding algorithm, with $d_t = z_t - z_{t-1}$ and $\xi_t = G_t - \nabla F(x_t)$,*

$$\mathbb{E}\left[A_T(F(y_T) - F(x^*)) + \frac{\beta}{2}\|z_T - x^*\|^2\right] \le \frac{\beta}{2}R_0^2 + \frac{\sigma^2}{\beta}\sum_{t=1}^T a_t^2. \tag{219}$$

Consequently

$$\mathbb{E}[F(y_T) - F(x^*)] \le \frac{\beta R_0^2}{T(T+1)} + \frac{\sigma^2(2T+1)}{3\beta}. \tag{220}$$

Proof. Smoothness, $y_t - x_t = (a_t/A_t)d_t$, and convexity at x_t give

$$\begin{aligned} A_t F(y_t) - A_{t-1}F(y_{t-1}) &\le a_t\big(F(x_t) + \langle \nabla F(x_t), z_t - x_t\rangle\big) + \frac{La_t^2}{2A_t}\|d_t\|^2 \\ &\le a_t F(x^*) + a_t\langle \nabla F(x_t), z_t - x^*\rangle + \frac{La_t^2}{2A_t}\|d_t\|^2. \end{aligned}$$

To verify the first line explicitly, convexity implies $F(y_{t-1}) \ge F(x_t) + \langle \nabla F(x_t), y_{t-1} - x_t\rangle$, and $A_{t-1}(y_{t-1} - x_t) = -a_t(z_{t-1} - x_t)$ by (217). Substitution into the smoothness inequality gives the stated expression. The second line is convexity with comparison point x^*.

The projection in (218) minimizes $a_t\langle G_t, z\rangle + \frac{\beta}{2}\|z - z_{t-1}\|^2$ over $\mathcal{X}$. Its first-order optimality condition is $\langle a_t G_t + \beta d_t, x^* - z_t\rangle \ge 0$. The elementary squared-distance identity therefore yields

$$a_t\langle G_t, z_t - x^*\rangle \le \frac{\beta}{2}\big(\|z_{t-1} - x^*\|^2 - \|z_t - x^*\|^2 - \|d_t\|^2\big).$$

Writing $\nabla F(x_t) = G_t - \xi_t$ and combining the last two displays proves the pathwise potential recurrence

$$\begin{aligned} &A_t(F(y_t) - F(x^*)) + \frac{\beta}{2}\|z_t - x^*\|^2 \\ &\quad\le A_{t-1}(F(y_{t-1}) - F(x^*)) + \frac{\beta}{2}\|z_{t-1} - x^*\|^2 - \frac{1}{2}\left(\beta - \frac{La_t^2}{A_t}\right)\|d_t\|^2 \\ &\qquad - a_t\langle \xi_t, z_{t-1} - x^*\rangle - a_t\langle \xi_t, d_t\rangle. \end{aligned} \tag{221}$$

Since $a_t^2/A_t = 2t/(t+1) \le 2$ and $\beta \ge 4L$, the coefficient of $\|d_t\|^2$ is at most $-\beta/4$. Young's inequality gives

$$-a_t\langle \xi_t, d_t\rangle \le \frac{\beta}{4}\|d_t\|^2 + \frac{a_t^2}{\beta}\|\xi_t\|^2.$$

The other noise term in (221) has conditional mean zero: z_{t-1} is fixed before the fresh sample is drawn. Conditional expectation thus bounds the next potential by the current potential plus $a_t^2\sigma^2/\beta$.

Summing from 1 to T telescopes; its initial value is $\frac{\beta}{2}\|X - x^*\|^2$ because $A_0 = 0$. This proves (219). Drop the nonnegative final squared distance and use $\sum_{t=1}^{T} t^2 = T(T+1)(2T+1)/6$ to obtain (220). For a random initial point, we first condition on the past and then average, which gives the same bound. □

For $R_0 > 0$ choose the deterministic tuning

$$\beta = \max\{4L, \sigma T^{3/2}/R_0\}. \tag{222}$$

When $\sigma > 0$, use both $\beta \le 4L + \sigma T^{3/2}/R_0$ and $\beta \ge \sigma T^{3/2}/R_0$ in (220) to obtain

$$\mathbb{E}[F(y_T) - F(x^*)] \le \frac{4LR_0^2}{T^2} + \frac{2\sigma R_0}{\sqrt{T}}. \tag{223}$$

The same display holds for $\sigma = 0$, using $\beta = 4L$. If $R_0 = 0$, $X = x^*$ almost surely, so returning X already solves the problem. For a convex objective with $\|x^* - x_0\| \le R$, take $R_0 = R$ and $\mathcal{X} = \{x : \|x - x_0\| \le R\}$, or simply $\mathcal{X} = \mathbb{R}^d$. The horizon

$$T = \max\left\{1, \left\lceil \max\left\{\sqrt{8LR^2/\eta}, 16\sigma^2R^2/\eta^2\right\}\right\rceil\right\} \tag{224}$$

makes both terms of (223) at most $\eta/2$. It uses exactly T fresh single-point calls and proves $O(1 + \sqrt{LR^2/\eta} + \sigma^2R^2/\eta^2)$. In the theorem's small-accuracy regime the additive one is absorbed by the deterministic term. This completes the smooth convex upper bound.

F.6 Accelerated restart for strong convexity

Suppose F is globally μ-strongly convex and L-smooth, and a stage starts at X with $\mathbb{E}[F(X) - F(x^*)] \le \delta$. Strong convexity implies

$$\mathbb{E}\|X - x^*\|^2 \le 2\delta/\mu. \tag{225}$$

Run (217)–(218) on $\mathbb{R}^d$, using $R_0 = \sqrt{2\delta/\mu}$ in (222). Equation (225) supplies the expected squared-distance bound required by Lemma 39, even when X is not confined to a ball of radius R_0. By (223), the stage output after T calls satisfies

$$\mathbb{E}[F(y_T) - F(x^*)] \le \frac{8\kappa\delta}{T^2} + \frac{2\sigma\sqrt{2\delta/\mu}}{\sqrt{T}}. \tag{226}$$

The fixed choice

$$T_\delta = \left\lceil \max\left\{\sqrt{32\kappa}, 128\sigma^2/(\mu\delta)\right\}\right\rceil \tag{227}$$

makes each term at most $\delta/4$. Thus this stage halves the expected gap and costs $O(\sqrt{\kappa} + \sigma^2/(\mu\delta))$. Restart for $J = \lceil \log_2(\Delta/\eta)\rceil$ stages at nominal gaps $\delta_j = 2^{-j}\Delta$, always with fresh future samples. The expected gap induction is the same as (206). Using (207), the total deterministic budget is

$$\sum_{j<J} T_{\delta_j} \le (1 + \sqrt{32\kappa})J + 256S = O\big(\sqrt{\kappa}\log(2\Delta/\eta) + S\big). \tag{228}$$

In particular, the stochastic term has no additional logarithmic factor.

It remains to identify the uniform deterministic term, including the endpoint. For $\kappa > 1$, write $u = \sqrt{\kappa}$, $b = \log(\Delta/\eta)$, and

$$E_\kappa = 1 + \frac{b}{\log((u+1)/(u-1))}, \qquad A_\kappa = 1 + \frac{b}{\log(u/(u-1))}, \qquad E_1 = 1. \tag{229}$$

The identity

$$\log\frac{u+1}{u-1} = \log\frac{u}{u-1} + \log(1 + 1/u), \qquad 0 < \log(1+1/u) \le \log\frac{u}{u-1}$$

implies $E_\kappa \le A_\kappa \le 2E_\kappa$. For $\kappa \ge 3$, $u^{-1} \le \log(u/(u-1)) \le (u-1)^{-1}$ and $u - 1 \ge (1 - 1/\sqrt{3})u$. Consequently $E_\kappa = \Theta(\sqrt{\kappa}\log(\Delta/\eta))$ uniformly in this range, and (228) is $O(E_\kappa + S)$.

For $1 < \kappa < 3$, put $\Lambda_D = \log(\kappa/(\kappa-1))$ and $\Lambda_A = \log(u/(u-1))$. Then $\Lambda_A = \Lambda_D + \log(1+1/u)$, $\Lambda_D > \log(3/2)$, and $\log(1+1/u) < \log 2 < 2\log(3/2)$. It follows that

$$E_\kappa \le A_\kappa \le D_\kappa \le 3A_\kappa \le 6E_\kappa. \tag{230}$$

Strong convexity implies the global PL inequality: minimizing the quadratic lower model $F(y) \ge F(x) + \langle \nabla F(x), y - x\rangle + \frac{\mu}{2}\|y-x\|^2$ over y gives $F(x) - F^* \le \|\nabla F(x)\|^2/(2\mu)$. Therefore the minibatch method (210)– (211) applies. Its budget $O(D_\kappa + \kappa S)$ is $O(E_\kappa + S)$ by (230) and $\kappa < 3$. At $\kappa = 1$, the one-step method (215) costs at most $1 + S/2$. This proves the upper bound $O(E_\kappa + S)$ for every $\kappa \ge 1$, with a universal constant and no singular endpoint convention in the algorithm.

Since the strongly convex algorithm queries each fresh sample at only one point, its evaluation budget is also a valid fresh-sample round budget for every finite $K \ge 1$. The same upper bound holds with a centered-noise regularity cap. If full sample-gradient smoothness L_s is prescribed, use $L_* = \min\{L, L_s\}$ and $\kappa_* = L_*/\mu$; the rate is $O(E_{\kappa_*} + S)$. Uniform component sampling transfers the same algorithm and budget to active finite sums, including every finite-sum strongly convex regime asserted in the paper.