

Minimax Oracle Call Complexity, Rounds and Admissible Noise in Zeroth-Order Optimization of Highly Smooth Functions

Timofei Loginov
MIRIAI

Darina Dvinskikh
HSE University

Nazarii Tupitsa
MBZUAI

Osman Osmanov
Innopolis University

Yuriy Dorn
AI Institute MSU

Alexander Gasnikov
Innopolis University

Abstract

We study zeroth-order optimization of convex and strongly convex functions with Hölder smoothness of order $\beta > 2$ from values corrupted by random noise with bias independent of the method’s random directions, a Lipschitz systematic error and a bounded adversarial error. We measure the number of function evaluations (calls), of sequential rounds of parallel calls, and the largest tolerable systematic and adversarial errors. For strongly convex functions with a bounded Hessian, we prove a lower bound of order $d^2 \varepsilon^{-\beta/(\beta-1)}$ on the number of calls, where d is the dimension and ε the accuracy, and find the sharp power of d under the stronger Hilbert–Schmidt smoothness for all $\beta > 2$. An accelerated finite-difference method attains them in $O(\sqrt{\kappa} \log(1/\varepsilon))$ rounds for condition number κ , like Nesterov’s method with exact gradients, and tolerates the optimal systematic and, up to logarithms, adversarial errors; rate-optimal methods need $\Omega(\log d)$ rounds. The same number of calls is attained when the Hessian grows at most linearly with the gradient and, for centered noise, without an upper bound on the Hessian, also in logarithmically many rounds. For convex functions, our lower bound improves the known one polynomially in d and is tight at the curvature of the hard instances; the systematic level is again optimal. These bounds extend to general norms and constraints; on the ℓ_1 ball, mirror descent needs calls nearly linear in d . Experiments confirm the theory.

1 INTRODUCTION

Zeroth-order optimization seeks a minimizer using only function evaluations (Conn et al., 2009; Larson et al., 2019). It applies when derivatives are unavailable, as in black-box attacks (Chen et al., 2017) and forward-only fine-tuning of language models (Malladi et al., 2023). We consider the noisy-value problem

$$\min_{x \in \mathbb{R}^d} f(x). \quad (1)$$

Each oracle call returns a corrupted scalar value, whereas the unavailable gradient has d coordinates. Dimension dependence is therefore a central complexity issue. Higher-order smoothness improves the accuracy dependence, but its effect on the sharp power of d has remained unresolved.

For strongly convex objectives with a bounded Hessian and smoothness order $\beta > 2$, the best known upper bound at accuracy ε was $O\left(d^2 \varepsilon^{-\frac{\beta}{\beta-1}}\right)$ oracle calls (Akhavan et al., 2024), whereas separable lower bounds gave $\Omega\left(d^{\frac{\beta}{\beta-1}} \varepsilon^{-\frac{\beta}{\beta-1}}\right)$ (Akhavan et al., 2020, 2024). Their ratio $d^{\frac{\beta-2}{\beta-1}}$ approaches d as β grows. Throughout, asymptotic constants are independent of d, T, ε ; other dependencies are explicit.

We close this gap for strongly convex and convex objectives with Hölder smoothness of order β , measured in the operator norm or in the Hilbert–Schmidt norm. Oracle values may contain centered or direction-oblivious random noise, a fixed Lipschitz systematic error, and a bounded adaptive error.

Related work. Higher-order finite-difference and kernel methods (Polyak and Tsybakov, 1990; Bach and Perchet, 2016; Akhavan et al., 2020; Novitskii and Gasnikov, 2022; Akhavan et al., 2024) attain the d^2 dependence. Matching bounds were previously known

only for $\beta = 3$ under Hilbert–Schmidt smoothness: Yu et al. (2024) obtained $d^{3/2}\varepsilon^{-3/2}$ for centered noise, unrestricted Hessian magnitude, and sufficiently large T . Under operator smoothness, Lobanov (2024) attains the d^2 rate up to logarithms in $O(\sqrt{\kappa}\log(1/\varepsilon))$ rounds, where κ is the condition number. Persistent errors and round lower bounds were studied in Gasnikov et al. (2022); Dvinskikh et al. (2023); Pasechnyuk-Vilensky et al. (2026); Duchi et al. (2018). Tables 2 and 3 give detailed comparisons.

Main result. Let $q = \lceil \beta \rceil - 1$ and set $s = 2$ in operator norm; in Hilbert–Schmidt norm, let $s = \frac{3}{2}$ for even q and $s = 2 - \frac{q-1}{2(\beta-1)}$ for odd q . For the class $\mathcal{F}_{\text{sc}}(\beta, 1; 1, 3, 1)$ defined in Section 2 under i.i.d. $\mathcal{N}(0, 1)$ noise, denote the Gaussian minimax expected suboptimality after T oracle calls by $\mathfrak{R}_T^G$. Then

$$\mathfrak{R}_T^G = \Theta\left(\min\left\{1, \left(\frac{d^s}{T}\right)^{\frac{\beta-1}{\beta}}\right\}\right), \quad d \geq 2, T \geq 1.$$

Consequently, operator smoothness requires $\Theta(d^2\varepsilon^{-\frac{\beta}{\beta-1}})$ oracle calls: higher-order smoothness improves the accuracy exponent but not the sharp d^2 factor. For Hilbert–Schmidt smoothness, $s = \frac{3}{2}$ for even q and $\frac{3}{2} < s \leq \frac{5}{3}$ for odd q .

Contributions.

- *Oracle calls.* Random-frame higher-order differences attain the factors d^2 under operator smoothness and d^s under Hilbert–Schmidt smoothness; dense hard instances give matching lower bounds.
- *Rounds and persistent errors.* A restarted Nesterov scheme attains the rate in $O(\sqrt{\kappa}\log(1/\varepsilon))$ rounds, and rate-optimal methods require $\Theta(\log d)$ rounds. For α -strong convexity, the sharp systematic-error level is $\Theta(\sqrt{\alpha\varepsilon})$; the adaptive threshold is determined up to a factor $\min\{\sqrt{d}, \sqrt{\log(4T)}\}$.
- *Broader curvature classes.* The same rate holds under gradient-dependent Hessian growth. With centered noise and unrestricted Hessian magnitude, it holds once T exceeds a polynomial in d .
- *Convexity and geometry.* We improve the convex lower bounds and match central differences in a curvature regime. An image-ball reduction extends the lower bounds to general constraints; on the ℓ_1 ball, the dimension dependence is linear up to $\log^2(2d)$.

Organization. Section 2 defines the model; Sections 3–5 develop the strongly convex theory; Sections 6 and 7 cover convexity and geometry; and Section 8 reports experiments. The supplement contains proofs, extensions, experimental details, computational scope, and detailed comparisons.

2 PROBLEM SETUP

This section defines the function classes, noisy value oracle, and complexity and robustness criteria for problem (1).

Notation. Fix $\beta > 2$ and set $q = \lceil \beta \rceil - 1 \geq 2$. For operator-norm smoothness let $\tau = 0$; for Hilbert–Schmidt smoothness let $\tau = \tau_\beta$, where

$$\tau_\beta = \frac{\beta-1}{2} \text{ for even } q, \quad \tau_\beta = \frac{q-1}{2} \text{ for odd } q. \quad (2)$$

In both cases, set $s = 2 - \frac{\tau}{\beta-1}$. Constants $c_\beta, C_\beta > 0$ depend only on β and are fixed within each formal statement. We use $\log^+ x = \max\{0, \log x\}$ and $\log_2^+ x = \log^+ x / \log 2$.

Smoothness and classes. Let $\gamma = \beta - q \in (0, 1]$. For the symmetric q -linear form $D^q f(x)$, set $\|A\|_{\text{op}} = \sup_{\|v_i\| \leq 1} |A[v_1, \dots, v_q]|$ and let $\|A\|_{\text{HS}}$ be the Euclidean norm of its coefficient array. The seminorm $[f]_{\beta, \text{op}}$ is the least L_β such that $\|D^q f(x) - D^q f(y)\|_{\text{op}} \leq L_\beta \|x - y\|^\gamma$; define $[f]_{\beta, \text{HS}}$ analogously. Since $\|A\|_{\text{op}} \leq \|A\|_{\text{HS}}$, the Hilbert–Schmidt assumption is stronger; $[f]_\beta$ denotes either seminorm.

For $0 < \alpha \leq L \leq \infty$ and $R > 0$, the class $\mathcal{F}_{\text{sc}} = \mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)$ consists of the q -times continuously differentiable f with $\alpha I \preceq \nabla^2 f \preceq LI$, $[f]_\beta \leq L_\beta$ and unique minimizer x^* in $B(R) = \{x : \|x\| \leq R\}$. Here $L = \infty$ allows arbitrary Hessian norms; $\kappa = L/\alpha$ is used when $L < \infty$. The class $\mathcal{F}_{\text{cvx}} = \mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$ is defined in the same way with $0 \preceq \nabla^2 f \preceq LI$ and some minimizer in $B(R)$. Methods start at the origin, so R bounds the initial distance to a minimizer; it does not constrain the queries or the output. These may lie anywhere in $\mathbb{R}^d$, except in Section 7.

Oracle calls, rounds, and noisy values. Before round k , the past and the randomness $\Theta_1, \dots, \Theta_{k-1}$ generate $\mathcal{F}_{k-1}$. The method draws fresh Θ_k and selects all queries of that round jointly from $(\mathcal{F}_{k-1}, \Theta_k)$; hence rounds measure adaptive depth. It makes T oracle calls in N rounds (at most n per round in Theorem 2), and its output may use the full transcript.

Definition 1 (Noisy values). Let $\mathcal{I}_k$ index round k . A query at X_t , $t \in \mathcal{I}_k$, returns $Y_t = f(X_t) + \eta(X_t) + \xi_t^c + \xi_t^o + \delta_t$. Let $\mathcal{E}_t$ be the σ -field generated by $\mathcal{F}_{k-1}$, Θ_k , the queries of round k , and the errors of its earlier queries. For levels $M, \sigma, \delta \geq 0$,

$$|\eta(x) - \eta(y)| \leq M \|x - y\| \quad \text{for all } x, y; \quad (L)$$

$$\mathbb{E}[\xi_t^c | \mathcal{E}_t] = 0, \quad \mathbb{E}[(\xi_t^c)^2 | \mathcal{E}_t] \leq \sigma^2; \quad (C)$$

$$\mathbb{E}[(\xi_t^o)^2 | \mathcal{F}_{k-1}] \leq \sigma^2, \quad (\xi_t^o)_{t \in \mathcal{I}_k} \perp\!\!\!\perp \Theta_k | \mathcal{F}_{k-1}; \quad (O)$$

$$|\delta_t| \leq \delta. \quad (A)$$

Table 1: Leading dependence on d , ε , and κ for strongly convex objectives.

Setting	Reference	Oracle calls lower	Oracle calls upper	Rounds lower	Rounds upper
$\alpha I \preceq \nabla^2 f \preceq LI$, $[f]_{\beta, \text{op}} \leq L_\beta$	Akhavan et al. (2024) Lobanov (2024) This paper (Thms. 1, 2)	$d^{\frac{\beta}{\beta-1}} \varepsilon^{-\frac{\beta}{\beta-1}}$ — $d^2 \varepsilon^{-\frac{\beta}{\beta-1}}$	$d^2 \varepsilon^{-\frac{\beta}{\beta-1}}$ $d^2 \varepsilon^{-\frac{\beta}{\beta-1}} *$ $d^2 \varepsilon^{-\frac{\beta}{\beta-1}}$	— — $\log d^\S$	— $\sqrt{\kappa} \log \frac{1}{\varepsilon}$ $\sqrt{\kappa} \log \frac{1}{\varepsilon}$
$\alpha I \preceq \nabla^2 f \preceq LI$, $[f]_{\beta, \text{HS}} \leq L_\beta$	Akhavan et al. (2024) This paper (Thms. 1, 2) [¶]	— $d^s \varepsilon^{-\frac{\beta}{\beta-1}}$	$d^2 \varepsilon^{-\frac{\beta}{\beta-1}}$ $d^s \varepsilon^{-\frac{\beta}{\beta-1}}$	— $\log d^\S$	— $\sqrt{\kappa} \log \frac{1}{\varepsilon}$
$\ \nabla^2 f\ _{\text{op}} \leq L_0 + L_1 \ \nabla f\ $	This paper (Sec. 3) [¶]	$d^s \varepsilon^{-\frac{\beta}{\beta-1}}$	$d^s \varepsilon^{-\frac{\beta}{\beta-1}}$	—	$\kappa_{\text{eff}} \log \frac{1}{\varepsilon}^\ddagger$
$\alpha I \preceq \nabla^2 f$, centered noise	Yu et al. (2024), $\beta = 3$, HS [†] This paper (Thm. 3) ^{¶¶}	$d^{\frac{3}{2}} \varepsilon^{-\frac{3}{2}}$ $d^s \varepsilon^{-\frac{\beta}{\beta-1}}$	$d^{\frac{3}{2}} \varepsilon^{-\frac{3}{2}}$ $d^s \varepsilon^{-\frac{\beta}{\beta-1}}$	— —	— $\log \frac{d}{\varepsilon}$

*Up to logarithmic factors. [†]Hilbert–Schmidt seminorm; asymptotic in T . [‡] $\kappa_{\text{eff}} = L_{\text{eff}}/\alpha$ (Section 3). [§]Rate-optimal methods, fixed d , $L \geq 3\alpha$. [¶] $s = 2$ for $[f]_{\beta, \text{op}}$; for $[f]_{\beta, \text{HS}}$, $s = \frac{3}{2}$ if q is even and $s = 2 - \frac{q-1}{2(\beta-1)}$ if q is odd. ^{¶¶}For T above a polynomial in d .

In (O), $\perp\!\!\!\perp$ denotes conditional independence given $\mathcal{F}_{k-1}$. Here η is fixed before the run, and δ_t may depend on the past, on Θ_k , and on the query. The noise is then *admissible at levels* (M, σ, δ) .

Thus η is a fixed systematic error; ξ^c is centered; ξ^o may be biased but is direction-oblivious within each round (Akhavan et al., 2020); and δ_t is adaptive but bounded. *Gaussian noise* means independent $\xi_t^c \sim \mathcal{N}(0, \sigma^2)$ and no other component. Methods know the class parameters and σ , but not M, δ .

Criteria and rates. A method is ε -accurate on a class at levels (M, σ, δ) if $\mathbb{E}[f(\hat{x}) - f^*] \leq \varepsilon$ for every f in the class and every noise admissible at these levels, where $f^* = \min_{x \in \mathbb{R}^d} f(x)$. The *admissible levels* $M_{\max}(\varepsilon, T)$ and $\delta_{\max}(\varepsilon, T)$ are the suprema of the M (with $\delta = 0$) and of the δ (with $M = 0$) for which some method with at most T oracle calls is ε -accurate. For strongly convex functions and $\sigma > 0$, the rate Φ_T and the radius h_T that balances the bias and the noise of the gradient estimate of Section 4 are

$$\Phi_T = \frac{L_\beta^{\frac{2}{\beta}}}{\alpha} d^{2-\frac{2+\tau}{\beta}} \left(\frac{\sigma^2}{T}\right)^{\frac{\beta-1}{\beta}}, \quad h_T = \left(\frac{d^{2+\tau} \sigma^2}{L_\beta^2 T}\right)^{\frac{1}{2\beta}}, \quad (3)$$

and $\Phi_1 := \Phi_T|_{T=1}$. Solving $\Phi_T = \varepsilon$ gives the *required number of oracle calls*

$$T_*(\varepsilon) = d^s \sigma^2 L_\beta^{\frac{2}{\beta-1}} (\alpha \varepsilon)^{-\frac{\beta}{\beta-1}}. \quad (4)$$

A gradient error of norm b causes function-value error at the scale $\frac{b^2}{\alpha}$ under strong convexity, so accuracy ε requires gradients with accuracy $\sqrt{\alpha \varepsilon}$. The radius at which the bias of the estimate equals $\sqrt{\alpha \varepsilon}$, and the

adversarial level with the same effect, are

$$h_\varepsilon = \left(\frac{\alpha \varepsilon d^\tau}{L_\beta^2}\right)^{\frac{1}{2(\beta-1)}}, \delta_\varepsilon = \frac{\sqrt{\alpha \varepsilon} h_\varepsilon}{\sqrt{d}} = \sigma d^{\frac{1}{2}} T_*(\varepsilon)^{-\frac{1}{2}}. \quad (5)$$

The last equality requires $\sigma > 0$. The scale δ_ε equals the standard-error scale associated with $T_*(\varepsilon)/d$ centered observations; the oblivious component of Definition 1 may remain dependent across oracle calls.

3 MAIN RESULTS

This section gives oracle-call, round, and persistent-error bounds for strongly convex objectives under bounded, gradient-dependent, and unrestricted curvature.

Theorem 1 (Oracle calls, rounds and admissible noise). *Let $\beta > 2$, $0 < \alpha < L < \infty$, $\kappa = \frac{L}{\alpha}$ and $\sigma > 0$, and let $\mathcal{F}_{\text{sc}}$ be defined with either seminorm. There are constants $0 < c_\beta \leq C_\beta$ such that the following holds.*

(a) (Upper bound.) *Let $T \geq C_\beta d \sqrt{\kappa} \log_2(C_\beta d \kappa (1 + \frac{LR^2}{\Phi_1}))$. The accelerated method of Procedure 1 with radius h_T and target $C_\beta \Phi_T$ uses at most T oracle calls in $N \leq (1 + 5\sqrt{\kappa})(1 + \log_2^+ \frac{LR^2}{\Phi_T})$ rounds, and for every $f \in \mathcal{F}_{\text{sc}}$ and every noise admissible at levels (M, σ, δ)*

$$\mathbb{E}[f(\hat{x}) - f^*] \leq C_\beta \left(\Phi_T + \frac{M^2}{\alpha} + \frac{T \delta^2}{d \sigma^2} \Phi_T \right). \quad (6)$$

(b) (Lower bound.) *Let $\rho = \min\{\alpha, \frac{L-\alpha}{2}\}$ and $\Phi_0 = \min\{\alpha R^2, \rho^{\frac{2(\beta-1)}{\beta-2}} \alpha^{-1} L_\beta^{-\frac{2}{\beta-2}}\}$. If $M \leq \alpha R$, then for every $T \geq 1$ every method with T oracle calls in any number of rounds satisfies $\sup_{f, \eta} \mathbb{E}[f(\hat{x}) -$*

Table 2: Leading oracle-call dependence on d , ε , and L for convex objectives.

Setting	Reference	Oracle calls lower bound	Oracle calls upper bound
$\nabla^2 f \preceq LI$, $[f]_{\beta, \text{op}} \leq L_\beta$	Akhavan et al. (2020, 2024)	$\min\{d^2 \varepsilon^{-2}, d^{\frac{\beta}{\beta-1}} \varepsilon^{-\frac{2\beta}{\beta-1}}\}^\parallel$	$d^2 \varepsilon^{-\frac{2\beta}{\beta-1}}$
	Bach and Perchet (2016)	—	$d^2 \varepsilon^{-\frac{2\beta}{\beta-1}}$
	Gasnikov et al. (2024a)	—	$d^2 L \varepsilon^{-3}$
	Lattimore and György (2023), Fokkema et al. (2024)	—	$d^p \varepsilon^{-2}$ for $p \in [3, \frac{7}{2}]^{*\dagger}$
	This paper (Thms. 4, 5)	$\min\{d^{\frac{5}{2}} \varepsilon^{-2}, d^2 \varepsilon^{-\frac{2(\beta+1)}{\beta}}, d^2 L \varepsilon^{-3}\}$	$d^2 \min\{\varepsilon^{-\frac{2\beta}{\beta-1}}, L \varepsilon^{-3}\}$
	This paper , $L \lesssim \varepsilon^{1-\frac{2}{\beta}} \ddagger$ (Sec. 6)	$d^2 L \varepsilon^{-3}$	$d^2 L \varepsilon^{-3}$
$\nabla^2 f \preceq LI$, $[f]_{\beta, \text{HS}} \leq L_\beta$	Bach and Perchet (2016)	—	$d^2 \varepsilon^{-\frac{2\beta}{\beta-1}}$
	This paper , q even (Sec. 6)	$d^2 \varepsilon^{-2} \S$	$d^{\frac{3}{2}} \varepsilon^{-\frac{2\beta}{\beta-1}}$
	This paper , q odd (Sec. 6)	$d^2 \varepsilon^{-2} \S$	$d^{2-\frac{q-1}{2(\beta-1)}} \varepsilon^{-\frac{2\beta}{\beta-1}}$
Constraints with an n -dimensional image ball	This paper (Sec. 7)	$\min\{n^{\frac{5}{2}} \varepsilon^{-2}, n^2 \varepsilon^{-\frac{2(\beta+1)}{\beta}}\}$	—
ℓ_1 ball, $[f]_{\beta, 1} \leq L_\beta$	This paper (Sec. 7)	$d \varepsilon^{-\frac{2(\beta+1)}{\beta}} \P$	$d \log^2(2d) \varepsilon^{-\frac{2\beta}{\beta-1}} *$

$a \lesssim b$ and $a \gtrsim b$ mean $a \leq Cb$ and $a \geq cb$ with $C, c > 0$ independent of d, ε and L . *Centered sub-Gaussian noise. $\dagger$ Up to logarithmic factors. $\ddagger$ For $d^{-\frac{1}{2}} L \lesssim \varepsilon \lesssim \min\{1, L\}$. $\S$ For $\varepsilon \lesssim d^{-\frac{1}{2} \lfloor \frac{q}{2} \rfloor}$, $L \gtrsim 1$. $\P$ For $d^{-\frac{3\beta}{4}} \lesssim \varepsilon \lesssim d^{-\frac{\beta}{2}}$. $\parallel$ For $L \gtrsim \varepsilon^{1-\frac{2}{\beta}}$.

$f^*] \geq c_\beta(\min\{\Phi_T, \Phi_0\} + \frac{M^2}{\alpha})$ under Gaussian noise, where f ranges over $\mathcal{F}_{\text{sc}}$ and η over the M -Lipschitz systematic errors.

(c) (Admissible levels.) For every $\varepsilon \in (0, LR^2)$, this method with target $\frac{\varepsilon}{2}$ and radius $h = c_\beta h_\varepsilon$ is ε -accurate with $C_\beta[T_\star(\varepsilon) + d\sqrt{\kappa}(1 + \log \frac{LR^2}{\varepsilon})]$ oracle calls in at most $C_\beta\sqrt{\kappa}(1 + \log \frac{LR^2}{\varepsilon})$ rounds whenever $M \leq c_\beta\sqrt{\alpha\varepsilon}$ and $\delta \leq c_\beta\delta_\varepsilon$. Conversely, $M_{\max}(\varepsilon, T) \leq \sqrt{2\alpha\varepsilon}$ for all T and $\varepsilon < \frac{\alpha R^2}{2}$, and $\delta_{\max}(\varepsilon, T) \leq C_\beta \min\{\sqrt{d}, \sqrt{\log(4T)}\} \delta_\varepsilon$ for all T and $\varepsilon \leq \varepsilon_0$, where $\varepsilon_0 > 0$ depends on the class parameters and d , but not on T .

When the bounds match. Let $\delta = 0$ and $M \leq \alpha R$; the latter is needed, since the output 0 has error at most $\frac{LR^2}{2}$. As $\Phi_T \leq \Phi_0$ if and only if $T \geq T_\star(\Phi_0)$, (a) and (b) give the minimax error $\Theta(\Phi_T + \frac{M^2}{\alpha})$ over $\mathcal{F}_{\text{sc}}$ and the noises admissible at levels $(M, \sigma, 0)$, jointly in all parameters, for $T \geq \max\{T_\star(\Phi_0), C_\beta d\sqrt{\kappa} \log_2(C_\beta d\kappa(1 + \frac{LR^2}{\Phi_1}))\}$, where $T_\star(\Phi_0) = \Theta(d^s)$ for fixed other parameters. Part (b) also applies below this threshold; for $L \geq 3\alpha$, Theorem 2(b) gives the corresponding small-budget lower bound for exact values. With $\alpha = 1$, $L = 3$, $R = L_\beta = \sigma = 1$ and $M = \delta = 0$ the minimax error is $\Theta(\min\{1, (\frac{d^s}{T})^{\frac{\beta-1}{\beta}}\})$, for all $d \geq 2$ and $T \geq 1$ (Lemma E.5).

Admissible errors. Part (c) compares errors at the statistically optimal budget $T = \Theta(T_\star(\varepsilon))$: the systematic level is sharp in every dimension and, for $\varepsilon \leq \varepsilon_0$, the adaptive level is characterized within $\min\{\sqrt{d}, \sqrt{\log(4T)}\}$. For fixed d , once $T \geq T_\star(\varepsilon)$ and $\log(4T) \geq d$, the same bounds give the factor $\sqrt{d}$. Both admissible levels also hold under two-point feedback. The adversarial converse extends to the high-probability criterion of Pasechnyuk-Vilensky et al. (2026). The two-phase method of Section 4 attains (6) in $O(\kappa(1 + \log^+ \frac{\alpha R^2}{\Phi_T}))$ rounds if $d\kappa$ replaces $d\sqrt{\kappa}$ in (a).

Rounds. Let $\mathfrak{R}_{N,n}(\mathcal{C}; \sigma)$ be the minimax risk of methods with N rounds of at most n oracle calls each, under Gaussian noise of variance σ^2 or exact values ($\sigma = 0$).

Theorem 2 (Lower bounds on the number of rounds). Let $\beta > 2$, $L_\beta, \alpha, R > 0$, $L \geq 3\alpha$, let $\mathcal{F}_{\text{sc}} = \mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)$ be defined with either seminorm, and let $N, n \geq 1$.

- (a) If $\sigma > 0$, then $\mathfrak{R}_{1,n}(\mathcal{F}_{\text{sc}}; \sigma) \geq c_1 n^{-\frac{2(\beta-1)}{2\beta+d}}$ with $c_1 > 0$ independent of n .
- (b) For every $\sigma \geq 0$, $\mathfrak{R}_{N,n}(\mathcal{F}_{\text{sc}}; \sigma) \geq c_N n^{-\nu_N}$ with $\nu_N = \frac{2(\beta-1)((\beta-1)^N - 1)}{d(\beta-2)}$, where $c_N > 0$ does not depend on n .
- (c) Let $\sigma > 0$ and fix d and all parameters except n . A method with N rounds of at most n oracle calls whose worst-case error on $\mathcal{F}_{\text{sc}}$ is $O(n^{-\frac{\beta-1}{\beta}} \log^k(n +$

Table 3: Admissible persistent-error scales.

Setting	Reference	Oblivious	$M_{\max}$	$\delta_{\max}$
$\alpha I \preceq \nabla^2 f \preceq LI$, $[f]_{\beta, \text{op}} \leq L_\beta$	Akhavan et al. (2020, 2024), Lobanov (2024) Gasnikov et al. (2024b) This paper (Thm. 1)	✓ — ✓	— — $\asymp \varepsilon^{\frac{1}{2}}$	— $\gtrsim \varepsilon^{\frac{\beta}{2(\beta-1)}} d^{-\frac{\beta}{\beta-1}} *$ $\gtrsim \varepsilon^{\frac{\beta}{2(\beta-1)}} d^{-\frac{1}{2}},$ $\lesssim \varepsilon^{\frac{\beta}{2(\beta-1)}} \sqrt{\frac{\log(4T)}{d}}$
$\alpha I \preceq \nabla^2 f \preceq LI$, $[f]_{\beta, \text{HS}} \leq L_\beta$	This paper (Thm. 1)	✓	$\asymp \varepsilon^{\frac{1}{2}}$	$\gtrsim \varepsilon^{\frac{\beta}{2(\beta-1)}} d^{-\frac{1}{4} \dagger},$ $\lesssim \varepsilon^{\frac{\beta}{2(\beta-1)}} d^{-\frac{1}{4}} \sqrt{\log(4T)} \dagger$
$\alpha I \preceq \nabla^2 f$, $[f]_{\beta, \text{op}} \leq L_\beta$, $\ \nabla^2 f\ _{\text{op}} \leq L_0 + L_1 \ \nabla f\ $	This paper (Sec. 3)	✓	$\asymp \varepsilon^{\frac{1}{2}}$	$\gtrsim \varepsilon^{\frac{\beta}{2(\beta-1)}} d^{-\frac{1}{2}},$ $\lesssim \varepsilon^{\frac{\beta}{2(\beta-1)}} \sqrt{\frac{\log(4T)}{d}}$
$\alpha I \preceq \nabla^2 f$, $[f]_{\beta, \text{op}} \leq L_\beta$	Yu et al. (2024), $\beta = 3$, $[f]_{\beta, \text{HS}} \leq L_\beta$ This paper (Thms. 1, 3)	✗ ✗	— $\lesssim \varepsilon^{\frac{1}{2}}$	— $\lesssim \varepsilon^{\frac{\beta}{2(\beta-1)}} \sqrt{\frac{\log(4T)}{d}}$
$0 \preceq \nabla^2 f \preceq LI$, $[f]_{\beta, \text{op}} \leq L_\beta$	Akhavan et al. (2020, 2024) Dvinskikh et al. (2023), $\sigma = 0$ Pasechnyuk-Vilensky et al. (2026), $\sigma = 0$ This paper (Sec. 6)	✓ — — ✓	— $\gtrsim \varepsilon d^{-\frac{1}{2}}$ — $\asymp \varepsilon$	— $\gtrsim \varepsilon^{\frac{3}{2}} d^{-\frac{1}{2} \dagger}$ $\gtrsim \varepsilon^{\min\{\frac{\beta}{\beta-1}, \frac{3}{2}\}} d^{-\frac{1}{2}},$ $\lesssim \varepsilon \sqrt{\frac{\log(4T)}{d}}$
ℓ_1 ball, $[f]_{\beta, 1} \leq L_\beta$	This paper (Sec. 7)	✗	$\gtrsim \varepsilon$	$\gtrsim \varepsilon^{\frac{\beta}{\beta-1}}$

$\lesssim, \gtrsim, \asymp$: up to constants independent of d, ε, T , for small ε and large T . ✓: oblivious noise allowed; ✗: centered noise assumed; —: not considered. *Fixed errors; Polyak–Łojasiewicz functions. †Fixed errors and queries in a ball. ‡For even q ; for odd q , $d^{-\frac{1}{2} + \frac{q-1}{4(\beta-1)}}$ instead.

1)) for some k (rate-optimal) has $N \geq N_\beta(d) := \max\{2, \lceil \log_{\beta-1}(1 + \frac{(\beta-2)d}{2\beta}) \rceil\}$.

The hard functions satisfy $\alpha I \preceq \nabla^2 f \preceq 3\alpha I$. Part (c) concerns fixed d and growing oracle calls per round; see Section Q on its constants and on finite budgets.

Gradient-dependent curvature. The gradient-dependent condition below controls the initial region while allowing unrestricted global curvature: $\|\nabla^2 f(x)\|_{\text{op}} \leq L_0 + L_1 \|\nabla f(x)\|$ (Zhang et al., 2020). It implies $\nabla^2 f \preceq L_{\text{eff}} I$ on $B(R)$ for $L_{\text{eff}} = \min\{L_0 e^{2L_1 R}, C_\beta(L_0(1 + L_1 R)^{q-2} + L_\beta R^{\beta-2})\}$, and above an explicit budget the projected two-phase method with L_{eff} in place of L satisfies (6) in $O(\kappa_{\text{eff}}(1 + \log^+ \frac{\alpha R^2}{\Phi_T}))$ rounds, $\kappa_{\text{eff}} = \frac{L_{\text{eff}}}{\alpha}$; the lower bounds hold with $L = L_0$ if $L_0 > \alpha$. The explicit budget and proof appear in Lemma I.1.

For the full class $\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, \infty, R)$, let $\mathfrak{R}_T(\mathcal{C}; \sigma)$ be the minimax error over $\mathcal{C}$ of methods with at most T oracle calls in any number of rounds, under centered noise (C) with $\eta = \xi^o = \delta_t = 0$.

Theorem 3 (Minimax rates under unrestricted curvature). *Let $\beta > 2$, $\alpha, L_\beta, R, \sigma > 0$, let $\mathcal{C} =$*

$\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, \infty, R)$ be defined with either seminorm, and set $K_q = \binom{d+q}{q}$.

- (a) *There is $C_0 = C_0(\beta, \alpha, L_\beta, R, \sigma)$ such that, if $T \geq C_0 d^2$ for $2 < \beta \leq 4$, or $T \geq C_0 d^q$ for $\beta > 4$, then $c_\beta \Phi_T \leq \mathfrak{R}_T(\mathcal{C}; \sigma) \leq C_\beta \Phi_T$, and the upper bound is attained in at most $C_\beta T/d$ rounds.*
- (b) *There are $C_1 = C_1(\beta, \alpha, L_\beta, R, \sigma)$ and $C_2 = C_2(\beta, \alpha, L_\beta, R)$ such that a batched method attains $\mathfrak{R}_T(\mathcal{C}; \sigma) \leq C_\beta \Phi_T$ in at most $C_2(1 + \log T)$ rounds whenever $T \geq C_1 K_q^{2+q/\gamma} \log(2 + d + T)$.*

For integer β , the threshold in (b) is at most a β -dependent constant times $d^{\beta^2-1} \log(2 + d + T)$.

Both methods assume centered noise and first fit the unrestricted low-order polynomial component of f ; every pilot oracle call is included in T . Part (a) uses $1 + d$ pilot rounds for $3 < \beta \leq 4$. For $\beta > 4$, and for $\beta > 3$ in part (b), global polynomial minimization is treated as exact computation in the oracle model. For $\beta = 3$ and Hilbert–Schmidt smoothness, (a) recovers the $d^{\frac{3}{2}} \varepsilon^{-\frac{3}{2}}$ oracle-call complexity of Yu et al. (2024) on their class and noise for all $T \geq C_0 d^2$. The converse bounds of Theorem 1(c) hold here with $L = 3\alpha$, since the hard

instances have $\nabla^2 f \preceq 3\alpha I$.

4 GRADIENT ESTIMATION

This section isolates the random-frame gradient estimate used by a standard strongly convex optimizer.

Higher-order finite differences. Let $m = \lceil \frac{q}{2} \rceil$, $t_j = \frac{j}{m}$ for $j \leq m$, and let $a_1, \dots, a_m$ solve $\sum_{j=1}^m a_j t_j^{2i+1} = \mathbf{1}\{i = 0\}$, $i = 0, \dots, m-1$. Then $\frac{1}{2h} \sum_j a_j (\varphi(ht_j) - \varphi(-ht_j)) = \varphi'(0)$ for every polynomial φ of degree at most $2m \geq q$. For $m = 1$ this is the central difference, and for $m = 2$ the weights are $(\frac{8}{3}, -\frac{1}{3})$. The constants $S_1 = \sum_j |a_j|$, $S_2 = \sum_j a_j^2$, $A_1 = \sum_j |a_j| t_j$ and $C_\sigma = (S_1 + (\frac{S_2}{2})^{\frac{1}{2}})^2$ measure the amplification of value errors. A *frame* is a uniformly random orthonormal basis of $\mathbb{R}^d$; differences along random directions are standard (Polyak and Tsybakov, 1990; Nesterov and Spokoiny, 2017; Gasnikov et al., 2024a).

The update in the second block below is Nesterov’s standard constant-momentum scheme (Nesterov, 2004, Section 2.2). The wrapper adds restarts, growing batches, and calls to the random-frame estimate.

Procedure 1 Random-frame finite differences with restarted Nesterov descent

Require: radius $h > 0$, target $\varepsilon > 0$; α, L, R, σ ; $\kappa = \frac{L}{\alpha}$
 Random-frame estimate $\bar{g}_{h,B}(z)$, one round of $2mdB$ oracle calls:

- 1: Draw B independent frames $(U_i^{(\ell)})_{i=1}^d$
 - 2: In one round query all $z \pm ht_j U_i^{(\ell)}$, receive $Y_{ij\ell}^\pm$
 - 3: $\hat{g}^{(\ell)} = \sum_{i=1}^d \frac{1}{2h} \sum_{j=1}^m a_j (Y_{ij\ell}^+ - Y_{ij\ell}^-) U_i^{(\ell)}$
 - 4: **return** $\bar{g} = \frac{1}{B} \sum_{\ell=1}^B \hat{g}^{(\ell)}$
 Restarted Nesterov scheme, J stages of N_0 rounds:
 - 5: $N_0 = \lceil 5\sqrt{\kappa} \rceil$, $\theta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, $x^{(0)} = 0$
 - 6: $\bar{V} = \frac{(L+\alpha)R^2}{2}$, $J = \lceil \log_2 \frac{\bar{V}}{\varepsilon} \rceil$
 - 7: **for** $j = 1, \dots, J$ **do**
 - 8: $B_j = \max\{1, \lceil \frac{12\sqrt{\kappa} C_\sigma d \sigma^2 2^j}{L h^2 \bar{V}} \rceil\}$
 - 9: $x_{-1} = x_0 = x^{(j-1)}$
 - 10: **for** $k = 0, \dots, N_0 - 1$ **do** $\triangleright$ one round each
 - 11: $y_k = x_k + \theta(x_k - x_{k-1})$, $x_{k+1} = y_k - L^{-1} \bar{g}_{h,B_j}(y_k)$
 - 12: $x^{(j)} = x_{N_0}$
 - 13: **return** $\hat{x} = x^{(J)}$
-

Proposition 1 (Random-frame error decomposition). *Let $[f]_\beta \leq L_\beta$ and $d \geq 2$, and let the center z and the radius h of round k be $\mathcal{F}_{k-1}$ -measurable. For every noise admissible at levels (M, σ, δ) , the random-frame estimate in Procedure 1 is $\bar{g} = \nabla f(z) + b + \zeta + \Delta$, where*

b is $\mathcal{F}_{k-1}$ -measurable, $\mathbb{E}[\zeta \mid \mathcal{F}_{k-1}] = 0$, and

$$\|b\| \leq C_\beta L_\beta d^{-\frac{\tau}{2}} h^{\beta-1} + A_1 M, \quad \|\Delta\| \leq \frac{\sqrt{d} S_1 \delta}{h},$$

$$\mathbb{E}[\|\zeta\|^2 \mid \mathcal{F}_{k-1}] \leq \frac{C_\beta L_\beta^2 h^{2\beta-2}}{B d^\tau} + \frac{4A_1^2 M^2}{B} + \frac{C_\sigma d \sigma^2}{B h^2},$$

The estimate relies only on higher-order smoothness, independently of the Hessian magnitude.

The oblivious noise multiplies each direction $U_i^{(\ell)}$ by a coefficient independent of it, and a uniformly random direction has mean zero. A systematic error is a frame sum with an odd $A_1 M$ -Lipschitz coefficient function on the sphere, so by the Poincaré inequality its second moment is at most $\frac{d}{d-1} A_1^2 M^2$. If those coefficients depended on the sampled frame, the shift could scale as $\sqrt{d} M$; a bounded adaptive error can align all d differences.

Descent. The second block of Procedure 1 restarts Nesterov’s constant-momentum scheme. Each stage contracts the error by a constant factor, while the growing batch keeps the estimation variance at the new target scale. Unlike general inexact oracles (Devolder et al., 2014; d’Aspremont, 2008; Cohen et al., 2018), a fixed gradient bias of norm b contributes only an absolute constant times $\frac{b^2}{\alpha}$. At radius h_T , smoothing bias and centered noise each contribute Φ_T ; systematic and adaptive errors contribute $\frac{M^2}{\alpha}$ and $\frac{d\sigma^2}{\alpha h_T^2}$, respectively.

At h_ε , the bias is $\sqrt{\alpha\varepsilon}$, which gives Theorem 1(c). The *two-phase* method (Procedure 2) takes projected steps $\frac{1}{4L}$ and then decreasing steps with averaging, and needs the Hessian bound only on $B(R)$.

Dimension dependence. Under Hilbert–Schmidt smoothness, ball averaging contracts the tensors that determine the finite-difference bias. A k -fold trace of $D^q f$ has Hölder constant at most $d^{\frac{k}{2}} L_\beta$ under the Hilbert–Schmidt bound but enters with the weight $\frac{h^{2k}}{d^k}$: for odd $q = 2k + 1$ this gains $d^{-\frac{k}{2}} = d^{-\frac{\tau}{2}}$, and for even $q = 2k$, where the k -fold trace is only γ -Hölder, one more difference at the radius $h d^{-\frac{1}{4}}$ gains $d^{-\frac{\beta-1}{4}} = d^{-\frac{\tau}{2}}$. The hard functions mirror this: a radial cutoff raises the Hilbert–Schmidt seminorm by $d^{\frac{1}{2} \lfloor \frac{\tau}{2} \rfloor}$, which matches odd q , and even q needs a ridge of width $h d^{-\frac{1}{4}}$. The same parity dependence appears in the matching hard instances.

5 LOWER BOUNDS

This section gives the proof ideas behind the oracle-call, persistent-error, and round lower bounds.

Every hard instance is a quadratic plus a smooth perturbation supported on a ball, with a cutoff $\chi = 1$ on

$[0, \frac{1}{4}]$ and $\chi = 0$ on $[1, \infty)$. A perturbation of slope a and radius h has seminorm scale $ah^{1-\beta}$ and Hessian scale a/h ; hence $a = c_\beta L_\beta h^{\beta-1}$, while $L - \alpha$ bounds h from above.

Oracle calls. For $\alpha < \mu \leq 2\alpha$ and $\omega \in \{\pm 1\}^d$, define the hard functions $f_\omega(x) = \frac{\mu}{2} \|x\|^2 + \frac{a}{\sqrt{d}} \langle \omega, x \rangle \chi(\frac{\|x\|^2}{h^2})$ for Theorem 1(b). Each coordinate of the minimizer $-\frac{a\omega}{\mu\sqrt{d}}$ carries one sign, and a wrong sign costs $\frac{a^2}{8\alpha d}$. Flipping ω_i changes a value by at most $\frac{2a|x_i|}{\sqrt{d}}$, and the squares of these changes sum over i to at most $\frac{4a^2h^2}{d}$ at every query. By Assouad's lemma a constant fraction of the signs are wrong unless $Ta^2h^2 \geq c_\beta d^2\sigma^2$; this spreading of the signal over all coordinates produces the factor d^2 . Balancing Ta^2h^2 with $d^2\sigma^2$ gives the error Φ_T . For small T this radius violates the curvature or the radius constraint, and the bound saturates at Φ_0 . Under Hilbert–Schmidt smoothness the cutoff, or for even q a ridge, changes the power of d (Section 4).

Persistent errors. A Lipschitz error can be observationally indistinguishable from a change of the objective: for any unit vector v , the instances $f_\pm(x) = \frac{\alpha}{2} \|x \mp \frac{M}{\alpha} v\|^2$ with $\eta_\pm(x) = \pm M \langle v, x \rangle$ give the same values, and one of them has error at least $\frac{M^2}{2\alpha}$ whatever the number of oracle calls. For the adversarial converse, a perturbation of slope of order $\sqrt{\alpha\varepsilon}$ and radius of order h_ε lies along a uniformly random unit vector v . With $\theta = \Theta(\min\{1, \sqrt{\frac{\log(4T)}{d}}\})$, an adversary at a level δ of order $\theta\sqrt{d}\delta_\varepsilon$ returns the unperturbed value outside the cone $\{|\langle v, x \rangle| > \theta \|x\|\}$, which the T queries of a run miss with probability at least $\frac{1}{2}$; then v and $-v$ cannot be told apart, as for convex Lipschitz functions in Pasechnyuk-Vilensky et al. (2026).

Rounds. For Theorem 2(a), the queries of one round hit a perturbation of radius r at a random center in $B(\frac{R}{2})$ with probability at most $(\frac{2r}{R})^d$ each. For (b), the minimizer is hidden behind $N + 1$ perturbations with nested random centers, as in Dorn et al. (2026, Appendix E) for exact values: with probability at least $\frac{3}{4}$ the queries of round k miss the ball where f differs from the sum of the first k perturbations, and the recurrence has scale $r_{k+1} = c_\beta n^{-\frac{1}{d}} r_k^{\beta-1}$, which gives ν_N , and (c) follows by comparing ν_N with $\frac{\beta-1}{\beta}$.

6 CONVEX FUNCTIONS

This section gives convex upper and lower bounds and identifies a curvature regime in which central differences are minimax.

Without strong convexity a gradient error of norm

b costs bR in value. Set $V = L$, or $V = \min\{L, L_\beta(2R)^{\beta-2}\}$ for operator smoothness when $2 < \beta \leq 3$. At the *gradient scale* $h = \Theta((\frac{\varepsilon d^{\frac{\beta}{2}}}{L_\beta R})^{\frac{1}{\beta-1}})$, the finite-difference bias is at most $\frac{\varepsilon}{R}$. At the *value scale* $h = \Theta((\frac{\varepsilon}{V})^{\frac{1}{2}})$, central differences estimate the gradient of $f_h(x) = \mathbb{E}f(x + hW)$, $W \sim \text{Unif}(B(1))$. Here only the oscillation of $f_h - f$ matters, and $|\text{tr} \nabla^2 f(x) - \text{tr} \nabla^2 f(y)| \leq dV$ controls it.

Theorem 4 (Convex upper bounds). *Let $\beta > 2$, $0 < L < \infty$, $f \in \mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$, and let the noise be admissible at levels (M, σ, δ) . An accelerated projected method on $B(R)$ with the estimate of Procedure 1 at radius h outputs $\hat{x} \in B(R)$ with $\mathbb{E}[f(\hat{x}) - f^*] \leq \varepsilon$ in $\lceil (\frac{32LR^2}{\varepsilon})^{\frac{1}{2}} \rceil$ rounds, as follows.*

- (a) (Gradient scale.) If $h = c_\beta(\frac{\varepsilon d^{\frac{\beta}{2}}}{L_\beta R})^{\frac{1}{\beta-1}}$, $M \leq c_\beta \frac{\varepsilon}{R}$ and $\delta \leq c_\beta \frac{\varepsilon h}{R\sqrt{d}}$, it uses $T \leq C_\beta \frac{\sigma^2}{\varepsilon^2} d^{2-\frac{\beta}{\beta-1}} (\frac{L_\beta R^\beta}{\varepsilon})^{\frac{2}{\beta-1}} + C_\beta d(1 + \frac{LR^2}{\varepsilon})$ oracle calls.
- (b) (Value scale.) With central differences ($m = 1$, $a_1 = t_1 = 1$) for every β , if $h = (\frac{\varepsilon}{8V})^{\frac{1}{2}}$, $M \leq \frac{\varepsilon}{16R}$, $\delta \leq \frac{\varepsilon h}{16R\sqrt{d}}$, and $\varepsilon \leq 8L_\beta(2R)^\beta$ when $V < L$, it uses $T \leq C \frac{d^2 \sigma^2 V R^2}{\varepsilon^3} + 32d(1 + \frac{LR^2}{\varepsilon})$ oracle calls, with an absolute constant C .

Part (a) has the leading term of Bach and Perchet (2016); (b) with $V = L$ is classical smoothing (Flaxman et al., 2005; Saha and Tewari, 2011) and uses central differences because ball averages preserve convexity. Both hold on a compact convex set Q , with a minimizer possibly on its boundary and the diameter of Q in place of $2R$ where it enters h , V , M and δ .

Theorem 5 (Convex lower bounds). *Let $\beta > 2$ and $0 < L \leq \infty$.*

- (a) *Let $d \geq 256$, $0 < \varepsilon \leq c'_\beta \min\{L_\beta R^\beta, LR^2\}$, and let the noise be Gaussian. Every method that is ε -accurate on $\mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$ with the operator seminorm, in any number of rounds, must use*

$$T \geq c_\beta \frac{\sigma^2}{\varepsilon^2} \min\left\{d^{\frac{5}{2}}, d^2 \left(\frac{L_\beta R^\beta}{\varepsilon}\right)^{\frac{2}{\beta}}, \frac{d^2 LR^2}{\varepsilon}\right\}.$$

- (b) *If $\varepsilon < \frac{LR^2}{2}$ and $M > \frac{2\varepsilon}{R}$, no method is ε -accurate on $\mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$. For every T covered by Theorem 4(b), $\frac{\varepsilon}{16R} \leq M_{\max}(\varepsilon, T) \leq \frac{2\varepsilon}{R}$.*

Part (a) extends Rajaraman and Han (2026) to Hölder derivatives of every order. The hard instances have curvature $L_\varepsilon = C_\beta L_\beta^{2/\beta} \varepsilon^{1-2/\beta}$; flattening them when $L < L_\varepsilon$ yields the last term. For $L \geq L_\varepsilon$ and $L_\beta R^\beta d^{-\beta/4} \leq \varepsilon \leq c'_\beta L_\beta R^\beta$, the lower bound improves the separable bound (Akhavan et al., 2020, 2024) by

at least $d^{(\beta-2)/\beta}$. At higher accuracy its leading term is $d^{5/2}\sigma^2\varepsilon^{-2}$, a factor $d^{1/2}$ larger than the previous bound. Under Hilbert–Schmidt smoothness a $d^2\sigma^2\varepsilon^{-2}$ lower bound holds in the regime of Lemma E.6.

For d above the threshold in Section Q and $L \leq C_\beta L_\varepsilon$, central differences attain the minimax complexity $d^2\sigma^2LR^2\varepsilon^{-3}$ over the stated accuracy range without requiring β or L_β (Lemma L.10(a)).

7 NORMS AND CONSTRAINTS

This section extends the convex lower bounds to general norms, ℓ_p and ℓ_1 geometries, and functional constraints.

Let f be convex with a minimizer in $\mathcal{K}$, and let $\|\cdot\|_{\mathcal{B}}$ be a norm on $\mathbb{R}^d$ with unit ball $\mathcal{B}$. Define $[f]_{\beta,\mathcal{B}}$ by replacing the Euclidean norm in $[f]_{\beta,\text{op}}$ for $x - y$ and the arguments of $D^q f$ by $\|\cdot\|_{\mathcal{B}}$.

Definition 2 (Image ball). A triple (P, z, r) with a linear $P : \mathbb{R}^d \rightarrow \mathbb{R}^n$, $z \in \mathbb{R}^n$ and $r > 0$ is an *image ball* of $\mathcal{K}$ if $\|Pu\| \leq \|u\|_{\mathcal{B}}$ for all u and $z + rB_2^n \subseteq P(\mathcal{K})$, with B_2^n the Euclidean unit ball.

For a Euclidean hard function g on $\mathbb{R}^n$, $f = g \circ P$ is convex, has $[f]_{\beta,\mathcal{B}} \leq [g]_{\beta,\text{op}}$ and a minimizer in $\mathcal{K}$, and a method for f induces a method for g with identical oracle calls, rounds, and noise. Thus the Euclidean lower bounds transfer to $\mathcal{K}$, as Dorn et al. (2026) showed for exact values.

Corollary 1 (Any norm and known constraints). *Let (P, z, r) be an image ball of $\mathcal{K}$ in $\mathbb{R}^n$, and let the noise be Gaussian. On convex functions with $[f]_{\beta,\mathcal{B}} \leq L_\beta$ and a minimizer in $\mathcal{K}$, Theorem 5(a) with $L = \infty$ holds with (n, r) in place of (d, R) , and every ε -accurate method needs $c_\beta n^2 \sigma^2 \varepsilon^{-2}$ oracle calls whenever $\varepsilon \leq c'_\beta L_\beta r^\beta$, for every $n \geq 1$. This remains true for restricted queries and outputs and for known functional constraints.*

For the ℓ_p ball $\mathcal{K} = B_p(R)$, $1 \leq p < 2$, and $[f]_{\beta,p} \leq L_\beta$ ($[f]_{\beta,\mathcal{B}}$ for the ℓ_p norm), coordinate projections are image balls with $r = Rn^{\frac{1}{2}-\frac{1}{p}}$, and the bound becomes $T \geq c_\beta \frac{\sigma^2}{\varepsilon^2} \min\{n^{\frac{5}{2}}, n^{3-\frac{2}{p}}(\frac{L_\beta R^\beta}{\varepsilon})^{\frac{2}{\beta}}\}$ for $256 \leq n \leq d$ and $\varepsilon \leq c'_\beta L_\beta r^\beta$. Let f be convex with $[f]_{\beta,1} \leq L_\beta$, a minimizer in $B_1(R)$ and $\|\nabla f\|_\infty \leq G_\infty$ there. Under centered, conditionally sub-Gaussian noise, entropic mirror descent (Nemirovski and Yudin, 1983; Beck and Teboulle, 2003) with coordinate finite differences and independently resampled signs is ε -accurate with $T \leq C_\beta d \log(2d) \varepsilon^{-2} (G_\infty^2 R^2 + \sigma^2 \log(2d) (\frac{L_\beta R^\beta}{\varepsilon})^{\frac{2}{\beta-1}})$ oracle calls for $M \leq c_\beta \frac{\varepsilon}{R}$ and $\delta \leq c_\beta \varepsilon (\frac{\varepsilon}{L_\beta R^\beta})^{\frac{1}{\beta-1}}$. If $G_\infty \geq C_\beta L_\beta R^{\beta-1} d^{-\frac{\beta-2}{2}}$, the noise is Gaussian and $L_\beta R^\beta d^{-\frac{3\beta}{4}} \leq \varepsilon \leq c'_\beta L_\beta R^\beta d^{-\frac{\beta}{2}}$, every method needs $c_\beta d (\frac{L_\beta R^\beta}{\varepsilon})^{\frac{2}{\beta}} \sigma^2 \varepsilon^{-2}$ oracle calls. The bounds are linear

in d up to $\log^2(2d)$, with accuracy exponents $2\beta/(\beta-1)$ and $2(\beta+1)/\beta$; the interval is nonempty above the threshold in Section Q.

For functional constraints, stochastic Mirror–Prox (Juditsky et al., 2011) with random-frame finite differences controls objective error and constraint violation under calibrated values, bounded multipliers, and a Lipschitz primal–dual operator. Each oracle call returns the objective and n_c constraint values; Section O gives the details.

8 EXPERIMENTS

Setup. We use smooth strongly convex test functions with known minimizers. Panels (a)–(d) evaluate: admissible systematic and adaptive errors at $d = 20$ and $T = 1.024 \cdot 10^6$; round complexity on quadratic–log–cosh objectives with $\kappa = 10, 100$ and $T = 1,310,720$; oracle calls needed for $\varepsilon = 2 \cdot 10^{-4}$ over $8 \leq d \leq 64$ (and $d = 128$ for the separable case); and central, fourth-order frame, and kernel estimators at equal budgets in $d = 16$. Noise is i.i.d. Gaussian with $\sigma = 0.03$ (0.1 in panel (b)). Each point averages at least 20 runs, and 40 for $d \leq 16$ in panel (c).

Results. Panel (a) scales M and δ by the thresholds in Theorem 1. The error doubles at normalized levels 1.07 and 1.30, close to the predictions 1.11 and 1.37. At fixed T , panel (b) comes within 3.5 times the statistical level in 16 and 32 accelerated rounds, whereas two-phase descent takes 256 and 2048 rounds to come within 1.7 times that level. Panel (c) tests (4): the matched operator and Hilbert–Schmidt designs give dimension exponents 2.03 ± 0.03 and 1.62 ± 0.03 , close to 2.00 and 1.59. At equal oracle budgets, panel (d) gives slopes near $T^{-4/5}$ for the higher-order estimates and $T^{-2/3}$ for central differences. Section P provides full protocols, five additional experiments, and code.

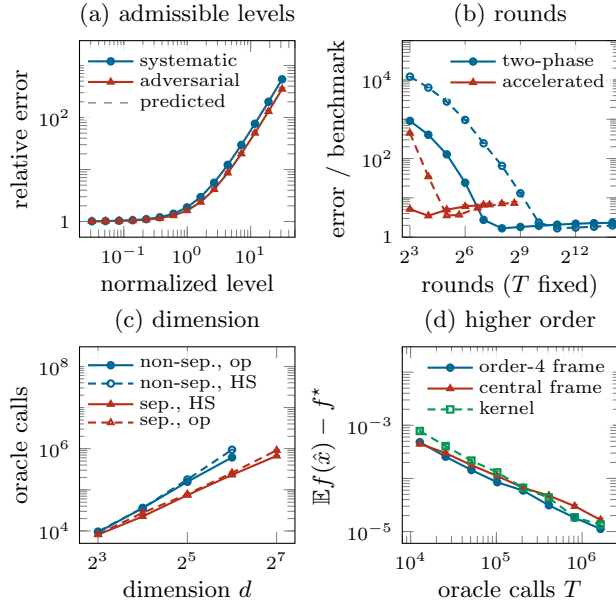


Figure 1: Checks of (a) admissible errors, (b) rounds, (c) dimension scaling, and (d) higher-order estimators. Bars: one standard error; op/HS denote the smoothness norm.

AI USE STATEMENT

Generative AI tools were used to improve the clarity and presentation of the manuscript and to help with the implementation and checking of the numerical experiments. The authors reviewed all AI-assisted changes, independently verified the reported results, and take full responsibility for the final content.

References

- Vadim Abronin, Alexander Gasnikov, and Darina Dvinskikh. Gradient-free methods for stochastic convex optimization with stochastic functional constraints, 2026. arXiv:2609.35946v2.
- Arya Akhavan and Alexandre B. Tsybakov. Gradient-free stochastic optimization for additive models. *Automation and Remote Control*, 86(9):817–830, 2025. URL <https://arxiv.org/abs/2503.02131>.
- Arya Akhavan, Massimiliano Pontil, and Alexandre B. Tsybakov. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. In *Advances in Neural Information Processing Systems*, volume 33, pages 9017–9027, 2020. URL https://papers.nips.cc/paper_files/paper/2020/hash/6646b06b90bd13dabc11ddba01270d23-Abstract.html.
- Arya Akhavan, Evgenii Chzhen, Massimiliano Pontil, and Alexandre B. Tsybakov. A gradient estimator via L1-randomization for online zero-order optimization with two point feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 7685–7696, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/329ef22fd8cb68223d5df09a037f7dd9-Abstract-Conference.html.
- Arya Akhavan, Evgenii Chzhen, Massimiliano Pontil, and Alexandre B. Tsybakov. Gradient-free optimization of highly smooth functions: improved analysis and a new algorithm. *Journal of Machine Learning Research*, 25(370):1–50, 2024. URL <https://jmlr.org/papers/v25/23-0733.html>.
- Arya Akhavan, Sirine Louati, and Alexandre B. Tsybakov. Gradient-free stochastic optimization of derivatives under strong convexity, 2026. arXiv:2607.07249.
- Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker’s Guide*. Springer, third edition, 2006. doi: 10.1007/3-540-29587-9.
- Francis Bach and Vianney Perchet. Highly-smooth zero-th order online optimization. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 257–283. PMLR, 2016. URL <https://proceedings.mlr.press/v49/bach16.html>.
- Keith Ball. An elementary introduction to modern convex geometry. In Silvio Levy, editor, *Flavors of Geometry*, volume 31 of *Mathematical Sciences Research Institute Publications*, pages 1–58. Cambridge University Press, Cambridge, 1997. URL <https://library.slmath.org/books/Book31/files/ball.pdf>.
- Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. doi: 10.1016/S0167-6377(02)00231-6.
- Alexandre Belloni, Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 240–265. PMLR, 2015. URL <https://proceedings.mlr.press/v40/Belloni15.html>.
- Peter Borwein and Tamás Erdélyi. *Polynomials and Polynomial Inequalities*, volume 161 of *Graduate Texts in Mathematics*. Springer, New York, 1995. doi: 10.1007/978-1-4612-0793-1.
- Alexandra Carpentier, Chloé Rouyer, Alexandre Tsybakov, and Arya Akhavan. Sharp optimal algorithm for derivative-free stochastic convex optimization in one dimension, 2026. arXiv:2607.12938.
- Hung Chen. Lower rate of convergence for locating a maximum of a function. *The Annals of Statistics*, 16(3):1330–1334, 1988. doi: 10.1214/aos/1176350965.
- Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017. doi: 10.1145/3128572.3140448.
- Michael B. Cohen, Jelena Diakonikolas, and Lorenzo Orecchia. On acceleration with noise-corrupted gradients. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1019–1028. PMLR, 2018. URL <https://proceedings.mlr.press/v80/cohen18a.html>.
- Andrew R. Conn, Katya Scheinberg, and Luis N. Vicente. *Introduction to Derivative-Free Optimization*, volume 8 of *MPS-SIAM Series on Optimization*.

SIAM, Philadelphia, 2009. doi: 10.1137/1.9780898718768.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

Feng Dai and Yuan Xu. *Approximation Theory and Harmonic Analysis on Spheres and Balls*. Springer Monographs in Mathematics. Springer, New York, 2013. doi: 10.1007/978-1-4614-6660-4. URL <https://arxiv.org/abs/1304.2585>. Chapter 1 is available as arXiv:1304.2585.

Alexandre d’Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008. doi: 10.1137/060676386.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1–2):37–75, 2014. doi: 10.1007/s10107-013-0677-5.

Jelena Diakonikolas and Cristóbal Guzmán. Lower bounds for parallel and randomized convex optimization. *Journal of Machine Learning Research*, 21(5):1–31, 2020. URL <https://jmlr.org/papers/v21/19-771.html>. Conference version: COLT 2019.

Jürgen Dippon. Accelerated randomized stochastic optimization. *The Annals of Statistics*, 31(4):1260–1281, 2003. doi: 10.1214/aos/1059655913.

Yuriy Dorn, Darina Dvinskikh, Timofei Loginov, Osman Osmanov, Aleksandr Shestakov, Nazarii Tupitsa, and Alexander Gasnikov. Lower bounds for gradient-free convex optimization and convex–concave saddle-point problems, 2026. arXiv:2610.00345.

John Duchi, Feng Ruan, and Chulhee Yun. Minimax bounds on stochastic batched convex optimization. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 3065–3162. PMLR, 2018. URL <https://proceedings.mlr.press/v75/duchi18a.html>.

John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015. doi: 10.1109/TIT.2015.2409256.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5 edition, 2019.

Darina Dvinskikh, Vladislav Tominin, Yaroslav Tominin, and Alexander Gasnikov. Gradient-free optimization for non-smooth saddle point problems under adversarial noise, 2023. arXiv:2202.06114v3.

Václav Fabian. Stochastic approximation of minima with improved asymptotic speed. *The Annals of Mathematical Statistics*, 38(1):191–200, 1967. doi: 10.1214/aoms/1177699070.

Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394, 2005. URL <https://www.microsoft.com/en-us/research/publication/online-convex-optimization-bandit-setting-gradient-descent-without-gradient/>.

Hidde Fokkema, Dirk van der Hoeven, Tor Lattimore, and Jack J. Mayo. Online Newton method for bandit convex optimisation: Extended abstract. In *Proceedings of the Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1713–1714. PMLR, 2024. URL <https://proceedings.mlr.press/v247/fokkema24a.html>. Extended abstract; full version: arXiv:2406.06506.

Alexander Gasnikov, Anton Novitskii, Vasilii Novitskii, Farshed Abdukhakimov, Dmitry Kamzolov, Aleksandr Beznosikov, Martin Takáč, Pavel Dvurechensky, and Bin Gu. The power of first-order smooth optimization for black-box non-smooth problems. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7241–7265. PMLR, 2022. URL <https://proceedings.mlr.press/v162/gasniko22a.html>. Extended version: arXiv:2201.12289.

Alexander Gasnikov, Darina Dvinskikh, Pavel Dvurechensky, Eduard Gorbunov, Aleksandr Beznosikov, and Aleksandr Lobanov. Randomized gradient-free methods in convex optimization. In Panos M. Pardalos and Oleg A. Prokopyev, editors, *Encyclopedia of Optimization*, pages 1–15. Springer, 2024a. doi: 10.1007/978-3-030-54621-2_859-1. arXiv:2211.13566.

Alexander V. Gasnikov, Pavel E. Dvurechensky, and Yurii E. Nesterov. Stochastic gradient methods with inexact oracle. *Trudy MIPT*, 8(1(29)):41–91, 2016a. In Russian. arXiv:1411.4218.

Alexander V. Gasnikov, Anastasia A. Lagunovskaya, Ilnura N. Usmanova, and Fedor A. Fedorenko. Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex. *Automation and Remote Control*, 77(11):2018–2034, 2016b. doi: 10.1134/S000511791610114.

- Alexander V. Gasnikov, Ekaterina A. Krymova, Anastasia A. Lagunovskaya, Il'nura N. Usmanova, and Fedor A. Fedorenko. Stochastic online optimization. Single-point and multi-point non-linear multi-armed bandits. Convex and strongly-convex case. *Automation and Remote Control*, 78(2):224–234, 2017. doi: 10.1134/S0005117917020035.
- Alexander V. Gasnikov, Aleksandr V. Lobanov, and Fedor S. Stonyakin. Highly smooth zeroth-order methods for solving optimization problems under the PL condition. *Computational Mathematics and Mathematical Physics*, 64(4):739–770, 2024b. doi: 10.1134/S0965542524700118. arXiv:2305.15828.
- Walter Gautschi. Some elementary inequalities relating to the gamma and incomplete gamma function. *Journal of Mathematics and Physics*, 38(1–4):77–81, 1959. doi: 10.1002/sapm195938177.
- 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, 2013. doi: 10.1137/110848876.
- Georges Glaeser. Racine carrée d’une fonction différentiable. *Annales de l’Institut Fourier*, 13(2): 203–210, 1963. doi: 10.5802/aif.146. URL https://www.numdam.org/item/AIF_1963__13_2_203_0/.
- Eduard Gorbunov, Pavel Dvurechensky, and Alexander Gasnikov. An accelerated method for derivative-free smooth stochastic convex optimization. *SIAM Journal on Optimization*, 32(2):1210–1238, 2022. doi: 10.1137/19M1259225.
- Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2.
- Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016. URL <https://jmlr.org/papers/v17/14-273.html>.
- Kevin G. Jamieson, Robert D. Nowak, and Benjamin Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*, volume 25, pages 2672–2680, 2012. URL <https://proceedings.neurips.cc/paper/2012/hash/e6d8545daa42d5ced125a4bf747b3688-Abstract.html>.
- Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011. doi: 10.1287/10-SSY011.
- J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952. doi: 10.1214/aoms/1177729392.
- David Kozak, Cesare Molinari, Lorenzo Rosasco, Luis Tenorio, and Silvia Villa. Zeroth-order optimization with orthogonal random directions. *Mathematical Programming*, 199:1179–1219, 2023. doi: 10.1007/s10107-022-01866-9.
- Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method, 2012. arXiv:1212.2002.
- Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1–2):365–397, 2012. doi: 10.1007/s10107-010-0434-y.
- Jeffrey Larson, Matt Menickelly, and Stefan M. Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019. doi: 10.1017/S0962492919000060.
- Tor Lattimore. Bandit convex optimisation, 2024. arXiv:2402.06535.
- Tor Lattimore and András György. A second-order method for stochastic bandit convex optimisation. In *Proceedings of the Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2067–2094. PMLR, 2023. URL <https://proceedings.mlr.press/v195/lattimore23a.html>.
- Yuanzhi Li and Andrej Risteski. Algorithms and matching lower bounds for approximately-convex optimization. In *Advances in Neural Information Processing Systems*, volume 29, pages 4745–4753, 2016. URL <https://proceedings.neurips.cc/paper/2016/hash/186fb23a33995d91ce3c2212189178c8-Abstract.html>.
- Aleksandr Lobanov. Improved iteration complexity in black-box optimization problems under higher order smoothness function condition, 2024. arXiv:2407.03507.

- Aleksandr Lobanov and Alexander Gasnikov. Power of generalized smoothness in stochastic convex optimization: First- and zero-order algorithms. arXiv:2501.18198, 2025.
- Aleksandr Lobanov, Nail Bashirov, and Alexander Gasnikov. The “black-box” optimization problem: Zero-order accelerated stochastic method via kernel approximation. *Journal of Optimization Theory and Applications*, 203(3):2451–2486, 2024. doi: 10.1007/s10957-024-02556-6. arXiv:2310.02371.
- Timofei Loginov, Alexander Gasnikov, Yuriy Dorn, Aleksandr Shestakov, Nazarii Tupitsa, Osman Osmanov, and Darina Dvinskikh. The complexity of convex optimization with mismatched geometry, 2026. arXiv:2609.32910.
- Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In *Advances in Neural Information Processing Systems*, volume 36, pages 53038–53075, 2023. doi: 10.52202/075280-2308.
- Edward James McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40(12):837–842, 1934.
- Alexander Mikhalev and Ivan V. Oseledets. Rectangular maximum-volume submatrices and their applications. *Linear Algebra and its Applications*, 538: 187–211, 2018. doi: 10.1016/j.laa.2017.10.014. URL <https://arxiv.org/abs/1502.07838>.
- Arkadi Nemirovski. On parallel complexity of nonsmooth convex optimization. *Journal of Complexity*, 10(4):451–463, 1994. doi: 10.1006/jcom.1994.1025.
- Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. doi: 10.1137/070704277.
- Arkadi S. Nemirovski and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. Wiley, Chichester and New York, 1983. URL https://www.isye.gatech.edu/~nemirovs/Nemirovski_i_Yudin_1983.pdf. Translated from the Russian by E. R. Dawson.
- Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, 2004. doi: 10.1007/978-1-4419-8853-9.
- Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, 186(1–2):157–183, 2021. doi: 10.1007/s10107-019-01449-1. URL <https://link.springer.com/article/10.1007/s10107-019-01449-1>.
- Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017. doi: 10.1007/s10208-015-9296-2.
- Vasilii Novitskii and Alexander Gasnikov. Improved exploitation of higher order smoothness in derivative-free optimization. *Optimization Letters*, 16(7):2059–2071, 2022. doi: 10.1007/s11590-022-01863-z. URL <https://link.springer.com/article/10.1007/s11590-022-01863-z>.
- Jeffrey S. Ovall. The Laplacian and mean and extreme values. *The American Mathematical Monthly*, 123(3): 287–291, 2016. doi: 10.4169/amer.math.monthly.123.3.287.
- Dmitry Pasechnyuk-Vilensky, Igor Pavlov, Martin Takáč, and Alexander Gasnikov. Information-theoretic upper bounds for deterministic noise in zeroth-order convex optimization, 2026. arXiv:2306.16371v5.
- Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. doi: 10.1137/0330046.
- Boris T. Polyak and Alexandre B. Tsybakov. Optimal order of accuracy of search algorithms in stochastic optimization. *Problems of Information Transmission*, 26(2):126–133, 1990. URL <https://www.mathnet.ru/eng/ppi605>. Russian original: Problemy Peredachi Informatsii 26(2):45–53.
- Nived Rajaraman and Yanjun Han. The price of hidden curvature: improved lower bounds for bandit convex optimization, 2026. arXiv:2607.18652v3.
- Marco Rando, Cesare Molinari, Silvia Villa, and Lorenzo Rosasco. Stochastic zeroth order descent with structured directions. *Computational Optimization and Applications*, 89:691–727, 2024. doi: 10.1007/s10589-024-00616-1.
- R. Tyrrell Rockafellar. *Convex Analysis*, volume 28 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1970.
- Ankan Saha and Ambuj Tewari. Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and*

Statistics, volume 15 of *Proceedings of Machine Learning Research*, pages 636–642. PMLR, 2011. URL <http://proceedings.mlr.press/v15/saha11a.html>.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 3–24. PMLR, 2013. URL <https://proceedings.mlr.press/v30/Shamir13.html>.

Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52): 1–11, 2017. URL <https://jmlr.org/papers/v18/16-632.html>.

Yaron Singer and Jan Vondrák. Information-theoretic lower bounds for convex optimization with erroneous oracles. In *Advances in Neural Information Processing Systems*, volume 28, pages 3204–3212, 2015.

Antonela Toma. Generating functions for the mean value of a function on a sphere and its associated ball in $\mathbb{R}^n$. *Journal of Inequalities and Applications*, 2008:656329, 2008. doi: 10.1155/2008/656329. URL <https://link.springer.com/article/10.1155/2008/656329>. Article 656329, 14 pages.

Joel A. Tropp. Freedman’s inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011. doi: 10.1214/ECP.v16-1624. URL <https://tropp.caltech.edu/papers/Tro11-Freedmans-Inequality.pdf>.

Paul Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000. doi: 10.1137/S0363012998338806. URL <https://pubs.siam.org/doi/10.1137/S0363012998338806>.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. doi: 10.1007/b13794. URL <https://link.springer.com/book/10.1007/b13794>.

Yining Wang, Sivaraman Balakrishnan, and Aarti Singh. Optimization of smooth functions with noisy observations: Local minimax rates. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Qian Yu, Yining Wang, Baihe Huang, Qi Lei, and Jason D. Lee. Stochastic zeroth-order optimization under strongly convexity and Lipschitz Hessian: Minimax sample complexity. In *Advances in Neural Information Processing Systems*, volume 37, pages 99564–99600, 2024. doi: 10.52202/079017-3159. URL https://papers.nips.cc/paper_files/paper/2024/hash/b444ad72520a5f5c467343be88e352ed-Abstract-Conference.html.

024/hash/b444ad72520a5f5c467343be88e352ed-Abstract-Conference.html.

Haihan Zhang, Wendao Wu, Chenheng Zhang, Yanyi Li, Chunyuan Zheng, Cong Fang, Haoxuan Li, and Zhouchen Lin. Sharp minimax rates for highly smooth strongly convex zeroth-order optimization, 2026. arXiv:2610.00276v1.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=BJgnXpVYwS>.

CHECKLIST

1. For all models and algorithms presented, check if you include:
 - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] Sections 2–7 and the algorithm descriptions specify the model and assumptions.
 - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] Calls and sequential rounds are counted, including the calls of the polynomial fit; for $\beta > 4$ (for $\beta > 3$ in the method with logarithmically many rounds) the methods without a Hessian bound use global polynomial minimization, whose cost is not analyzed.
 - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] The supplementary code lists its dependencies (NumPy; Matplotlib only for an optional preview), the seeds and the commands that reproduce the figure.
2. For any theoretical claim, check if you include:
 - (a) Statements of the full set of assumptions of all theoretical results. [Yes] Each result specifies its function class, noise model, and budget regime.
 - (b) Complete proofs of all theoretical results. [Yes] Proofs of the new results are in the supplementary material; standard results are cited.
 - (c) Clear explanations of any assumptions. [Yes] The two tensor seminorms, the components of the noise and the calibration of constraint values are defined and discussed.
3. For all figures and tables that present empirical results, check if you include:

- (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] The supplementary material contains the code, the data tables behind the figure and the run log.
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] Appendix P gives every objective, parameter, budget and tuning rule; there are no data splits.
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] Error bars are one standard error over independent runs; Appendix P gives the number of runs for each experiment.
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] All experiments run on a CPU; Appendix P reports the processor, the run time and the software versions.
- 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator If your work uses existing assets. [Yes] NumPy is cited.
 - (b) The license information of the assets, if applicable. [Yes] NumPy is distributed under a permissive, BSD-compatible license.
 - (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] The code and the synthetic data are supplied.
 - (d) Information about consent from data providers/curators. [Not Applicable] All experimental data are generated synthetically.
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable] The data contain only synthetic numerical values.
- 5. If you used crowdsourcing or conducted research with human subjects, check if you include:
 - (a) The full text of instructions given to participants and screenshots. [Not Applicable] No human subjects or crowdsourcing are involved.
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable] No human subjects or crowdsourcing are involved.
 - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable] No human subjects or crowdsourcing are involved.

Supplementary Material

CONTENTS OF THE SUPPLEMENTARY MATERIAL

A GUIDE TO THE PROOFS

The notation and the noise model are those of Section 2. Constants c_β, C_β depend only on β and on the fixed one-dimensional stencil and cutoff functions; other dependences are stated where they occur. Expectations are over the randomness of the method and of the noise. For unconstrained problems the loss is $f(\hat{x}) - \min_{x \in \mathbb{R}^d} f(x)$; over a known compact set Q it is $f(\hat{x}) - \min_{x \in Q} f(x)$ with $\hat{x} \in Q$; for functional constraints the objective gap and the constraint violation are bounded separately.

Statement in the main text	Proof
Proposition 1 (stencil estimate)	Sections C and D, Lemma D.3
Procedure 1, random-frame estimate and restarted Nesterov descent	Section F.5: Lemmas F.5 and F.6
Theorem 1(a) and the achievable levels in (c)	Section F: Lemmas F.3 and F.6
Theorem 1(b) and the converse levels in (c)	Section E: Lemmas E.3, E.4 and E.7; Lemma H.1
Matching bounds and normalized minimax error (Section 3)	Lemmas E.3 to E.5
Theorem 2 (rounds)	Section G: Lemmas G.3 to G.5
Hessian growing with the gradient (Section 3)	Section I: Lemma I.1
Theorem 3(a), no Hessian bound	Section J: Lemmas J.11 and J.17; lower bound from Lemmas E.3 and E.4
Theorem 3(b), logarithmic rounds	Section K: Lemma K.2
Converse levels on the classes of Lemma I.1 ($L = L_0$) and Theorem 3 ($L = 3\alpha$)	Lemma E.7; Lemma H.1; proof of Lemma I.1 in Section I
Two-point feedback (Section 3)	Lemmas H.3 and H.4
Theorem 4 (convex methods)	Section L: Lemmas L.6 and L.9
Theorem 5 and the fixed class (Section 6)	Sections L.5 and M: Lemmas L.10 and L.11; Lemma E.7
Convex lower bound for $d < 256$ and Hilbert–Schmidt smoothness	Lemma E.6
Convex adversarial levels (Table 3)	Sections H.3 and H.4
Corollary 1, ℓ_p and ℓ_1 balls	Section N: Lemmas N.3, N.4, N.6, N.9 and N.10
Functional constraints (Section 7)	Section O: Lemmas O.5 and O.9
Comparison with other works (Tables 1 to 3)	Section M.4, Section R
Experiments (Section 8)	Section P
Comparison with concurrent work	Section S

Lemmas D.4 and D.5 show that two features of Definition 1 cannot be dropped. Section K also gives a convex method without a Hessian bound (Lemma K.3). The sizes of the constants quoted in Section Q are discussed in Lemma G.7, after Lemma K.2 and in Lemma N.11.

B PRELIMINARIES

We write $B(z, r) = z + B(r)$ for the closed Euclidean ball of radius r around z , and $e_1, \dots, e_d$ for the standard basis of $\mathbb{R}^d$.

B.1 Tensors

A k -tensor on $\mathbb{R}^d$ is an array $A = (A_{i_1 \dots i_k})$; it is symmetric if its entries do not change under permutations of the indices. We write $\langle A, B \rangle_{\text{HS}} = \sum A_{i_1 \dots i_k} B_{i_1 \dots i_k}$, so that $\|A\|_{\text{HS}}^2 = \langle A, A \rangle_{\text{HS}}$, and $A[v_1, \dots, v_k]$ for the multilinear form. For $j \leq k$, $A[v_1, \dots, v_j]$ is the $(k-j)$ -tensor obtained by contracting the first j indices. For a k -tensor A and an l -tensor A' , $A \otimes A'$ is the $(k+l)$ -tensor with entries $A_{i_1 \dots i_k} A'_{i_{k+1} \dots i_{k+l}}$, and $v^{\otimes j} = v \otimes \dots \otimes v$ (j factors). The identity matrix I is a 2-tensor, and $I^{\otimes m}$ has the entries $\prod_{r=1}^m \mathbf{1}\{i_{2r-1} = i_{2r}\}$.

Lemma B.1 (Elementary tensor facts). *Let A be a k -tensor with $k \geq 1$, and $v \in \mathbb{R}^d$.*

$$(a) \|A\|_{\text{op}} \leq \|A\|_{\text{HS}} \leq d^{(k-1)/2} \|A\|_{\text{op}}, \text{ and } \|A[v]\|_{\text{HS}} \leq \|v\| \|A\|_{\text{HS}}.$$

$$(b) \|I^{\otimes m}\|_{\text{HS}} = d^{m/2}, \text{ and } |\langle A, I^{\otimes m} \otimes B \rangle_{\text{HS}}| \leq d^{m/2} \|A\|_{\text{HS}} \|B\|_{\text{HS}} \text{ for every } (k-2m)\text{-tensor } B.$$

Proof. Exactly d^m entries of $I^{\otimes m}$ equal one and the others vanish, and $\|I^{\otimes m} \otimes B\|_{\text{HS}} = \|I^{\otimes m}\|_{\text{HS}} \|B\|_{\text{HS}}$. With this, the first inequality in (a), the contraction bound and (b) follow from the Cauchy–Schwarz inequality. The second inequality in (a) follows by induction on k from $\|A\|_{\text{HS}}^2 = \sum_{i=1}^d \|A[e_i]\|_{\text{HS}}^2$ and $\|A[e_i]\|_{\text{op}} \leq \|A\|_{\text{op}}$. $\square$

B.2 Taylor remainders

For $f \in C^q$ and $z, v \in \mathbb{R}^d$ let $T_q f(z; v) = \sum_{j=0}^q D^j f(z) [v^{\otimes j}] / j!$ be the Taylor polynomial of degree q , and $T_{q-j} D^j f(z; v)$ the Taylor polynomial of degree $q-j$ of the tensor-valued map $v \mapsto D^j f(z + v)$.

Lemma B.2 (Taylor remainders). *Let $\beta = q + \gamma$ with $q \geq 1$ and $\gamma \in (0, 1]$, let $f \in C^q(\mathbb{R}^d)$, and let $z, v \in \mathbb{R}^d$.*

(a) $|f(z + v) - T_q f(z; v)| \leq [f]_{\beta, \text{op}} \|v\|^\beta / q!$.

(b) For $1 \leq j \leq q$, $\|D^j f(z + v) - T_{q-j} D^j f(z; v)\|_{\text{HS}} \leq [f]_{\beta, \text{HS}} \|v\|^{\beta-j} / (q-j)!$.

(c) The bound (b) holds with op in place of HS on both sides. For $j = 1$ the left-hand side is then the Euclidean norm of $\nabla f(z + v) - T_{q-1} \nabla f(z; v)$.

Proof. For $j = q$, (b) and (c) are the definitions of the seminorms. For $0 \leq j < q$, with $D^0 f = f$, the integral form of the remainder is

$$D^j f(z + v) - T_{q-j} D^j f(z; v) = \int_0^1 \frac{(1-s)^{q-j-1}}{(q-j-1)!} (D^q f(z + sv) - D^q f(z)) [v^{\otimes (q-j)}] ds.$$

A contraction with v multiplies the operator norm, and by Lemma B.1(a) the Hilbert–Schmidt norm, by at most $\|v\|$. Since $\int_0^1 (1-s)^{q-j-1} s^\gamma ds \leq 1/(q-j)$, this gives (a) ($j = 0$), (b) and (c). $\square$

B.3 Spherical and ball averages

Throughout, U is uniform on $\mathbb{S}^{d-1}$ and W is uniform on the unit ball $B(1)$. For a continuous scalar, vector or tensor valued function g , a center $z \in \mathbb{R}^d$ and $s \geq 0$ let

$$S_g(s) = \mathbb{E} g(z + sU), \quad B_g(s) = \mathbb{E} g(z + sW).$$

As a function of the center we write $\mathcal{A}_s g(z) = B_g(s)$ for the ball average. Since $\|W\|$ has density dr^{d-1} on $[0, 1]$, $B_g(s) = d \int_0^1 r^{d-1} S_g(sr) dr$. The Laplacian $\Delta = \sum_i \partial_i^2$ acts entrywise. For continuous a on $[0, s_0]$ define the operator

$$(J_d a)(s) = \int_0^s t^{1-d} \int_0^t u^{d-1} a(u) du dt. \quad (\text{B.1})$$

It is positive ($a \geq 0$ implies $J_d a \geq 0$), and $J_d(s^{2j}) = s^{2j+2} / ((2j+2)(d+2j))$, so that $(J_d^j 1)(s) = s^{2j} / (2^j j! \prod_{i=0}^{j-1} (d+2i))$.

Lemma B.3 (Finite Pizzetti formula). *Let $k \geq 1$ and let g be $2k$ times continuously differentiable near $B(z, s_0)$. For $0 \leq s \leq s_0$,*

$$S_g(s) = \sum_{j=0}^{k-1} \frac{s^{2j} \Delta^j g(z)}{2^j j! \prod_{i=0}^{j-1} (d+2i)} + (J_d^k S_{\Delta^k g})(s), \quad (\text{B.2})$$

where empty products equal one, and

$$B_g(s) = \sum_{j=0}^{k-1} \frac{s^{2j} \Delta^j g(z)}{2^j j! \prod_{i=1}^j (d+2i)} + d \int_0^1 r^{d-1} (J_d^k S_{\Delta^k g})(sr) dr. \quad (\text{B.3})$$

The norm of the last term is at most $s^{2k} \sup_{\|x-z\| \leq s} \|\Delta^k g(x)\| / (2^k k! \prod_{i=1}^k (d+2i))$.

Proof. For $g \in C^2$ the divergence theorem gives $S'_g(s) = s^{1-d} \int_0^s u^{d-1} S_{\Delta g}(u) du$ (for $d = 1$ differentiate $S_g(s) = (g(z+s) + g(z-s))/2$), that is, $S_g = g(z) + J_d S_{\Delta g}$. Iterating k times gives (B.2). Averaging over the radius multiplies the coefficient of s^{2j} by $d/(d+2j)$, which gives (B.3), and positivity of J_d bounds the remainder. $\square$

Lemma B.3 is a finite form, with an explicit remainder, of Pizzetti's formula (Ovall, 2016; Toma, 2008).

Lemma B.4 (Gradients of ball averages). *Let $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be continuous and $s > 0$. Then $\mathcal{A}_s g$ is continuously differentiable and*

$$\nabla \mathcal{A}_s g(z) = \frac{d}{s} \mathbb{E}[g(z + sU)U].$$

If g is Lipschitz, then $\mathcal{A}_s g$ is Lipschitz with the same constant. If $g \in C^1$, then $\nabla \mathcal{A}_s g(z) = B_{\nabla g}(s)$.

Proof. The identity is Flaxman et al. (2005, Lemma 1), and its right-hand side is continuous in z . The other two claims follow from $\mathcal{A}_s g(z) = \mathbb{E}g(z + sW)$, by averaging and by differentiation under the integral. $\square$

We use the Poincaré inequality on the sphere: for $d \geq 2$ and a differentiable function ϕ on a neighborhood of $\mathbb{S}^{d-1}$ with $\mathbb{E}\phi(U) = 0$,

$$\mathbb{E}\phi(U)^2 \leq \frac{1}{d-1} \mathbb{E} \|\nabla_{\mathbb{S}} \phi(U)\|^2, \quad (\text{B.4})$$

where $\nabla_{\mathbb{S}}$ is the tangential part of the gradient. This is the statement that the first nonzero eigenvalue of the Laplace–Beltrami operator on $\mathbb{S}^{d-1}$ equals $d-1$ (Dai and Xu, 2013, Chapter 1). Applied to each component, it holds for vector and tensor valued ϕ , with the sum of the squared component gradients on the right. By mollification it also gives $\mathbb{E}\phi(U)^2 - (\mathbb{E}\phi(U))^2 \leq \ell^2/(d-1)$ for every ℓ -Lipschitz $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$. In particular, if $e : \mathbb{R}^d \rightarrow \mathbb{R}$ is odd and ℓ -Lipschitz and $(U_i)_{i=1}^d$ is a uniformly random orthonormal basis, then $\mathbb{E}e(U) = 0$ and, by orthonormality,

$$\mathbb{E} \left\| \sum_{i=1}^d e(U_i) U_i \right\|^2 = d \mathbb{E}e(U)^2 \leq \frac{d \ell^2}{d-1}. \quad (\text{B.5})$$

B.4 Testing inequalities

For probability measures P, Q let $\text{TV}(P, Q) = \sup_A |P(A) - Q(A)|$ and $\text{KL}(P\|Q) = \int \log(dP/dQ) dP$ ($+\infty$ without absolute continuity).

Lemma B.5 (Testing and Gaussian observations).

- (a) *Every test between P and Q has error probabilities summing to at least $1 - \text{TV}(P, Q)$, and $\text{TV}(P, Q) \leq \sqrt{\text{KL}(P\|Q)/2}$.*
- (b) *Let an algorithm interact with f or with g through independent $N(0, \sigma^2)$ noise, and let P_f, P_g be the laws of its seed and transcript. If $f - g$ is bounded, then*

$$\text{KL}(P_f\|P_g) = \frac{1}{2\sigma^2} \mathbb{E}_f \sum_{t=1}^T (f(X_t) - g(X_t))^2.$$

Part (a) is the two-point testing bound and Pinsker's inequality (Tsybakov, 2009, Section 2.4 and Lemma 2.5). Part (b) is the chain rule for the Kullback–Leibler divergence (Cover and Thomas, 2006, Theorem 2.5.3): given the seed and the past, X_t is the same under both hypotheses, and the conditional divergence of Y_t is $(f(X_t) - g(X_t))^2/(2\sigma^2)$.

We use Assouad's lemma in the following form (Tsybakov, 2009, Section 2.7.2). Let $\{P_\omega : \omega \in \{-1, 1\}^d\}$ be laws of transcripts and let $\omega^{(j)}$ denote ω with its j -th coordinate flipped. For any estimator $\hat{\omega}$ of ω ,

$$\frac{1}{2^d} \sum_{\omega} \mathbb{E}_{\omega} \sum_{j=1}^d \mathbf{1}\{\hat{\omega}_j \neq \omega_j\} \geq \frac{d}{2} \left(1 - \frac{1}{d2^d} \sum_{\omega, j} \text{TV}(P_{\omega}, P_{\omega^{(j)}}) \right), \quad (\text{B.6})$$

which follows from Lemma B.5(a) applied to each pair $(\omega, \omega^{(j)})$.

B.5 Spherical caps

Lemma B.6 (Caps). *Let $d \geq 2$, $0 < \theta < 1$, w uniform on $\mathbb{S}^{d-1}$, and $x \neq 0$. Then*

$$\mathbb{P}(|\langle w, x \rangle| > \theta \|x\|) \leq 2e^{-d\theta^2/2} \leq 2e^{-\theta^2(d-1)/2}.$$

By Ball (1997, Lemma 2.2), each of the caps $\{w : \langle w, \pm x \rangle \geq \theta \|x\|\}$ has probability at most $e^{-d\theta^2/2}$.

C THE FINITE-DIFFERENCE ESTIMATE WITH EXACT VALUES

This appendix analyzes one frame of Procedure 1 with exact values; Section D adds the noise of Definition 1 and proves Proposition 1 and Lemma D.3.

Throughout, $\beta > 2$, $q = \lceil \beta \rceil - 1 \geq 2$, $\gamma = \beta - q$, $m = \lceil q/2 \rceil$ and $t_j = j/m$, and L_β bounds $[f]_{\beta, \text{op}}$ (with $\tau = 0$) or $[f]_{\beta, \text{HS}}$ (with $\tau = \tau_\beta$ from (2)). The weights $a_1, \dots, a_m$ solve

$$\sum_{j=1}^m a_j t_j^{2i+1} = \mathbf{1}\{i = 0\}, \quad i = 0, 1, \dots, m-1. \quad (\text{C.1})$$

The matrix (t_j^{2i+1}) is a Vandermonde matrix in the distinct numbers t_j^2 times the diagonal matrix $\text{diag}(t_j)$, so the solution exists, is unique and depends only on β . The constants

$$S_1 = \sum_j |a_j|, \quad S_2 = \sum_j a_j^2, \quad A_1 = \sum_j |a_j| t_j, \quad (\text{C.2})$$

$$c_i = \frac{1}{(q-i)!} \sum_j |a_j| t_j^\beta \quad (0 \leq i \leq q) \quad (\text{C.3})$$

also depend only on β , and $A_1 \leq S_1$ since $t_j \leq 1$. The values of S_1, S_2, A_1 for $m \leq 4$ are in Table 4. We use the notation $U, W, S_g, B_g, \mathcal{A}_s$ and J_d of Section B.3. For a function g , a center z , a radius $h > 0$ and $u \in \mathbb{R}^d$ let

$$D_{h,u}g(z) = \frac{1}{2h} \sum_{j=1}^m a_j (g(z + ht_j u) - g(z - ht_j u)). \quad (\text{C.4})$$

With exact values, one frame $(U_1, \dots, U_d)$ of Procedure 1 returns $\sum_{i=1}^d D_{h,U_i} f(z) U_i$.

C.1 Exactness and the mean

Lemma C.1 (Exactness). *For every polynomial φ on $\mathbb{R}$ of degree at most $2m$ and every $h > 0$, $\frac{1}{2h} \sum_j a_j (\varphi(ht_j) - \varphi(-ht_j)) = \varphi'(0)$. If $[f]_{\beta, \text{op}} \leq L_\beta$ and $\|u\| = 1$, then $|D_{h,u}f(z) - \langle \nabla f(z), u \rangle| \leq c_0 L_\beta h^{\beta-1}$.*

Proof. The symmetric difference annihilates even powers of s and maps s^{2i+1} , $0 \leq i \leq m-1$, to $h^{2i} \sum_j a_j t_j^{2i+1} = \mathbf{1}\{i = 0\}$. For the second claim apply the first to $\varphi(s) = T_q f(z; su)$, which has degree $q \leq 2m$ and $\varphi'(0) = \langle \nabla f(z), u \rangle$. By Lemma B.2(a) the remainders at $\pm ht_j u$ are at most $L_\beta (ht_j)^\beta / q!$, and $\frac{1}{2h} \sum_j |a_j| 2L_\beta (ht_j)^\beta / q! = c_0 L_\beta h^{\beta-1}$. $\square$

Lemma C.2 (The mean). *For a continuous $g : \mathbb{R}^d \rightarrow \mathbb{R}$ and $h > 0$ define the surrogate $S_h g$, and for $f \in C^1$ the vector $G_h(z)$, by*

$$S_h g(z) = \sum_{j=1}^m a_j t_j \mathbb{E} g(z + ht_j W), \quad G_h(z) = \nabla S_h f(z) = \sum_{j=1}^m a_j t_j B_{\nabla f}(ht_j). \quad (\text{C.5})$$

Then $S_h g$ is continuously differentiable and $d\mathbb{E}[D_{h,U}g(z)U] = \nabla S_h g(z)$. Consequently, for a uniformly random orthonormal basis (U_i) , $\mathbb{E} \sum_i D_{h,U_i}g(z)U_i = \nabla S_h g(z)$. For $g = f$ this says that the noiseless part of one frame has mean $G_h(z)$.

Proof. By the symmetry $U \mapsto -U$ and Lemma B.4,

$$d \mathbb{E}[D_{h,U}g(z)U] = \frac{d}{h} \sum_j a_j \mathbb{E}[g(z + ht_j U)U] = \sum_j a_j t_j \nabla \mathcal{A}_{ht_j} g(z) = \nabla \mathcal{S}_h g(z),$$

and for $g = f \in C^1$ the last claim of Lemma B.4 gives the second expression for G_h . Each column of a uniformly random orthonormal basis is uniform on $\mathbb{S}^{d-1}$. $\square$

Since $\sum_j a_j t_j = 1$ by (C.1), G_h is an average of ∇f over balls with signed weights of total mass one.

C.2 Bias

Lemma C.3 (Bias). *For all z and $h > 0$:*

- (a) *if $[f]_{\beta, \text{op}} \leq L_\beta$, then $\|G_h(z) - \nabla f(z)\| \leq c_1 L_\beta h^{\beta-1}$;*
- (b) *if $[f]_{\beta, \text{HS}} \leq L_\beta$ and $q = 2k + 1$ is odd, then $\|G_h(z) - \nabla f(z)\| \leq \frac{q! c_0}{2^k k!} L_\beta d^{-k/2} h^{\beta-1}$;*
- (c) *if $[f]_{\beta, \text{HS}} \leq L_\beta$ and $q = 2k$ is even, then $\|G_h(z) - \nabla f(z)\| \leq C_\beta L_\beta d^{-(\beta-1)/4} h^{\beta-1}$.*

Hence $\|G_h(z) - \nabla f(z)\| \leq C_b L_\beta d^{-\tau/2} h^{\beta-1}$ in both geometries, where C_b depends only on β , and $C_b = c_1$ in the operator case.

Part (a) is the argument of Akhavan et al. (2024, Lemma 6), written for the stencil.

Proof. In (a) and (b) we expand the ball averages in (C.5) around z . Terms proportional to s^{2i} with $1 \leq i \leq m-1$ cancel at $s = ht_j$ after summation with the weights $a_j t_j$, because $\sum_j a_j t_j (ht_j)^{2i} = h^{2i} \sum_j a_j t_j^{2i+1} = 0$, and the constant term gives $\sum_j a_j t_j \nabla f(z) = \nabla f(z)$.

(a) *Operator norm.* By Lemma B.2(c), $\nabla f(z+v) = T_{q-1} \nabla f(z; v) + \rho(v)$ with $\|\rho(v)\| \leq L_\beta \|v\|^{\beta-1}/(q-1)!$. The ball average of the term of degree r of $T_{q-1} \nabla f(z; sW)$ vanishes for odd r and is proportional to s^r for even r ; since $r \leq q-1 \leq 2m-1$, only $r=0$ survives. As $\mathbb{E}\|W\|^{\beta-1} \leq 1$, the remainder contributes at most $\sum_j |a_j| t_j L_\beta (ht_j)^{\beta-1}/(q-1)! = c_1 L_\beta h^{\beta-1}$.

(b) *Hilbert–Schmidt norm, q odd.* Let $g = \nabla f \in C^{2k}$ and $H = \Delta^k g$. For a unit vector e , $\langle e, H(x) - H(y) \rangle = \langle D^q f(x) - D^q f(y), e \otimes I^{\otimes k} \rangle_{\text{HS}}$, so Lemma B.1(b) gives

$$\|H(x) - H(y)\| \leq d^{k/2} L_\beta \|x - y\|^\gamma. \quad (\text{C.6})$$

By (B.2) with k steps, $S_g(s)$ is a polynomial in s^2 of degree k with constant term $\nabla f(z)$, plus $J_d^k[S_H - H(z)](s)$. Ball averaging keeps the polynomial a polynomial in s^2 , and since $k = m-1$, the weights leave only $\nabla f(z)$. By positivity of J_d , (C.6) and $(J_d^k 1)(s) \leq s^{2k}/(2^k k! d^k)$,

$$\|J_d^k[S_H - H(z)](s)\| \leq d^{k/2} L_\beta s^\gamma (J_d^k 1)(s) \leq \frac{L_\beta d^{-k/2} s^{\beta-1}}{2^k k!},$$

and ball averaging preserves this bound. Summation over j with $s = ht_j$ gives (b).

(c) *Hilbert–Schmidt norm, q even.* Here $m = k$, and by Lemma B.1(b) the scalar function $\Delta^k f$ is γ -Hölder with constant $d^{k/2} L_\beta$. Let $E_h = \mathcal{S}_h f - f$. By (B.3) with k steps, the terms with $1 \leq i \leq k-1$ cancel and the term $i=0$ gives $f(z)$, so $E_h(z) = \sum_j a_j t_j \mathcal{R}(ht_j; z)$, where $\mathcal{R}(s; z)$ is the remainder of (B.3). For two centers z, z' , $\mathcal{R}(s; z) - \mathcal{R}(s; z')$ is a positive linear functional, of total mass $s^{2k}/(2^k k! \prod_{i=1}^k (d+2i))$, applied to $\Delta^k f(z + \cdot) - \Delta^k f(z' + \cdot)$. Hence

$$|E_h(z) - E_h(z')| \leq C_1 L_\beta d^{-k/2} h^{2k} \|z - z'\|^\gamma, \quad C_1 = \frac{1}{2^k k!} \sum_j |a_j| t_j^{2k+1}. \quad (\text{C.7})$$

Fix a unit vector e and let $h' = hd^{-1/4}$. Since $[f]_{\beta, \text{op}} \leq L_\beta$ and $[S_h f]_{\beta, \text{op}} \leq A_1 L_\beta$, Lemma C.1 along the line $z + se$ gives $|D_{h',e} v(z) - \partial_e v(z)| \leq c_0 L_\beta h'^{\beta-1}$ for $v = f$ and $\leq c_0 A_1 L_\beta h'^{\beta-1}$ for $v = S_h f$. By (C.7),

$$|D_{h',e} E_h(z)| \leq \frac{1}{2h'} \sum_j |a_j| |E_h(z + h't_j e) - E_h(z - h't_j e)| \leq S_1 C_1 L_\beta d^{-k/2} h^{2k} h'^{\gamma-1}.$$

Since $D_{h',e}$ is linear and $\partial_e S_h f = \langle e, G_h \rangle$,

$$|\langle e, G_h(z) - \nabla f(z) \rangle| \leq c_0(1 + A_1) L_\beta h'^{\beta-1} + S_1 C_1 L_\beta d^{-k/2} h^{2k} h'^{\gamma-1} = (c_0(1 + A_1) + S_1 C_1) L_\beta d^{-(\beta-1)/4} h^{\beta-1},$$

because $2k + \gamma - 1 = \beta - 1$. Take the supremum over e .

Finally, $\tau = 0$ in (a), $\tau_\beta = k$ in (b) and $\tau_\beta = (\beta - 1)/2$ in (c). $\square$

C.3 Second moment

Fix z and h and let $e_f(u) = D_{h,u} f(z) - \langle \nabla f(z), u \rangle$ for $u \in \mathbb{R}^d$, an odd C^q function. For a uniformly random orthonormal basis (U_i) let $Z = \sum_i e_f(U_i) U_i$ be the noiseless error of the estimate. By orthonormality, Lemma C.2 and $d \mathbb{E}[UU^\top] = I$,

$$\mathbb{E} \|Z\|^2 = d \mathbb{E} e_f(U)^2, \quad \mathbb{E} Z = d \mathbb{E}[e_f(U)U] = G_h(z) - \nabla f(z). \quad (\text{C.8})$$

Lemma C.4 (Derivatives of the stencil error). *For $\|u\| = 1$: if $[f]_{\beta, \text{op}} \leq L_\beta$, then $\|\nabla e_f(u)\| \leq c_1 L_\beta h^{\beta-1}$; if $[f]_{\beta, \text{HS}} \leq L_\beta$, then $\|D^j e_f(u)\|_{\text{HS}} \leq c_j L_\beta h^{\beta-1}$ for $1 \leq j \leq q$.*

Proof. Differentiating (C.4) j times in u ,

$$D^j e_f(u) = \frac{h^{j-1}}{2} \sum_l a_l t_l^j (D^j f(z + ht_l u) - (-1)^j D^j f(z - ht_l u)) - \mathbf{1}\{j = 1\} \nabla f(z).$$

Replace $D^j f(z \pm v)$ by its Taylor polynomial of degree $q - j$. Its term of degree r survives the signed sum only if $j + r$ is odd, with weight $h^{j+r-1} \sum_l a_l t_l^{j+r}$. This weight vanishes for $3 \leq j + r \leq q$ by (C.1), and the term $j + r = 1$ cancels $\nabla f(z)$. The two Taylor remainders are bounded by Lemma B.2(b),(c). $\square$

Lemma C.5 (Means of odd derivatives). *Let $[f]_{\beta, \text{HS}} \leq L_\beta$, let $3 \leq j < q$ be odd and $k = \lfloor (q - j)/2 \rfloor$. Then $\|\mathbb{E} D^j e_f(U)\|_{\text{HS}} \leq C_\beta L_\beta h^{\beta-1} d^{-k/2}$.*

Proof. For odd $j \geq 3$ the formula in the proof of Lemma C.4 gives $\mathbb{E} D^j e_f(U) = h^{j-1} \sum_l a_l t_l^j S_{D^j f}(ht_l)$. If $k \geq 1$, expand $S_{D^j f}$ by (B.2) with k steps; this uses $2k \leq q - j$ derivatives of $D^j f$. The polynomial terms are proportional to s^{2r} , $r \leq k - 1$, and receive the weights $h^{j+2r-1} \sum_l a_l t_l^{j+2r}$, which vanish because $j + 2r$ is odd and $3 \leq j + 2r \leq q$. In the remainder (for $k = 0$, in $S_{D^j f}$ itself), the tensor $\Delta^k D^j f$ is a k -fold trace of $D^{j+2k} f$, where $q - j - 2k \in \{0, 1\}$. Subtract from it its Taylor polynomial at z of degree $q - j - 2k$. A linear term has zero spherical mean, and the constant term $\Delta^k D^j f(z)$ is multiplied by $(J_d^k 1)(s) \propto s^{2k}$ and receives the weight $\sum_l a_l t_l^{j+2k} = 0$, because $j + 2k \in \{q - 1, q\}$ is odd and at least 3. By Lemma B.2(b) and Lemma B.1(b), the rest has Hilbert–Schmidt norm at most $d^{k/2} L_\beta \|v\|^{\beta-j-2k}$ at $z + v$. Positivity of J_d bounds the remaining term by

$$d^{k/2} L_\beta s^{\beta-j-2k} (J_d^k 1)(s) \leq L_\beta s^{\beta-j} d^{-k/2} / (2^k k!),$$

and $s = ht_l$ gives the claim. $\square$

Proposition C.6 (Noiseless error). *There is a constant C_Z , depending only on β , with $C_Z = 2c_1^2$ in the operator case, such that $\mathbb{E} \|Z\|^2 \leq C_Z L_\beta^2 d^{-\tau} h^{2\beta-2}$.*

Proof. For $d = 1$ the error $Z = e_f(1)$ is deterministic, since $e_f(-1) \cdot (-1) = e_f(1)$, and $|e_f(1)| \leq c_0 L_\beta h^{\beta-1}$ by Lemma C.1. Let $d \geq 2$. We apply the Poincaré inequality (B.4) once in the operator case, and repeatedly along the derivatives of e_f in the Hilbert–Schmidt case.

Operator norm. Since e_f is odd, $\mathbb{E} e_f(U) = 0$, and (B.4) with Lemma C.4 gives $\mathbb{E} \|Z\|^2 = d \mathbb{E} e_f(U)^2 \leq \frac{d}{d-1} c_1^2 L_\beta^2 h^{2\beta-2} \leq 2c_1^2 L_\beta^2 h^{2\beta-2}$.

Hilbert–Schmidt norm. Write $\mathbb{E} \|Z\|^2 = \|\mathbb{E}Z\|^2 + \mathbb{E} \|Z - \mathbb{E}Z\|^2$; the first term is at most $C_b^2 L_\beta^2 d^{-\tau} h^{2\beta-2}$ by (C.8) and Lemma C.3. The vector $\mathbf{a}_* = \mathbb{E}Z = d \mathbb{E}[e_f(U)U]$ is the coefficient of the $L^2(\mathbb{S}^{d-1})$ projection of e_f onto linear functions, and $Z - \mathbb{E}Z = \sum_i (e_f(U_i) - \langle \mathbf{a}_*, U_i \rangle) U_i$. Hence, with $\mathbf{a}_0 = \mathbb{E} \nabla e_f(U)$ and $\Lambda_j = \mathbb{E} \|D^j e_f(U)\|_{\text{HS}}^2$,

$$\mathbb{E} \|Z - \mathbb{E}Z\|^2 = d \mathbb{E} (e_f(U) - \langle \mathbf{a}_*, U \rangle)^2 \leq d \mathbb{E} (e_f(U) - \langle \mathbf{a}_0, U \rangle)^2 \leq \frac{d \Lambda_2}{(d-1)^2},$$

where the last step applies (B.4) to the odd function $e_f - \langle \mathbf{a}_0, \cdot \rangle$ and then to $\nabla e_f - \mathbf{a}_0$. For $j < q$, (B.4) applied to $D^j e_f - \mathbb{E} D^j e_f$ gives $\Lambda_j \leq \|\mathbb{E} D^j e_f(U)\|_{\text{HS}}^2 + \Lambda_{j+1}/(d-1)$, and $\mathbb{E} D^j e_f(U) = 0$ for even j , since $D^j e_f$ is then odd. Iterating,

$$\mathbb{E} \|Z - \mathbb{E}Z\|^2 \leq \sum_{\substack{3 \leq j \leq q \\ j \text{ odd}}} \frac{d \|\mathbb{E} D^j e_f(U)\|_{\text{HS}}^2}{(d-1)^j} + \frac{d \Lambda_q}{(d-1)^q}.$$

By Lemmas C.4 and C.5 and $d/(d-1)^j \leq 2^j d^{1-j}$, the j -th term is at most $C_\beta L_\beta^2 h^{2\beta-2} d^{-(j-1+\lfloor (q-j)/2 \rfloor)}$, and the last is at most $C_\beta L_\beta^2 h^{2\beta-2} d^{-(q-1)}$. Both exponents are at least $\lceil q/2 \rceil \geq \tau_\beta$, since $j-1 + \lfloor (q-j)/2 \rfloor = \lfloor (q+j-2)/2 \rfloor \geq \lfloor (q+1)/2 \rfloor$ for $j \geq 3$, and $q-1 \geq \lceil q/2 \rceil$ for $q \geq 2$. $\square$

D THE RANDOM-FRAME ESTIMATE UNDER THE UNIFIED NOISE

This appendix proves Lemma D.3, which decomposes the estimate of one round of Procedure 1 under the noise of Definition 1. For $d \geq 2$, its parts (i)–(iii) are Proposition 1. Lemmas D.4 and D.5 show that two features of Definition 1 cannot be dropped: in (C) the σ -algebra $\mathcal{E}_t$ contains the error components of the earlier queries of the round, and in (L) the function η is fixed before the run.

D.1 Setting

We use the notation of Section C, in particular $D_{h,u}$, $\mathcal{S}_h$, G_h , e_f , the constants (C.2), C_b of Lemma C.3 and C_Z of Lemma C.6.

One round. Fix a round k . Its center z and radius $h > 0$ are $\mathcal{F}_{k-1}$ -measurable, and Θ_k consists of the $B \geq 1$ independent frames $U^{(\ell)} = (U_1^{(\ell)}, \dots, U_d^{(\ell)})$ of Procedure 1. The query at $z \pm h t_j U_i^{(\ell)}$ is labeled $(\ell, i, j, \pm)$, and the queries are ordered lexicographically; any order that does not depend on Θ_k would do. Its observed value is $Y_{ij\ell}^\pm$, with error components $\xi_{ij\ell}^{c,\pm}$, $\xi_{ij\ell}^{o,\pm}$ and $\delta_{ij\ell}^\pm$, and ξ_k^o is the vector of the oblivious errors of the round. All query points are functions of (z, h, Θ_k) . Some results of Section N use, in addition, the conditional sub-Gaussian condition

$$\mathbb{E}[\exp(s \xi_t^c) \mid \mathcal{E}_t] \leq \exp(s^2 \sigma^2 / 2) \quad (s \in \mathbb{R}) \quad \text{for all queries } t, \quad (\text{SG})$$

which implies (C). Independent $N(0, \sigma^2)$ errors satisfy it.

Localization. Since h may be small with positive probability, the random vectors below need not be integrable. Let $E_n = \{1/n \leq h \leq n, \|z\| \leq n\} \in \mathcal{F}_{k-1}$; these events increase to the whole space. On E_n all query points lie in $B(2n)$, where f and η are bounded, $1/h \leq n$, and all error components are square integrable by (C), (O) and (A). A statement such as $\mathbb{E}[\zeta \mid \mathcal{F}_{k-1}] = 0$ means that $\mathbb{E}[\mathbf{1}_{E_n} \zeta \mid \mathcal{F}_{k-1}] = 0$ for every n , and a conditional second moment is the increasing limit of those of $\mathbf{1}_{E_n} \zeta$. This agrees with the usual definitions for integrable variables, since $\mathbf{1}_{E_n} \mathbb{E}[X \mid \mathcal{F}_{k-1}] = \mathbb{E}[\mathbf{1}_{E_n} X \mid \mathcal{F}_{k-1}]$. We prove every statement on a fixed E_n without further mention. For a random vector X let $\mathbf{N}(X) = (\mathbb{E}[\|X\|^2 \mid \mathcal{F}_{k-1}])^{1/2}$; then $\mathbf{N}(X + X') \leq \mathbf{N}(X) + \mathbf{N}(X')$ almost surely.

Lemma D.1 (Independence within a round).

- (a) Θ_k is independent of $\mathcal{F}_{k-1} \vee \sigma(\xi_k^o)$. Hence, conditionally on $\mathcal{F}_{k-1} \vee \sigma(\xi_k^o)$, the frames are independent uniformly random orthonormal bases, and every column $U_i^{(\ell)}$ is uniform on $\mathbb{S}^{d-1}$.
- (b) Let $\mathcal{H}$ be a σ -algebra independent of Θ_k , let X be $\mathcal{H}$ -measurable, and let ϕ be jointly measurable with $\phi(X, \Theta_k)$ integrable. Then $\mathbb{E}[\phi(X, \Theta_k) \mid \mathcal{H}] = \Phi(X)$ almost surely, where $\Phi(x) = \mathbb{E}\phi(x, \Theta_k)$.

Proof. (a) Θ_k is independent of $\mathcal{F}_{k-1}$, and by (O) it is conditionally independent of ξ_k° given $\mathcal{F}_{k-1}$. Hence, for $F \in \mathcal{F}_{k-1}$ and bounded measurable ϕ, ψ , $\mathbb{E}[\mathbf{1}_F \phi(\xi_k^\circ) \psi(\Theta_k)] = \mathbb{E}[\mathbf{1}_F \phi(\xi_k^\circ)] \mathbb{E}[\psi(\Theta_k)]$, which proves (a). Part (b) is the substitution rule for conditional expectations (Durrett, 2019, Section 4.1). $\square$

D.2 The systematic error

Lemma D.2 (The systematic error). *Let η be M -Lipschitz, $z \in \mathbb{R}^d$, $h > 0$, and $e_\eta(u) = D_{h,u}\eta(z)$ for $u \in \mathbb{R}^d$.*

- (a) e_η is odd and $(A_1 M)$ -Lipschitz on $\mathbb{R}^d$.
- (b) $\mathcal{S}_h \eta$ is $(A_1 M)$ -Lipschitz, so $\|\nabla \mathcal{S}_h \eta(z)\| \leq A_1 M$.
- (c) If $d \geq 2$, then $\mathbb{E} e_\eta(U) = 0$ and $\mathbb{E} e_\eta(U)^2 \leq A_1^2 M^2 / (d-1)$.
- (d) If $d \geq 2$, then for a uniformly random orthonormal basis (U_i) of $\mathbb{R}^d$

$$\mathbb{E} \left\| \sum_{i=1}^d e_\eta(U_i) U_i \right\|^2 = d \mathbb{E} e_\eta(U)^2 \leq \frac{d}{d-1} A_1^2 M^2.$$

If $d = 1$, then $\sum_i e_\eta(U_i) U_i = e_\eta(1)$ is deterministic.

Proof. (a) Oddness is clear, and $|e_\eta(u) - e_\eta(u')| \leq \frac{1}{2h} \sum_j |a_j| 2M h t_j \|u - u'\| = A_1 M \|u - u'\|$. (b) $\mathcal{S}_h \eta = \sum_j a_j t_j \mathcal{A}_{ht_j} \eta$, and each $\mathcal{A}_{ht_j} \eta$ is M -Lipschitz by Lemma B.4. (c) and (d) are (B.5) with $e = e_\eta$ and $\ell = A_1 M$. For $d = 1$, $e_\eta(U_1) U_1 = e_\eta(1)$ for both $U_1 = \pm 1$, by oddness. $\square$

D.3 The batched estimate

Proposition D.3 (Batched stencil estimate under the unified noise). *Let $f \in C^q$ satisfy $[f]_{\beta, \text{op}} \leq L_\beta$ (with $\tau = 0$) or $[f]_{\beta, \text{HS}} \leq L_\beta$ (with $\tau = \tau_\beta$), and let the noise be admissible at levels (M, σ, δ) . In round k let z and $h > 0$ be $\mathcal{F}_{k-1}$ -measurable, and let $\bar{g}$ be the output of Procedure 1 with $B \geq 1$ frames. Then*

$$\bar{g} = \nabla f(z) + b(z) + \zeta + \Delta,$$

where $b(z)$ is given in (i), ζ and Δ are defined in (D.2) below, and the following hold.

- (i) $b(z) = G_h(z) - \nabla f(z) + \nabla \mathcal{S}_h \eta(z)$ is $\mathcal{F}_{k-1}$ -measurable and $\|b(z)\| \leq C_b L_\beta d^{-\tau/2} h^{\beta-1} + A_1 M$.
- (ii) $\mathbb{E}[\zeta \mid \mathcal{F}_{k-1}] = 0$, and for $d \geq 2$

$$\mathbb{E}[\|\zeta\|^2 \mid \mathcal{F}_{k-1}] \leq \frac{1}{B} \left(C_f L_\beta^2 d^{-\tau} h^{2\beta-2} + C_\eta M^2 + C_\sigma \frac{d\sigma^2}{h^2} \right),$$

with $C_f = 2C_Z$, $C_\eta = 4A_1^2$ and $C_\sigma = (S_1 + (S_2/2)^{1/2})^2$. For $d = 1$, $\mathbb{E}[\|\zeta\|^2 \mid \mathcal{F}_{k-1}] \leq C_\sigma \sigma^2 / (Bh^2)$.

- (iii) $\|\Delta\| \leq \sqrt{d} S_1 \delta / h$ almost surely.

- (iv) The vector $\zeta + \Delta = \bar{g} - \nabla f(z) - b(z)$ is $\mathcal{F}_k$ -measurable; if $\delta = 0$, then $\Delta = 0$ and ζ is $\mathcal{F}_k$ -measurable. The vector $\tilde{\zeta} = \zeta + \Delta - \mathbb{E}[\Delta \mid \mathcal{F}_{k-1}]$ is $\mathcal{F}_k$ -measurable, and

$$\mathbb{E}[\tilde{\zeta} \mid \mathcal{F}_{k-1}] = 0, \quad \mathbf{N}(\tilde{\zeta}) \leq \mathbf{N}(\zeta) + \sqrt{d} S_1 \delta / h, \quad \|\mathbb{E}[\Delta \mid \mathcal{F}_{k-1}]\| \leq \sqrt{d} S_1 \delta / h.$$

In particular, the vectors $\tilde{\zeta}$ of different rounds form a martingale difference sequence with respect to $(\mathcal{F}_k)_k$.

When $\delta > 0$, the vector ζ need not be $\mathcal{F}_k$ -measurable, and the vectors ζ of different rounds need not be orthogonal. An adversarial error may cancel a centered error, which then stays hidden, and a centered error of a later round, conditioned only on the past and on the errors of its own round, may be correlated with the hidden one. The later appendices therefore use ζ only within one round, through $\mathbb{E}[\langle \zeta, x \rangle \mid \mathcal{F}_{k-1}] = 0$ for $\mathcal{F}_{k-1}$ -measurable random vectors x and through (ii). Sums over rounds use the decomposition $\bar{g} = \nabla f(z) + b(z) + \mathbb{E}[\Delta \mid \mathcal{F}_{k-1}] + \tilde{\zeta}$ with the adapted martingale differences $\tilde{\zeta}$ of (iv).

Table 4: Constants of Lemma D.3; $C_\eta = 4A_1^2$ and $C_\sigma = (S_1 + (S_2/2)^{1/2})^2$.

m	(a_j)	S_1	S_2	A_1	C_η	C_σ
1	1	1	1	1	4	2.914
2	$\frac{8}{3}, -\frac{1}{3}$	3	$\frac{65}{9}$	$\frac{5}{3}$	$\frac{100}{9}$	24.01
3	$\frac{9}{2}, -\frac{9}{10}, \frac{1}{10}$	$\frac{11}{2}$	$\frac{2107}{100}$	$\frac{11}{5}$	$\frac{484}{25}$	76.49
4	$\frac{32}{5}, -\frac{8}{5}, \frac{32}{105}, -\frac{1}{35}$	$\frac{25}{3}$	$\frac{480841}{11025}$	$\frac{93}{35}$	$\frac{34596}{1225}$	169.1

Proof. Condition on $\mathcal{F}_{k-1}$. Let $e_\eta(u) = D_{h,u}\eta(z)$, and for frame ℓ put

$$Y^{(\ell)} = \sum_i (e_f + e_\eta)(U_i^{(\ell)})U_i^{(\ell)} - b(z),$$

with $b(z)$ as in (i). For the other three error components put

$$\omega_i^{(\ell)} = \frac{1}{2h} \sum_j a_j (\xi_{ij\ell}^{o,+} - \xi_{ij\ell}^{o,-}), \quad \chi_i^{(\ell)} = \frac{1}{2h} \sum_j a_j (\xi_{ij\ell}^{c,+} - \xi_{ij\ell}^{c,-}), \quad \varrho_i^{(\ell)} = \frac{1}{2h} \sum_j a_j (\delta_{ij\ell}^+ - \delta_{ij\ell}^-),$$

and let $\Xi^{o,(\ell)} = \sum_i \omega_i^{(\ell)} U_i^{(\ell)}$, $\Xi^{c,(\ell)} = \sum_i \chi_i^{(\ell)} U_i^{(\ell)}$ and $\Delta^{(\ell)} = \sum_i \varrho_i^{(\ell)} U_i^{(\ell)}$. A bar denotes the average over the B frames. Since $\sum_i \langle \nabla f(z), U_i^{(\ell)} \rangle U_i^{(\ell)} = \nabla f(z)$,

$$\bar{g} = \nabla f(z) + b(z) + \zeta + \Delta, \tag{D.1}$$

$$\zeta = \bar{Y} + \bar{\Xi}^o + \bar{\Xi}^c, \quad \Delta = \frac{1}{B} \sum_{\ell=1}^B \Delta^{(\ell)}. \tag{D.2}$$

(i) and (iii). The surrogate gradients are continuous in (z, h) , so $b(z)$ is $\mathcal{F}_{k-1}$ -measurable. Its bound follows from Lemma C.3 and Lemma D.2(b). Since $|\varrho_i^{(\ell)}| \leq S_1 \delta/h$, orthonormality gives $\|\Delta^{(\ell)}\| \leq \sqrt{d} S_1 \delta/h$, hence (iii).

The frame part. By Lemma C.2, applied to f and to η , $\mathbb{E}Y^{(\ell)} = 0$, and the $Y^{(\ell)}$ are independent. For $d \geq 2$, Lemma C.6, Lemma D.2(d) and $d/(d-1) \leq 2$ give

$$\mathbb{E}[\|\bar{Y}\|^2 \mid \mathcal{F}_{k-1}] \leq B^{-1}(2C_Z L_\beta^2 d^{-\tau} h^{2\beta-2} + 4A_1^2 M^2). \tag{D.3}$$

For $d = 1$, $(e_f + e_\eta)(U_1)U_1 = (e_f + e_\eta)(1)$ is deterministic, hence $Y^{(\ell)} = 0$.

Random noise. Conditionally on $\mathcal{F}_{k-1} \vee \sigma(\xi_k^o)$, the coefficients $\omega_i^{(\ell)}$ are fixed and the frames are independent and uniform (Lemma D.1). So the vectors $\Xi^{o,(\ell)}$ have mean zero and are orthogonal in conditional L^2 across frames. Since $\mathbb{N}(\omega_i^{(\ell)}) \leq S_1 \sigma/h$, orthonormality yields

$$\mathbb{E}[\|\bar{\Xi}^o\|^2 \mid \mathcal{F}_{k-1}] \leq dS_1^2 \sigma^2 / (Bh^2). \tag{D.4}$$

The centered part is $\bar{\Xi}^c = \sum_t w_t \xi_t^c$, a sum over the queries $t = (\ell, i, j, \pm)$ of the round with $w_t = \pm a_j U_i^{(\ell)} / (2hB)$. Both w_t and the centered errors of the earlier queries are $\mathcal{E}_t$ -measurable, so by (C) the terms $w_t \xi_t^c$ are centered and pairwise orthogonal, and $\mathbb{E}[\|\bar{\Xi}^c\|^2 \mid \mathcal{F}_{k-1}] \leq \sigma^2 \sum_t \|w_t\|^2 = dS_2 \sigma^2 / (2Bh^2)$. The two parts need not be orthogonal to each other; Minkowski's inequality gives

$$\mathbb{E}[\|\bar{\Xi}^o + \bar{\Xi}^c\|^2 \mid \mathcal{F}_{k-1}] \leq (S_1 + \sqrt{S_2/2})^2 d\sigma^2 / (Bh^2). \tag{D.5}$$

Cross terms. Since $\langle \bar{Y}, w_t \rangle$ is $\mathcal{E}_t$ -measurable, (C) gives $\mathbb{E} \langle \bar{Y}, w_t \xi_t^c \rangle = 0$. For the oblivious part, condition on $\mathcal{F}_{k-1} \vee \sigma(\xi_k^o)$. Different frames give zero by independence, and within frame ℓ

$$\mathbb{E} \langle Y^{(\ell)}, U_i^{(\ell)} \rangle = \mathbb{E}(e_f + e_\eta)(U_i^{(\ell)}) - \langle b(z), \mathbb{E}U_i^{(\ell)} \rangle = 0$$

by oddness. Thus $\bar{Y}$ is orthogonal to both noise parts, and adding (D.3) and (D.5) proves (ii). The same computations give $\mathbb{E}[\zeta \mid \mathcal{F}_{k-1}] = 0$.

(iv). The vector $\zeta + \Delta = \bar{g} - \nabla f(z) - b(z)$ is a function of the observed values, Θ_k and $\mathcal{F}_{k-1}$ -measurable quantities, hence $\mathcal{F}_k$ -measurable. Subtracting the $\mathcal{F}_{k-1}$ -measurable vector $\mathbb{E}[\Delta \mid \mathcal{F}_{k-1}]$ gives $\tilde{\zeta}$, which is centered, and

$$\mathbf{N}(\Delta - \mathbb{E}[\Delta \mid \mathcal{F}_{k-1}])^2 = \mathbb{E}[\|\Delta\|^2 \mid \mathcal{F}_{k-1}] - \|\mathbb{E}[\Delta \mid \mathcal{F}_{k-1}]\|^2 \leq dS_1^2\delta^2/h^2$$

proves (iv) by Minkowski's inequality. Martingale differences adapted to $(\mathcal{F}_k)_k$ are orthogonal across rounds by the tower property. $\square$

Remark D.4 (The σ -algebra in the centered-noise condition). *The modified condition.* Replace $\mathcal{E}_t$ in (C) by the smaller σ -algebra $\mathcal{G}_t$ generated by $\mathcal{F}_{k-1}$, Θ_k , the queries of round k , and the observed values of the queries $s < t$ of round k . The decomposition, the bias and the centering in the proof of Lemma D.3 use only $\sigma(\mathcal{F}_{k-1}, \Theta_k) \subset \mathcal{G}_t$, so (i), (iii), (iv) and $\mathbb{E}[\zeta \mid \mathcal{F}_{k-1}] = 0$ remain valid. The variance argument loses the orthogonality of ξ_s^c and ξ_t^c for $s < t$, because $\xi_s^c = Y_s - f(X_s) - \eta(X_s) - \xi_s^o - \delta_s$ need not be $\mathcal{G}_t$ -measurable. Bounding $\mathbf{N}(\Xi^{c,(\ell)}) \leq \sqrt{d}S_1\sigma/h$ by Minkowski's inequality, as for the oblivious part, gives only

$$\mathbb{E}[\|\zeta\|^2 \mid \mathcal{F}_{k-1}] \leq \frac{1}{B}(C_f L_\beta^2 d^{-\tau} h^{2\beta-2} + C_\eta M^2) + (1 + B^{-1/2})^2 S_1^2 \frac{d\sigma^2}{h^2}.$$

The factor $1/B$ in front of $d\sigma^2/h^2$ cannot be kept with a constant depending only on β : an oblivious error can hide a centered error from the observation, and a later centered error of the round, which may depend on all frames, can be correlated with the hidden part. The following example loses a factor of order $\log^2 B$.

Example: the vectors. Let $k = 1$, $z = 0$, $h = 1$, $f \equiv 0$, $\eta \equiv 0$ and $\delta = 0$. Fix a node j_* with $a_{j_*} \neq 0$, and let only the queries $t_{j_*} U_i^{(\ell)}$ carry errors. In the sequence space $\mathbf{H} = \ell^2(\mathbb{Z})$ with unit vectors $\mathbf{e}_n$ put

$$\vartheta_n = \frac{1}{\pi(n + \frac{1}{2})} \quad (n \in \mathbb{Z}), \quad \mathbf{u}_\ell = \sum_n \vartheta_{n-\ell} \mathbf{e}_n, \quad \ell = 1, \dots, B.$$

These vectors are orthonormal. First, $\sum_n \vartheta_n^2 = 8\pi^{-2} \sum_{n \geq 0} (2n+1)^{-2} = 1$. Second, for $r \neq 0$,

$$\sum_n \vartheta_n \vartheta_{n+r} = \frac{1}{\pi^2 r} \sum_n \left(\frac{1}{n + \frac{1}{2}} - \frac{1}{n+r + \frac{1}{2}} \right) = 0;$$

the series converges absolutely, and its partial sum over $|n| \leq K$ reduces to $2|r|$ terms of order $1/K$. Let P_ℓ be the orthogonal projection onto $\text{span}\{\mathbf{u}_1, \dots, \mathbf{u}_{\ell-1}\}$, and put

$$\mathbf{p}_\ell = (I - P_\ell)\mathbf{e}_\ell, \quad \alpha_\ell = \langle \mathbf{u}_\ell, \mathbf{p}_\ell \rangle, \quad r_\ell = \alpha_\ell + (\alpha_\ell^2 + 1 - \|\mathbf{p}_\ell\|^2)^{1/2} \geq 0.$$

The vectors

$$\mathbf{y}_\ell = r_\ell \mathbf{u}_\ell - \mathbf{p}_\ell, \quad \mathbf{v}_\ell = \mathbf{y}_\ell + \mathbf{e}_\ell = r_\ell \mathbf{u}_\ell + P_\ell \mathbf{e}_\ell$$

satisfy $\|\mathbf{y}_\ell\|^2 = (r_\ell - \alpha_\ell)^2 - \alpha_\ell^2 + \|\mathbf{p}_\ell\|^2 = 1$, $\mathbf{y}_\ell \perp \mathbf{u}_1, \dots, \mathbf{u}_{\ell-1}$ (both $\mathbf{u}_\ell$ and $\mathbf{p}_\ell$ are orthogonal to these vectors), and $\mathbf{v}_{\ell'} \in \text{span}\{\mathbf{u}_1, \dots, \mathbf{u}_{\ell'}\}$. Hence $\langle \mathbf{y}_\ell, \mathbf{v}_{\ell'} \rangle = 0$ for $\ell' < \ell$.

Example: the errors. Let $\mathbf{o}^{(1)}, \dots, \mathbf{o}^{(B)}$ be independent $N(0, I_d)$ vectors, independent of Θ_1 , and write $U^{(\ell)}$ also for the orthogonal matrix with columns $U_i^{(\ell)}$. For $1 \leq p \leq d$ and $\mathbf{x} \in \mathbf{H}$ let $\mathcal{I}_p(\mathbf{x}) = \sum_n x_n \gamma_{p,n}$, where $\gamma_{p,n} = (U^{(n)} \mathbf{o}^{(n)})_p$ for $1 \leq n \leq B$ and the other $\gamma_{p,n}$ are independent $N(0, 1)$ variables, independent of everything else. Conditionally on Θ_1 all $\gamma_{p,n}$ are independent $N(0, 1)$, by rotation invariance, so

$$\text{Cov}(\mathcal{I}_p(\mathbf{x}), \mathcal{I}_{p'}(\mathbf{x}') \mid \Theta_1) = \mathbf{1}\{p = p'\} \langle \mathbf{x}, \mathbf{x}' \rangle.$$

At the query $t_{j_*} U_i^{(\ell)}$ set

$$\xi^o = \sigma \mathbf{o}_i^{(\ell)}, \quad \xi^c = \sigma \sum_p U_{pi}^{(\ell)} \mathcal{I}_p(\mathbf{y}_\ell),$$

and let all other errors vanish. Since $\mathbf{o}_i^{(\ell)} = \sum_p U_{pi}^{(\ell)} (U^{(\ell)} \mathbf{o}^{(\ell)})_p = \sum_p U_{pi}^{(\ell)} \mathcal{I}_p(\mathbf{e}_\ell)$, the observed error is $\nu_{\ell,i} = \sigma \sum_p U_{pi}^{(\ell)} \mathcal{I}_p(\mathbf{v}_\ell)$. Condition (O) holds, since the oblivious errors are functions of $\mathbf{o}^{(1)}, \dots, \mathbf{o}^{(B)}$. Conditionally on

Θ_1 , the error ξ^c at (ℓ, i) and the observed errors $\nu_{\ell', i'}$ of the earlier queries are jointly Gaussian with covariance $\sigma^2 \langle U_i^{(\ell)}, U_{i'}^{(\ell')} \rangle \langle \mathbf{y}_\ell, \mathbf{v}_{\ell'} \rangle$. This covariance vanishes for $\ell' < \ell$ by the orthogonality above, and for $\ell' = \ell, i' < i$ because $U_i^{(\ell)} \perp U_{i'}^{(\ell)}$. So, given $\mathcal{G}_t$, ξ^c is $N(0, \sigma^2)$: (C) holds with $\mathcal{G}_t$ in place of $\mathcal{E}_t$. With $\mathcal{E}_t$ it cannot hold for large B , since otherwise (ii) would apply.

Example: the second moment. Here $b(z) = 0$ and $\Delta = 0$, so $\zeta = \bar{g}$. Since $\sum_i U_{p'i}^{(\ell)} U_{pi}^{(\ell)} = \mathbf{1}\{p = p'\}$, the p -th coordinate of $\bar{g}$ equals $\frac{a_{j*} \sigma}{2B} \mathcal{I}_p(\mathbf{v}_\Sigma)$, where $\mathbf{v}_\Sigma = \sum_\ell \mathbf{v}_\ell \in \text{span}\{\mathbf{u}_1, \dots, \mathbf{u}_B\}$, and therefore $\mathbb{E} \|\zeta\|^2 = a_{j*}^2 \sigma^2 d \|\mathbf{v}_\Sigma\|^2 / (4B^2)$. The inner product $\langle \mathbf{v}_{\ell'}, \mathbf{u}_\ell \rangle$ equals 0 for $\ell' < \ell$, r_ℓ for $\ell' = \ell$, and $\langle \mathbf{e}_{\ell'}, \mathbf{u}_\ell \rangle = \vartheta_{\ell' - \ell}$ for $\ell' > \ell$, since then $P_{\ell'} \mathbf{u}_\ell = \mathbf{u}_\ell$. Hence, with $\sum_{n=1}^K (n + \frac{1}{2})^{-1} \geq \log \frac{2K+3}{3}$,

$$\langle \mathbf{v}_\Sigma, \mathbf{u}_\ell \rangle = r_\ell + \sum_{n=1}^{B-\ell} \vartheta_n \geq \frac{1}{\pi} \log \frac{2(B-\ell) + 3}{3}.$$

Keeping the at least $B/2$ indices with $2(B-\ell) + 3 \geq B + 2$, we obtain

$$\mathbb{E} \|\zeta\|^2 = \frac{a_{j*}^2 \sigma^2 d}{4B^2} \sum_{\ell=1}^B \langle \mathbf{v}_\Sigma, \mathbf{u}_\ell \rangle^2 \geq \frac{a_{j*}^2}{8\pi^2} \log^2 \left(\frac{B+2}{3} \right) \frac{d\sigma^2}{Bh^2}.$$

For any constant K the right-hand side exceeds $Kd\sigma^2/(Bh^2)$ once B is large, although $L_\beta = M = 0$.

Example: exact values. The lower bound above is crude; the norm is explicit. Since $\alpha_\ell = \vartheta_0 = 2/\pi$ and $\|\mathbf{p}_\ell\|^2 = 1 - \sum_{n=1}^{\ell-1} \vartheta_n^2$,

$$\|\mathbf{v}_\Sigma\|^2 = \sum_{\ell=1}^B \left(r_\ell + \sum_{n=1}^{B-\ell} \vartheta_n \right)^2, \quad r_\ell = \frac{2}{\pi} + \left(\frac{4}{\pi^2} + \sum_{n=1}^{\ell-1} \vartheta_n^2 \right)^{1/2}.$$

A direct computation gives $\|\mathbf{v}_\Sigma\|^2/B = 3.77, 5.59, 7.83, 10.47, 11.94$ for $B = 16, 64, 256, 1024, 2048$; it is 2 when the centered and oblivious errors are independent. For $m = 1$ the bound of (ii), $C_\sigma d\sigma^2/(Bh^2)$ with $C_\sigma = (1 + 2^{-1/2})^2$, is violated from $B = 1797$ on; the displayed lower bound shows this only for B of order 10^7 .

D.4 Lipschitz noise chosen after the frames

Remark D.5 (Lipschitz noise must not see the frames). Condition (L) requires η to be fixed before the run. Suppose instead that η may be chosen after Θ_k is revealed. We construct such an η that shifts the estimate of one frame by order $\sqrt{d}M$. Let $\varkappa_d = d \mathbb{E} |\langle U, e_1 \rangle| = d \Gamma(d/2) / (\sqrt{\pi} \Gamma((d+1)/2))$. The Cauchy–Schwarz inequality and Gautschi’s inequality (Gautschi, 1959) give $\sqrt{2d/\pi} \leq \varkappa_d \leq \sqrt{d}$.

Let $B = 1$, $\xi^c = \xi^o = 0$ and $\delta = 0$, fix a unit vector v , and let $\sigma_i = \text{sgn} \langle U_i, v \rangle$ (nonzero almost surely). On the finite set $\mathcal{P} = \{z + ht_j s U_i : 1 \leq i \leq d, 1 \leq j \leq m, s = \pm 1\}$ of query points put $\eta(z + x) = M \|x\| \text{sgn} \langle x, v \rangle$, that is, $\eta(z + ht_j s U_i) = M ht_j s \sigma_i$. This assignment is $\sqrt{2}M$ -Lipschitz on $\mathcal{P}$. For two points on the same line, $x = rU_i$ and $y = r'U_i$, the values differ by $M|r - r'| = M\|x - y\|$. For $x = rU_i$ and $y = r'U_{i'}$ with $i \neq i'$, orthogonality gives $\|x - y\| = (r^2 + r'^2)^{1/2}$, and the values differ by at most $M(|r| + |r'|) \leq \sqrt{2}M\|x - y\|$. The McShane extension (McShane, 1934) $\eta(x) = \min_{x' \in \mathcal{P}} (\eta(x') + \sqrt{2}M\|x - x'\|)$ is $\sqrt{2}M$ -Lipschitz on $\mathbb{R}^d$ and agrees with the assignment on $\mathcal{P}$. Since $\sum_j a_j t_j = 1$, we get $e_\eta(U_i) = M\sigma_i$. The f -part of the estimate still has mean $G_h(z)$, so

$$\mathbb{E} \bar{g} - G_h(z) = M \sum_i \mathbb{E} [\text{sgn} \langle U_i, v \rangle U_i] = M \varkappa_d v, \quad \|\mathbb{E} \bar{g} - G_h(z)\| \geq \sqrt{2/\pi} \sqrt{d} M.$$

For a $\sqrt{2}M$ -Lipschitz η fixed before the frames the same norm is at most $\sqrt{2}A_1 M$ by Lemma D.3(i). After rescaling η to be M -Lipschitz, the lower bound becomes $\sqrt{d/\pi} M$.

E LOWER BOUNDS

This appendix proves the lower bounds for Gaussian noise and for Lipschitz noise. Lemmas E.3 and E.4 give the bound of Theorem 1(b) for every budget (Section E.4), and Lemma E.5 the resulting minimax rate for normalized parameters. Lemma E.6 gives the factor d^2 for convex objectives used in Lemma N.3, and Lemma E.7 bounds the admissible level of Lipschitz noise in every ℓ_p geometry; its Euclidean case enters Theorems 1 and 5.

Algorithms and minimax risk. The lower bounds use independent $N(0, \sigma^2)$ noise or exact values and allow adaptive randomized algorithms with queries and output anywhere in $\mathbb{R}^d$. Such an algorithm draws a random seed Θ_0 at the start; each query is a measurable function of Θ_0 and of the earlier observations, and the output is a measurable function of Θ_0 and of all observations. Algorithms with rounds (Section 2) are the special case $\Theta_0 = (\Theta_1, \Theta_2, \dots)$, since the Gaussian noise is independent of all Θ_k . For a class $\mathcal{C}$ of functions on $\mathbb{R}^d$, the Gaussian minimax risk $\mathfrak{R}_T^G(\mathcal{C})$ is the infimum of $\sup_{f \in \mathcal{C}} \mathbb{E}[f(\hat{x}) - f^*]$ over these algorithms with at most T calls under Gaussian noise. It is nonincreasing in T , since extra calls can be ignored.

E.1 Two localized profiles

Fix a smooth $\chi : \mathbb{R} \rightarrow [0, 1]$ with $\chi(s) = 1$ for $s \leq 1/4$ and $\chi(s) = 0$ for $s \geq 1$, for instance $\chi(s) = \phi(1 - s)/(\phi(1 - s) + \phi(s - 1/4))$ with $\phi(t) = e^{-1/t} \mathbf{1}\{t > 0\}$; all derivatives of χ are bounded. For a unit vector v let

$$\psi_v(x) = \langle v, x \rangle \chi(\|x\|^2). \quad (\text{E.1})$$

Lemma E.1 (Radial profile). *Let $q \geq 2$, $\gamma \in (0, 1]$, $\beta = q + \gamma$. There are constants $K_\beta, K_2 \geq 1$, depending only on β and χ , such that for all d , all unit v and all x, y :*

$$\begin{aligned} \|D^q \psi_v(x) - D^q \psi_v(y)\|_{\text{op}} &\leq K_\beta \|x - y\|^\gamma, \\ \|D^q \psi_v(x) - D^q \psi_v(y)\|_{\text{HS}} &\leq K_\beta d^{\lfloor q/2 \rfloor / 2} \|x - y\|^\gamma, \\ \sup_x \|\nabla^2 \psi_v(x)\|_{\text{op}} &\leq K_2. \end{aligned}$$

Proof. Let $s(x) = \|x\|^2$, so that $Ds(x)[u] = 2\langle x, u \rangle$, $D^2s[u, w] = 2\langle u, w \rangle$ and $D^3s = 0$. By the chain rule, $D^j(\chi \circ s)$ is a sum, over the partitions of $\{1, \dots, j\}$ into i pairs and $j - 2i$ singletons, of index permutations of $2^{j-i} \chi^{(j-i)}(\|x\|^2) x^{\otimes(j-2i)} \otimes I^{\otimes i}$. These terms vanish unless $\|x\| \leq 1$, and by Lemma B.1 a term with i pairs has Hilbert–Schmidt norm at most $C_j d^{i/2}$ and operator norm at most C_j , where C_j depends only on j and χ .

By the product rule, $D^q \psi_v$ is the sum of $\langle v, x \rangle D^q(\chi \circ s)$ and of the terms in which v fills one slot of $D^{q-1}(\chi \circ s)$; each of its terms has at most $\lfloor q/2 \rfloor$ identity factors. A further derivative in a unit direction e replaces $\langle v, x \rangle$ by $\langle v, e \rangle$, turns a coefficient $\chi^{(r)}(\|x\|^2)$ into $2\langle x, e \rangle \chi^{(r+1)}(\|x\|^2)$, or replaces a factor x by e ; none of these creates an identity factor. Hence, with B_q, E_q depending only on q and χ ,

$$\sup_x \|D^q \psi_v(x)\|_{\text{HS}} \leq B_q d^{\lfloor q/2 \rfloor / 2}, \quad \sup_{x, \|e\|=1} \|D^{q+1} \psi_v(x)[e]\|_{\text{HS}} \leq E_q d^{\lfloor q/2 \rfloor / 2}.$$

Integration along segments gives $\|D^q \psi_v(x) - D^q \psi_v(y)\|_{\text{HS}} \leq d^{\lfloor q/2 \rfloor / 2} \min\{2B_q, E_q\} \|x - y\|$, and $\min\{A, Bt\} \leq A^{1-\gamma} (Bt)^\gamma$ gives the Hilbert–Schmidt bound. On unit vectors every factor $\langle x, u \rangle$ or $\langle u, w \rangle$ of these tensors is at most one in absolute value, which gives the operator bound and the Hessian bound without a power of d . $\square$

The second profile replaces the linear factor by a narrow ridge. Let $\vartheta : \mathbb{R} \rightarrow \mathbb{R}$ be smooth and odd, equal to its argument on $[-1/2, 1/2]$ and vanishing outside $[-1, 1]$ (for instance $\vartheta(s) = s\chi(s^2)$), and let $L_\vartheta = \sup |\vartheta'|$. For a unit vector v put $\tilde{\psi}_v(x) = \vartheta(d^{1/4} \langle v, x \rangle) \chi(\|x\|^2)$, a ridge of width $d^{-1/4}$.

Lemma E.2 (Ridge profile). *For every integer $j \geq 0$ there is $\tilde{K}_j$, independent of d and v , with $\sup_x \|D^j \tilde{\psi}_v(x)\|_{\text{HS}} \leq \tilde{K}_j d^{j/4}$. Consequently*

$$\|D^q \tilde{\psi}_v(x) - D^q \tilde{\psi}_v(y)\|_{\text{HS}} \leq \tilde{K}_\beta d^{\beta/4} \|x - y\|^\gamma, \quad \tilde{K}_\beta = (2\tilde{K}_q)^{1-\gamma} \tilde{K}_{q+1}^\gamma.$$

Proof. In a product-rule term of order j let l derivatives fall on the ridge, giving $d^{l/4} \vartheta^{(l)}(d^{1/4} \langle v, x \rangle) v^{\otimes l}$, and the rest on $\chi(\|x\|^2)$, giving terms with i identity factors and Hilbert–Schmidt norm $O(d^{i/2})$ as in the proof of Lemma E.1, where $l + 2i \leq j$. Since $d^{l/4} d^{i/2} \leq d^{j/4}$, each term is $O(d^{j/4})$. The Hölder bound follows as there, from $\min\{2\tilde{K}_q d^{q/4}, \tilde{K}_{q+1} d^{(q+1)/4} t\} \leq (2\tilde{K}_q d^{q/4})^{1-\gamma} (\tilde{K}_{q+1} d^{(q+1)/4} t)^\gamma$. $\square$

E.2 The radial family

Let $0 < \alpha < L$, $\rho = \min\{\alpha, (L - \alpha)/2\}$ and $\mu = \alpha + \rho$, and let $\nu = 0$ for the operator seminorm and $\nu = \lfloor q/2 \rfloor$ for the Hilbert–Schmidt seminorm. For $\omega \in \{-1, 1\}^d$ put $v_\omega = \omega/\sqrt{d}$, and for $a, h > 0$ let

$$f_\omega(x) = \frac{\mu}{2} \|x\|^2 + ah\psi_{v_\omega}(x/h) = \frac{\mu}{2} \|x\|^2 + a \langle v_\omega, x \rangle \chi(\|x\|^2/h^2), \quad a = \frac{L_\beta h^{\beta-1}}{K_\beta d^{\nu/2}}. \quad (\text{E.2})$$

Theorem E.3 (Radial lower bound). *Let $\beta > 2$, $L_\beta, R, \sigma > 0$ and $0 < \alpha < L$, let the class $\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)$ be defined with the operator or the Hilbert–Schmidt seminorm, and let ρ and ν be as above. Let K_β, K_2 be the constants of Lemma E.1, $\bar{K}_2 = \max\{1, K_2\}$ and*

$$H = \min\left\{\left(\frac{\rho K_\beta d^{\nu/2}}{\bar{K}_2 L_\beta}\right)^{\frac{1}{\beta-2}}, \left(\frac{\alpha R K_\beta d^{\nu/2}}{L_\beta}\right)^{\frac{1}{\beta-1}}\right\}. \quad (\text{E.3})$$

For every $T \geq 1$,

$$\mathfrak{R}_T^G(\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)) \geq \min\left\{c_\beta \frac{L_\beta^{2/\beta}}{\alpha} d^{2-\frac{\nu+2}{\beta}} \left(\frac{\sigma^2}{T}\right)^{\frac{\beta-1}{\beta}}, \frac{3a_H^2}{64\alpha}\right\}, \quad c_\beta = \frac{3}{64} 16^{-\frac{\beta-1}{\beta}} K_\beta^{-\frac{2}{\beta}}, \quad (\text{E.4})$$

where $a_H = L_\beta H^{\beta-1}/(K_\beta d^{\nu/2}) = \min\{(\rho/\bar{K}_2)^{\frac{\beta-1}{\beta-2}} (K_\beta d^{\nu/2}/L_\beta)^{\frac{1}{\beta-2}}, \alpha R\}$. The first term is the smaller one if and only if $T \geq K_\beta^2 d^{\nu+2} \sigma^2 / (16 L_\beta^2 H^{2\beta})$.

Proof. We show that the functions (E.2) belong to the class, bound the error from below by the number of wrong signs of an estimate of ω , and apply Assouad’s lemma (B.6).

Step 1: membership. By Lemma E.1 and scaling, the q -th derivative of the perturbation $ah\psi_{v_\omega}(\cdot/h)$ has Hölder constant at most $K_\beta d^{\nu/2} ah^{1-\beta} = L_\beta$ in the chosen norm, and the quadratic does not contribute since $q \geq 2$; hence $[f_\omega]_\beta \leq L_\beta$. Choose

$$h = \min\{H, h_\star\}, \quad h_\star^{2\beta} = \frac{K_\beta^2 d^{\nu+2} \sigma^2}{16 L_\beta^2 T}, \quad \text{so that} \quad a^2 h^2 \leq \frac{d^2 \sigma^2}{16 T}. \quad (\text{E.5})$$

Since $h \leq H$, we have $\bar{K}_2 a/h \leq \rho$ and $a \leq \alpha R$. The Hessian of the perturbation has operator norm at most $K_2 a/h \leq \rho$, so $\alpha I \preceq \nabla^2 f_\omega \preceq (\alpha + 2\rho)I \preceq LI$. The point $x_\omega^\star = -(a/\mu)v_\omega$ has norm $a/\mu \leq \rho h/\mu \leq h/2$, so $\chi = 1$ and $\chi' = 0$ at $\|x_\omega^\star\|^2/h^2$, and $\nabla f_\omega(x_\omega^\star) = \mu x_\omega^\star + av_\omega = 0$. By strong convexity $x_\omega^\star$ is the minimizer, and $\|x_\omega^\star\| \leq a/\alpha \leq R$. Thus $f_\omega \in \mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)$.

Step 2: from the error to signs. For an output $\hat{x}$ let $\hat{\omega}_j = -\text{sgn}(\hat{x}_j)$. A wrong sign means $|\hat{x}_j - x_{\omega,j}^\star| \geq a/(\mu\sqrt{d})$, and strong convexity with $\mu \leq 2\alpha$ gives

$$f_\omega(\hat{x}) - f_\omega(x_\omega^\star) \geq \frac{\alpha}{2} \|\hat{x} - x_\omega^\star\|^2 \geq \frac{a^2}{8\alpha d} \sum_{j=1}^d \mathbf{1}\{\hat{\omega}_j \neq \omega_j\}. \quad (\text{E.6})$$

Step 3: information. Let P_ω be the law of the seed and the transcript under f_ω , let $\omega^{(j)}$ be ω with its j -th coordinate flipped, and let $X_{t,j}$ be the j -th coordinate of the t -th query X_t . By Lemma B.5(b),

$$\text{KL}(P_\omega \| P_{\omega^{(j)}}) = \frac{2a^2}{d\sigma^2} \mathbb{E}_\omega \sum_t X_{t,j}^2 \chi(\|X_t\|^2/h^2)^2.$$

Since $\sum_j X_{t,j}^2 \chi(\|X_t\|^2/h^2)^2 \leq h^2$ for every query,

$$\frac{1}{d2^d} \sum_{\omega,j} \text{KL}(P_\omega \| P_{\omega^{(j)}}) \leq \frac{2Ta^2 h^2}{d^2 \sigma^2}.$$

Pinsker’s and Jensen’s inequalities bound the average total variation by $ah\sqrt{T}/(d\sigma) \leq 1/4$. By (B.6) the average number of wrong signs is at least $3d/8$, and (E.6) gives

$$\max_\omega \mathbb{E}_\omega [f_\omega(\hat{x}) - f_\omega(x_\omega^\star)] \geq 3a^2/(64\alpha).$$

Since $a = L_\beta h^{\beta-1}/(K_\beta d^{\nu/2})$ is increasing in h , substituting $h = h_\star$ and $h = H$ gives the two terms of (E.4), and $h_\star \leq H$ is the stated condition on T . $\square$

E.3 The ridge family

Theorem E.4 (Ridge lower bound). *Let $\beta > 2$, $L_\beta, R, \sigma > 0$ and $0 < \alpha < L$, and let the class $\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)$ be defined with the Hilbert–Schmidt seminorm. Let $\rho = \min\{\alpha, (L - \alpha)/2\}$, let $\tilde{K}_\beta, \tilde{K}_2$ be the constants of Lemma E.2, $L_\vartheta = \sup |\vartheta'|$, $\tilde{K}'_2 = \max\{1, \tilde{K}_2\}$ and*

$$H_0 = \min\{(\rho \tilde{K}_\beta / (\tilde{K}'_2 L_\beta))^{1/(\beta-2)}, (\alpha R \tilde{K}_\beta / L_\beta)^{1/(\beta-1)}\}.$$

For every $T \geq 1$,

$$\mathfrak{R}_T^G(\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)) \geq \min\left\{\tilde{c}_\beta \frac{L_\beta^{2/\beta}}{\alpha} d^{\frac{3}{2} - \frac{3}{2\beta}} \left(\frac{\sigma^2}{T}\right)^{\frac{\beta-1}{\beta}}, \frac{3\tilde{a}_H^2}{64\alpha}\right\}, \quad \tilde{c}_\beta = \frac{3}{64}(16L_\vartheta^2)^{-\frac{\beta-1}{\beta}} \tilde{K}_\beta^{-\frac{2}{\beta}},$$

where $\tilde{a}_H = L_\beta H_0^{\beta-1} / \tilde{K}_\beta = \min\{(\rho / \tilde{K}'_2)^{\frac{\beta-1}{\beta-2}} (\tilde{K}_\beta / L_\beta)^{\frac{1}{\beta-2}}, \alpha R\}$. The first term is the smaller one if and only if $T \geq \tilde{K}_\beta^2 d^{3/2} \sigma^2 / (16L_\vartheta^2 L_\beta^2 H_0^{2\beta})$.

Proof. We follow the proof of Lemma E.3; only the functions, their membership and the information bound change. With μ, v_ω as there, $h > 0$ and the ridge width $w = h d^{-1/4}$, let

$$\tilde{f}_\omega(x) = \frac{\mu}{2} \|x\|^2 + A \vartheta\left(\frac{\langle v_\omega, x \rangle}{w}\right) \chi(\|x\|^2 / h^2), \quad A = \frac{L_\beta w^\beta}{\tilde{K}_\beta}, \quad a = \frac{A}{w}.$$

Membership. The perturbation equals $A \tilde{\psi}_{v_\omega}(x/h)$. By Lemma E.2, its q -th derivative has Hilbert–Schmidt Hölder constant at most $A h^{-\beta} \tilde{K}_\beta d^{\beta/4} = A \tilde{K}_\beta w^{-\beta} = L_\beta$, so $[\tilde{f}_\omega]_{\beta, \text{HS}} \leq L_\beta$, and its Hessian has operator norm at most $A h^{-2} \tilde{K}_2 d^{1/2} = \tilde{K}_2 a / w$. Choose $w = \min\{H_0, w_\star\}$ with

$$w_\star^{2\beta} = \tilde{K}_\beta^2 d^{3/2} \sigma^2 / (16L_\vartheta^2 L_\beta^2 T).$$

Since $w \leq H_0$, we have $\tilde{K}'_2 a / w \leq \rho$ and $a \leq \alpha R$, hence $\alpha I \preceq \nabla^2 \tilde{f}_\omega \preceq L I$. The point $-(a/\mu)v_\omega$ has norm at most $w/2$; there $\vartheta' = 1$, $\chi = 1$ and $\chi' = 0$, so the gradient vanishes, and this point is the minimizer, of norm at most $a/\alpha \leq R$. Thus $\tilde{f}_\omega \in \mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)$, and (E.6) holds with $a = A/w$.

Information. Since ϑ is L_ϑ -Lipschitz,

$$\left| \tilde{f}_\omega(x) - \tilde{f}_{\omega^{(j)}}(x) \right| \leq \frac{2L_\vartheta a}{\sqrt{d}} |x_j| \chi(\|x\|^2 / h^2),$$

and Step 3 of the proof of Lemma E.3 bounds the average total variation by $L_\vartheta a h \sqrt{T} / (d\sigma)$. The choice of w gives $L_\vartheta^2 a^2 h^2 \leq d^2 \sigma^2 / (16T)$, so this bound is at most $1/4$, and the error is at least $3a^2 / (64\alpha)$ with $a^2 = L_\beta^2 w^{2\beta-2} / \tilde{K}_\beta^2$; the two values $w = w_\star$ and $w = H_0$ give the two terms. $\square$

E.4 Consequences for the strongly convex and convex classes

Let $\beta > 2$, $L_\beta, R, \sigma > 0$ and $0 < \alpha < L$, let the seminorm be either one, let τ be as in (2) and $\rho = \min\{\alpha, (L - \alpha)/2\}$, and put

$$\Phi_0 = \min\left\{\alpha R^2, \frac{\rho^{2(\beta-1)/(\beta-2)}}{\alpha L_\beta^{2/(\beta-2)}}\right\}, \quad T_{\text{lb}} = T_\star(\Phi_0) = \frac{d^{2-\tau/(\beta-1)} \sigma^2 L_\beta^{2/(\beta-1)}}{(\alpha \Phi_0)^{\beta/(\beta-1)}}, \quad (\text{E.7})$$

so that $\Phi_T \leq \Phi_0$ if and only if $T \geq T_{\text{lb}}$. For fixed $\beta, \alpha, L, R, L_\beta, \sigma$, T_{lb} is of order $d^{2-\tau/(\beta-1)} \leq d^2$.

Proof of Theorem 1(b). In the operator case and in the Hilbert–Schmidt case with odd q , Lemma E.3 with $\nu = \tau$ applies, because then $\tau = \lfloor q/2 \rfloor$ or $\tau = 0$ and its exponent $2 - (\nu + 2)/\beta$ is that of (3); in the Hilbert–Schmidt case with even q , Lemma E.4 applies, because $\tau_\beta = (\beta - 1)/2$ and $\frac{3}{2} - \frac{3}{2\beta} = 2 - (2 + \tau_\beta)/\beta$. Since $d^{\nu/2} \geq 1$, both a_H^2/α and $\tilde{a}_H^2/\alpha$ are at least a constant depending only on β times Φ_0 . Hence $\mathfrak{R}_T^G(\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)) \geq c_\beta \min\{\Phi_T, \Phi_0\}$ for every $T \geq 1$. The hard functions satisfy $\alpha I \preceq \nabla^2 f \preceq (\alpha + 2\rho)I \preceq 3\alpha I$; for $L = 3\alpha$ one has $\rho = \alpha$, and the bound holds on $\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, 3\alpha, R)$. For the term M^2/α , Lemma E.7(b) with $p = 2$ and the same Gaussian errors gives $M^2/(2\alpha)$: its quadratic instances have Hessian αI , zero Hölder seminorm and minimizer norm $M/\alpha \leq R$. The maximum of the two bounds is at least half their sum. $\square$

With the parameters normalized, the two bounds of Theorem 1 match for every dimension and every budget.

Corollary E.5 (Normalized minimax rate). *Let $\beta > 2$, $\alpha = 1$, $L = 3$, $R = L_\beta = \sigma = 1$, let the seminorm be either one, and put $p = (\beta - 1)/\beta$. There are constants $0 < c_\beta \leq C_\beta$ such that for all integers $d \geq 2$ and $T \geq 1$*

$$c_\beta \min\left\{1, \left(\frac{d^s}{T}\right)^p\right\} \leq \mathfrak{R}_T^G(\mathcal{F}_{\text{sc}}(\beta, 1; 1, 3, 1)) \leq C_\beta \min\left\{1, \left(\frac{d^s}{T}\right)^p\right\},$$

and the upper bound is attained under every noise admissible at levels $(0, 1, 0)$. Consequently there is $\varepsilon_\beta > 0$ such that for $0 < \varepsilon \leq \varepsilon_\beta$ the least number of calls of an ε -accurate method under Gaussian noise is of order $d^s \varepsilon^{-\beta/(\beta-1)}$, uniformly in $d \geq 2$. Here $s = 2$ for the operator seminorm, $s = \frac{3}{2}$ under Hilbert–Schmidt smoothness with even q , and $\frac{3}{2} < s \leq \frac{5}{3}$ with odd q .

Proof. Here $\rho = 1$, $\Phi_0 = 1$ and $\Phi_T = d^{2-(2+\tau)/\beta} T^{-p} = (d^s/T)^p$, so the lower bound is Theorem 1(b). For the upper bound, $\Phi_1 \geq 1$, so the budget condition of Theorem 1(a) holds if $T \geq D_\beta d \log(2d)$ for a constant D_β depending only on β , and in particular if $T \geq D'_\beta d^s$, since $s \geq \frac{3}{2}$; then the error is at most $C_\beta \Phi_T$ and $\Phi_T \leq 1$. For $T < D'_\beta d^s$ the method that returns 0 without calls has error $f(0) - f^* \leq \frac{3}{2} \|x^*\|^2 \leq \frac{3}{2}$, while $\min\{1, (d^s/T)^p\} \geq (D'_\beta)^{-p}$. For the number of calls, take $\varepsilon_\beta = c_\beta/2$: an error $\varepsilon \leq \varepsilon_\beta$ excludes the first term of the lower bound and forces $T \geq d^s (c_\beta/\varepsilon)^{1/p}$, while $T = \lceil d^s (C_\beta/\varepsilon)^{1/p} \rceil \leq 2d^s (C_\beta/\varepsilon)^{1/p}$ oracle calls suffice. For odd q , $\tau = (q-1)/2$ and $q < \beta \leq q+1$ give $\frac{3}{2} < s \leq 2 - \frac{q-1}{2q} \leq \frac{5}{3}$. $\square$

Corollary E.6 (Convex objectives: the factor d^2). *Let $\beta > 2$ and $L_\beta, R, \sigma > 0$, and let $\mathcal{C}$ be the class of convex q times continuously differentiable functions f on $\mathbb{R}^d$ with $[f]_{\beta, \text{op}} \leq L_\beta$ and a minimizer in $B(R)$. There are $c_\beta, c'_\beta > 0$ such that every algorithm whose expected error is at most ε for every $f \in \mathcal{C}$ under independent $N(0, \sigma^2)$ noise, where $\varepsilon \leq c'_\beta L_\beta R^\beta$, uses $T \geq c_\beta d^2 \sigma^2 / \varepsilon^2$ calls. For the Hilbert–Schmidt seminorm (that is, with $[f]_{\beta, \text{HS}} \leq L_\beta$ in the definition of $\mathcal{C}$) the same holds for $\varepsilon \leq c'_\beta L_\beta R^\beta d^{-\lfloor q/2 \rfloor / 2}$. All hard functions have $\nabla^2 f \preceq C_\beta L_\beta R^{\beta-2} I$.*

For the operator seminorm and $d \geq 256$, Theorem 5(a) is stronger, up to constants by the factor $\min\{d^{1/2}, (L_\beta R^\beta / \varepsilon)^{2/\beta}\}$. Lemma E.6 also covers $d < 256$ and the Hilbert–Schmidt seminorm, and it is the bound transferred in Lemma N.3(b).

Proof of Lemma E.6. The functions. Take ν as in Lemma E.3 and the curvature $\mu = 2\bar{K}_2 a/h$ in (E.2). The perturbation has Hessian of operator norm at most $K_2 a/h \leq \mu/2$, so $(\mu/2)I \preceq \nabla^2 f_\omega \preceq (3\mu/2)I$; in particular f_ω is convex. As in Step 1 of the proof of Lemma E.3, the point $-(a/\mu)v_\omega$, of norm $h/(2\bar{K}_2) \leq h/2$, is the minimizer; its norm is at most R if $h \leq 2\bar{K}_2 R$. Then also $a/h = L_\beta h^{\beta-2}/(K_\beta d^{\nu/2}) \leq L_\beta (2\bar{K}_2 R)^{\beta-2}/K_\beta$, which gives the Hessian bound with $C_\beta = 3\bar{K}_2 (2\bar{K}_2)^{\beta-2}/K_\beta$. As in (E.6), strong convexity with modulus $\mu/2$ gives

$$f_\omega(\hat{x}) - \min f_\omega \geq \frac{a^2}{4\mu d} \sum_j \mathbf{1}\{\hat{\omega}_j \neq \omega_j\}.$$

Large budgets. Choose h by $ah = d\sigma/(4\sqrt{T})$, that is, $h^\beta = K_\beta d^{\nu/2+1}\sigma/(4L_\beta\sqrt{T})$. Then $h \leq 2\bar{K}_2 R$ if and only if $T \geq T_1 := (K_\beta d^{\nu/2+1}\sigma/(4L_\beta(2\bar{K}_2 R)^\beta))^2$. For such T , Step 3 of the proof of Lemma E.3 bounds the average total variation by $1/4$, and

$$\mathfrak{R}_T^G(\mathcal{C}) \geq \frac{3a^2}{32\mu} = \frac{3ah}{64\bar{K}_2} = \frac{3d\sigma}{256\bar{K}_2\sqrt{T}}.$$

All budgets. Put $\varepsilon_1 = 3d\sigma/(256\bar{K}_2\sqrt{T_1}) = c''_\beta L_\beta R^\beta d^{-\nu/2}$ with $c''_\beta = 3(2\bar{K}_2)^\beta/(64\bar{K}_2 K_\beta)$. An algorithm with $T < T_1$ calls has at most T_1 calls, so by monotonicity its worst-case error is at least $\mathfrak{R}_{T_1}^G(\mathcal{C}) \geq \varepsilon_1$. Hence worst-case error $\varepsilon < \varepsilon_1$ forces $T \geq T_1$, and then $\varepsilon \geq 3d\sigma/(256\bar{K}_2\sqrt{T})$, that is, $T \geq (3/(256\bar{K}_2))^2 d^2 \sigma^2 / \varepsilon^2$. The corollary follows with $c_\beta = (3/(256\bar{K}_2))^2$ and $c'_\beta = c''_\beta/2$. $\square$

E.5 Lipschitz noise

The hard instances for Lipschitz noise are pairs (f_+, η_+) , (f_-, η_-) with $f_+ + \eta_+ = f_- + \eta_-$: a Lipschitz noise cannot be distinguished from a Lipschitz perturbation of the objective. Here $1 \leq p \leq \infty$, the Lipschitz noise is measured in $\|\cdot\|_p$, that is, $|\eta(x) - \eta(y)| \leq M \|x - y\|_p$, and $B_p(R) = \{x : \|x\|_p \leq R\}$.

Proposition E.7 (Lipschitz noise: lower bounds). *Let $1 \leq p \leq \infty$, $\beta > 2$, $L_\beta \geq 0$, $R > 0$ and $M > 0$. Let the centered, oblivious and adversarial components of the noise be given by a rule that does not depend on (f, η) , for instance exact values, or independent $N(0, \sigma^2)$ noise and no other component. Consider any algorithm, with any number of calls and rounds and with queries and output anywhere in $\mathbb{R}^d$.*

- (a) (Convex objectives.) *Let $L > 0$ and $M \leq LRd^{-|1-2/p|}$. There are convex quadratic functions f_+, f_- with $0 \leq \nabla^2 f_\pm(x)[v, v] \leq L \|v\|_p^2$ for all x, v and minimizers in $B_p(R)$, and linear functions $\eta_\pm$ with $|\eta_\pm(x) - \eta_\pm(y)| \leq M \|x - y\|_p$, such that*

$$\max_{\pm} \mathbb{E}[f_\pm(\hat{x}) - \min f_\pm] \geq \frac{MR}{2}.$$

- (b) (Strongly convex objectives.) *Let $\alpha > 0$, $M \leq \alpha R$ and $L \geq \alpha d^{2/p-1}$. There are quadratic functions f_+, f_- with $\alpha \|v\|_p^2 \leq \nabla^2 f_\pm(x)[v, v] \leq L \|v\|_p^2$ for all x, v and minimizers in $B_p(R)$, and linear functions $\eta_\pm$ with $|\eta_\pm(x) - \eta_\pm(y)| \leq M \|x - y\|_p$, such that*

$$\max_{\pm} \mathbb{E}[f_\pm(\hat{x}) - \min f_\pm] \geq \frac{M^2}{2\alpha}.$$

For $p = 2$ the pair is $f_\pm(x) = \frac{\alpha}{2} \|x \mp (M/\alpha)v\|^2$, $\eta_\pm(x) = \pm M \langle v, x \rangle$ with a unit vector v .

In both parts $D^q f_\pm$ is constant, so $[f_\pm]_{\beta,p} = [f_\pm]_{\beta,\text{op}} = [f_\pm]_{\beta,\text{HS}} = 0$, where $[f]_{\beta,p}$ is the Hölder seminorm of $D^q f$ with respect to $\|\cdot\|_p$ (Section [N](#)). Consequently, if the Lipschitz noise is measured in $\|\cdot\|_p$ and $\sigma, \delta \geq 0$ are arbitrary, then for every number of calls T :

- if $\varepsilon < LR^2 d^{-|1-2/p|}/2$, then $M_{\max}(\varepsilon, T) \leq 2\varepsilon/R$ on every class that contains the convex quadratic functions f with $\nabla^2 f[v, v] \leq L \|v\|_p^2$ for all v and a minimizer in $B_p(R)$;
- if $L \geq \alpha d^{2/p-1}$ and $\varepsilon < \alpha R^2/2$, then $M_{\max}(\varepsilon, T) \leq \sqrt{2\alpha\varepsilon}$ on every class that contains the quadratic functions f with $\alpha \|v\|_p^2 \leq \nabla^2 f[v, v] \leq L \|v\|_p^2$ for all v and a minimizer in $B_p(R)$.

Such classes are the convex q times continuously differentiable f with $[f]_{\beta,p} \leq L_\beta$, $\nabla^2 f(x)[v, v] \leq L \|v\|_p^2$ for all x, v and a minimizer in $B_p(R)$ in the first case, the analogous class with $\alpha \|v\|_p^2 \leq \nabla^2 f(x)[v, v] \leq L \|v\|_p^2$ in the second case, and, for $p = 2$, $\mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$ and $\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)$ with either seminorm.

Proof. Two indistinguishable instances. For $\mu > 0$ and $w \in \mathbb{R}^d$ put $f_\pm(x) = \frac{\mu}{2} \|x \mp w\|^2$ and $\eta_\pm(x) = \pm \mu \langle w, x \rangle$. Then $f_\pm + \eta_\pm = \frac{\mu}{2} (\|x\|^2 + \|w\|^2)$ for both signs. As the remaining noise does not depend on (f, η) , induction over the queries shows that the joint law of all queries, observations and the output $\hat{x}$ is the same under both instances. Since $\min f_\pm = f_\pm(\pm w) = 0$, the parallelogram law gives

$$\frac{1}{2} \sum_{\pm} \mathbb{E} f_\pm(\hat{x}) = \frac{\mu}{2} \mathbb{E} [\|\hat{x}\|^2 + \|w\|^2] \geq \frac{\mu}{2} \|w\|^2.$$

The functions $f_\pm$ have Hessian μI and constant $D^q f_\pm$, and by Hölder's inequality $\eta_\pm$ is $\mu \|w\|_{p^*}$ -Lipschitz with respect to $\|\cdot\|_p$, where $p^* = p/(p-1)$.

Choice of μ and w . With $\mathbf{1} = (1, \dots, 1)$, take

	$p \leq 2$	$p \geq 2$
(a)	$w = Rd^{-1/p} \mathbf{1}, \quad \mu = Md^{2/p-1}/R$	$w = Re_1, \quad \mu = M/R$
(b)	$w = (M/\alpha)d^{-1/p} \mathbf{1}, \quad \mu = \alpha d^{2/p-1}$	$w = (M/\alpha)e_1, \quad \mu = \alpha$

In each case $\mu \|w\|_{p^*} = M$; moreover $\|w\|_p = R$ and $\mu \|w\|^2/2 = MR/2$ in (a), and $\|w\|_p = M/\alpha \leq R$ and $\mu \|w\|^2/2 = M^2/(2\alpha)$ in (b). Since $\|v\| \leq \|v\|_p \leq d^{1/p-1/2} \|v\|$ for $p \leq 2$ and $d^{1/p-1/2} \|v\| \leq \|v\|_p \leq \|v\|$ for $p \geq 2$, the assumption on M in (a) gives $\mu \|v\|^2 \leq L \|v\|_p^2$, and the assumption on L in (b) gives $\alpha \|v\|_p^2 \leq \mu \|v\|^2 \leq L \|v\|_p^2$.

Admissible levels. Let the centered, oblivious and adversarial components vanish; this is admissible at levels (M, σ, δ) for all $\sigma, \delta \geq 0$. In the convex case let $M > 2\varepsilon/R$ and $M' = \min\{M, LRd^{-|1-2/p|}\}$. Then $M'R/2 > \varepsilon$ because $\varepsilon < LR^2d^{-|1-2/p|}/2$, an M' -Lipschitz noise is admissible at level M , and the functions of (a) with M' belong to every class of the first kind; so no method is ε -accurate at level M on such a class. The strongly convex case is the same, with $M > \sqrt{2\alpha\varepsilon}$, $M' = \min\{M, \alpha R\}$ and (b). $\square$

For $p = 2$ the bound in (b) is attained by the exact minimizer of the observed function: if $\hat{x}$ minimizes $f + \eta$, then $\|\nabla f(\hat{x})\| \leq M$ because η is M -Lipschitz, and strong convexity gives $f(\hat{x}) - f^* \leq M^2/(2\alpha)$.

F STRONGLY CONVEX OBJECTIVES: ORACLE CALLS, ROUNDS AND LIPSCHITZ NOISE

This appendix proves the upper bound of Theorem 1(a) and the achievable levels in Theorem 1(c). Both methods of Procedure 2 call the estimate of Procedure 1 once per round. The two-phase method is analyzed in Lemma F.3; it is projected onto $B(R)$ and uses an upper bound on the Hessian only there, which Section I needs. The accelerated method is analyzed in Section F.5. The optimization lemmas hold for any oracle with a bounded bias and conditionally centered noise.

Procedure 2 Strongly convex methods with random-frame estimates

Require: radius h ; α, L, R ; $\kappa = L/\alpha$; choose (a) with N, B , or (b) with $J, (B_j)_{j=1}^J$ from Lemma F.6

- 1: **(a) two phases** (N even): $z_1 = 0, k_0 = 16\kappa, P = \text{Proj}_{B(R)}$ $\triangleright$ one round each
 - 2: **for** $k = 1, \dots, N$ **do**
 - 3: step $\nu_k = \frac{1}{4L}$ if $k \leq \frac{N}{2}$, else $\nu_k = \frac{4}{\alpha(k-N/2+k_0)}$
 - 4: $z_{k+1} = P(z_k - \nu_k \bar{g}_{h,B}(z_k))$
 - 5: **return** $\hat{x} = \sum_{k=N/2+1}^N w_k z_k / \sum_{k=N/2+1}^N w_k, w_k = k - \frac{N}{2} + k_0 - 1$
 - 6: **(b) accelerated** (J stages): $x^{(0)} = 0, \theta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, N_0 = \lceil 5\sqrt{\kappa} \rceil$
 - 7: **for** stages $j = 1, \dots, J$ **do** $\triangleright$ the number of frames B_j doubles
 - 8: $x_{-1} = x_0 = x^{(j-1)}$
 - 9: **for** $k = 0, \dots, N_0 - 1$ **do**
 - 10: $y_k = x_k + \theta(x_k - x_{k-1}), x_{k+1} = y_k - \bar{g}_{h,B_j}(y_k)/L$
 - 11: $x^{(j)} = x_{N_0}$
 - 12: **return** $\hat{x} = x^{(J)}$
-

Setting. We use the class $\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)$ with either seminorm, $\kappa = L/\alpha$, and the noise of Definition 1. The methods also know d, T and the stencil constants. Since $h_T^{2\beta} = d^{2+\tau} \sigma^2 / (L_\beta^2 T)$ by (3),

$$L_\beta^2 d^{-\tau} h_T^{2\beta-2} = \frac{d^2 \sigma^2}{T h_T^2} = \alpha \Phi_T, \quad (\text{F.1})$$

and $\Phi_T = \Phi_1 T^{-(\beta-1)/\beta}$, where Φ_1 is the value of Φ_T at $T = 1$.

The estimate. Consider round k of Procedure 1 with an $\mathcal{F}_{k-1}$ -measurable center z , a deterministic radius $h > 0$ and $B \geq 1$ frames; it makes $2mdB$ calls. By Lemma D.3, it returns $\bar{g} = \nabla f(z) + b(z) + \zeta + \Delta$, where:

- (E1) $b(z)$ is $\mathcal{F}_{k-1}$ -measurable and $\|b(z)\| \leq b_f + b_\eta$, with $b_f := C_b L_\beta d^{-\tau/2} h^{\beta-1}$ and $b_\eta := A_1 M$;
- (E2) $\mathbb{E}[\zeta \mid \mathcal{F}_{k-1}] = 0$ and $\mathbb{E}[\|\zeta\|^2 \mid \mathcal{F}_{k-1}] \leq v_h^2/B$, with $v_h^2 := C_f L_\beta^2 d^{-\tau} h^{2\beta-2} + C_\eta M^2 + C_\sigma d \sigma^2 / h^2$ (also for $d = 1$);

(E3) $\|\Delta\| \leq b_\delta := \sqrt{d} S_1 \delta / h$ almost surely.

The constants are those of Lemma D.3. Only the sum $\zeta + \Delta$ is $\mathcal{F}_k$ -measurable. As explained after Lemma D.3, we use ζ only through (E2) and through $\mathbb{E}\langle \zeta, w \rangle = 0$ for square-integrable $\mathcal{F}_{k-1}$ -measurable random vectors w .

F.1 Projected descent in two phases

The first phase contracts the initial distance with a constant step, and the second uses decreasing steps and weighted averaging, as in Lacoste-Julien et al. (2012, Section 3.2), with an offset that keeps the steps small. The following computation is shared by all two-phase methods of the paper.

Lemma F.1 (Two-phase recursion). *Let $a, c > 0$, $\mathcal{B}, \mathcal{V} \geq 0$ and $k_0 \geq 1$, and let numbers $\Delta_k, F_k \geq 0$ and steps $\mu_k > 0$ satisfy, for the indices k specified below,*

$$\Delta_{k+1} \leq (1 - a\mu_k)\Delta_k - c\mu_k F_k + \mu_k \mathcal{B} + \mu_k^2 \mathcal{V}. \quad (\text{F.2})$$

(a) Let (F.2) hold for $k = 1, \dots, N$ with $\mu_k = 2/(a(k + k_0))$, and let $w_k = k + k_0 - 1$ and $\mathcal{W} = \sum_{k=1}^N w_k$. Then

$$\frac{1}{\mathcal{W}} \sum_{k=1}^N w_k F_k \leq \frac{ak_0^2 \Delta_1}{cN^2} + \frac{\mathcal{B}}{c} + \frac{4\mathcal{V}}{acN}.$$

(b) Let (F.2) hold for $k = 1, \dots, N_1 + N_2$, with a constant step $\mu_k = \mu$ for $k \leq N_1$, where $a\mu \leq \min\{1, 2/k_0\}$, and $\mu_{N_1+k} = 2/(a(k + k_0))$ for $1 \leq k \leq N_2$. If $N_2 \geq k_0$, then, with w_k and $\mathcal{W}$ as in (a) for $N = N_2$,

$$\frac{1}{\mathcal{W}} \sum_{k=1}^{N_2} w_k F_{N_1+k} \leq \frac{ak_0^2 \Delta_1}{cN_2^2} e^{-a\mu N_1} + \frac{2\mathcal{B}}{c} + \frac{6\mathcal{V}}{acN_2}.$$

Proof. (a) Since $1/(c\mu_k) = a(w_k + 1)/(2c)$, dividing (F.2) by $c\mu_k$ gives

$$F_k \leq \frac{a}{2c} ((w_k - 1)\Delta_k - (w_k + 1)\Delta_{k+1}) + \frac{\mathcal{B}}{c} + \frac{\mu_k \mathcal{V}}{c}.$$

Multiply by w_k and sum. The distance terms telescope because $w_{k+1}(w_{k+1} - 1) = w_k(w_k + 1)$, and at most $\frac{a}{2c} k_0(k_0 - 1)\Delta_1$ remains. Moreover $w_k \mu_k \leq 2/a$ and $\mathcal{W} = N(N + 2k_0 - 1)/2 \geq N^2/2$.

(b) In the first phase $0 \leq 1 - a\mu < 1$, so the recursion can be iterated; dropping $F_k \geq 0$ and summing the geometric series gives $\Delta_{N_1+1} \leq (1 - a\mu)^{N_1} \Delta_1 + \mathcal{B}/a + \mu \mathcal{V}/a$, and $\mu \mathcal{V}/a \leq 2\mathcal{V}/(a^2 k_0)$. Insert this into (a), applied to the second phase, and use $(1 - a\mu)^{N_1} \leq e^{-a\mu N_1}$ and $N_2 \geq k_0$. $\square$

Lemma F.2 (Two-phase projected descent with a biased oracle). *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be closed and convex, and let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be twice continuously differentiable and α -strongly convex, with its minimizer x^* in $\mathcal{K}$ and $\nabla^2 f(x) \preceq LI$ for $x \in \mathcal{K}$; put $\kappa = L/\alpha$. Let $(\mathcal{F}_k)_{k \geq 0}$ be a filtration and let $z_1 \in \mathcal{K}$ be $\mathcal{F}_0$ -measurable with $\|z_1 - x^*\| \leq R$. Suppose that at every $\mathcal{F}_{k-1}$ -measurable center $z_k \in \mathcal{K}$ an oracle returns an $\mathcal{F}_k$ -measurable vector*

$$g_k = \nabla f(z_k) + b_k + \zeta_k + \Delta_k, \quad \|b_k\| \leq b, \quad \|\Delta_k\| \leq b_D \text{ a.s.}, \quad \mathbb{E}[\zeta_k \mid \mathcal{F}_{k-1}] = 0, \quad \mathbb{E}[\|\zeta_k\|^2 \mid \mathcal{F}_{k-1}] \leq v^2.$$

Fix integers $N_1 \geq 0$ and $N_2 \geq 16\kappa$, let $k_0 = 16\kappa$, and run, with the Euclidean projection $\text{Proj}_{\mathcal{K}}$ onto $\mathcal{K}$,

$$z_{k+1} = \text{Proj}_{\mathcal{K}}(z_k - \nu_k g_k), \quad \nu_k = \begin{cases} \frac{1}{4L}, & k = 1, \dots, N_1, \\ \frac{4}{\alpha(k - N_1 + k_0)}, & k = N_1 + 1, \dots, N_1 + N_2. \end{cases}$$

Output

$$\hat{x} = \mathcal{W}^{-1} \sum_{k=N_1+1}^{N_1+N_2} w_k z_k, \quad w_k = k - N_1 + k_0 - 1, \quad \mathcal{W} = \sum_{k=N_1+1}^{N_1+N_2} w_k.$$

Then

$$\mathbb{E}[f(\hat{x}) - f^*] \leq \frac{\alpha R^2}{2} e^{-N_1/(8\kappa)} + \frac{7(b + b_D)^2}{\alpha} + \frac{48v^2}{\alpha N_2}. \quad (\text{F.3})$$

Proof. Write $D_k = \mathbb{E} \|z_k - x^*\|^2$, $F_k = \mathbb{E}[f(z_k) - f^*]$, $u_k = b_k + \Delta_k$ and $\bar{b} = b + b_D$, so that $\|u_k\| \leq \bar{b}$. By induction, $z_k \in \mathcal{K}$ is $\mathcal{F}_{k-1}$ -measurable and square integrable, so $\mathbb{E} \langle \zeta_k, z_k - x^* \rangle = 0$.

For $z \in \mathcal{K}$ the segment from x^* to z lies in $\mathcal{K}$, so $\nabla f(z) = Q(z - x^*)$ with $Q = \int_0^1 \nabla^2 f(x^* + t(z - x^*)) dt$ and $\alpha I \preceq Q \preceq LI$. Hence $Q^2 \preceq LQ$, and strong convexity gives

$$\|\nabla f(z)\|^2 \leq L \langle \nabla f(z), z - x^* \rangle, \quad \frac{3}{2} \langle \nabla f(z), z - x^* \rangle \geq f(z) - f^* + \alpha \|z - x^*\|^2, \quad (\text{F.4})$$

the second by adding $\langle \nabla f(z), z - x^* \rangle \geq f(z) - f^* + \frac{\alpha}{2} \|z - x^*\|^2$ to half of $\langle \nabla f(z), z - x^* \rangle \geq \alpha \|z - x^*\|^2$. Since $x^* \in \mathcal{K}$, $D_{k+1} \leq \mathbb{E} \|z_k - \nu_k g_k - x^*\|^2$, and $\|g_k\|^2 \leq 2 \|\nabla f(z_k)\|^2 + 4\bar{b}^2 + 4 \|\zeta_k\|^2$. Every step satisfies $\nu_k \leq 1/(4L)$, since $4/(\alpha(k_0 + 1)) < 1/(4L)$, so the coefficient of $\mathbb{E} \langle \nabla f(z_k), z_k - x^* \rangle \geq 0$ is at most $-2\nu_k + 2L\nu_k^2 \leq -\frac{3}{2}\nu_k$. Young's inequality $2\bar{b} \|z_k - x^*\| \leq \alpha \|z_k - x^*\|^2 / 2 + 2\bar{b}^2 / \alpha$ and (F.4) give

$$D_{k+1} \leq \left(1 - \frac{\alpha\nu_k}{2}\right) D_k - \nu_k F_k + \frac{2\nu_k \bar{b}^2}{\alpha} + 4\nu_k^2 (\bar{b}^2 + v^2). \quad (\text{F.5})$$

This is (F.2) with $a = \alpha/2$, $c = 1$, $\mathcal{B} = 2\bar{b}^2/\alpha$, $\mathcal{V} = 4(\bar{b}^2 + v^2)$ and the steps of Lemma F.1(b) with $\mu = 1/(4L)$, for which $a\mu = 2/k_0 = 1/(8\kappa) \leq 1$. Since $D_1 \leq R^2$, $N_2 \geq k_0$ and $N_2 \geq 16$, Lemma F.1(b) and the convexity of f give (F.3). $\square$

F.2 Oracle calls and rounds without acceleration

Theorem F.3 (Strongly convex objectives: oracle calls and rounds). *Let $f \in \mathcal{F}_{\text{sc}}$ and let the noise be admissible at levels (M, σ, δ) . Let $N \geq 32\kappa$ be even, $B \geq 1$ an integer and $h > 0$. Run Lemma F.2 with $\mathcal{K} = B(R)$, $z_1 = 0$ and $N_1 = N_2 = N/2$, where the oracle of round k is the estimate of Procedure 1 with center z_k , radius h and B frames; this is variant (a) of Procedure 2. The method uses N rounds and $T' = 2mdNB$ calls, and*

$$\begin{aligned} \mathbb{E}[f(\hat{x}) - f^*] &\leq \frac{\alpha R^2}{2} e^{-N/(16\kappa)} + \frac{1}{\alpha} \left[(21C_b^2 + 3C_f) L_\beta^2 d^{-\tau} h^{2\beta-2} + \frac{192mC_\sigma d^2 \sigma^2}{T' h^2} \right] \\ &\quad + (21A_1^2 + 3C_\eta) \frac{M^2}{\alpha} + 21S_1^2 \frac{d\delta^2}{\alpha h^2}. \end{aligned} \quad (\text{F.6})$$

(a) Let $\sigma > 0$ and $T \geq 2mdN$. With $B = \lfloor T/(2mdN) \rfloor$ and $h = h_T$ the method uses at most T calls, and

$$\mathbb{E}[f(\hat{x}) - f^*] \leq \frac{\alpha R^2}{2} e^{-N/(16\kappa)} + C_\Phi \Phi_T + C_M \frac{M^2}{\alpha} + C_\delta \frac{d\delta^2}{\alpha h_T^2}, \quad (\text{F.7})$$

where

$$C_\Phi = 21C_b^2 + 3C_f + 384mC_\sigma, \quad C_M = 21A_1^2 + 3C_\eta, \quad C_\delta = 21S_1^2.$$

Moreover $d\delta^2/(\alpha h_T^2) = (T\delta^2/(d\sigma^2))\Phi_T$.

(b) Let $\sigma > 0$ and

$$N_T = 2 \lceil \max\{16\kappa, 8\kappa \log(\alpha R^2/(2\Phi_T))\} \rceil.$$

If $T \geq 2mdN_T$, then (a) with $N = N_T$ uses $N_T \leq 2 + 32\kappa + 16\kappa \log^+(\alpha R^2/(2\Phi_T))$ rounds, at most T calls, and

$$\mathbb{E}[f(\hat{x}) - f^*] \leq (C_\Phi + 1)\Phi_T + C_M \frac{M^2}{\alpha} + C_\delta \frac{d\delta^2}{\alpha h_T^2}.$$

The condition $T \geq 2mdN_T$, which is equivalent to $B \geq 1$, holds whenever

$$T \geq 8md + 64md\kappa \left[1 + \log(64md\kappa) + \log^+ \frac{\alpha R^2}{2\Phi_1} \right]. \quad (\text{F.8})$$

The same holds if the upper bound $\nabla^2 f \preceq LI$ is assumed only on $B(R)$.

For $\delta = 0$, $L > \alpha$ and $T \geq T_{\text{lb}}$ (see (E.7)), Theorem 1(b) shows that the bound in (b) is optimal up to a factor depending only on β ; the same holds for the accelerated bound (F.12) below.

Proof. Apply Lemma F.2 with $b = b_f + b_\eta$, $b_D = b_\delta$, $v^2 = v_h^2/B$ and $N_1 = N_2 = N/2$. Its hypotheses follow from (E1)–(E3), since each center is measurable before its round, and they involve the Hessian only on $\mathcal{K} = B(R)$; the estimate itself uses no Hessian bound. Using $(b_f + b_\eta + b_\delta)^2 \leq 3(b_f^2 + b_\eta^2 + b_\delta^2)$ and $96/(NB) = 192md/T' \leq 3$ gives (F.6).

For (a), $\lfloor x \rfloor \geq x/2$ for $x \geq 1$ gives $T/2 \leq T' \leq T$; substitute (F.1). For (b), the definition of N_T makes the initial-error term at most Φ_T . Moreover $\Phi_T \geq \Phi_1/T$ and $\log T \leq T/(64md\kappa) + \log(64md\kappa) - 1$ imply $2mdN_T \leq T/2 + 4md + 32md\kappa[1 + \log(64md\kappa) + \log^+(\alpha R^2/(2\Phi_1))]$, so (F.8) suffices. $\square$

F.3 An accelerated method with inexact gradients

The method of the next lemma is the constant-momentum scheme of Nesterov (2004, Section 2.2), run with inexact gradients.

Lemma F.4 (Accelerated method with inexact gradients). *Let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ be differentiable and μ -strongly convex with an L_φ -Lipschitz gradient. Put $\kappa_\varphi = L_\varphi/\mu$, $\omega = \kappa_\varphi^{-1/2}$, $\theta = (1 - \omega)/(1 + \omega)$, and let x_φ^* be the minimizer of φ and $\varphi^* = \varphi(x_\varphi^*)$. Let $(\mathcal{F}_k)_{k \geq 0}$ be a filtration, let x_0 be $\mathcal{F}_0$ -measurable and square integrable, $x_{-1} = x_0$, and for $k \geq 0$*

$$y_k = x_k + \theta(x_k - x_{k-1}), \quad x_{k+1} = y_k - \frac{g_k}{L_\varphi}, \quad g_k = \nabla\varphi(y_k) + u_k + \zeta_k,$$

where g_k is $\mathcal{F}_{k+1}$ -measurable, $\|u_k\| \leq b$ almost surely, $\mathbb{E}[\zeta_k \mid \mathcal{F}_k] = 0$ and $\mathbb{E}[\|\zeta_k\|^2 \mid \mathcal{F}_k] \leq v^2$. Let $\mathcal{V}_0 = \varphi(x_0) - \varphi^* + \frac{\mu}{2} \|x_0 - x_\varphi^*\|^2$. Then for every $K \geq 0$

$$\mathbb{E}[\varphi(x_K) - \varphi^*] \leq \left(1 - \frac{1}{2\sqrt{\kappa_\varphi}}\right)^K \mathbb{E}\mathcal{V}_0 + \frac{3b^2}{\mu} + \frac{3\sqrt{\kappa_\varphi}v^2}{L_\varphi}. \quad (\text{F.9})$$

Proof. Write $x^* = x_\varphi^*$, $\nabla_k = \nabla\varphi(y_k)$ and $e_k = u_k + \zeta_k$. Put $v_0 = x_0$, $m_k = (1 - \omega)v_k + \omega y_k$ and $v_{k+1} = m_k - (\omega/\mu)g_k$. By induction, $(1 + \omega)y_k = x_k + \omega v_k$, so $m_k - x^* = y_k - x^* + (1 - \omega)(y_k - x_k)/\omega$. The vectors y_k , v_k and m_k are $\mathcal{F}_k$ -measurable and square integrable. For $\mathcal{V}_k = \varphi(x_k) - \varphi^* + (\mu/2) \|v_k - x^*\|^2$, the estimate-sequence argument of Nesterov (2004, Section 2.2), with the error retained, gives

$$\mathcal{V}_{k+1} \leq (1 - \omega)\mathcal{V}_k - \omega \langle e_k, v_{k+1} - x^* \rangle. \quad (\text{F.10})$$

Indeed, smoothness gives $\varphi(x_{k+1}) \leq \varphi(y_k) - \|\nabla_k\|^2/(2L_\varphi) + \|e_k\|^2/(2L_\varphi)$. Expand $\|v_{k+1} - x^*\|^2$, bound $\|m_k - x^*\|^2$ by convexity of the squared norm, and use $\omega^2/\mu = 1/L_\varphi$, the formula for m_k , strong convexity at y_k and convexity between y_k and x_k . This leaves

$$\mathcal{V}_{k+1} \leq (1 - \omega)\mathcal{V}_k + \frac{\|g_k\|^2 - \|\nabla_k\|^2 + \|e_k\|^2}{2L_\varphi} - \omega \langle e_k, m_k - x^* \rangle.$$

The numerator equals $2 \langle e_k, g_k \rangle$, and $g_k/L_\varphi = \omega(m_k - v_{k+1})$; this proves (F.10).

Since $v_{k+1} - x^* = (m_k - x^* - (\omega/\mu)\nabla_k) - (\omega/\mu)e_k$, where the first vector is $\mathcal{F}_k$ -measurable, conditional centering and the Cauchy–Schwarz inequality give

$$-\omega \mathbb{E} \langle \zeta_k, v_{k+1} - x^* \rangle = \frac{\mathbb{E} \langle \zeta_k, u_k \rangle + \mathbb{E} \|\zeta_k\|^2}{L_\varphi} \leq \frac{bv + v^2}{L_\varphi}.$$

Pathwise, Young’s inequality gives $-\omega \langle u_k, v_{k+1} - x^* \rangle \leq (\omega/2)\mathcal{V}_{k+1} + \omega b^2/\mu$. Hence $(1 - \omega/2)\mathbb{E}\mathcal{V}_{k+1} \leq (1 - \omega)\mathbb{E}\mathcal{V}_k + c$, where $c = \omega b^2/\mu + (bv + v^2)/L_\varphi$. With $\rho = (1 - \omega)/(1 - \omega/2) \leq 1 - \omega/2$, iteration yields

$$\mathbb{E}\mathcal{V}_K \leq \rho^K \mathbb{E}\mathcal{V}_0 + 2c/\omega \leq (1 - \omega/2)^K \mathbb{E}\mathcal{V}_0 + 3b^2/\mu + 3v^2/(\omega L_\varphi),$$

where we used $2bv \leq b^2 + v^2$ and $\omega L_\varphi \geq \mu$. The claim follows because the function gap is at most $\mathcal{V}_K$. $\square$

F.4 Restarts with growing batches

Restarting an accelerated method in stages is a standard way to obtain linear convergence for strongly convex problems (cf. the multistage method of [Ghadimi and Lan, 2013](#)). In the next lemma the batch size doubles from stage to stage, and the bias and the residual variance are unknown to the method.

Lemma F.5 (Restarts with growing batches). *Let $\varphi, \mu, L_\varphi, \kappa_\varphi$ be as in Lemma F.4, let $x_0 \in \mathbb{R}^d$ be deterministic, and let $\bar{\mathcal{V}}_0$ be a known number with $\bar{\mathcal{V}}_0 \geq \varphi(x_0) - \varphi^* + \frac{\mu}{2} \|x_0 - x_\varphi^*\|^2$. Let $0 < \varepsilon < \bar{\mathcal{V}}_0$, let $v_{\text{sam}}^2 \geq 0$ be a known number, and put*

$$J = \left\lceil \log_2 \frac{\bar{\mathcal{V}}_0}{\varepsilon} \right\rceil, \quad N_0 = \lceil 5\sqrt{\kappa_\varphi} \rceil, \quad \varepsilon_j = 2^{-j} \bar{\mathcal{V}}_0, \quad B_j = \max \left\{ 1, \left\lceil \frac{12\sqrt{\kappa_\varphi} v_{\text{sam}}^2}{L_\varphi \varepsilon_j} \right\rceil \right\}.$$

Let $x^{(0)} = x_0$. For $j = 1, \dots, J$ (stage j), run N_0 iterations of the method of Lemma F.4 started at $x_0 = x_{-1} = x^{(j-1)}$, each iteration using a batch of B_j samples, and let $x^{(j)}$ be its last iterate. Number the iterations of all stages consecutively by $i = 0, 1, \dots, JN_0 - 1$. Let $(\mathcal{F}_i)_{i \geq 0}$ be a filtration such that the gradient of iteration i is $\mathcal{F}_{i+1}$ -measurable and, if iteration i belongs to stage j , its error is $u_i + \zeta_i$ with

$$\|u_i\| \leq b \text{ almost surely}, \quad \mathbb{E}[\zeta_i \mid \mathcal{F}_i] = 0, \quad \mathbb{E}[\|\zeta_i\|^2 \mid \mathcal{F}_i] \leq v_{\text{sam}}^2/B_j + r^2,$$

where $b, r \geq 0$ are unknown. Then

$$\mathbb{E}[\varphi(x^{(J)}) - \varphi^*] \leq \frac{3}{4}\varepsilon + \frac{4b^2}{\mu} + \frac{4\sqrt{\kappa_\varphi} r^2}{L_\varphi}. \quad (\text{F.11})$$

The procedure uses $JN_0 \leq (1 + \log_2(\bar{\mathcal{V}}_0/\varepsilon))(1 + 5\sqrt{\kappa_\varphi})$ iterations (rounds) and $\sum_{j=1}^J N_0 B_j \leq JN_0 + 288v_{\text{sam}}^2/(\mu\varepsilon)$ samples. In particular, if $r = 0$ and $b^2 \leq \mu\varepsilon/16$, then $\mathbb{E}[\varphi(x^{(J)}) - \varphi^*] \leq \varepsilon$.

Proof. Let $E_j = \mathbb{E}[\varphi(x^{(j)}) - \varphi^*]$ and $G = 3b^2/\mu + 3\sqrt{\kappa_\varphi} r^2/L_\varphi$. The starting point of each stage is measurable before the stage, so Lemma F.4 applies to every stage. In it, $(1 - 1/(2\sqrt{\kappa_\varphi}))^{N_0} \leq e^{-5/2} < 1/8$, the batch size bounds the sampling term by $\varepsilon_j/4$, and, after the first stage, strong convexity bounds the initial potential by twice the starting gap. Thus

$$E_1 \leq \bar{\mathcal{V}}_0/8 + G + \varepsilon_1/4, \quad E_j \leq E_{j-1}/4 + G + \varepsilon_j/4 \quad (j \geq 2).$$

Unrolling and using $\varepsilon_j = 2^{J-j}\varepsilon_J$ gives $E_J \leq \varepsilon_J/4 + (4/3)G + \varepsilon_J/2$, which is (F.11) since $\varepsilon_J \leq \varepsilon$.

The round count is JN_0 . Moreover $B_j \leq 1 + 12\sqrt{\kappa_\varphi} v_{\text{sam}}^2 2^j / (L_\varphi \bar{\mathcal{V}}_0)$, $\sum_{j=1}^J 2^j < 4\bar{\mathcal{V}}_0/\varepsilon$ and $N_0 \leq 6\sqrt{\kappa_\varphi}$. Therefore $\sum_j N_0 B_j \leq JN_0 + 288v_{\text{sam}}^2/(\mu\varepsilon)$. For $r = 0$ and $b^2 \leq \mu\varepsilon/16$, the error bound is at most ε . $\square$

F.5 Accelerated rounds with random-frame estimates

The batch sizes are chosen from the stochastic part $v_{\text{sam}}^2 = C_\sigma d\sigma^2/h^2$ of the variance in (E2). The remaining part $r^2 = C_f L_\beta^2 d^{-\tau} h^{2\beta-2} + C_\eta M^2$ is not divided by the batch size in the analysis, which keeps the batch sizes independent of the unknown level M .

Theorem F.6 (Accelerated method with explicit constants). *Let $\beta > 2$, $L_\beta, \alpha, R > 0$, $L \geq \alpha$, $\kappa = L/\alpha$ and $\sigma > 0$, and let $\mathcal{F}_{\text{sc}} = \mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)$ be defined with either seminorm. Put*

$$A = 1152mC_\sigma, \quad C_\Phi^{\text{acc}} = A + 12C_b^2 + 4C_f, \quad C_M^{\text{acc}} = 12A_1^2 + 4C_\eta, \quad C_\delta^{\text{acc}} = 12S_1^2,$$

and $\bar{\mathcal{V}}_0 = (L + \alpha)R^2/2$.

- (a) Fixed budget. For $T \geq 1$, put $h = h_T$, $\varepsilon = A\Phi_T$ and $v_{\text{sam}}^2 = C_\sigma d\sigma^2/h_T^2$. If $\bar{\mathcal{V}}_0 \leq \varepsilon$, output $\hat{x} = 0$ and set $J = 0$. Otherwise run Lemma F.5 with $\varphi = f$, $\mu = \alpha$, $L_\varphi = L$, $x_0 = 0$ and these $\bar{\mathcal{V}}_0, \varepsilon, v_{\text{sam}}^2$, that is, variant
- (b) of Procedure 2 with

$$N_0 = \lceil 5\sqrt{\kappa} \rceil, \quad J = \lceil \log_2(\bar{\mathcal{V}}_0/\varepsilon) \rceil, \quad \varepsilon_j = 2^{-j}\bar{\mathcal{V}}_0, \quad B_j = \max\{1, \lceil 12\sqrt{\kappa} v_{\text{sam}}^2 / (L\varepsilon_j) \rceil\};$$

each gradient estimate is Procedure 1 at the center y_k , with radius h_T and B_j frames. If $T \geq 4mdN_0J$, the method uses at most T calls and $N = N_0J \leq (1 + 5\sqrt{\kappa})(1 + \log_2^+(LR^2/(A\Phi_T)))$ rounds. For every $f \in \mathcal{F}_{\text{sc}}$ and noise admissible at levels (M, σ, δ) ,

$$\mathbb{E}[f(\hat{x}) - f^*] \leq C_{\Phi}^{\text{acc}} \Phi_T + C_M^{\text{acc}} \frac{M^2}{\alpha} + C_{\delta}^{\text{acc}} \frac{d\delta^2}{\alpha h_T^2}. \quad (\text{F.12})$$

A sufficient condition for $T \geq 4mdN_0J$ is

$$T \geq 48md\sqrt{\kappa} \left[\log_2(51md\sqrt{\kappa}) + \log_2^+ \frac{\bar{V}_0}{A\Phi_1} \right]. \quad (\text{F.13})$$

(b) Target accuracy and admissible errors. Let $\sigma \geq 0$, $0 < \varepsilon < LR^2$ and

$$h = \left(\frac{\alpha \varepsilon d^{\tau}}{8(12C_b^2 + 4C_f)L_{\beta}^2} \right)^{\frac{1}{2\beta-2}}.$$

Run Lemma F.5 as in (a), but with target $\varepsilon/2$, radius h and $v_{\text{sam}}^2 = C_{\sigma}d\sigma^2/h^2$. The method uses N_0J rounds, where $J = \lceil \log_2((L + \alpha)R^2/\varepsilon) \rceil \leq 2 + \log_2(LR^2/\varepsilon)$, and

$$T \leq 2mdN_0J + \frac{1152mC_{\sigma}d^2\sigma^2}{\alpha \varepsilon h^2} \leq C_{\beta} \left(T_{\star}(\varepsilon) + d\sqrt{\kappa} \left[1 + \log \frac{LR^2}{\varepsilon} \right] \right) \quad (\text{F.14})$$

calls. It satisfies $\mathbb{E}[f(\hat{x}) - f^*] \leq \varepsilon$ for every $f \in \mathcal{F}_{\text{sc}}$ and noise admissible at levels (M, σ, δ) with

$$M \leq \left(\frac{\alpha \varepsilon}{8C_M^{\text{acc}}} \right)^{1/2}, \quad \delta \leq h \left(\frac{\alpha \varepsilon}{8C_{\delta}^{\text{acc}}d} \right)^{1/2} = c_{\beta} \delta_{\varepsilon},$$

where c_{β} depends only on β . The batch sizes do not depend on M or δ .

The accelerated method is not projected: its centers, queries and output may lie outside $B(R)$, as Definition 1 allows, and the proof uses the Hessian and estimator bounds on all of $\mathbb{R}^d$.

Proof of Lemma F.6(a) and of the upper bound in Theorem 1(a). Iteration i of Lemma F.5 is round $k = i + 1$, so the filtration of the lemma is $(\mathcal{F}_{k-1})_{k \geq 1}$ of Definition 1. By (E1)–(E3), the oracle of a round k of stage j , with center y , has error $u + \zeta$, where $u = b(y) + \Delta$ satisfies $\|u\| \leq b_f + b_{\eta} + b_{\delta}$, $\mathbb{E}[\zeta \mid \mathcal{F}_{k-1}] = 0$, and $\mathbb{E}[\|\zeta\|^2 \mid \mathcal{F}_{k-1}] \leq v_{\text{sam}}^2/B_j + r^2$, where $r^2 = C_f L_{\beta}^2 d^{-\tau} h_T^{2\beta-2} + C_{\eta} M^2$. By (F.1),

$$\frac{b_f^2}{\alpha} = C_b^2 \Phi_T, \quad \frac{b_{\delta}^2}{\alpha} = S_1^2 \frac{d\delta^2}{\alpha h_T^2}, \quad \frac{v_{\text{sam}}^2}{\alpha} = C_{\sigma} \frac{T}{d} \Phi_T, \quad \frac{r^2}{\alpha} = C_f \Phi_T + C_{\eta} \frac{M^2}{\alpha}. \quad (\text{F.15})$$

Smoothness bounds the initial potential by $\bar{V}_0$. If $\bar{V}_0 \leq \varepsilon$, then $f(0) - f^* \leq \varepsilon = A\Phi_T$, which implies (F.12). Otherwise Lemma F.5, $\sqrt{\kappa}/L \leq 1/\alpha$ and $(b_f + b_{\eta} + b_{\delta})^2 \leq 3(b_f^2 + b_{\eta}^2 + b_{\delta}^2)$ prove (F.12). The same lemma bounds the number of frames by $N_0J + 288v_{\text{sam}}^2/(\alpha A\Phi_T) = N_0J + T/(4md)$, so the method makes at most $2mdN_0J + T/2 \leq T$ calls when $T \geq 4mdN_0J$. Since $\bar{V}_0 \leq LR^2$, $J \leq 1 + \log_2^+(LR^2/(A\Phi_T))$.

A direct computation with $N_0 \leq 6\sqrt{\kappa}$, $\Phi_T \geq \Phi_1/T$ and $\log T \leq T/a + \log a - 1$ for $a = 48md\sqrt{\kappa}/\log 2$ shows that (F.13) implies $T \geq 4mdN_0J$.

Theorem 1(a) follows with $C_{\beta} \geq \max\{51m, C_{\Phi}^{\text{acc}}, C_M^{\text{acc}}, C_{\delta}^{\text{acc}}\}$: since $A \geq 1$ and $\bar{V}_0 \leq LR^2$, its budget condition implies (F.13), and $d\delta^2/(\alpha h_T^2) = (T\delta^2/(d\sigma^2))\Phi_T$ by (F.1). $\square$

Proof of Lemma F.6(b) and of the achievable levels in Theorem 1(c). As in the proof of (a), with target $\varepsilon/2$ and radius h , the bound (F.11) is at most $3\varepsilon/8$ plus $(12C_b^2 + 4C_f)L_{\beta}^2 d^{-\tau} h^{2\beta-2}/\alpha = \varepsilon/8$, the contribution of the smoothing bias and of its residual variance, plus $C_M^{\text{acc}} M^2/\alpha + C_{\delta}^{\text{acc}} d\delta^2/(\alpha h^2) \leq \varepsilon/4$; the sum is less than ε . The sample bound of Lemma F.5 gives (F.14), since h/h_{ε} depends only on β and $T_{\star}(\varepsilon) = d^2\sigma^2(\alpha\varepsilon)^{-1}h_{\varepsilon}^{-2}$ by (4) and (5); for the same reason $h(\alpha\varepsilon/d)^{1/2} = (h/h_{\varepsilon})\delta_{\varepsilon}$. The converse bounds in Theorem 1(c) are proved in Lemma E.7 and Section H. $\square$

G LOWER BOUNDS ON THE NUMBER OF ROUNDS

This appendix bounds from below the number of rounds needed on $\mathcal{F}_{\text{sc}}$ by algorithms that request at most n values per round, with exact values or with Gaussian noise. Section G.2 treats one round and Section G.3 treats N rounds, both with the tilts of Section G.1; Section G.4 deduces the necessary number of rounds (Lemma G.5), and Section G.5 compares it with the upper bound. For $\beta = 2$ and noisy values, Duchi et al. (2018) proved the one-round bound $n^{-2/(d+4)}$ on smooth strongly convex functions, which is the exponent of Lemma G.3 at $\beta = 2$.

Setting. Throughout this appendix $\beta > 2$, $q = \lceil \beta \rceil - 1 \geq 2$ and $\gamma = \beta - q \in (0, 1]$; the numbers $L_\beta, \alpha, R > 0$ and $L \geq 3\alpha$ are fixed, and $\mathcal{F}_{\text{sc}} = \mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)$ is defined with the operator seminorm (the Hilbert–Schmidt seminorm is treated in Lemma G.6). We write vol for the Lebesgue measure on $\mathbb{R}^d$ and $f^\star = \min f$.

Definition G.1 (Rounds with a bounded number of calls). Let $N, n \geq 1$ be integers. An *algorithm with N rounds of at most n calls* draws independent random elements $\Theta_1, \dots, \Theta_N$ with laws fixed by the algorithm. In round k it makes at most n calls at points of $\mathbb{R}^d$ that are measurable functions of $\Theta_1, \dots, \Theta_k$ and of the answers of rounds $1, \dots, k-1$. The output $\hat{x}$ is a measurable function of $\Theta_1, \dots, \Theta_N$ and of all answers. The answer to a call at X_t is either the exact value $Y_t = f(X_t)$, or $Y_t = f(X_t) + \xi_t$ with independent $N(0, \sigma^2)$ variables ξ_t , independent of $\Theta_1, \dots, \Theta_N$ (the *Gaussian model*). For a class $\mathcal{C}$ and $\sigma \geq 0$ let

$$\mathfrak{R}_{N,n}(\mathcal{C}; \sigma) = \inf_{f \in \mathcal{C}} \sup \mathbb{E}[f(\hat{x}) - f^\star],$$

where the infimum is over all algorithms with N rounds of at most n calls, with exact values if $\sigma = 0$ and in the Gaussian model if $\sigma > 0$.

This is the round structure of Section 2, and the Gaussian model is the oracle of Definition 1 with only the centered component. Calls whose answers are ignored do not change the law of the output, so in the proofs we may assume that every round has exactly n calls. An algorithm with N rounds and T calls in total has at most T calls per round, so every lower bound on $\mathfrak{R}_{N,n}$ below also holds, with $n = T$, for such algorithms.

G.1 Tilts

Let χ be the cutoff and $K_\beta, K_2 \geq 1$ the constants of Lemma E.1. For $z, v \in \mathbb{R}^d$ and $r > 0$ the *tilt* with center z , slope v and radius r is

$$P_{z,v,r}(x) = -\langle v, x - z \rangle \chi(\|x - z\|^2 / r^2). \quad (\text{G.1})$$

Lemma G.2 (Translated and rescaled tilts). *Let $z, v \in \mathbb{R}^d$ and $r > 0$. The function $P_{z,v,r}$ is infinitely differentiable and vanishes where $\|x - z\| \geq r$; where $\|x - z\| \leq r/2$ it satisfies $P_{z,v,r}(x) = -\langle v, x - z \rangle$ and $\nabla P_{z,v,r}(x) = -v$. Moreover*

$$[P_{z,v,r}]_{\beta, \text{op}} \leq K_\beta \|v\| r^{1-\beta}, \quad \sup_{x \in \mathbb{R}^d} \|\nabla^2 P_{z,v,r}(x)\|_{\text{op}} \leq K_2 \|v\| / r. \quad (\text{G.2})$$

Proof. The support and the local form follow from $\chi = 0$ on $[1, \infty)$ and $\chi = 1, \chi' = 0$ on $(-\infty, 1/4]$. For $v \neq 0$ let $u = v / \|v\|$; with the profile ψ_u of (E.1),

$$P_{z,v,r}(x) = -\|v\| r \psi_u((x - z)/r), \quad D^k P_{z,v,r}(x) = -\|v\| r^{1-k} D^k \psi_u((x - z)/r), \quad (\text{G.3})$$

and the operator-norm bounds of Lemma E.1 give (G.2). $\square$

G.2 One round

Theorem G.3 (One round with Gaussian noise). *Let $\sigma > 0$ and let $n \geq 1$ be an integer. Put*

$$\bar{r} = \min \left\{ \frac{R}{2}, \left(\frac{\alpha K_\beta}{K_2 L_\beta} \right)^{\frac{1}{\beta-2}} \right\}, \quad r = \min \left\{ \bar{r}, \left(\frac{K_\beta^2 \sigma^2 R^d}{16 \cdot 2^d n L_\beta^2} \right)^{\frac{1}{2\beta+d}} \right\}.$$

Then

$$\mathfrak{R}_{1,n}(\mathcal{F}_{\text{sc}}; \sigma) \geq \frac{3L_\beta^2}{64\alpha K_\beta^2} r^{2(\beta-1)}. \quad (\text{G.4})$$

Consequently $\mathfrak{R}_{1,n}(\mathcal{F}_{\text{sc}}; \sigma) \geq c_1 n^{-2(\beta-1)/(2\beta+d)}$ for every $n \geq 1$, where

$$c_1 = \frac{3L_\beta^2}{64\alpha K_\beta^2} \min\left\{\bar{r}, \left(\frac{K_\beta^2 \sigma^2 R^d}{16 \cdot 2^d L_\beta^2}\right)^{\frac{1}{2\beta+d}}\right\}^{2(\beta-1)}$$

does not depend on n , and $2(\beta-1)/(2\beta+d) < (\beta-1)/\beta$ for every $d \geq 1$.

Proof. We hide a sign perturbation at a uniformly random point of $B(R/2)$, bound the information that one round gives about the sign, and reduce the error to a test of the sign.

Step 1: hard functions. Let $a = L_\beta r^{\beta-1}/K_\beta$. For $z \in B(R/2)$ and $\omega \in \{-1, 1\}$ let

$$f_{z,\omega}(x) = \alpha \|x - z\|^2 + P_{z, -\omega a e_1, r}(x) = \alpha \|x - z\|^2 + \omega a \langle e_1, x - z \rangle \chi(\|x - z\|^2 / r^2).$$

By Lemma G.2, $[f_{z,\omega}]_{\beta, \text{op}} \leq K_\beta a r^{1-\beta} = L_\beta$, and since $r \leq \bar{r}$ the Hessian of the tilt has operator norm at most $K_2 a / r = K_2 L_\beta r^{\beta-2} / K_\beta \leq \alpha$. Hence $\alpha I \preceq \nabla^2 f_{z,\omega} \preceq 3\alpha I \preceq L I$, and $a \leq \alpha r / K_2 \leq \alpha r$. The point $x_{z,\omega}^* = z - \omega a e_1 / (2\alpha)$ satisfies $\|x_{z,\omega}^* - z\| = a / (2\alpha) \leq r/2$, so by Lemma G.2 $\nabla f_{z,\omega}(x_{z,\omega}^*) = 2\alpha(x_{z,\omega}^* - z) + \omega a e_1 = 0$. By strong convexity $x_{z,\omega}^*$ is the minimizer, and $\|x_{z,\omega}^*\| \leq R/2 + r/2 \leq 3R/4$. Thus $f_{z,\omega} \in \mathcal{F}_{\text{sc}}$.

Step 2: information. Fix an algorithm with one round of n calls at $X_1, \dots, X_n$, which are functions of Θ_1 . Let Z be uniform on $B(R/2)$ and ω uniform on $\{-1, 1\}$, independent of each other, of Θ_1 and of the noise, and let $P_{z,\omega}$ be the joint law of Θ_1 and of the answers when the objective is $f_{z,\omega}$. Since

$$|f_{z,1}(x) - f_{z,-1}(x)| = 2a |\langle e_1, x - z \rangle| \chi(\|x - z\|^2 / r^2) \leq 2ar \mathbf{1}\{\|x - z\| < r\}$$

and the law of $(X_1, \dots, X_n)$ does not depend on (z, ω) , Lemma B.5(b) gives

$$\text{KL}(P_{z,1} \| P_{z,-1}) \leq \frac{2a^2 r^2}{\sigma^2} \sum_{i=1}^n \mathbb{P}(\|X_i - z\| < r).$$

For every fixed x , $\mathbb{P}(\|x - Z\| < r) \leq \text{vol}(B(r)) / \text{vol}(B(R/2)) = (2r/R)^d$, and Z is independent of the queries. Let $\mathbb{E}_Z$ denote the average over z distributed as Z . By Pinsker's inequality (Lemma B.5(a)) and Jensen's inequality,

$$\mathbb{E}_Z \text{TV}(P_{Z,1}, P_{Z,-1}) \leq \left(\frac{1}{2} \mathbb{E}_Z \text{KL}(P_{Z,1} \| P_{Z,-1}) \right)^{1/2} \leq \frac{ar}{\sigma} \left(n \left(\frac{2r}{R} \right)^d \right)^{1/2} = \left(\frac{2^d n L_\beta^2 r^{2\beta+d}}{K_\beta^2 \sigma^2 R^d} \right)^{1/2} \leq \frac{1}{4}$$

by the choice of r .

Step 3: from the error to a test. Given z , let $\hat{\omega} = 1$ if $\hat{x}_1 < z_1$ and $\hat{\omega} = -1$ otherwise. The first coordinate of $x_{z,\omega}^*$ is $z_1 - \omega a / (2\alpha)$, so $\hat{\omega} \neq \omega$ implies $\|\hat{x} - x_{z,\omega}^*\| \geq a / (2\alpha)$, and then $f_{z,\omega}(\hat{x}) - \min f_{z,\omega} \geq a^2 / (8\alpha)$ by strong convexity. For fixed z , $\hat{\omega}$ is a test between $P_{z,1}$ and $P_{z,-1}$, so by Lemma B.5(a) and Step 2, $\mathbb{P}(\hat{\omega} \neq \omega) \geq \frac{1}{2}(1 - \mathbb{E}_Z \text{TV}(P_{Z,1}, P_{Z,-1})) \geq \frac{3}{8}$. The worst-case error over $\mathcal{F}_{\text{sc}}$ is at least the average over (Z, ω) , so

$$\sup_{f \in \mathcal{F}_{\text{sc}}} \mathbb{E}[f(\hat{x}) - f^*] \geq \frac{a^2}{8\alpha} \cdot \frac{3}{8} = \frac{3L_\beta^2 r^{2\beta-2}}{64\alpha K_\beta^2}.$$

The algorithm was arbitrary, which proves (G.4).

Step 4: the power of n . Since $n^{-1/(2\beta+d)} \leq 1$, $r \geq \min\{\bar{r}, (K_\beta^2 \sigma^2 R^d / (16 \cdot 2^d L_\beta^2))^{1/(2\beta+d)}\} n^{-1/(2\beta+d)}$, which gives the bound with c_1 . Finally $2(\beta-1)/(2\beta+d) < (\beta-1)/\beta$ is equivalent to $d > 0$. $\square$

The queries of one round cannot concentrate near a perturbation at a uniformly random location.

G.3 Several rounds: nested tilts

The construction adapts the smooth nested perturbations of Dorn et al. (2026, Appendix E) to rounds and to a fixed Hölder seminorm.

Theorem G.4 (Several rounds). *Let $N, n \geq 1$ be integers and put*

$$r_1 = \min\left\{R, \left(\frac{\alpha K_\beta}{8K_2 L_\beta}\right)^{\frac{1}{\beta-2}}\right\}, \quad \vartheta_N = \frac{L_\beta}{2\alpha K_\beta(N+1)},$$

and define recursively

$$\varrho_i = \vartheta_N r_i^{\beta-1} \quad (1 \leq i \leq N+1), \quad r_{i+1} = (4Nn)^{-1/d} \varrho_i \quad (1 \leq i \leq N). \quad (\text{G.5})$$

Then, for every $\sigma \geq 0$,

$$\mathfrak{R}_{N,n}(\mathcal{F}_{\text{sc}}; \sigma) \geq \mathfrak{R}_{N,n}(\mathcal{F}_{\text{sc}}; 0) \geq \frac{3\alpha}{8} \varrho_{N+1}^2, \quad \varrho_{N+1} = r_1 (\vartheta_N r_1^{\beta-2})^{\frac{(\beta-1)^{N+1}-1}{\beta-2}} (4Nn)^{-\frac{(\beta-1)((\beta-1)^N-1)}{d(\beta-2)}}. \quad (\text{G.6})$$

In particular $\mathfrak{R}_{N,n}(\mathcal{F}_{\text{sc}}; \sigma) \geq c_N n^{-\nu_N}$ for every $n \geq 1$, where

$$\nu_N = \frac{2(\beta-1)((\beta-1)^N-1)}{d(\beta-2)}, \quad c_N = \frac{3\alpha}{8} r_1^2 (\vartheta_N r_1^{\beta-2})^{\frac{2((\beta-1)^{N+1}-1)}{\beta-2}} (4N)^{-\nu_N}, \quad (\text{G.7})$$

and c_N does not depend on n .

Proof. We hide the minimizer behind $N+1$ tilts with nested random centers $Z_1, Z_2, \dots$. With probability at least $3/4$, every query of round k misses the ball around Z_k in which f differs from the sum F_k of the first k tilts. The algorithm then behaves as if it saw only F_k in round k , and it learns nothing about the sign of the last tilt.

Step 1: scales. By the choice of r_1 , and since $K_2 \geq 1$ and $N \geq 1$,

$$\vartheta_N r_1^{\beta-2} \leq \frac{1}{16K_2(N+1)} \leq \frac{1}{32}. \quad (\text{G.8})$$

If $r_i \leq r_1$, then $\varrho_i/r_i = \vartheta_N r_i^{\beta-2} \leq 1/32$ and $r_{i+1} = (4Nn)^{-1/d} \varrho_i \leq \varrho_i \leq r_i/32 \leq r_1$. By induction, $r_{i+1} \leq \varrho_i \leq r_i/32$ for $1 \leq i \leq N$ and $\varrho_{N+1} \leq r_{N+1}/32$; hence $\varrho_{i+1} \leq \varrho_i/32$, and for $1 \leq i \leq N+1$

$$\sum_{j=i}^{N+1} \varrho_j \leq \frac{32}{31} \varrho_i \leq \frac{r_i}{31}. \quad (\text{G.9})$$

Step 2: hard functions. Let $W_1, \dots, W_N$ be uniform on $B(1)$ and ω uniform on $\{-1, 1\}$, all independent. Put $V_i = \varrho_i W_i$ for $i \leq N$, $V_{N+1} = \omega \varrho_{N+1} e_1$, $Z_0 = 0$ and $Z_i = Z_{i-1} + V_i$ for $1 \leq i \leq N+1$, and

$$F_k(x) = \alpha \|x\|^2 + \sum_{j=1}^k P_{Z_{j-1}, 2\alpha V_j, r_j}(x) \quad (0 \leq k \leq N+1), \quad f = F_{N+1}. \quad (\text{G.10})$$

Since $\|V_j\| \leq \varrho_j$, Lemma G.2 bounds the seminorm of the j -th tilt by $2\alpha K_\beta \varrho_j r_j^{1-\beta} = 2\alpha K_\beta \vartheta_N = L_\beta/(N+1)$, and the quadratic contributes nothing, so $[f]_{\beta, \text{op}} \leq L_\beta$. The Hessian of the j -th tilt has operator norm at most $2\alpha K_2 \varrho_j/r_j = 2\alpha K_2 \vartheta_N r_j^{\beta-2} \leq \alpha/(8(N+1))$ by (G.8), so $\alpha I \preceq \nabla^2 f \preceq 3\alpha I \preceq LI$. By (G.9), $\|Z_{N+1} - Z_{j-1}\| \leq \sum_{l=j}^{N+1} \varrho_l \leq r_j/31$, so by Lemma G.2 the j -th tilt has gradient $-2\alpha V_j$ at Z_{N+1} , and $\nabla f(Z_{N+1}) = 2\alpha(Z_{N+1} - \sum_{j=1}^{N+1} V_j) = 0$. Thus every realization of f has the unique minimizer Z_{N+1} , of norm at most $r_1/31 \leq R$, and lies in $\mathcal{F}_{\text{sc}}$.

Step 3: localization. For $1 \leq k \leq N$,

$$f(x) = F_k(x) \quad \text{whenever} \quad \|x - Z_k\| \geq r_{k+1}. \quad (\text{G.11})$$

The difference $f - F_k$ is the sum of the tilts $j = k+1, \dots, N+1$. The tilt $j = k+1$ has center Z_k and radius r_{k+1} . For $j \geq k+2$, (G.9) gives $\|Z_{j-1} - Z_k\| \leq r_{k+1}/31$, hence $\|x - Z_{j-1}\| \geq \frac{30}{31} r_{k+1} > r_{k+1}/32 \geq r_{k+2} \geq r_j$. All these tilts vanish at x .

Step 4: a truncated run. Fix an algorithm with N rounds of n calls and exact values, with $\Theta_1, \dots, \Theta_N$ independent of $(W_1, \dots, W_N, \omega)$. We allow its queries in round k to depend also on $V_1, \dots, V_{k-1}$ and its output on $V_1, \dots, V_N$,

which only enlarges the class of algorithms. In the *truncated run*, with the same $\Theta_1, \dots, \Theta_N$ and $V_1, \dots, V_N$, the calls of round k are answered with the values of F_k instead of f ; let $\tilde{X}_{k,1}, \dots, \tilde{X}_{k,n}$ be its queries in round k and $\tilde{x}$ its output. Since F_s depends only on $V_1, \dots, V_s$, induction on k shows that $\tilde{X}_{k,i}$ is a function of $(\Theta_1, \dots, \Theta_k, V_1, \dots, V_{k-1})$ and $\tilde{x}$ a function of $(\Theta_1, \dots, \Theta_N, V_1, \dots, V_N)$.

For $1 \leq k \leq N$ let E_k be the event that $\|\tilde{X}_{k,i} - Z_k\| < r_{k+1}$ for some $i \leq n$. Given $\Theta_1, \dots, \Theta_N, W_1, \dots, W_{k-1}$, which determine $\tilde{X}_{k,i}$ and Z_{k-1} , the vector W_k is uniform on $B(1)$ and $Z_k = Z_{k-1} + \varrho_k W_k$; hence each query satisfies $\|\tilde{X}_{k,i} - Z_k\| < r_{k+1}$ with conditional probability at most $(r_{k+1}/\varrho_k)^d = 1/(4Nn)$. A union bound over the Nn queries gives

$$\mathbb{P}\left(\bigcup_{k=1}^N E_k\right) \leq \frac{1}{4}. \quad (\text{G.12})$$

On $\mathcal{G} = \bigcap_{k=1}^N E_k^c$ the actual and the truncated runs coincide: if they agree before round k , the queries of round k agree, and on E_k^c the answers agree by (G.11). In particular $\hat{x} = \tilde{x}$ on $\mathcal{G}$.

Step 5: the error with exact values. By strong convexity, $f(\hat{x}) - f^* \geq \frac{\alpha}{2} \|\hat{x} - Z_{N+1}\|^2$, where $Z_{N+1} = Z_N + \omega \varrho_{N+1} e_1$. The variables $\mathbf{1}_{\mathcal{G}}$, $\tilde{x}$ and Z_N are functions of $(\Theta_1, \dots, \Theta_N, W_1, \dots, W_N)$, which is independent of ω , and $\frac{1}{2} \sum_{o \in \{-1,1\}} \|y - o \varrho_{N+1} e_1\|^2 = \|y\|^2 + \varrho_{N+1}^2$ for every $y \in \mathbb{R}^d$. Conditioning on these variables and using (G.12),

$$\mathbb{E}[f(\hat{x}) - f^*] \geq \frac{\alpha}{2} \mathbb{E}[\mathbf{1}_{\mathcal{G}} \|\tilde{x} - Z_N - \omega \varrho_{N+1} e_1\|^2] \geq \frac{\alpha}{2} \varrho_{N+1}^2 \mathbb{P}(\mathcal{G}) \geq \frac{3\alpha}{8} \varrho_{N+1}^2.$$

The left-hand side averages the error over realizations of f in $\mathcal{F}_{\text{sc}}$, so the worst-case error is at least $\frac{3\alpha}{8} \varrho_{N+1}^2$. As the algorithm was arbitrary, $\mathfrak{R}_{N,n}(\mathcal{F}_{\text{sc}}; 0) \geq \frac{3\alpha}{8} \varrho_{N+1}^2$.

Step 6: Gaussian noise. An algorithm with exact values can simulate a given algorithm for the Gaussian model with the same rounds and calls: it draws the noise variables as part of Θ_k and adds them to the exact answers. Its queries and output have the same joint law, hence the same error on every f , so $\mathfrak{R}_{N,n}(\mathcal{F}_{\text{sc}}; 0) \leq \mathfrak{R}_{N,n}(\mathcal{F}_{\text{sc}}; \sigma)$.

Step 7: the endpoint. By (G.5), $\log \varrho_{i+1} = (\beta-1) \log \varrho_i + \log \vartheta_N - \frac{\beta-1}{d} \log(4Nn)$ and $\log \varrho_1 = \log \vartheta_N + (\beta-1) \log r_1$. Solving this linear recursion gives

$$\varrho_{N+1} = \vartheta_N^{\frac{(\beta-1)^{N+1}-1}{\beta-2}} r_1^{(\beta-1)^{N+1}} (4Nn)^{-\frac{(\beta-1)((\beta-1)^N-1)}{d(\beta-2)}},$$

which is (G.6) because $(\beta-1)^{N+1} = 1 + (\beta-2) \frac{(\beta-1)^{N+1}-1}{\beta-2}$. Squaring gives $\frac{3\alpha}{8} \varrho_{N+1}^2 = c_N n^{-\nu_N}$. $\square$

The shifts $\varrho_i = \vartheta_N r_i^{\beta-1}$ are much smaller than the radii r_i , so the radii shrink faster than geometrically. This is forced by the fixed Hölder seminorm: by (G.3), a tilt with slope of norm $2\alpha\varrho$ and radius r has seminorm $2\alpha\varrho r^{1-\beta} [\psi_u]_{\beta, \text{op}}$, where $[\psi_u]_{\beta, \text{op}} > 0$ does not depend on u , so shifts proportional to the radii would give tilts with seminorms growing like $r_i^{2-\beta}$. For $\sigma > 0$ and $N = 1$, Lemma G.3 is stronger than Lemma G.4, whose exponent is $\nu_1 = 2(\beta-1)/d > 2(\beta-1)/(2\beta+d)$; with exact values, Lemma G.4 is the only bound of this appendix.

G.4 Rounds needed for the sequential rate

Corollary G.5 (A necessary number of rounds). *Let $\sigma > 0$, let $d, \beta, L_\beta, \alpha, R$ and $L \geq 3\alpha$ be fixed, and let $N \geq 1$ be an integer such that $N = 1$ or*

$$(\beta-1)^N < 1 + \frac{d(\beta-2)}{2\beta}. \quad (\text{G.13})$$

Then for every $k \geq 0$,

$$\lim_{n \rightarrow \infty} n^{\frac{\beta-1}{\beta}} (\log(n+1))^{-k} \mathfrak{R}_{N,n}(\mathcal{F}_{\text{sc}}; \sigma) = \infty.$$

Consequently, if there are $C, k \geq 0$ such that for all sufficiently large n some algorithm with N rounds of at most n calls satisfies $\sup_{f \in \mathcal{F}_{\text{sc}}} \mathbb{E}[f(\hat{x}) - f^*] \leq C n^{-(\beta-1)/\beta} (\log(n+1))^k$, then

$$N \geq N_\beta(d) := \max \left\{ 2, \left\lceil \log_{\beta-1} \left(1 + \frac{d(\beta-2)}{2\beta} \right) \right\rceil \right\}. \quad (\text{G.14})$$

Proof. If $N = 1$, Lemma G.3 gives $\mathfrak{R}_{1,n}(\mathcal{F}_{\text{sc}}; \sigma) \geq c_1 n^{-2(\beta-1)/(2\beta+d)}$, with an exponent smaller than $(\beta - 1)/\beta$. Otherwise (G.13) is equivalent to $\nu_N < (\beta - 1)/\beta$ for the exponent ν_N of (G.7), and Lemma G.4 gives $\mathfrak{R}_{N,n}(\mathcal{F}_{\text{sc}}; \sigma) \geq c_N n^{-\nu_N}$. In both cases the risk is at least a positive constant, independent of n , times n^{-e} with $e < (\beta - 1)/\beta$, and $n^{(\beta-1)/\beta-e}(\log(n+1))^{-k} \rightarrow \infty$.

Under the assumption of the second claim, $\mathfrak{R}_{N,n}(\mathcal{F}_{\text{sc}}; \sigma) \leq C n^{-(\beta-1)/\beta}(\log(n+1))^k$ for all large n , so neither $N = 1$ nor (G.13) holds. Thus $N \geq 2$ and $N \geq \log_{\beta-1}(1 + d(\beta - 2)/(2\beta))$, which gives (G.14) because N is an integer. $\square$

For $\beta = 3$, for instance, $N_\beta(d) = \max\{2, \lceil \log_2(1 + d/6) \rceil\}$, which is 2 for $d \leq 18$ and 5 for $d = 100$. For fixed β , $N_\beta(d)$ grows like $\log d / \log(\beta - 1)$, and as $\beta \downarrow 2$ the logarithm in (G.14) tends to $d/4$. By the remark after Lemma G.1, the corollary also holds with n replaced by the total number of calls T .

G.5 Discussion

Remark G.6 (Hilbert–Schmidt seminorm). For a q -tensor A , $\|A\|_{\text{HS}} \leq d^{(q-1)/2} \|A\|_{\text{op}}$ (Lemma B.1(a)). Hence the class $\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)$ defined with the Hilbert–Schmidt seminorm contains the class defined with the operator seminorm and the constant $L_\beta d^{-(q-1)/2}$, and its minimax risk is at least as large. So Lemmas G.3 and G.4 hold for the Hilbert–Schmidt seminorm with L_β replaced by $L_\beta d^{-(q-1)/2}$ in all parameters. For fixed d this changes only the constants c_1 and c_N , and Lemma G.5 holds unchanged.

Remark G.7 (Size of the constants). The constants of Lemma G.4 are small: by (G.8), the factor $(\vartheta_N r_1^{\beta-2})^{2((\beta-1)^{N+1}-1)/(\beta-2)}$ in (G.7) is at most $32^{-2((\beta-1)^{N+1}-1)/(\beta-2)}$, which is below 10^{-45} for $\beta = 3$ and $N = 3$.

H ADVERSARIAL AND LIPSCHITZ ERRORS

This appendix proves the bound on $\delta_{\max}$ in Theorem 1(c) and the adversarial levels for convex objectives stated at the end of Section 6. Achievable levels follow from the descent results of Sections F and L: for strongly convex objectives they are Lemma F.6(b). The impossibility bounds use a hiding argument in which the adversarial error is a fixed function of the query point: Lemma H.1 for strongly convex and Lemma H.2 for convex objectives. Section H.5 treats two-point feedback and a probability criterion.

H.1 Setting

Admissible levels. Let $\mathcal{C}$ be a class, $\varepsilon > 0$, $T \geq 1$ and $M, \sigma \geq 0$. Then $\delta_{\max}(\varepsilon, T)$ is the supremum of the $\delta \geq 0$ for which some algorithm with at most T oracle calls, in any number of rounds, is ε -accurate on $\mathcal{C}$ at levels (M, σ, δ) . The level $M_{\max}(\varepsilon, T)$ is defined in the same way, with σ and δ fixed. For $M = 0$, respectively $\delta = 0$, these are the admissible levels of Section 2. Noise admissible at some levels is admissible at all larger levels, so both quantities are nonincreasing in the levels held fixed.

Classes and scales. Let $\beta > 2$, $q = \lceil \beta \rceil - 1$ and $m = \lceil q/2 \rceil$, and let $[f]_\beta$ and τ be as in Section 2. We use $\mathcal{F}_{\text{sc}} = \mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)$ with $0 < \alpha < L$ and $\kappa = L/\alpha$, and $\mathcal{F}_{\text{cvx}} = \mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$. For $\varepsilon > 0$, the quantities $T_\star(\varepsilon)$, h_ε and δ_ε of (4) and (5) satisfy

$$T_\star(\varepsilon) = d^2 \sigma^2 (\alpha \varepsilon)^{-1} h_\varepsilon^{-2}, \quad L_\beta^2 d^{-\tau} h_\varepsilon^{2\beta-2} = \alpha \varepsilon, \quad \frac{d \delta_\varepsilon^2}{\alpha h_\varepsilon^2} = \varepsilon, \quad (\text{H.1})$$

and for $\sigma > 0$, (3) gives $\Phi_{T_\star(\varepsilon)} = \varepsilon$ and $h_T = h_{\Phi_T}$.

H.2 Strongly convex objectives

Constants. Put $\rho = \min\{\alpha, (L - \alpha)/2\}$; let χ , K_β and K_2 be as in Lemma E.1, and ϑ , L_ϑ , $\tilde{K}_\beta$ and $\tilde{K}_2$ as in Lemma E.2.

- In the operator case, and in the Hilbert–Schmidt case with odd q , put $r_\beta = (\sqrt{32} K_\beta)^{\frac{1}{\beta-1}}$, $\ell_\star = 1$, $K_\star = \max\{1, K_2\}$ and $\nu_\star = \tau$.

- In the Hilbert–Schmidt case with even q , put $r_\beta = (\sqrt{32}\tilde{K}_\beta)^{\frac{1}{\beta-1}}$, $\ell_* = L_\vartheta$, $K_* = \max\{1, \tilde{K}_2\}$ and $\nu_* = 0$.

Define

$$\varepsilon_0 = \min\left\{\frac{\alpha R^2}{32}, \frac{1}{\alpha}\left(\frac{\rho r_\beta}{\sqrt{32}K_*}\right)^{\frac{2(\beta-1)}{\beta-2}} \frac{d^{\nu_*/(\beta-2)}}{L_\beta^{2/(\beta-2)}}\right\}, \quad C_{\text{hid}} = 2\sqrt{32}\ell_*r_\beta. \quad (\text{H.2})$$

Theorem H.1 (Hiding adversarial errors, strongly convex case). *Let $\beta > 2$, $0 < \alpha < L$, $R, L_\beta > 0$ and $\mathcal{C} = \mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L, R)$ with either seminorm, and let ε_0 and C_{hid} be as in (H.2). For every $\varepsilon \in (0, \varepsilon_0]$, every $T \geq 1$ and all $\sigma \geq 0$,*

$$\delta_{\max}(\varepsilon, T) \leq C_{\text{hid}} \min\{\sqrt{d}, \sqrt{\log(4T)}\} \delta_\varepsilon.$$

The adversarial error that achieves this is a fixed function of the query point, and the other noise components vanish.

By Lemma F.6(b), the level $c_\beta \delta_\varepsilon$ is admissible with at most $C_\beta(T_*(\varepsilon) + d\sqrt{\kappa}[1 + \log(LR^2/\varepsilon)])$ oracle calls, so the bound is attained up to the factor $C_\beta \min\{\sqrt{d}, \sqrt{\log(4T)}\}$.

Proof of Lemma H.1. Hard functions. Put $a = \sqrt{32\alpha\varepsilon}$, $\mu = \alpha + \rho$, $f_0(x) = \mu \|x\|^2/2$ and $\bar{h} = r_\beta h_\varepsilon$, and let $v \in \mathbb{S}^{d-1}$. In the operator case and in the Hilbert–Schmidt case with odd q , let

$$f_v(x) = f_0(x) + a \langle v, x \rangle \chi(\|x\|^2/\bar{h}^2), \quad a = \frac{L_\beta \bar{h}^{\beta-1}}{K_\beta d^{\tau/2}}.$$

In the Hilbert–Schmidt case with even q , let $\bar{w} = \bar{h}d^{-1/4}$ and

$$f_v(x) = f_0(x) + a\bar{w} \vartheta(\langle v, x \rangle / \bar{w}) \chi(\|x\|^2/\bar{h}^2), \quad a = \frac{L_\beta \bar{w}^{\beta-1}}{\tilde{K}_\beta}.$$

In both cases the second expression for a equals $\sqrt{32\alpha\varepsilon}$, by (H.1) and the choice of r_β , and the condition $\varepsilon \leq \varepsilon_0$ is equivalent to $a \leq \alpha R$ together with $K_*a/\bar{h} \leq \rho$ in the first case, or $K_*a/\bar{w} \leq \rho$ in the second. As in the membership proofs of Lemmas E.3 and E.4, $[f_v]_\beta \leq L_\beta$, $\alpha I \preceq \nabla^2 f_v \preceq (\alpha + 2\rho)I \preceq LI$ and $x_v^* = -(a/\mu)v \in B(R)$. Both families satisfy $f_{-v} - f_0 = -(f_v - f_0)$, $x_{-v}^* = -x_v^*$ and

$$|f_v(x) - f_0(x)| \leq \ell_* a |\langle v, x \rangle| \mathbf{1}\{\|x\| \leq \bar{h}\} \quad (x \in \mathbb{R}^d). \quad (\text{H.3})$$

Hiding. Let $\theta = \min\{1, \sqrt{2\log(4T)/(d-1)}\}$ for $d \geq 2$ and $\theta = 1$ for $d = 1$, let $A_v = \{x : |\langle v, x \rangle| > \theta \|x\|\}$, and put $\delta_{\text{hid}} = \ell_* a \theta \bar{h}$. The oracle returns $Y_v(x) = f_0(x)$ if $|f_v(x) - f_0(x)| \leq \delta_{\text{hid}}$ and $Y_v(x) = f_v(x)$ otherwise. Its error $Y_v - f_v$ is a Borel function of the query point bounded by δ_{hid} , hence admissible at level δ_{hid} for any round structure; the other noise components are zero. By (H.3), $Y_v = f_0$ outside A_v .

Let v be uniform on $\mathbb{S}^{d-1}$, independent of the seed Θ of the algorithm. Run the algorithm on the exact values of f_0 , and let $X_1^0, \dots, X_T^0$ be its queries and $\tilde{x}$ its output; they depend only on Θ . On $\Omega_v = \{X_t^0 \notin A_v \text{ for every } t \leq T\}$, the run on Y_v has the same queries and answers, by induction over rounds, so $\hat{x} = \tilde{x}$. If $\theta = 1$, then A_v is empty. Otherwise Lemma B.6, conditionally on Θ , gives $\mathbb{P}(X_t^0 \in A_v) \leq 2e^{-(d-1)\theta^2/2} = 1/(2T)$. In either case a union bound gives $\mathbb{P}(\Omega_v) \geq 1/2$.

Error. The event Ω_v does not change under $v \mapsto -v$, and $(\|y - x\|^2 + \|y + x\|^2)/2 \geq \|x\|^2$. Applied with $x = x_v^*$ and $y = \tilde{x}$, strong convexity and $\mu \leq 2\alpha$ give, with $\mathbb{E}_v$ the average over v ,

$$\mathbb{E}_v \mathbb{E}[f_v(\hat{x}) - f_v(x_v^*)] \geq \frac{\alpha}{2} \frac{a^2}{\mu^2} \mathbb{P}(\Omega_v) \geq \frac{a^2}{16\alpha} = 2\varepsilon,$$

so some instance has error larger than ε . Finally $\theta \leq 2 \min\{1, \sqrt{\log(4T)/d}\}$ for every $d \geq 1$ and $a\bar{h} = \sqrt{32}r_\beta\sqrt{d}\delta_\varepsilon$, hence $\delta_{\max}(\varepsilon, T) \leq \delta_{\text{hid}} \leq C_{\text{hid}} \min\{\sqrt{d}, \sqrt{\log(4T)}\} \delta_\varepsilon$. $\square$

H.3 Convex objectives: achievable levels

For convex objectives the gradient must be known to accuracy of order ε/R instead of $\sqrt{\alpha\varepsilon}$, and both methods of Theorem 4 are ε -accurate for adversarial levels of order $\varepsilon h/(R\sqrt{d})$, the gradient accuracy times the radius divided by $\sqrt{d}$. At the gradient scale $h \asymp (\varepsilon d^{\tau/2}/(L_\beta R))^{1/(\beta-1)}$, and at the value scale $h = (\varepsilon/(8V))^{1/2}$ with V as in Section 6. For $\tau = 0$ and $u = \varepsilon/(L_\beta R^\beta) \leq 1$ the first radius is of order $Ru^{1/(\beta-1)}$. For $\beta \in (2, 3]$, $V \leq L_\beta(2R)^{\beta-2}$ gives a value-scale radius of at least $(2^{\beta+1})^{-1/2}Ru^{1/2}$, which is larger for $\beta \in (2, 3)$ and $u < 1$, so there the value scale tolerates the larger level up to constants. For $\beta > 3$ only $V = L$ is available, and the value scale tolerates the larger level if and only if $L \lesssim \varepsilon/h^2$, with h the gradient-scale radius.

H.4 Convex objectives: an impossibility bound

For minimization over a ball with queries in the ball, the obstruction of Pasechnyuk-Vilensky et al. (2026), built from linear functions, gives a bound of the same order for every class that contains linear functions. Linear functions have no minimizer on $\mathbb{R}^d$, so for $\mathcal{F}_{\text{cvx}}$ we combine the localized functions of Lemma E.6 with the hiding argument of Lemma H.1.

Proposition H.2 (An impossibility bound for convex objectives). *Let $\beta > 2$, $L, R, L_\beta > 0$ and $\mathcal{C} = \mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$, with either the operator seminorm ($\nu = 0$) or the Hilbert–Schmidt seminorm ($\nu = \lfloor q/2 \rfloor$). Let K_β, K_2 be the constants of Lemma E.1, $\bar{K}_2 = \max\{1, K_2\}$, and*

$$\varepsilon_c = \min\left\{\frac{L_\beta(2\bar{K}_2 R)^\beta}{32\bar{K}_2 K_\beta d^{\nu/2}}, \left(\frac{L}{96\bar{K}_2^2}\right)^{\frac{\beta}{\beta-2}} \left(\frac{32\bar{K}_2 K_\beta d^{\nu/2}}{L_\beta}\right)^{\frac{2}{\beta-2}}\right\}.$$

For every $\varepsilon \in (0, \varepsilon_c]$, every $T \geq 1$, every number of rounds and all $M, \sigma \geq 0$,

$$\delta_{\max}(\varepsilon, T) \leq 64\bar{K}_2 \varepsilon \min\left\{1, \sqrt{\frac{\log(4T)}{d}}\right\} = 64\bar{K}_2 \min\{\sqrt{d}, \sqrt{\log(4T)}\} \frac{\varepsilon}{\sqrt{d}}.$$

Proof. Put $\bar{h} = (32\bar{K}_2 K_\beta d^{\nu/2} \varepsilon / L_\beta)^{1/\beta}$, $a = L_\beta \bar{h}^{\beta-1} / (K_\beta d^{\nu/2})$ and $\mu = 2\bar{K}_2 a / \bar{h}$, so that $a\bar{h} = 32\bar{K}_2 \varepsilon$ and $\mu = 64\bar{K}_2^2 \varepsilon / \bar{h}^2$. For $v \in \mathbb{S}^{d-1}$ let $f_0(x) = \mu \|x\|^2 / 2$ and $f_v(x) = f_0(x) + a \langle v, x \rangle \chi(\|x\|^2 / \bar{h}^2)$. As in the proof of Lemma E.6, with v in place of v_ω and $\bar{h}$ in place of h , $[f_v]_\beta \leq L_\beta$, $(\mu/2)I \preceq \nabla^2 f_v \preceq (3\mu/2)I$, and f_v has the minimizer $x_v^* = -(a/\mu)v$, of norm $\bar{h}/(2\bar{K}_2)$. The two conditions in $\varepsilon \leq \varepsilon_c$ are $\|x_v^*\| \leq R$ and $3\mu/2 \leq L$, so $f_v \in \mathcal{C}$.

The proof of Lemma H.1 applies with the same θ, A_v, Y_v and $\delta_{\text{hid}} = a\theta\bar{h}$. Strong convexity with modulus $\mu/2$ in its last step gives

$$\mathbb{E}_v \mathbb{E}[f_v(\hat{x}) - f_v(x_v^*)] \geq \frac{\mu}{4} \frac{a^2}{\mu^2} \frac{1}{2} = \frac{a\bar{h}}{16\bar{K}_2} = 2\varepsilon.$$

Since $a\bar{h} = 32\bar{K}_2 \varepsilon$ and $\theta \leq 2 \min\{1, \sqrt{\log(4T)/d}\}$, the asserted bound follows. $\square$

H.5 Two-point feedback and the probability criterion

Remark H.3 (Two-point oracles). In two-point feedback a call specifies two points x, x' and returns $f(x) + \eta(x) + \xi + \delta_t$ and $f(x') + \eta(x') + \xi' + \delta'_t$, where ξ, ξ' are stochastic (centered plus oblivious) errors and the adversarial errors δ_t, δ'_t are bounded by δ and chosen as in (A).

(i) *Achievability.* Suppose that the stochastic part is shared, $\xi = \xi'$, and that one call queries $z + ht_j U_i$ and $z - ht_j U_i$. The stochastic error then cancels in every difference of the stencil, and the Lipschitz and adversarial parts are as before, so the estimate is that of Lemma D.3 at levels $(M, 0, \delta)$. Hence Lemma F.6(b) and Theorem 4 hold with $\sigma = 0$, with T counting oracle calls and md oracle calls per frame. In particular, the levels $c_\beta \sqrt{\alpha\varepsilon}$ and $c_\beta \delta_\varepsilon$ are attained with $C_\beta md \sqrt{\kappa} [1 + \log(LR^2/\varepsilon)]$ oracle calls, and δ_ε does not depend on σ . Without sharing, a call consists of two one-point queries of the same round, and the one-point results apply.

(ii) *Impossibility.* The two instances of Lemma E.7 have the same function $f_\pm + \eta_\pm$, so its proof applies to every feedback that depends on the objective only through $f + \eta$ and whose other noise components do not depend on (f, η) , in particular to two-point feedback. The hard instances of Lemma H.1 and Lemma H.2 have $M = \sigma = 0$ and an adversarial error that is a fixed function of the point. Feedback that returns these values at r points per call is simulated by r one-point queries, so the proofs apply with rT queries: for $d \geq 2$ let

$\theta = \min\{1, \sqrt{2\log(4rT)/(d-1)}\}$, and $\theta = 1$ for $d = 1$. If $\theta = 1$, the cone A_v is empty and $\mathbb{P}(\Omega_v^c) = 0$. Otherwise $\theta^2 = 2\log(4rT)/(d-1)$, and the union bound for the event Ω_v of the proof of Lemma H.1 and Lemma B.6 give

$$\mathbb{P}(\Omega_v^c) \leq 2rTe^{-(d-1)\theta^2/2} = 1/2.$$

Both bounds hold with $\log(4rT)$ in place of $\log(4T)$. For $r = 2$, $\log(8T) \leq \frac{3}{2}\log(4T)$, so the constants grow by at most $\sqrt{3/2}$.

Remark H.4 (The probability criterion of Pasechnyuk-Vilensky et al. (2026)). Pasechnyuk-Vilensky et al. (2026) require $\mathbb{P}\{f(\hat{x}) - \inf f \leq \varepsilon\} \geq 1 - \varpi$ for a fixed $\varpi \in (0, 1/2)$ (called β there). The hiding argument also applies to this criterion, with $r \geq 1$ points per call as in Lemma H.3(ii). For $d \geq 2$ take $\theta = \min\{1, \sqrt{2\log(4rT/(1-2\varpi))/(d-1)}\}$, and $\theta = 1$ for $d = 1$; then $\mathbb{E}_v\mathbb{P}(\Omega_v^c) \leq (1-2\varpi)/2$, by the same two cases ($A_v = \emptyset$ if $\theta = 1$, the union bound otherwise), where $\mathbb{P}$ refers to the randomness of the algorithm for fixed v and $\mathbb{E}_v$ averages over v uniform on $\mathbb{S}^{d-1}$.

On Ω_v the outputs for v and $-v$ coincide, and one point x cannot solve both instances. Otherwise strong convexity would give $\|x - x_v^*\|^2, \|x + x_v^*\|^2 \leq 2\varepsilon/\alpha$ in Lemma H.1, and $\leq 4\varepsilon/\mu$ in Lemma H.2, and the parallelogram identity would give $\|x_v^*\|^2 \leq 2\varepsilon/\alpha$, respectively $\leq 4\varepsilon/\mu$; in these constructions, however, $\|x_v^*\|^2 \geq 8\varepsilon/\alpha$, respectively $= 16\varepsilon/\mu$. Hence, for each v , the success probabilities for v and $-v$ sum to at most $1 + \mathbb{P}(\Omega_v^c)$. Averaging over v , the mean success probability is at most $(1 + \mathbb{E}_v\mathbb{P}(\Omega_v^c))/2 \leq (3 - 2\varpi)/4 < 1 - \varpi$, so some instance is solved with probability less than $1 - \varpi$. Therefore the levels of Lemma H.1 and Lemma H.2, with $\log(4rT/(1-2\varpi))$ in place of $\log(4T)$, are not admissible for this criterion. This covers the deterministic oracle of Pasechnyuk-Vilensky et al. (2026) with unrestricted queries.

I HESSIAN GROWING WITH THE GRADIENT

This appendix states and proves Lemma I.1, which Section 3 summarizes. Throughout, f belongs to $\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, \infty, R)$, with either seminorm, and satisfies

$$\|\nabla^2 f(x)\|_{\text{op}} \leq L_0 + L_1 \|\nabla f(x)\| \quad (x \in \mathbb{R}^d) \quad (\text{I.1})$$

with $L_0 > 0$ and $L_1 \geq 0$, the condition of Zhang et al. (2020). At the minimizer $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succeq \alpha I$, so $\alpha \leq L_0$. The method is the projected two-phase method of Lemma F.3, which needs an upper bound on the Hessian only on $B(R)$; the two lemmas below provide it.

Theorem I.1 (Hessian growing with the gradient). *Let $\beta > 2$, $\alpha, L_\beta, R, L_0, \sigma > 0$, $L_1 \geq 0$, and let $f \in \mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, \infty, R)$, with either seminorm, satisfy (I.1). Put $L_{\text{eff}} = \min\{L_0 e^{2L_1 R}, C_\beta(L_0(1+L_1 R)^{q-2} + L_\beta R^{\beta-2})\}$ and $\kappa_{\text{eff}} = L_{\text{eff}}/\alpha$. If $T \geq C_\beta d \kappa_{\text{eff}}(1 + \log(d\kappa_{\text{eff}}) + \log^+(\alpha R^2/\Phi_1))$, the projected two-phase method with L_{eff} in place of L and radius h_T satisfies (6) for every noise admissible at levels (M, σ, δ) , in at most $C_\beta \kappa_{\text{eff}}(1 + \log^+(\alpha R^2/\Phi_T))$ rounds. If $L_0 > \alpha$, the lower bounds of Theorem 1(b), (c) hold with $L = L_0$.*

Lemma I.2 (Growth along rays). *For every x , with $s = \|x - x^*\|$,*

$$\|\nabla f(x)\| \leq \frac{L_0}{L_1} (e^{L_1 s} - 1) \quad (L_0 s \text{ if } L_1 = 0), \quad \|\nabla^2 f(x)\|_{\text{op}} \leq L_0 e^{L_1 s}.$$

Proof. Write $x = x^* + su$ with $\|u\| = 1$ and put $g(t) = \|\nabla f(x^* + tu)\|$. Integration of the Hessian along the segment and (I.1) give $g(t) \leq w(t) := L_0 t + L_1 \int_0^t g$. Then $w(0) = 0$ and $w' \leq L_0 + L_1 w$, and multiplication by $e^{-L_1 t}$ and integration bound w , hence g , by the first claim. Inserting it into (I.1) gives the second. $\square$

Lemma I.3 (Polynomial growth). *For every x , with $s = \|x - x^*\|$: if $q = 2$, then $\|\nabla^2 f(x)\|_{\text{op}} \leq L_0 + L_\beta s^{\beta-2}$; if $q \geq 3$, then, with $n = q - 2$ and $C_n = (n + 1)n^n$,*

$$\|\nabla^2 f(x)\|_{\text{op}} \leq e C_n L_0 \max\{1, L_1 s\}^n + \frac{C_n + 1}{n!} L_\beta s^{\beta-2}.$$

Proof. Both seminorms bound $[f]_{\beta, \text{op}}$. If $q = 2$, then $\|\nabla^2 f(x^*)\|_{\text{op}} \leq L_0$ by (I.1), and $\|\nabla^2 f(x) - \nabla^2 f(x^*)\|_{\text{op}} \leq L_\beta s^{\beta-2}$. Let $q \geq 3$ and $s > 0$, write $x = x^* + su$ with $\|u\| = 1$, fix unit vectors v, w and put $\varphi(t) = D^2 f(x^* + tu)[v, w]$. Then $\varphi \in C^n$, and $\varphi^{(n)}(t) = D^q f(x^* + tu)[v, w, u, \dots, u]$ is γ -Hölder with constant L_β . By

Lemma B.2(a) in dimension one, the Taylor polynomial P of degree n of φ at 0 satisfies $|\varphi(t) - P(t)| \leq St^{n+\gamma}$ for $t \geq 0$, with $S = L_\beta/n!$, and $n + \gamma = \beta - 2$. Let $r = \min\{s, 1/L_1\}$ ($r = s$ if $L_1 = 0$). On $[0, r]$, Lemma I.2 gives $|\varphi| \leq eL_0$, hence $|P| \leq eL_0 + Sr^{n+\gamma}$. Lagrange interpolation at the nodes ir/n , $0 \leq i \leq n$, whose basis polynomials are bounded at $s \geq r$ by $(ns/r)^n$, gives $|P(s)| \leq (n+1)(ns/r)^n(eL_0 + Sr^{n+\gamma})$. Since $(s/r)^n r^{n+\gamma} \leq s^{n+\gamma}$ and $s/r = \max\{1, L_1 s\}$,

$$|\varphi(s)| \leq |P(s)| + Ss^{n+\gamma} \leq eC_n L_0 \max\{1, L_1 s\}^n + (C_n + 1)Ss^{\beta-2}.$$

The supremum over v, w is the operator norm of the symmetric matrix $\nabla^2 f(x)$. For $s = 0$ the claim follows from (I.1). $\square$

For $x \in B(R)$ we have $\|x - x^*\| \leq 2R$, so the two lemmas give $\nabla^2 f \preceq L_{\text{eff}} I$ on $B(R)$ with

$$L_{\text{eff}} = \min\left\{L_0 e^{2L_1 R}, eC_n L_0 \max\{1, 2L_1 R\}^n + \frac{C_n + 1}{n!} L_\beta (2R)^{\beta-2}\right\} \quad (\text{I.2})$$

for $q \geq 3$, and with the second term replaced by $L_0 + L_\beta (2R)^{\beta-2}$ for $q = 2$. Since $\max\{1, 2t\} \leq 2(1+t)$, this is at most the value $\min\{L_0 e^{2L_1 R}, C_\beta(L_0(1+L_1 R)^{q-2} + L_\beta R^{\beta-2})\}$ used in Lemma I.1, and both are at least $L_0 \geq \alpha$.

Proof of Lemma I.1. Upper bound. Apply Lemma F.3(b) with L_{eff} in place of L and $\kappa_{\text{eff}} = L_{\text{eff}}/\alpha \geq 1$ in place of κ . By the remark after that theorem, it needs the Hessian bound only on $B(R)$, which (I.2) provides; the estimate of Procedure 1 uses no bound on the Hessian, even at queries outside $B(R)$. The budget condition (F.8), the number of rounds and the bound $(C_\Phi + 1)\Phi_T + C_M M^2/\alpha + C_\delta d\delta^2/(\alpha h_T^2)$ take the form stated in Lemma I.1 after enlarging C_β , since $d\delta^2/(\alpha h_T^2) = (T\delta^2/(d\sigma^2))\Phi_T$.

Lower bounds. A function with $\nabla^2 f \preceq L_0 I$ satisfies (I.1) for every $L_1 \geq 0$. If $L_0 > \alpha$, the class therefore contains $\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, L_0, R)$, on which Theorem 1(b) and the converse bounds of Theorem 1(c) hold. $\square$

Remark I.4 (The power of $L_1 R$). For $4 < \beta \leq 5$, so that $q = 4$ and $n = 2$, the power n in Lemma I.3 cannot be lowered. For $k > 0$ let $f_k(x) = \frac{1}{2}\|x\|^2 + \frac{k^2}{12}x_1^4$. Then $\nabla^2 f_k(x) = I + k^2 x_1^2 e_1 e_1^\top \succeq I$, $D^4 f_k$ is constant, so $[f_k]_\beta = 0$, and the minimizer is 0. With $t = k|x_1|$, $\|\nabla^2 f_k(x)\|_{\text{op}} = 1 + t^2 \leq 2 + 3t + t^3 \leq 2 + 3k\|\nabla f_k(x)\|$, so (I.1) holds with $L_0 = 2$ and $L_1 = 3k$, while at $x = Re_1$ the Hessian has norm $1 + (L_1 R/3)^2$. Hence no bound of the form $C \max\{1, L_1 R\}^a + C'$ on $B(R)$ with $a < 2$ and C, C' independent of L_1 holds on these classes.

Remark I.5 (Rounds). The accelerated method of Lemma F.6 is not projected: its extrapolated points may leave $B(R)$, where (I.2) does not bound the Hessian. Whether a round bound proportional to $\sqrt{\kappa_{\text{eff}}} \log(1/\varepsilon)$ suffices under (I.1) is open.

J POLYNOMIAL-PILOT METHODS FOR STRONGLY CONVEX OBJECTIVES

This appendix proves Theorem 3, except for the method with logarithmically many rounds, which is treated in Section K. In this class, the curvature of f and the coefficients of its Taylor polynomial of degree q may have arbitrary magnitude. The methods below estimate this polynomial from function values and then minimize the remainder. The upper bounds are Lemma J.11 ($\beta > 4$, $T \geq C_0 d^q$) and Lemma J.17 ($2 < \beta \leq 4$, $T \geq C_0 d^2$), where C_0 depends only on $\beta, \alpha, R, L_\beta, \sigma$; Section J.6 combines them with the lower bound of Theorem 1(b).

J.1 Setting

Throughout this appendix $\beta > 2$, $q = \lceil \beta \rceil - 1 \geq 2$, $\gamma = \beta - q \in (0, 1]$, $m = \lceil q/2 \rceil$ and $\alpha, L_\beta, R > 0$. We write $B(r) = \{x : \|x\| \leq r\}$, $\bar{R} = R + 1$, $f^* = \min f$, $K_j = \binom{d+j}{j}$ for the dimension of the space of polynomials of degree at most j on $\mathbb{R}^d$, and $p_\beta = (\beta - 1)/\beta$, so that Φ_T of (3) is proportional to T^{-p_β} . As in Section 2, either L_β bounds $[f]_{\beta, \text{op}}$ and $\tau = 0$, or L_β bounds $[f]_{\beta, \text{HS}}$ and $\tau = \tau_\beta$. In both cases $[f]_{\beta, \text{op}} \leq L_\beta$, and only the stencil bounds (J.2) depend on the choice.

Definition J.1 (Class, centered noise and minimax risk). The class $\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, \infty, R)$ is the class of Section 2 with $L = \infty$: $\nabla^2 f \succeq \alpha I$, $[f]_\beta \leq L_\beta$ and $x^* \in B(R)$, while derivatives of orders $2, \dots, q$ may have arbitrary

magnitude. The noise is *centered at level* $\sigma \geq 0$ if $\eta \equiv 0$, $\xi^\circ \equiv 0$, $\delta_t \equiv 0$ and the centered component satisfies (C); a query X_t then returns $Y_t = f(X_t) + \xi_t$ with

$$\mathbb{E}[\xi_t \mid \mathcal{E}_t] = 0, \quad \mathbb{E}[\xi_t^2 \mid \mathcal{E}_t] \leq \sigma^2.$$

For a class $\mathcal{C}$, $\mathfrak{R}_T(\mathcal{C}; \sigma) = \inf \sup \mathbb{E}[f(\hat{x}) - f^*]$, where the infimum is over all algorithms that know $d, T, \beta, \alpha, R, L_\beta, \sigma$ and make at most T oracle calls in any number of rounds, and the supremum is over $f \in \mathcal{C}$ and all noises centered at level σ .

A function with $\nabla^2 f \succeq \alpha I$ is α -strongly convex, so its minimizer exists and is unique. Adding a polynomial of degree at most q to f adds a constant to $D^q f$ and changes neither $[f]_{\beta, \text{op}}$ nor $[f]_{\beta, \text{HS}}$. Exact values and independent $N(0, \sigma^2)$ errors that are independent of $\Theta_1, \Theta_2, \dots$ are centered noise at level σ . For centered noise the error $\xi_t = Y_t - f(X_t)$ is a function of the query and the answer. Hence $\mathcal{F}_{k-1}$ contains the errors of the rounds before round k , and $\mathcal{E}_t$ contains $\mathcal{F}_{k-1}$, Θ_k and the errors of the earlier queries of round k .

The stencil under centered noise. Let the center z of round k be bounded and $\mathcal{F}_{k-1}$ -measurable, and let the radius $h > 0$ be deterministic. We write $\bar{g}_h(z) = \bar{g}_{h,1}(z)$ for one round of Procedure 1 with one frame ($2md$ oracle calls). Lemma D.3(i), (ii) and (iv), with $M = \delta = 0$, $\xi^\circ \equiv 0$ and $B = 1$, give

$$\bar{g}_h(z) = \nabla f(z) + b(z) + \zeta, \quad \|b(z)\| \leq b_h, \quad \mathbb{E}[\zeta \mid \mathcal{F}_{k-1}] = 0, \quad \mathbb{E}[\|\zeta\|^2 \mid \mathcal{F}_{k-1}] \leq \varsigma_h^2, \quad (\text{J.1})$$

where $b(z) = G_h(z) - \nabla f(z)$ is $\mathcal{F}_{k-1}$ -measurable, ζ is $\mathcal{F}_k$ -measurable, and

$$b_h = C_b L_\beta d^{-\tau/2} h^{\beta-1}, \quad \varsigma_h^2 = C_f L_\beta^2 d^{-\tau} h^{2\beta-2} + C_\sigma \frac{d\sigma^2}{h^2}; \quad (\text{J.2})$$

for $d = 1$ the variance bound holds a fortiori. Since $\mathbb{E}[\|b(z) + \zeta\|^2 \mid \mathcal{F}_{k-1}] = \|b(z)\|^2 + \mathbb{E}[\|\zeta\|^2 \mid \mathcal{F}_{k-1}]$, for every bounded $\mathcal{F}_{k-1}$ -measurable random vector w and every event $G \in \mathcal{F}_{k-1}$,

$$\mathbb{E}[\langle \zeta, w \rangle \mathbf{1}_G] = 0, \quad \mathbb{E}[\|\bar{g}_h(z) - \nabla f(z)\|^2 \mathbf{1}_G] \leq (b_h^2 + \varsigma_h^2) \mathbb{P}(G). \quad (\text{J.3})$$

With B frames, (J.1) and (J.3) hold with ς_h^2/B in place of ς_h^2 , and so does Lemma J.9 below.

Conventions. Minimizers and projections of known continuous functions over compact sets are chosen by Lemma J.2, so they are Borel functions of the data; computation is not charged (Lemma J.18). We write $\|X\|_{L^v} = (\mathbb{E}|X|^v)^{1/v}$ and $\text{clip}_{[-b,b]}(x) = \max\{-b, \min\{b, x\}\}$ for $b > 0$. In statements, C_β and c_β depend only on β . In proofs, C and c denote positive numbers that depend only on $\beta, \alpha, R, L_\beta, \sigma$ and on integers fixed in the proof as functions of β ; they never depend on d or T , and their values may change from line to line.

J.2 Tools

Lemma J.2 (Borel choice of minimizers). *Let X be a metric space, $Y \subset \mathbb{R}^n$ compact, $\mathcal{K} \subseteq X \times Y$ closed with nonempty sections $\mathcal{K}_x = \{y : (x, y) \in \mathcal{K}\}$, and $\psi : \mathcal{K} \rightarrow \mathbb{R}$ continuous. There is a Borel map $\pi : X \rightarrow Y$ with $\pi(x) \in \mathcal{K}_x$ and $\psi(x, \pi(x)) = \min_{y \in \mathcal{K}_x} \psi(x, y)$ for every x .*

Proof. Extend ψ to a continuous function on $X \times Y$ by Tietze's theorem. For closed $F \subseteq Y$, the set $\{x : \mathcal{K}_x \cap F \neq \emptyset\}$ is the projection of the closed set $\mathcal{K} \cap (X \times F)$ along the compact factor Y , hence closed. Every open subset of Y is a countable union of closed sets, so $x \mapsto \mathcal{K}_x$ is weakly measurable with nonempty compact values. The measurable maximum theorem (Aliprantis and Border, 2006, Theorem 18.19), applied to $-\psi$ on the Borel σ -algebra of X , gives π . $\square$

The next lemma concerns the median of block means (Nemirovski and Yudin, 1983; Hsu and Sabato, 2016), for errors that are martingale differences rather than independent.

Lemma J.3 (Medians of block means under martingale noise).

- (a) Let r be odd, $j = (r + 1)/2$, and let $Z_1, \dots, Z_r$ be real random variables adapted to a filtration $(\mathcal{G}_i)_{i=0}^r$ with $\mathbb{E}[Z_i^2 | \mathcal{G}_{i-1}] \leq v$ for all i . Let med be their median. Then

$$|\text{med}|^{2j} \leq \sum_S \prod_{i \in S} Z_i^2,$$

the sum being over all $S \subseteq \{1, \dots, r\}$ with $|S| = j$. Hence $\mathbb{E}|\text{med}|^{2j} \leq \binom{r}{j} v^j$, and $\mathbb{P}(|\text{med}| > u) \leq \binom{r}{j} (v/u^2)^j \leq 2^r (v/u^2)^j$ for $u > 0$.

- (b) Let $\text{med}_1, \dots, \text{med}_J$ be as in (a), each with its own filtration, and $E_0 = \max_i |\text{med}_i|$. For $u > 0$ and $0 < a < 2j$,

$$\mathbb{P}(E_0 > u) \leq J 2^r \left(\frac{v}{u^2} \right)^j, \quad \mathbb{E}[E_0^a \mathbf{1}\{E_0 > u\}] \leq J 2^r v^j \frac{2j}{2j-a} u^{a-2j}.$$

- (c) Let the noise be centered at level σ , and let the queries of round k be made in an order that is fixed before the round. Let $I_1, \dots, I_r$ be disjoint sets of queries of the round such that every query of I_i precedes every query of I_{i+1} , let w_t be deterministic numbers, and put $Z_i = \sum_{t \in I_i} w_t \xi_t$. Let $\mathcal{G}_i$ be the σ -algebra generated by $\mathcal{F}_{k-1}$, Θ_k and the errors of all queries of the round up to the last query of I_i , and $\mathcal{G}_0 = \mathcal{F}_{k-1} \vee \sigma(\Theta_k)$. Then (Z_i) is adapted to $(\mathcal{G}_i)$, $\mathbb{E}[Z_i | \mathcal{G}_{i-1}] = 0$ and $\mathbb{E}[Z_i^2 | \mathcal{G}_{i-1}] \leq \sigma^2 \sum_{t \in I_i} w_t^2$.

Proof. (a) At least j of the Z_i have absolute value at least $|\text{med}|$, which gives the pointwise bound. For $S = \{i_1 < \dots < i_j\}$, conditioning successively on $\mathcal{G}_{i_j-1}, \dots, \mathcal{G}_{i_1-1}$ gives $\mathbb{E} \prod_{i \in S} Z_i^2 \leq v^j$; the integrands are nonnegative. Summation over S and Markov's inequality give the moment and tail bounds.

- (b) The first bound is a union bound, and the second follows from it and $\mathbb{E}[E_0^a \mathbf{1}\{E_0 > u\}] = u^a \mathbb{P}(E_0 > u) + \int_u^\infty a t^{a-1} \mathbb{P}(E_0 > t) dt$.

- (c) For $t \in I_i$ we have $\mathcal{G}_{i-1} \subseteq \mathcal{E}_t$, so the tower property gives $\mathbb{E}[\xi_t | \mathcal{G}_{i-1}] = 0$ and $\mathbb{E}[\xi_t^2 | \mathcal{G}_{i-1}] \leq \sigma^2$. For $t < t'$ in I_i , ξ_t is $\mathcal{E}_{t'}$ -measurable, so $\mathbb{E}[\xi_t \xi_{t'} | \mathcal{G}_{i-1}] = 0$. Expanding the square proves the claim. $\square$

Lemma J.4 (Maximum-volume design). *Let $\phi = (\phi_1, \dots, \phi_K)$ be a basis of the polynomials of degree at most q on $\mathbb{R}^d$, $K = K_q$, let $\mathcal{D} \subset \mathbb{R}^d$ be a closed ball of positive radius, and let $J \geq K$. Let $x_1, \dots, x_J \in \mathcal{D}$ maximize $\det \mathbf{G}$, where $\mathbf{G} = \sum_{i=1}^J \phi(x_i) \phi(x_i)^\top$. Then $\mathbf{G} \succ 0$ and $\sup_{x \in \mathcal{D}} \phi(x)^\top \mathbf{G}^{-1} \phi(x) \leq K/(J - K + 1)$.*

Proof. The maximum exists by compactness, and it is positive because the vectors $\phi(x)$, $x \in \mathcal{D}$, span $\mathbb{R}^K$: a polynomial that vanishes on a ball is zero. Fix $x \in \mathcal{D}$ and append the row $\phi(x)^\top$ to the $J \times K$ matrix with rows $\phi(x_i)^\top$. By maximality, replacing one of the first J rows by the last one does not increase $\det \mathbf{G}$, so the first J rows form a dominant submatrix, and [Mikhalev and Oseledets \(2018, Theorem 4.2\)](#) gives $c \in \mathbb{R}^J$ with $\phi(x) = \sum_i c_i \phi(x_i)$ and $\|c\|^2 \leq K/(J - K + 1)$. The smallest $\|c\|^2$ among such c equals $\phi(x)^\top \mathbf{G}^{-1} \phi(x)$. $\square$

Lemma J.5 (Polynomials on balls). *Let p be a polynomial of degree at most q on $\mathbb{R}^d$ and $\rho > 0$.*

- (a) $\|\nabla^2 p(x)\|_{\text{op}} \leq q^4 \rho^{-2} \sup_{B(2\rho)} |p|$ for $\|x\| \leq \rho$.

- (b) $\sup_{B(\rho)} |p| \leq C_q^{\text{pol}} (\rho/a)^q \sup_{B(a)} |p|$ for $0 < a \leq \rho$, where $C_q^{\text{pol}} = \sum_{j=0}^q q^{2j}/j!$.

Proof. By Markov's inequality ([Borwein and Erdélyi, 1995](#), Chapter 5), a univariate polynomial φ of degree at most q satisfies $\sup_{[-t,t]} |\varphi'| \leq (q^2/t) \sup_{[-t,t]} |\varphi|$ for $t > 0$. (a) For a unit vector u , apply it twice to $\varphi(t) = p(x + tu)$ on $[-\rho, \rho]$; the points $x + tu$ lie in $B(2\rho)$, and $\varphi''(0) = u^\top \nabla^2 p(x) u$. (b) For a unit vector v , applied j times to $\varphi(t) = p(x + tv)$ on $[-a, a]$ it gives $|\varphi^{(j)}(0)| \leq (q^2/a)^j \sup_{B(a)} |p|$. Insert this into the Taylor expansion of φ at 0 and use $(\rho/a)^j \leq (\rho/a)^q$. $\square$

For data $y \in \mathbb{R}^J$ at the nodes of the design of [Lemma J.4](#) let $\Pi y = \mathbf{c}^\top \phi$, with $\mathbf{c} = \mathbf{G}^{-1} \sum_i y_i \phi(x_i)$, be the least-squares fit by a polynomial of degree at most q . It is linear in y and reproduces polynomials of degree at most q .

Lemma J.6 (Stability of the fit). *Let the design of Lemma J.4 be taken on $\mathcal{D} = B(s)$, $s > 0$, with $J = 2K_q$, and put $\Gamma_0 = \sqrt{2K_q}$. If $\max_i |y_i| \leq \bar{y}$, then $\sup_{\mathcal{D}} |\Pi y| \leq \Gamma_0 \bar{y}$. If $s = 2\bar{R}$, then also $\|\nabla^2 \Pi y(x)\|_{\text{op}} \leq \Gamma_2 \bar{y}$ for $\|x\| \leq \bar{R}$, where $\Gamma_2 = \Gamma_0 q^4 / \bar{R}^2$.*

Proof. The fitted values $(\Pi y(x_i))_i$ are the orthogonal projection of y onto a subspace of $\mathbb{R}^J$, so $\mathbf{c}^\top \mathbf{G} \mathbf{c} = \sum_i \Pi y(x_i)^2 \leq J \bar{y}^2$. By the Cauchy–Schwarz inequality and Lemma J.4, $|\Pi y(x)|^2 \leq (\mathbf{c}^\top \mathbf{G} \mathbf{c}) \phi(x)^\top \mathbf{G}^{-1} \phi(x) \leq 2K_q \bar{y}^2$ on $\mathcal{D}$. The second claim follows from Lemma J.5(a) with $\rho = \bar{R}$. $\square$

J.3 Pilot fit, guard and corrected residual method

The pilot. Fix a radius $s > 0$, the design $x_1, \dots, x_J$ of Lemma J.6 on $B(s)$ ($J = 2K_q$ nodes), an odd integer $r \geq 3$ and an integer $\ell \geq 1$. In one round, query every node $r\ell$ times, node after node, and split the $r\ell$ values at x_i into r consecutive blocks of ℓ values. Let $\tilde{f}_i$ be the median of the r block means at x_i , let $P = \Pi \tilde{f}$ be the fit of these medians, and put $\nu_i = \tilde{f}_i - f(x_i)$ and $E_0 = \max_i |\nu_i|$. The pilot uses $n_0 = 2K_q r \ell$ oracle calls in one round, at points fixed in advance. Outside Section K we take $s = 2\bar{R}$. Let $\varrho_j = L_\beta (2\bar{R})^{\beta-j} / (q-j)!$ for $j = 0, 2$, and for a tolerance $\theta \geq 1$ put

$$\Upsilon_\theta = \varrho_0 + \Gamma_0(\varrho_0 + \theta), \quad \Lambda_\theta = \varrho_2 + \Gamma_2(\varrho_0 + \theta). \quad (\text{J.4})$$

Since $\bar{R} \geq 1$ and $\theta \geq 1$, both numbers are at most $C\sqrt{K_q}\theta$.

Lemma J.7 (Accuracy of the pilot). *Let $f \in C^q$ with $[f]_{\beta, \text{op}} \leq L_\beta$, and let $\mathbf{t} = T_q f(0; \cdot)$ be its Taylor polynomial of degree q at 0 (Section B.2). On every realization, $\sup_{B(s)} |P - \mathbf{t}| \leq \Gamma_0(L_\beta s^\beta / q! + E_0)$. If $s = 2\bar{R}$, then on every realization*

$$\sup_{B(R)} |f - P| \leq \varrho_0 + \Gamma_0(\varrho_0 + E_0), \quad \sup_{x \in B(R)} \|\nabla^2(f - P)(x)\|_{\text{op}} \leq \varrho_2 + \Gamma_2(\varrho_0 + E_0); \quad (\text{J.5})$$

in particular, on the event $\{E_0 \leq \theta\}$ the residual $g = f - P$ satisfies $|g| \leq \Upsilon_\theta$ and $\|\nabla^2 g\|_{\text{op}} \leq \Lambda_\theta$ on $B(R)$. If the noise is centered at level σ , then for $u > 0$

$$\mathbb{P}(E_0 > u) \leq J 2^r \left(\frac{\sigma^2}{\ell u^2} \right)^{(r+1)/2}, \quad \mathbb{E}[E_0 \mathbf{1}\{E_0 > u\}] \leq 2J 2^r \left(\frac{\sigma^2}{\ell} \right)^{(r+1)/2} u^{-r}. \quad (\text{J.6})$$

Proof. The fit is linear and reproduces $\mathbf{t}$, so $P - \mathbf{t} = \Pi((f - \mathbf{t})(x_i))_i + \Pi \nu$. By Lemma B.2(a), $|f - \mathbf{t}| \leq L_\beta \|x\|^\beta / q!$, which is at most $L_\beta s^\beta / q!$ at the nodes, and Lemma J.6 gives the first claim. For $s = 2\bar{R}$ write $f - P = (f - \mathbf{t}) - (P - \mathbf{t})$. On $B(R)$, $|f - \mathbf{t}| \leq \varrho_0$ and, by Lemma B.2(c) with $j = 2$ and $\nabla^2 \mathbf{t}(x) = T_{q-2} D^2 f(0; x)$, $\|\nabla^2(f - \mathbf{t})\|_{\text{op}} \leq \varrho_2$; Lemma J.6 bounds $P - \mathbf{t}$ and its Hessian. This gives (J.5).

For the tails, ν_i is the median of the r block means of the errors at x_i . By Lemma J.3(c) with weights $1/\ell$, these means satisfy the assumption of Lemma J.3(a) with $v = \sigma^2/\ell$. Lemma J.3(b) with $j = (r+1)/2$ and $a = 1$, where $2j/(2j-1) \leq 2$, gives (J.6). $\square$

Lemma J.8 (Polynomial-value guard). *Let P be continuous, let x_P minimize P over $B(R)$, let $z \in B(R)$ and $\Upsilon > 0$. Output $\hat{x} = z$ if $P(z) \leq P(x_P) + 2\Upsilon$, and $\hat{x} = x_P$ otherwise. If $\sup_{B(R)} |f - P| \leq \Upsilon$, then $f(\hat{x}) \leq f(z)$. On every realization, $f(\hat{x}) - f^* \leq 2\Upsilon + 2 \sup_{B(R)} |f - P|$.*

Proof. If $\sup_{B(R)} |f - P| \leq \Upsilon$ and z is rejected, then $f(z) - f(x_P) \geq P(z) - P(x_P) - 2\Upsilon > 0$. In both cases $\hat{x} \in B(R)$ and $P(\hat{x}) \leq P(x_P) + 2\Upsilon \leq P(x^*) + 2\Upsilon$, because $x^* \in B(R)$. Comparing f with P at $\hat{x}$ and at x^* gives the second claim. $\square$

A corrected residual method. Let P be a polynomial of degree at most q that is $\mathcal{F}_{k_1}$ -measurable, where k_1 is the last round of the pilot, let $g = f - P$, and fix a radius $h > 0$. At a center $x \in B(R)$ the residual oracle makes one round of Procedure 1 with radius h and returns $\hat{v}(x) = \bar{g}_h(x) - \nabla P(x)$. By (J.1), $\hat{v}(x) = \nabla g(x) + b(x) + \zeta$: the residual gradient is estimated with the error of $\bar{g}_h$, and no coefficient of P enters the bounds. With step sizes

$\mu_k > 0$ and a starting point $z_1 \in B(R)$ that is measurable with respect to the rounds before the first iteration, iteration k uses two rounds:

$$\begin{aligned} y_k &\in \arg \min_{x \in B(R)} \left\{ P(x) + \langle \hat{v}(z_k), x \rangle + \frac{\|x - z_k\|^2}{2\mu_k} \right\}, \\ z_{k+1} &= \text{Proj}_{B(R)}(y_k - \mu_k(\hat{v}(y_k) - \hat{v}(z_k))). \end{aligned} \quad (\text{J.7})$$

The first line treats the known polynomial exactly; the second corrects the step with a fresh estimate at y_k , as in the forward-backward-forward method of [Tseng \(2000\)](#). By Lemma [J.2](#), y_k is a Borel function of P , $\hat{v}(z_k)$, z_k and μ_k . Hence z_k is measurable with respect to the rounds before iteration k , y_k with respect to these and the first round of iteration k , and both lie in $B(R)$. If $\|\nabla^2 g\|_{\text{op}} \leq \Lambda$ on $B(R)$ and $\mu_k \Lambda \leq 1/2$, the objective of the first line has Hessian $\nabla^2 f - \nabla^2 g + \mu_k^{-1} I \succeq (\alpha + \Lambda)I$ on $B(R)$, so y_k is unique although P need not be convex.

Lemma J.9 (One corrected step). *Let $f \in C^q$ with $[f]_\beta \leq L_\beta$, $\nabla^2 f \succeq \alpha I$ for some $\alpha \geq 0$, and a minimizer $x^* \in B(R)$. Let $G \in \mathcal{F}_{k_1}$ be an event on which $\|\nabla^2 g(x)\|_{\text{op}} \leq \Lambda$ for all $x \in B(R)$, where $\Lambda > 0$ is a number. Let $\Delta_k = \mathbb{E}[\|z_k - x^*\|^2 \mathbf{1}_G]$, $F_k = \mathbb{E}[(f(y_k) - f^*)\mathbf{1}_G]$ and $V_h = b_h^2 + \varsigma_h^2$. If $\mu_k \Lambda \leq 1/2$, then*

$$\Delta_{k+1} \leq \Delta_k - 2\mu_k F_k + 4\mu_k R b_h + 8\mu_k^2 V_h,$$

and if moreover $\alpha > 0$ and $\alpha\mu_k \leq 1$, then

$$\Delta_{k+1} \leq \left(1 - \frac{\alpha\mu_k}{4}\right) \Delta_k - 2\mu_k F_k + \frac{2\mu_k b_h^2}{\alpha} + 8\mu_k^2 V_h.$$

Proof. Omit the index k and write $u = x^*$, $w = z - y$, $c = \mu(\hat{v}(y) - \hat{v}(z))$, $e_z = \hat{v}(z) - \nabla g(z)$ and $e_y = \hat{v}(y) - \nabla g(y)$. Steps 1 and 2 hold on every realization in G .

Step 1: the first line of (J.7). The optimality condition at y over the convex set $B(R) \ni u$ gives $\langle w, y - u \rangle \geq \mu \langle \nabla P(y) + \hat{v}(z), y - u \rangle$. Add $\langle c, y - u \rangle$ to both sides. Since $\nabla P + \nabla g = \nabla f$ and $\nabla^2 f \succeq \alpha I$,

$$\langle w + c, y - u \rangle \geq \mu \langle \nabla f(y) + e_y, y - u \rangle \geq \mu \left(f(y) - f(u) + \frac{\alpha}{2} \|y - u\|^2 + \langle e_y, y - u \rangle \right);$$

the error e_z has canceled.

Step 2: the second line. Since $y - c - u = (z - u) - (w + c)$, we have $\|y - c - u\|^2 = \|z - u\|^2 - 2\langle w + c, y - u \rangle - \|w\|^2 + \|c\|^2$. The projection onto $B(R) \ni u$ is nonexpansive, so Step 1 gives

$$\|z_{k+1} - u\|^2 \leq \|z - u\|^2 - 2\mu(f(y) - f(u)) - \alpha\mu\|y - u\|^2 - 2\mu\langle e_y, y - u \rangle - \|w\|^2 + \|c\|^2.$$

On G , ∇g is Λ -Lipschitz on $B(R)$, and $\mu\Lambda \leq 1/2$ gives $\|c\|^2 \leq 2\mu^2\Lambda^2\|w\|^2 + 4\mu^2(\|e_y\|^2 + \|e_z\|^2) \leq \frac{1}{2}\|w\|^2 + 4\mu^2(\|e_y\|^2 + \|e_z\|^2)$.

Step 3: expectations. Multiply by $\mathbf{1}_G$ and take expectations. The centers z and y are bounded and measurable with respect to the past of their rounds, and G belongs to that past. By [\(J.1\)](#) and [\(J.3\)](#), $e_y = b(y) + \zeta_y$ with $\mathbb{E}[\langle \zeta_y, y - u \rangle \mathbf{1}_G] = 0$, and $\mathbb{E}[(\|e_y\|^2 + \|e_z\|^2)\mathbf{1}_G] \leq 2V_h$. For the first claim, bound $-2\mu\langle b(y), y - u \rangle \leq 4\mu R b_h$ and drop the remaining nonpositive terms. For the second, bound $-2\mu\langle b(y), y - u \rangle \leq \frac{\alpha\mu}{2}\|y - u\|^2 + \frac{2\mu b_h^2}{\alpha}$; this leaves the term $-\mathbb{E}[(\frac{\alpha\mu}{2}\|y - u\|^2 + \frac{1}{2}\|w\|^2)\mathbf{1}_G]$, and $\|z - u\|^2 \leq 2\|y - u\|^2 + 2\|w\|^2$ with $\alpha\mu \leq 1$ gives $\frac{\alpha\mu}{2}\|y - u\|^2 + \frac{1}{2}\|w\|^2 \geq \frac{\alpha\mu}{4}\|z - u\|^2$. $\square$

Both the restarted method below and the metric method of Section [J.5](#) lead to the recursion [\(F.2\)](#), which is solved by Lemma [F.1](#).

Proposition J.10 (Restarted corrected method). *Let $f \in \mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, \infty, R)$, let G and Λ be as in Lemma [J.9](#), and put $k_0 = \max\{7, 16\Lambda/\alpha\}$. Run $N_1 \geq 0$ iterations of [\(J.7\)](#) from $z_1 = 0$ with the step $8/(\alpha(k_0 + 1))$, and then $N_2 \geq 1$ iterations with the steps $8/(\alpha(k + k_0))$, $k = 1, \dots, N_2$. Output $\bar{y} = \mathcal{W}^{-1} \sum_{k=1}^{N_2} w_k y_{N_1+k}$, where $w_k = k + k_0 - 1$ and $\mathcal{W} = \sum_k w_k$. The method uses $2(N_1 + N_2)$ rounds of $2md$ oracle calls each, and if $N_2 \geq k_0$, then*

$$\mathbb{E}[(f(\bar{y}) - f^*)\mathbf{1}_G] \leq \frac{\alpha k_0^2 R^2}{8N_2^2} e^{-2N_1/(k_0+1)} + \frac{2b_h^2}{\alpha} + \frac{96(b_h^2 + \varsigma_h^2)}{\alpha N_2}.$$

Procedure 3 Pilot fit and corrected residual method ($\beta > 4$)

Require: budget T ; $d, \beta, \alpha, R, L_\beta, \sigma$; the stencil of Procedure 1

- 1: $q \leftarrow \lceil \beta \rceil - 1$, $\omega \leftarrow (q - 2)/(4q)$, $\theta \leftarrow T^\omega$
 - 2: choose an odd $r \geq 3$ with $\omega(r - 1) \geq p_\beta + 3$, and $\ell \leftarrow \max\{1, \lceil 16\sigma^2 \rceil\}$
 - 3: run the pilot of Section J.3 on $B(2R)$ (one round, $n_0 = 2K_q r \ell$ oracle calls)
 - 4: $P \leftarrow$ fit of the medians of block means
 - 5: $\Lambda \leftarrow \Lambda_\theta$, $\Upsilon \leftarrow \Upsilon_\theta$ from (J.4), $k_0 \leftarrow \max\{7, 16\Lambda/\alpha\}$
 - 6: $N_{\text{it}} \leftarrow \lfloor (T - n_0)/(4md) \rfloor$, $N_1 \leftarrow \lfloor N_{\text{it}}/2 \rfloor$, $N_2 \leftarrow N_{\text{it}} - N_1$
 - 7: $h \leftarrow (d^{1+\tau}\sigma^2/(L_\beta^2 N_2))^{1/(2\beta)}$
 - 8: run Lemma J.10 with the residual oracle $\hat{v}(x) = \bar{g}_h(x) - \nabla P(x)$; obtain $\bar{y}$
 - 9: **return** the output of the guard of Lemma J.8 with candidate $\bar{y}$ and level Υ
-

Proof. All steps are at most $8/(\alpha(k_0 + 1))$, hence $\mu_k \Lambda \leq 1/2$ and $\alpha\mu_k \leq 1$. Lemma J.9 gives (F.2) with $a = \alpha/4$, $c = 2$, $\mathcal{B} = 2b_h^2/\alpha$ and $\mathcal{V} = 8(b_h^2 + \varsigma_h^2)$, and the steps are those of Lemma F.1(b), with $a\mu = 2/(k_0 + 1) \leq \min\{1, 2/k_0\}$. Since $\Delta_1 \leq R^2$ and, by convexity, $f(\bar{y}) - f^* \leq \mathcal{W}^{-1} \sum_k w_k(f(y_{N_1+k}) - f^*)$, Lemma F.1(b) gives the claim. $\square$

J.4 Budget of order d^q for $\beta > 4$

Theorem J.11 (Polynomial-pilot upper bound, $\beta > 4$). *Let $\beta > 4$, $\alpha, R, L_\beta, \sigma > 0$, and let the noise be centered at level σ . There is C_0 , depending only on $\beta, \alpha, R, L_\beta, \sigma$, such that for every $d \geq 1$ and every $T \geq C_0 d^q$, Procedure 3 uses at most T oracle calls in $1 + 2N_{\text{it}}$ rounds and satisfies*

$$\mathbb{E}[f(\hat{x}) - f^*] \leq C_\beta \Phi_T \quad \text{for every } f \in \mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, \infty, R),$$

for the operator seminorm with $\tau = 0$ and for the Hilbert–Schmidt seminorm with $\tau = \tau_\beta$.

Proof. Write $p = p_\beta$. Here $q \geq 4$, so $\omega = (q - 2)/(4q) \in (0, 1/4)$ and $\theta = T^\omega \geq 1$. The integer r depends only on β , ℓ only on σ , and $\sigma^2/\ell \leq 1/16$. Let $G = \{E_0 \leq \theta\}$; it is determined by the pilot round.

Step 1: oracle calls. Each factor of $K_q = \prod_{i=1}^q (d + i)/i$ is at most $d + 1$, so $K_q \leq 2^q d^q$ and $n_0 \leq 2^{q+1} r \ell d^q$. Let $C_0 \geq \max\{2^{q+2} r \ell, 32m\}$. Then $n_0 \leq T/2$ and $T/(8md) \geq 4$, so $N_{\text{it}} \geq T/(16md) \geq 2$, $N_2 \geq T/(32md)$ and $N_1 \geq N_{\text{it}}/4$. The method uses at most $n_0 + 4mdN_{\text{it}} \leq T$ oracle calls.

Step 2: the radius. Put $X = L_\beta^2 d^{-\tau} h^{2\beta-2}$. The choice of h gives $X = d\sigma^2/(N_2 h^2) = L_\beta^{2/\beta} d^{-\tau/\beta} (d\sigma^2/N_2)^p$, and $N_2 \geq T/(32md)$ gives $X \leq (32m)^p \alpha \Phi_T$, because $2p - \tau/\beta = 2 - (2 + \tau)/\beta$. By (J.2), $b_h^2 = C_b^2 X$ and $\varsigma_h^2 = (C_f + C_\sigma N_2)X$. Hence

$$\frac{2b_h^2}{\alpha} + \frac{96(b_h^2 + \varsigma_h^2)}{\alpha N_2} \leq C_1 (32m)^p \Phi_T, \quad C_1 = 98C_b^2 + 96C_f + 96C_\sigma. \quad (\text{J.8})$$

Step 3: the event G . By Lemma J.7, on G the residual $g = f - P$ satisfies $\|\nabla^2 g\|_{\text{op}} \leq \Lambda_\theta$ and $|g| \leq \Upsilon_\theta$ on $B(R)$. Since $\Lambda_\theta \leq C\sqrt{K_q}\theta$, we have $k_0 + 1 \leq Cd^{q/2}T^\omega$, and $T \geq C_0 d^q$ gives $d \leq T^{1/q}$. Hence

$$\frac{N_{\text{it}}}{k_0 + 1} \geq c \frac{T^{1-\omega}}{d^{1+q/2}} \geq c T^{\frac{1}{2} - \frac{1}{q} - \omega} = c T^\omega,$$

because $\frac{1}{2} - \frac{1}{q} = 2\omega$. After increasing C_0 , $N_2 \geq k_0$ and $2N_1/(k_0 + 1) \geq cT^\omega$. On G the guard does not increase the value (Lemma J.8), so Lemma J.10 with $\Lambda = \Lambda_\theta$ and (J.8) give

$$\mathbb{E}[(f(\hat{x}) - f^*)\mathbf{1}_G] \leq \frac{\alpha R^2}{8} e^{-cT^\omega} + C_1 (32m)^p \Phi_T.$$

Step 4: the complement. On G^c we have $E_0 > \theta \geq 1$, so Lemma J.8 and (J.5) give $f(\hat{x}) - f^* \leq 2\Upsilon_\theta + 2\varrho_0 + 2\Gamma_0(\varrho_0 + E_0) \leq C\sqrt{K_q}E_0$. By (J.6) with $\sigma^2/\ell \leq 1/16$, $\mathbb{E}[E_0\mathbf{1}_{G^c}] \leq J2^{-r}\theta^{-r} \leq 2K_q\theta^{-r}$. Since $K_q \leq 2^q d^q \leq 2^q T$ and $\omega r \geq p + 3$,

$$\mathbb{E}[(f(\hat{x}) - f^*)\mathbf{1}_{G^c}] \leq CK_q^{3/2}\theta^{-r} \leq CT^{3/2-\omega r} \leq CT^{-p-1}.$$

Step 5: conclusion. The exponent $2 - (2 + \tau)/\beta$ of d in Φ_T is positive, because $\tau \leq (\beta - 1)/2$; hence $\Phi_T \geq \alpha^{-1} L_\beta^{2/\beta} \sigma^{2p} T^{-p}$. After increasing C_0 , the remainders of Steps 3 and 4 are at most Φ_T in total, and $\mathbb{E}[f(\hat{x}) - f^*] \leq (C_1(32m)^p + 1)\Phi_T$, where the constant depends only on β . The pilot uses only $[f]_{\beta, \text{op}} \leq L_\beta$, which holds in both cases, and τ enters only through (J.2). $\square$

J.5 Budget of order d^2 for $2 < \beta \leq 4$

For $q \in \{2, 3\}$ the fit of Section J.4 needs of order d^q values, and its curvature bound Λ_θ grows like $\sqrt{K_q}$, which is too large when T is of order d^2 . We replace it by an estimate of $H_0 = \nabla^2 f(0)$ whose error in operator norm grows more slowly than d . For $2 < \beta \leq 3$ the quadratic model built from this estimate is accurate on $B(R)$, and the method of Section J.4 applies. For $3 < \beta \leq 4$ the cubic part of f is not estimated. Instead, convexity bounds the variation of $\nabla^2 f$ relative to H_0 (Lemma J.14), and projected steps in a metric calibrated by the estimate of H_0 replace the corrected method.

Lemma J.12 (Robust estimate of a symmetric matrix). *Let $M_\star$ be a deterministic symmetric $d \times d$ matrix, $\sigma_0 > 0$, and $v \geq 2$ an integer. Suppose that observations $Y_1, Y_2, \dots$ are made in an order fixed in advance, each referring to a pair (i, j) with $i \leq j$, and that $Y_n = (M_\star)_{i_n j_n} + \xi_n$, where (ξ_n) is adapted to a filtration $(\mathcal{H}_n)$ with $\mathbb{E}[\xi_n | \mathcal{H}_{n-1}] = 0$ and $\mathbb{E}[\xi_n^2 | \mathcal{H}_{n-1}] \leq \sigma_0^2$. From $2v$ observations of every pair one can compute a symmetric matrix $\hat{M}$ with*

$$(\mathbb{E} \left\| \hat{M} - M_\star \right\|_{\text{op}}^v)^{1/v} \leq C_v \sigma_0 \sqrt{d \log(2d)},$$

where C_v depends only on v .

Proof. Let $\ell_d = \log(2d)$ and $b = \sigma_0 \sqrt{d/\ell_d}$. For every pair, let c_{ij} be the median of the first $2v - 1$ observations, let Y'_{ij} be the last one, and put

$$\hat{M}_{ij} = \hat{M}_{ji} = c_{ij} + \text{clip}_{[-b, b]}(Y'_{ij} - c_{ij}).$$

We write $\hat{M} - M_\star$ as a bias matrix plus a matrix martingale.

Step 1: one entry. By Lemma J.3(a) with $r = 2v - 1$, $\omega_{ij} = c_{ij} - (M_\star)_{ij}$ satisfies $\mathbb{E}\omega_{ij}^2 \leq \binom{2v-1}{v} \sigma_0^{2v}$. Let $\mathcal{H}'_{ij}$ be the σ -algebra before Y'_{ij} and $\tilde{\xi}$ the error of Y'_{ij} . Then $\hat{M}_{ij} - (M_\star)_{ij} = \omega_{ij} + \text{clip}_{[-b, b]}(\tilde{\xi} - \omega_{ij}) = \varpi_{ij} + \chi_{ij}$, where ϖ_{ij} is the conditional mean given $\mathcal{H}'_{ij}$. Clipping is 1-Lipschitz, and a 1-Lipschitz function of a real random variable has at most its variance; applied to the conditional law of $\tilde{\xi}$, this gives $\mathbb{E}[\chi_{ij}^2 | \mathcal{H}'_{ij}] \leq \sigma_0^2$. Moreover $|\chi_{ij}| \leq 2b$. Since $\mathbb{E}[\tilde{\xi} | \mathcal{H}'_{ij}] = 0$,

$$\varpi_{ij} = \mathbb{E}[\text{clip}_{[-b, b]}(\tilde{\xi} - \omega_{ij}) - (\tilde{\xi} - \omega_{ij}) | \mathcal{H}'_{ij}].$$

If $|\omega_{ij}| \leq b/2$, then $|\text{clip}_{[-b, b]}(x) - x| = (|x| - b)_+$ and $(|\tilde{\xi} - \omega_{ij}| - b)_+ \leq (|\tilde{\xi}| - b/2)_+ \leq 2\tilde{\xi}^2/b$, so $|\varpi_{ij}| \leq 2\sigma_0^2/b$. Otherwise $|\varpi_{ij}| \leq |\omega_{ij}| + b \leq 3|\omega_{ij}|$.

Step 2: the bias matrix. The entries ϖ_{ij} with $|\omega_{ij}| \leq b/2$ form a matrix of operator norm at most $2d\sigma_0^2/b$. The others form a matrix of operator norm at most $F = (\sum_{i,j} 9\omega_{ij}^2 \mathbf{1}\{|\omega_{ij}| > b/2\})^{1/2}$, the sum running over ordered pairs. Since $|\omega|^v \mathbf{1}\{|\omega| > b/2\} \leq (2/b)^v |\omega|^{2v}$, Minkowski's inequality in $L^{v/2}$ gives $\|F\|_{L^v}^2 \leq 36 \binom{2v-1}{v}^{2/v} d^2 \sigma_0^4 / b^2$. Hence the bias matrix ϖ satisfies $\|\varpi\|_{\text{op}} \leq C_v d \sigma_0^2 / b$.

Step 3: the martingale part. Let $E_{ii} = e_i e_i^\top$ and $E_{ij} = e_i e_j^\top + e_j e_i^\top$ for $i < j$, and $\mathbf{Y} = \sum_{i \leq j} \chi_{ij} E_{ij}$, summed in the order of the observations Y'_{ij} . This is a martingale of symmetric matrices whose increments have operator norm at most $2b$, and its predictable quadratic variation is at most $\sigma_0^2 \sum_{i \leq j} E_{ij}^2 = d\sigma_0^2 I$. The matrix Freedman inequality (Tropp, 2011, Theorem 1.2), applied to $\mathbf{Y}$ and to $-\mathbf{Y}$, gives

$$\mathbb{P}(\|\mathbf{Y}\|_{\text{op}} \geq u) \leq 2d \exp\left(-\frac{u^2}{2d\sigma_0^2 + 4bu/3}\right) \leq 2de^{-u^2/(4d\sigma_0^2)} + 2de^{-3u/(8b)}.$$

With $u_0 = 2\sigma_0 \sqrt{d\ell_d} + 3b\ell_d$, the right-hand side at $u = u_0 + s$, $s \geq 0$, is at most $e^{-s^2/(4d\sigma_0^2)} + e^{-3s/(8b)}$, and integrating the tail gives $\|\mathbf{Y}\|_{\text{op}} \leq C_v(u_0 + \sigma_0 \sqrt{d} + b)$.

Step 4: conclusion. We have $\hat{M} - M_\star = \varpi + \mathbf{Y}$, $d\sigma_0^2/b = b\ell_d = \sigma_0 \sqrt{d\ell_d}$ and $\sigma_0 \sqrt{d} + b \leq C\sigma_0 \sqrt{d\ell_d}$. $\square$

Lemma J.13 (Hessian pilot). *Let $2 < \beta \leq 4$, $f \in C^q$ with $[f]_{\beta, \text{op}} \leq L_\beta$, $H_0 = \nabla^2 f(0)$, let the noise be centered at level $\sigma > 0$, and let $v \geq 2$ be an integer. For $\rho > 0$ and $i < j$ consider the difference quotients*

$$\frac{Y(\rho e_i) - 2Y(0) + Y(-\rho e_i)}{\rho^2}, \quad \frac{Y(\rho(e_i + e_j)) - Y(\rho(e_i - e_j)) - Y(\rho(e_j - e_i)) + Y(-\rho(e_i + e_j))}{4\rho^2},$$

where $Y(x)$ is a fresh value at x each time, and let $\hat{H}$ be the matrix of Lemma J.12 computed from them with $\rho = (\sigma/L_\beta)^{1/\beta}(\log(2d)/d)^{1/(2\beta)}$. This uses at most $8vd^2$ oracle calls in one round, and

$$(\mathbb{E} \|\hat{H} - H_0\|_{\text{op}}^v)^{1/v} \leq C_v d^{a'}, \quad a' = \frac{3}{4} + \frac{1}{2\beta} < 1,$$

where C_v depends only on $v, \beta, L_\beta, \sigma$.

Proof. The query points do not depend on any observation, so they form one round with an order fixed in advance; there are $2vd(d+1)/2$ quotients of at most 4 oracle calls each. The error of a quotient is a combination of at most four errors of consecutive queries whose squared coefficients sum to at most $6/\rho^4$. Hence Lemma J.3(c) gives the assumptions of Lemma J.12 with $\sigma_0^2 = 6\sigma^2/\rho^4$.

The means of the quotients are

$$(M_\star)_{ii} = \int_{-1}^1 (1 - |s|) \partial_{ii} f(s\rho e_i) ds, \quad (M_\star)_{ij} = \frac{1}{4} \int_{-1}^1 \int_{-1}^1 \partial_{ij} f(\rho(se_i + te_j)) ds dt.$$

For $q = 2$, $\|\nabla^2 f(y) - H_0\|_{\text{op}} \leq L_\beta \|y\|^{\beta-2}$. For $q = 3$, Lemma B.2(c) gives $\|\nabla^2 f(y) - H_0 - D^3 f(0)[y]\|_{\text{op}} \leq L_\beta \|y\|^{\beta-2}$, and the linear term averages to zero, because both weightings are symmetric under $y \mapsto -y$. The points have norm at most $\sqrt{2}\rho$ and $\beta - 2 \leq 2$, so every entry of $M_\star - H_0$ is at most $2L_\beta \rho^{\beta-2}$ in absolute value and $\|M_\star - H_0\|_{\text{op}} \leq 2dL_\beta \rho^{\beta-2}$. With Lemma J.12,

$$(\mathbb{E} \|\hat{H} - H_0\|_{\text{op}}^v)^{1/v} \leq C_v \frac{\sqrt{6}\sigma}{\rho^2} \sqrt{d \log(2d)} + 2dL_\beta \rho^{\beta-2}.$$

For the given ρ both terms are constants times $L_\beta^{2/\beta} \sigma^{1-2/\beta} d^a (\log(2d))^{1-a}$ with $a = \frac{1}{2} + \frac{1}{\beta} < 1$, and $\log(2d) \leq 2\sqrt{d}$ gives $d^a (\log(2d))^{1-a} \leq 2d^{(1+a)/2} = 2d^{a'}$. $\square$

The next lemma, used for $3 < \beta \leq 4$, is a Hölder version of an inequality of Glaeser (1963, Lemme I); a related bound is Nesterov (2021, Lemma 3).

Lemma J.14 (Relative curvature). *Let $3 < \beta \leq 4$, so that $q = 3$, and let $f \in C^3$ with $\nabla^2 f \succeq \alpha I$ and $[f]_{\beta, \text{op}} \leq L_\beta$. Put $\varkappa = \beta - 2 \in (1, 2]$, $v = 1 - 1/\varkappa \in (0, 1/2]$, $H_0 = \nabla^2 f(0)$ and, for a unit vector w , $g_w(x) = w^\top \nabla^2 f(x)w$.*

(a) $\|\nabla g_w(x)\| \leq 2L_\beta^{1/\varkappa} g_w(x)^v$ for all x .

(b) For $\varrho > 0$ and $\|x\| \leq \varrho$, $C_\varrho^{-1} H_0 \preceq \nabla^2 f(x) \preceq C_\varrho H_0$ with $C_\varrho = [1 + \frac{2\varrho}{\varkappa} (L_\beta/\alpha)^{1/\varkappa}]^\varkappa$.

(c) For $\|x\| \leq R$, $\frac{1}{2}H_0 - \bar{c}I \preceq \nabla^2 f(x) \preceq \frac{3}{2}H_0 + \bar{c}I$ with $\bar{c} = 3L_\beta R^\varkappa$.

Proof. (a) The gradient $\nabla g_w(x) = D^3 f(x)[w, w, \cdot]$ is $(\varkappa - 1)$ -Hölder with constant L_β , so for a unit vector u and $s \geq 0$,

$$g_w(x + su) \leq g_w(x) + s \langle \nabla g_w(x), u \rangle + \frac{L_\beta}{\varkappa} s^\varkappa.$$

If $\nabla g_w(x) \neq 0$, take $u = -\nabla g_w(x)/\|\nabla g_w(x)\|$ and $s = (g_w(x)/L_\beta)^{1/\varkappa}$. Since $g_w > 0$, this gives $s \|\nabla g_w(x)\| \leq (1 + 1/\varkappa)g_w(x) \leq 2g_w(x)$, which is (a).

(b) By (a), $\|\nabla(g_w^{1/\varkappa})\| = \varkappa^{-1} g_w^{-v} \|\nabla g_w\| \leq \frac{2}{\varkappa} L_\beta^{1/\varkappa}$, so $|g_w(x)^{1/\varkappa} - g_w(0)^{1/\varkappa}| \leq \frac{2\varrho}{\varkappa} L_\beta^{1/\varkappa}$ for $\|x\| \leq \varrho$. Since $g_w \geq \alpha$, this gives $g_w(x) \leq C_\varrho g_w(0)$ and $g_w(0) \leq C_\varrho g_w(x)$ for every unit w .

(c) Let $\|x\| \leq R$ and $h_w = g_w(0)$. By the Hölder bound and (a),

$$|g_w(x) - h_w| \leq |\langle \nabla g_w(0), x \rangle| + \frac{L_\beta}{\varkappa} R^\varkappa \leq 2L_\beta^{1/\varkappa} R h_w^v + L_\beta R^\varkappa.$$

With $z = h_w/(L_\beta R^\varkappa)$ we have $2L_\beta^{1/\varkappa} R h_w^v = L_\beta R^\varkappa \cdot 2z^v$, and $2z^v \leq z/2 + 2$ for $z \geq 0$, because $z^v \leq \max\{1, \sqrt{z}\}$ and $2\sqrt{z} \leq z/2 + 2$. Hence $|g_w(x) - h_w| \leq h_w/2 + 3L_\beta R^\varkappa$ for every unit w , which is (c). $\square$

Lemma J.15 (Spectral initialization). *Let f be as in Lemma J.14 with minimizer $x^* \in B(R)$. Let $A \succeq \alpha I$ be a symmetric matrix with eigenvalues $a_1 \geq \dots \geq a_d$ and orthonormal eigenvectors $o_1, \dots, o_d$, let $E_A = \|A - H_0\|_{\text{op}}$, write $x_i = \langle x, o_i \rangle$, and let $\mathcal{C}_A = \{x : |x_i| \leq R \ \forall i\}$, which contains x^* . Put $R_d = R\sqrt{d}$ and $\Xi_d = 1 + 2\sqrt{2}L_\beta^{1/\varkappa} R_d + (L_\beta/\varkappa)R_d^\varkappa$. Starting from $x_0 = 0$, make d updates*

$$x_{t+1,i} = \text{clip}_{[-R,R]}(x_{t,i} - a_i^{-1} \hat{g}_i(x_t)), \quad i = 1, \dots, d,$$

where $\hat{g}(x_t) = \nabla f(x_t) + \zeta_t$ is any gradient estimate and $\hat{g}_i = \langle \hat{g}, o_i \rangle$. Let u be the projection of x_d onto $B(R)$ in the norm $\|y\|_A = (y^\top A y)^{1/2}$. Then, on every realization,

$$\|u - x^*\|_A^2 \leq 80R^2 d^3 \Xi_d^2 + 12R^2 d + 12R^2 d E_A^2 + 4d \sum_{t < d} \|\zeta_t\|^2.$$

Proof. We bound the entries of $\nabla^2 f - A$ on $\mathcal{C}_A$ in the rows of the large eigenvalues, show by induction that these coordinates contract, and add the coordinates with small a_i at the end.

Step 1: entries of the Hessian. For $i \neq j$, polarization with $w_\pm = (o_i \pm o_j)/\sqrt{2}$ gives

$$D^3 f(0)[o_i, o_j, \cdot] = \frac{1}{2}(\nabla g_{w_+}(0) - \nabla g_{w_-}(0)), \quad g_{w_\pm}(0) \leq w_\pm^\top A w_\pm + E_A = \frac{a_i + a_j}{2} + E_A.$$

With Lemma J.14(a) this gives, also for $i = j$, $\|D^3 f(0)[o_i, o_j, \cdot]\| \leq 2L_\beta^{1/\varkappa}(\max\{a_i, a_j\} + E_A)^v$. By the Hölder continuity of $D^3 f$, $\|\nabla^2 f(y) - H_0 - D^3 f(0)[y]\|_{\text{op}} \leq L_\beta \|y\|^\varkappa/\varkappa$, and every $y \in \mathcal{C}_A$ has $\|y\| \leq R_d$. Hence, in the basis (o_i) ,

$$|(\nabla^2 f(y) - A)_{ij}| \leq E_A + 2L_\beta^{1/\varkappa} R_d(\max\{a_i, a_j\} + E_A)^v + \frac{L_\beta}{\varkappa} R_d^\varkappa \quad (y \in \mathcal{C}_A).$$

Let $J_A = (1 + E_A + 2d\Xi_d)^2$ and let k_A be the number of i with $a_i > J_A$. For $i \leq k_A$ and any j , $a := \max\{a_i, a_j\} > J_A \geq 1$ gives $E_A \leq \sqrt{a}$, $a + E_A \leq 2a$ and, since $v \leq \frac{1}{2}$, $(a + E_A)^v \leq \sqrt{2a}$. Hence

$$|(\nabla^2 f(y) - A)_{ij}| \leq \Xi_d \sqrt{\max\{a_i, a_j\}} \quad (i \leq k_A, y \in \mathcal{C}_A). \quad (\text{J.9})$$

Step 2: one update. Let $e_t = x_t - x^*$. Both x_t and x^* lie in the convex set $\mathcal{C}_A$, so $\nabla f(x_t) = A e_t + \mathcal{M}_t e_t$ with $\mathcal{M}_t = \int_0^1 (\nabla^2 f(x^* + s e_t) - A) ds$, whose entries obey (J.9) in the rows $i \leq k_A$. Clipping fixes x_i^* and is 1-Lipschitz, so $|e_{t+1,i}| \leq a_i^{-1} |(\mathcal{M}_t e_t)_i + \zeta_{t,i}|$. Hence, with $\bar{z}_i^t = \sqrt{a_i} |e_{t,i}|$, for $i \leq k_A$

$$\bar{z}_i^{t+1} \leq \Xi_d \sum_{j \leq k_A} \frac{\bar{z}_j^t}{\sqrt{\min\{a_i, a_j\}}} + \Xi_d \sum_{j > k_A} |e_{t,j}| + \|\zeta_t\|,$$

where for $j > k_A$ we used $a_j \leq a_i$, and $a_i > 1$.

Step 3: induction. Let $\bar{B}_0 = 0$ and $\bar{B}_{t+1} = \bar{B}_t/2 + 2Rd\Xi_d + \|\zeta_t\|$. We show by induction that $\bar{z}_i^t \leq \bar{B}_t$ for $i \leq \min\{t, k_A\}$. Let $i \leq \min\{t+1, k_A\}$. In the bound of Step 2, the indices $j \leq \min\{t, k_A\}$ satisfy $\bar{z}_j^t \leq \bar{B}_t$ and $\Xi_d/\sqrt{\min\{a_i, a_j\}} \leq \Xi_d/\sqrt{J_A} \leq 1/(2d)$, so they contribute at most $\bar{B}_t/2$. The other indices satisfy $j \geq i$, hence $a_j \leq a_i$. Their terms are at most $\Xi_d |e_{t,j}| \leq 2R\Xi_d$ (for $j \leq k_A$ use $\bar{z}_j^t/\sqrt{a_j} = |e_{t,j}|$), and there are at most d of them.

Step 4: conclusion. After d updates, $\bar{B}_d \leq 4Rd\Xi_d + \sum_{t < d} 2^{-(d-1-t)} \|\zeta_t\|$. By the Cauchy-Schwarz inequality, the indices $i \leq k_A$ contribute at most $d\bar{B}_d^2 \leq 32R^2 d^3 \Xi_d^2 + 4d \sum_t \|\zeta_t\|^2$ to $\|x_d - x^*\|_A^2$. The other indices have $a_i \leq J_A \leq 3(1 + E_A^2 + 4d^2 \Xi_d^2)$ and $|e_{d,i}| \leq 2R$, so they contribute at most $12R^2 d + 12R^2 d E_A^2 + 48R^2 d^3 \Xi_d^2$. Finally, the projection onto $B(R) \ni x^*$ in the norm $\|\cdot\|_A$ does not increase the distance to x^* . $\square$

Lemma J.16 (Projected steps in a calibrated metric). *Let $f \in \mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, \infty, R)$ and let k_2 be a round. Let a nonempty compact convex set $\mathcal{T} \subseteq B(R)$, a symmetric matrix $\mathbf{S}$ and a point $x_1 \in \mathcal{T}$ be $\mathcal{F}_{k_2}$ -measurable, with $\mathbf{S} \succeq \Lambda \mathbf{I}$ for a number $\Lambda > 0$. Let $\mu_{\mathbf{S}} \in (0, 1]$ be a number and $G \in \mathcal{F}_{k_2}$ an event on which $x^* \in \mathcal{T}$ and $\mu_{\mathbf{S}} \mathbf{S} \preceq \nabla^2 f(x) \preceq \frac{3}{2} \mathbf{S}$ for all $x \in B(R)$. Fix a radius $h > 0$, let $k_0 = \max\{7, 24/\mu_{\mathbf{S}}\}$, and after round k_2 run, with one round of Procedure 1 (one frame, radius h) per iteration,*

$$x_{k+1} = \text{Proj}_{\mathcal{T}}^{\mathbf{S}}(x_k - \mu_k \mathbf{S}^{-1} \bar{g}_h(x_k)), \quad k = 1, \dots, N_1 + N_2,$$

where $\text{Proj}_{\mathcal{T}}^{\mathbf{S}}$ is the projection onto $\mathcal{T}$ in the norm $\|y\|_{\mathbf{S}} = (y^\top \mathbf{S} y)^{1/2}$, $\mu_k = 8/(\mu_{\mathbf{S}}(k_0 + 1))$ for $k \leq N_1$ and $\mu_{N_1+k} = 8/(\mu_{\mathbf{S}}(k + k_0))$ for $1 \leq k \leq N_2$. Output $\hat{x} = \mathcal{W}^{-1} \sum_{k=1}^{N_2} w_k x_{N_1+k}$, where $w_k = k + k_0 - 1$ and $\mathcal{W} = \sum_k w_k$. Then $\hat{x} \in \mathcal{T}$, and if $N_2 \geq k_0$,

$$\mathbb{E}[(f(\hat{x}) - f^*) \mathbf{1}_G] \leq \frac{\mu_{\mathbf{S}} k_0^2}{4N_2^2} e^{-2N_1/(k_0+1)} \mathbb{E}[\|x_1 - x^*\|_{\mathbf{S}}^2 \mathbf{1}_G] + \frac{8b_h^2}{\mu_{\mathbf{S}} \Lambda} + \frac{48(b_h^2 + \varsigma_h^2)}{\mu_{\mathbf{S}} \Lambda N_2}.$$

Proof. The centers lie in $\mathcal{T}$ and are measurable with respect to the past of their rounds, so $\hat{x} \in \mathcal{T}$. Let $\Delta_k = \mathbb{E}[\|x_k - x^*\|_{\mathbf{S}}^2 \mathbf{1}_G]$, $F_k = \mathbb{E}[(f(x_k) - f^*) \mathbf{1}_G]$, $V_h = b_h^2 + \varsigma_h^2$, and $\|y\|_{\mathbf{S}^{-1}} = (y^\top \mathbf{S}^{-1} y)^{1/2}$.

Step 1: two inequalities on G . As in the proof of Lemma F.2, on G and for $x \in B(R)$ we have $\nabla f(x) = \bar{H}(x - x^*)$ with $\bar{H} = \int_0^1 \nabla^2 f(x^* + s(x - x^*)) ds$, now with $\mu_{\mathbf{S}} \mathbf{S} \preceq \bar{H} \preceq \frac{3}{2} \mathbf{S}$. Hence $\bar{H}^{1/2} \mathbf{S}^{-1} \bar{H}^{1/2} \preceq \frac{3}{2} \mathbf{I}$, and

$$\|\nabla f(x)\|_{\mathbf{S}^{-1}}^2 \leq \frac{3}{2} \langle \nabla f(x), x - x^* \rangle, \quad \langle \nabla f(x), x - x^* \rangle \geq f(x) - f^* + \frac{\mu_{\mathbf{S}}}{2} \|x - x^*\|_{\mathbf{S}}^2.$$

Step 2: one step. Write $\bar{g}_h(x_k) = \nabla f(x_k) + b_k + \zeta_k$ as in (J.1). On G the projection is nonexpansive in $\|\cdot\|_{\mathbf{S}}$ and $x^* \in \mathcal{T}$, so

$$\|x_{k+1} - x^*\|_{\mathbf{S}}^2 \leq \|x_k - x^*\|_{\mathbf{S}}^2 - 2\mu_k \langle \bar{g}_h(x_k), x_k - x^* \rangle + \mu_k^2 \|\bar{g}_h(x_k)\|_{\mathbf{S}^{-1}}^2.$$

Multiply by $\mathbf{1}_G$ and take expectations. We use Step 1, (J.3), $\|y\|_{\mathbf{S}^{-1}}^2 \leq \|y\|^2 / \Lambda$, $\|\bar{g}_h\|_{\mathbf{S}^{-1}}^2 \leq 2\|\nabla f\|_{\mathbf{S}^{-1}}^2 + 2\|b_k + \zeta_k\|_{\mathbf{S}^{-1}}^2$ and

$$2\mu_k \|b_k\|_{\mathbf{S}^{-1}} \|x_k - x^*\|_{\mathbf{S}} \leq \frac{\mu_k \mu_{\mathbf{S}}}{4} \|x_k - x^*\|_{\mathbf{S}}^2 + \frac{4\mu_k}{\mu_{\mathbf{S}}} \|b_k\|_{\mathbf{S}^{-1}}^2.$$

All steps satisfy $\mu_k \leq 8/(\mu_{\mathbf{S}}(k_0 + 1)) \leq 1/3$, so the terms with $\langle \nabla f(x_k), x_k - x^* \rangle \geq 0$ have total coefficient $-2\mu_k + 3\mu_k^2 \leq -\mu_k$. This gives (F.2) with $a = \mu_{\mathbf{S}}/4$, $c = 1$, $\mathcal{B} = 4b_h^2/(\mu_{\mathbf{S}} \Lambda)$ and $\mathcal{V} = 2V_h/\Lambda$, and the steps are those of Lemma F.1(b), with $a\mu = 2/(k_0 + 1) \leq \min\{1, 2/k_0\}$. That lemma and the convexity of f on $\mathcal{T}$ prove the claim. $\square$

The method for $2 < \beta \leq 4$. Procedure 4 collects the two constructions. Let $a' = \frac{3}{4} + \frac{1}{2\beta}$ be as in Lemma J.13, $\omega = (1 - a')/4$ and $\theta = T^\omega$. With fresh values for every quotient, Lemma J.13 with moment parameter v uses $n_H(v) = 2v(3d + 4\binom{d}{2}) = 2vd(2d + 1)$ oracle calls. The numbers $S_d \leq Cd^{a'}$ and, for $\beta > 3$, $F_d \leq Cd^{\beta+1}$ are the explicit moment bounds of Step 1 of the corresponding part of the proof of Lemma J.17; they depend only on d and the known parameters.

Theorem J.17 (Quadratic-pilot upper bound, $2 < \beta \leq 4$). *Let $2 < \beta \leq 4$, $\alpha, R, L_\beta, \sigma > 0$, and let the noise be centered at level σ . There is C_0 , depending only on $\beta, \alpha, R, L_\beta, \sigma$, such that for every $d \geq 1$ and every $T \geq C_0 d^2$, Procedure 4 uses at most T oracle calls, in $1 + 2N_{\text{it}}$ rounds for $\beta \leq 3$ and in $1 + d + N_{\text{it}}$ rounds for $\beta > 3$, and satisfies*

$$\mathbb{E}[f(\hat{x}) - f^*] \leq C_\beta \Phi_T \quad \text{for every } f \in \mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, \infty, R),$$

for the operator seminorm with $\tau = 0$ and for the Hilbert–Schmidt seminorm with $\tau = \tau_\beta$.

The algorithm does not use the events G of the proofs. For $\beta > 3$ every output lies in the trust set $\mathcal{T}$, and (J.11) below takes the place of the polynomial-value guard.

Proof for $2 < \beta \leq 3$. Here $q = 2$, $m = 1$, $p = p_\beta$, and v is the integer of Procedure 4, which depends only on β . We follow the proof of Lemma J.11.

Step 1: the pilot. It uses $n_0 \leq C_v d^2$ oracle calls in one round. By Lemma B.2(a), $|f(x) - T_2 f(0; x)| \leq L_\beta \|x\|^\beta / 2$, so the mean of $(Y(e_i) - Y(-e_i))/2$ is within $L_\beta/2$ of $\partial_i f(0)$, and its error has conditional variance at most $\sigma^2/2$.

Procedure 4 Quadratic startup without a Hessian bound, $2 < \beta \leq 4$

Require: $d, \beta, \alpha, L_\beta, R, \sigma, T$, with the onset of Lemma J.17; the stencil of Procedure 1

- 1: **if** $2 < \beta \leq 3$ **then**
 - 2: Choose an integer $v \geq 2$ with $\omega(v-1) \geq a'/2 + p_\beta + 2$
 - 3: In one round form $\hat{H}$ by Lemma J.13 with parameter v , the median $\hat{c}$ of $2v-1$ values at zero, and, for each i , the median $\hat{g}_i$ of $2v-1$ fresh differences $(Y(e_i) - Y(-e_i))/2$
 - 4: $P(x) \leftarrow \hat{c} + \langle \hat{g}, x \rangle + x^\top \hat{H}x/2$, $n_0 \leftarrow n_H(v) + (2v-1)(2d+1)$, $\Lambda \leftarrow S_d\theta$, $\Upsilon \leftarrow S_d\theta$
 - 5: $N_{\text{it}} \leftarrow \lfloor (T - n_0)/(4md) \rfloor$, $N_1 \leftarrow \lfloor N_{\text{it}}/2 \rfloor$, $N_2 \leftarrow N_{\text{it}} - N_1$, $h \leftarrow (d^{1+\tau}\sigma^2/(L_\beta^2 N_2))^{1/(2\beta)}$
 - 6: Run Lemma J.10 from zero with this P, Λ, h, N_1, N_2 ; **return** its averaged output after the guard of Lemma J.8 at level Υ
 - 7: **else**
 - 8: Choose an integer $v \geq 2$ with $\omega(v-1) \geq (\beta+1)/2 + p_\beta + 2$
 - 9: In one round form $\hat{H}$ by Lemma J.13 with parameter $2v$; raise every eigenvalue below α to α , obtaining A with ordered eigenbasis (o_i)
 - 10: Run the d updates of Lemma J.15 from zero. In update t , estimate coordinate i by the median of $2v-1$ fresh directional stencils (C.4) at x_t , with direction o_i and radius 1; query all coordinates in one round
 - 11: $u \leftarrow \text{Proj}_{B(R)}^A x_d$, $n_0 \leftarrow n_H(2v) + 2m(2v-1)d^2$, $\Lambda_0 \leftarrow S_d\theta$, $K_{\text{tr}} \leftarrow F_d\theta$
 - 12: $\mathcal{T} \leftarrow \{x \in B(R) : \|x - u\|_A^2 \leq K_{\text{tr}}\}$, $\Lambda \leftarrow \alpha + 3L_\beta R^{\beta-2} + \Lambda_0$, $\mathbf{S} \leftarrow A + \Lambda I$, $\mu_{\mathbf{S}} \leftarrow \alpha/(4\Lambda)$
 - 13: $N_{\text{it}} \leftarrow \lfloor (T - n_0)/(2md) \rfloor$, $N_1 \leftarrow \lfloor N_{\text{it}}/2 \rfloor$, $N_2 \leftarrow N_{\text{it}} - N_1$, $h \leftarrow (d^{1+\tau}\sigma^2/(L_\beta^2 N_2))^{1/(2\beta)}$
 - 14: **return** the weighted average of Lemma J.16, run from $x_1 = u$ on $\mathcal{T}$ with $\mathbf{S}, \mu_{\mathbf{S}}, h, N_1, N_2$
 - 15: **end if**
-

By Lemma J.3 and Minkowski's inequality, $\|\hat{c} - f(0)\|_{L^{2v}} \leq C\sigma$ and $\|\|\hat{g} - \nabla f(0)\|\|_{L^v} \leq \sqrt{d}(L_\beta/2 + C\sigma)$. On $B(R)$,

$$|f - P| \leq \frac{L_\beta R^\beta}{2} + |\hat{c} - f(0)| + R \|\hat{g} - \nabla f(0)\| + \frac{R^2}{2} \|\hat{H} - H_0\|_{\text{op}}, \quad \|\nabla^2(f - P)\|_{\text{op}} \leq L_\beta R^{\beta-2} + \|\hat{H} - H_0\|_{\text{op}}.$$

With Lemma J.13 and $\sqrt{d} \leq d^{a'}$, Minkowski's inequality gives an explicit number $S_d \leq Cd^{a'}$ such that

$$Z = S_d^{-1} \max \left\{ 1, \sup_{B(R)} |f - P|, \sup_{B(R)} \|\nabla^2(f - P)\|_{\text{op}} \right\}$$

satisfies $\mathbb{E}Z^v \leq 1$.

Step 2: the event $G = \{Z \leq \theta\}$. It is determined by the pilot round, and on G the residual satisfies $\|\nabla^2(f - P)\|_{\text{op}} \leq S_d\theta = \Lambda$ and $|f - P| \leq S_d\theta = \Upsilon$ on $B(R)$. As in Steps 1 and 2 of the proof of Lemma J.11, after increasing C_0 the oracle calls are at most T , $N_{\text{it}} \geq T/(16md)$, $N_2 \geq T/(32md)$, and (J.8) holds. Now $k_0 + 1 \leq Cd^{a'}T^\omega$ and $d \leq \sqrt{T}$, so

$$\frac{N_{\text{it}}}{k_0 + 1} \geq c \frac{T^{1-\omega}}{d^{1+a'}} \geq c T^{1-\omega-\frac{1+a'}{2}} = c T^\omega,$$

because $\frac{1-a'}{2} = 2\omega$. As in Step 3 there, $\mathbb{E}[(f(\hat{x}) - f^*)\mathbf{1}_G] \leq \frac{\alpha R^2}{8} e^{-cT^\omega} + C_1(32m)^p \Phi_T$.

Step 3: the complement. On G^c the guard gives $f(\hat{x}) - f^* \leq 2S_d\theta + 2S_dZ \leq 4S_dZ$. Since $Z\mathbf{1}\{Z > \theta\} \leq Z^v\theta^{1-v}$,

$$\mathbb{E}[(f(\hat{x}) - f^*)\mathbf{1}_{G^c}] \leq 4S_d\theta^{1-v} \leq CT^{a'/2-\omega(v-1)} \leq CT^{-p-2}.$$

Both remainders are absorbed into Φ_T as in Step 5 of the proof of Lemma J.11. $\square$

Proof for $3 < \beta \leq 4$. Here $q = 3$, $m = 2$, $p = p_\beta$, $\varkappa, v, \bar{c}$ are as in Lemma J.14, and v is the integer of Procedure 4. A preliminary stage estimates H_0 and a point u near x^* ; Lemma J.16 then runs on a trust set around u , in a metric calibrated by the estimate of H_0 .

Step 1: the preliminary stage. Lemma J.13 with $2v$ in place of v gives $\hat{H}$ in one round. By Weyl's inequality the smallest eigenvalue of $\hat{H}$ is at least $\alpha - \|\hat{H} - H_0\|_{\text{op}}$, so $\|A - \hat{H}\|_{\text{op}} \leq \|\hat{H} - H_0\|_{\text{op}}$ and $E_A \leq 2\|\hat{H} - H_0\|_{\text{op}}$; the eigenvectors are chosen by Lemma J.2. In the updates of Lemma J.15, each value of the stencil (C.4) with

$h = 1$ in the direction o_i has mean within $c_0 L_\beta$ of $\langle \nabla f(x_t), o_i \rangle$ (Lemma C.1, with c_0 from (C.3)), and its error has conditional second moment at most $S_2 \sigma^2 / 2$ (Lemma J.3(c)). By Lemma J.3(a) and Minkowski's inequality, $\mathbb{E}[\|\zeta_t\|^{2v} | \mathcal{F}]^{1/(2v)} \leq C\sqrt{d}$ given the past $\mathcal{F}$ of the round. The stage uses $n_0 \leq 16vd^2 + 2m(2v-1)d^2$ oracle calls in $1+d$ rounds. Since $\Xi_d^2 \leq Cd^\beta$ and $d^3\Xi_d^2 \leq Cd^{\beta+1}$, Lemmas J.13 and J.15 and Minkowski's inequality give explicit numbers S_d and F_d with

$$\|E_A\|_{L^{2v}} \leq S_d \leq Cd^{a'}, \quad \|W_A\|_{L^v} \leq F_d \leq Cd^{\beta+1}, \quad W_A = \|u - x^*\|_A^2. \quad (\text{J.10})$$

Step 2: the trust set. Let $\Lambda_0 = S_d \theta$, $K_{\text{tr}} = F_d \theta$ and $\mathcal{T} = \{x \in B(R) : \|x - u\|_A^2 \leq K_{\text{tr}}\}$, a nonempty compact convex set. On the event $G = \{E_A \leq \Lambda_0, W_A \leq K_{\text{tr}}\}$ we have $x^* \in \mathcal{T}$. Let C_R be the number C_ϱ of Lemma J.14(b) at $\varrho = R$, and $\|y\|_{H_0} = (y^\top H_0 y)^{1/2}$. For every $x \in \mathcal{T}$ and on every realization, Lemma J.14(b), $H_0 \preceq A + E_A I$ and $\|x - x^*\|_A^2 \leq 2K_{\text{tr}} + 2W_A$ give

$$f(x) - f^* \leq \frac{C_R}{2} \|x - x^*\|_{H_0}^2 \leq C_R (K_{\text{tr}} + W_A + 2R^2 E_A). \quad (\text{J.11})$$

By (J.10) and Markov's inequality, $\mathbb{P}(G^c) \leq \theta^{-2v} + \theta^{-v} \leq 2\theta^{-v}$. Hölder's inequality and (J.11) then give $\mathbb{E}[(f(x) - f^*) \mathbf{1}_{G^c}] \leq CF_d \theta^{1-v}$ for every $\mathcal{T}$ -valued random point x .

Step 3: the calibrated metric. Let $\Lambda = \alpha + \bar{c} + \Lambda_0$, $\mathbf{S} = A + \Lambda I$ and $\mu_{\mathbf{S}} = \alpha/(4\Lambda) \leq \frac{1}{4}$. On G , Lemma J.14(c) and $A - \Lambda_0 I \preceq H_0 \preceq A + \Lambda_0 I$ give, for $x \in B(R)$, $\nabla^2 f(x) \preceq \frac{3}{2}A + (\frac{3}{2}\Lambda_0 + \bar{c})I \preceq \frac{3}{2}\mathbf{S}$ and $\nabla^2 f(x) \succeq \frac{1}{2}A - (\frac{1}{2}\Lambda_0 + \bar{c})I$. The convex combination of the last bound and $\nabla^2 f \succeq \alpha I$ with weights $\lambda = \alpha/(2\Lambda)$ and $1 - \lambda$ gives

$$\nabla^2 f(x) \succeq \frac{\lambda}{2}A + ((1 - \lambda)\alpha - \lambda(\frac{1}{2}\Lambda_0 + \bar{c}))I \succeq \mu_{\mathbf{S}}\mathbf{S},$$

because $\lambda/2 = \mu_{\mathbf{S}}$ and $(1 - \lambda)\alpha - \lambda(\frac{1}{2}\Lambda_0 + \bar{c}) \geq \alpha - \lambda\Lambda = \frac{\alpha}{2} \geq \mu_{\mathbf{S}}\Lambda$.

Step 4: optimization. After increasing C_0 , $n_0 \leq T/2$, $N_{\text{it}} \geq T/(8md)$ and $N_2 \geq T/(16md)$, so the choice of h gives, as in Step 2 of the proof of Lemma J.11, $b_h^2 + \varsigma_h^2/N_2 \leq (C_b^2 + C_f + C_\sigma)(16m)^p \alpha \Phi_T$. Since $\mu_{\mathbf{S}}\Lambda = \alpha/4$, the last two terms of Lemma J.16 are at most $C_2 \Phi_T$, where C_2 depends only on β . Here $k_0 = \max\{7, 96\Lambda/\alpha\} \leq Cd^{a'} T^\omega$ and $d \leq \sqrt{T}$, so $N_{\text{it}}/(k_0 + 1) \geq cT^\omega$ as for $\beta \leq 3$, and $N_2 \geq k_0$ after increasing C_0 . On G , $\|u - x^*\|_{\mathbf{S}}^2 = W_A + \Lambda \|u - x^*\|^2 \leq F_d \theta + 4\Lambda R^2$, so the first term of Lemma J.16 is a power of T times e^{-cT^ω} .

Step 5: conclusion. The output lies in $\mathcal{T}$, so by Step 2 and $d \leq \sqrt{T}$ the contribution of G^c is at most $CF_d \theta^{1-v} \leq CT^{(\beta+1)/2 - \omega(v-1)} \leq CT^{-p-2}$. Both remainders are absorbed into Φ_T after increasing C_0 , as in the proof of Lemma J.11. The preliminary stage uses only $[f]_{\beta, \text{op}} \leq L_\beta$. $\square$

J.6 The minimax rate

Proof of Theorem 3. Upper bound. For $2 < \beta \leq 4$ it is Lemma J.17, and for $\beta > 4$ it is Lemma J.11. Both methods use at most $C_\beta T/d$ rounds (Lemma J.19).

Lower bound. Apply Theorem 1(b) with $L = 3\alpha$ and $M = 0$. Its proof (Section E.4) gives $c_\beta > 0$ and C'_0 , depending only on $\beta, \alpha, R, L_\beta, \sigma$, such that every algorithm with $T \geq C'_0 d^2$ oracle calls, in any number of rounds, has worst-case error at least $c_\beta \Phi_T$ over the functions with $\alpha I \preceq \nabla^2 f \preceq 3\alpha I$, $[f]_\beta \leq L_\beta$ and minimizer in $B(R)$, under independent $N(0, \sigma^2)$ errors. These functions belong to $\mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, \infty, R)$, and these errors are centered noise. Taking $C_0 \geq C'_0$ in Lemmas J.11 and J.17, and since $d^q \geq d^2$, both bounds hold under the stated budget condition.

Logarithmically many rounds. This is Lemma K.2; the size of its pilot is estimated after its proof. $\square$

J.7 Remarks

Remark J.18 (Computation). The methods of this appendix use exact minimizers of known functions over balls, and the oracle model does not charge this computation. The maximum-volume design depends only on d, q, R and is computed once; any design with the conclusion of Lemma J.4 can replace it. For $\beta > 4$ the guard needs a global minimizer of a polynomial of degree $q \geq 4$ over a ball, which is a nonconvex problem. On the event where the analysis takes place, the objective of the first line of (J.7) has Hessian at least $(\alpha + \Lambda)I$ on the ball; off that event

an exact global minimizer is still used. For $2 < \beta \leq 3$ the polynomial P is quadratic, and both minimizations are trust-region subproblems. For $3 < \beta \leq 4$ the method uses eigendecompositions and projections onto $B(R)$ and onto the intersection of $B(R)$ with an ellipsoid, in the norms $\|\cdot\|_A$ and $\|\cdot\|_S$ of Lemmas J.15 and J.16, and minimizes no polynomial of degree three.

Remark J.19 (Rounds). The pilots for $\beta > 4$ and for $2 < \beta \leq 3$ are single rounds, because their query points are fixed in advance; for $3 < \beta \leq 4$ the preliminary stage uses $1 + d$ rounds. The method of Lemma J.10 uses two rounds of $2md$ oracle calls per iteration, and that of Lemma J.16 one. Hence the number of rounds is of order $T/(md)$: apart from the $2md$ oracle calls of one stencil, the methods are sequential.

Remark J.20 (Oblivious noise). The stencil tolerates oblivious noise (Lemma D.3), but the analysis of the one-round pilots does not. Condition (O) bounds the second moment of an oblivious error only given $\mathcal{F}_{k-1}$, so the oblivious errors at all nodes of the pilot round may equal $\pm W$ for one random variable W with $\mathbb{E}W^2 = \sigma^2$. The medians of block means then have only the tail of W , Lemma J.3(a) does not apply, and the event $\{E_0 > \theta\}$ is no longer negligible in the proof of Lemma J.11. A bias of order σ alone would be harmless, since the proofs use only $E_0 \leq \theta$. Lemma J.12 uses the centering both in its bias bound and in its martingale part. We therefore treat centered noise only.

K LOGARITHMIC-ROUND POLYNOMIAL CORRECTION

The methods of Section J have a small polynomial startup, but their optimization steps are sequential. Here a more accurate pilot, whose residual has curvature bounded independently of d , allows epochs with doubling numbers of frames: the statistical rate is the same, the number of rounds is logarithmic, and the startup is larger. All startup calls are counted in the budget. We also treat convex objectives without a Hessian bound (Section K.3).

The notation is that of Section J.1, $K = K_q$, and $L_\beta, \sigma, R > 0$. In Sections K.1 to K.3 the noise is centered at level σ (Lemma J.1).

K.1 A pilot with dimension-free residual curvature

Let $\Gamma_0 = \sqrt{2K}$ and C_q^{pol} be as in Lemmas J.5 and J.6, and define

$$H_q = C_q^{\text{pol}}(2\bar{R})^q, \quad a = \min \left\{ 1, \left(\frac{q!}{2L_\beta \Gamma_0 H_q} \right)^{1/\gamma} \right\}, \quad \theta = \frac{a^q}{2\Gamma_0 H_q}, \quad \ell = \max\{1, \lceil 16\sigma^2/\theta^2 \rceil\}. \quad (\text{K.1})$$

The pilot is that of Section J.3 with design radius $s = a$, this ℓ and an odd $r \geq 3$. Let P be its fit, E_0 the largest absolute error of its medians, $G = \{E_0 \leq \theta\}$, and, with ϱ_0, ϱ_2 as in Section J.3,

$$U = \varrho_0 + 1, \quad \Lambda = \varrho_2 + \frac{q^4}{\bar{R}^2}. \quad (\text{K.2})$$

Unlike Υ_θ and Λ_θ in Section J.3, U and Λ do not depend on d .

Lemma K.1 (A rescaled polynomial pilot). *For every $f \in C^q(\mathbb{R}^d)$ with $[f]_{\beta, \text{op}} \leq L_\beta$, this pilot uses $Q_{\text{pil}} = 2Kr\ell$ calls in one round. On G ,*

$$\sup_{B(R)} |f - P| \leq U, \quad \sup_{B(R)} \|\nabla^2(f - P)\|_{\text{op}} \leq \Lambda.$$

If f also has a global minimizer in $B(R)$, then for every candidate in $B(R)$ the output $\hat{x}$ of the guard of Lemma J.8, with level U , satisfies $\mathbb{E}[(f(\hat{x}) - f^*)\mathbf{1}_{G^c}] \leq 4UK2^{-r}$.

Proof. Let $\mathfrak{t} = T_q f(0; \cdot)$. By Lemma J.7 with $s = a$, $\sup_{B(a)} |P - \mathfrak{t}| \leq \Gamma_0(L_\beta a^\beta/q! + E_0)$. Since $a \leq 1 \leq 2\bar{R}$, Lemma J.5(b) with $\rho = 2\bar{R}$ multiplies this bound by $H_q a^{-q}$, and (K.1) gives $H_q a^{-q} \Gamma_0 L_\beta a^\beta/q! = H_q \Gamma_0 L_\beta a^\gamma/q! \leq \frac{1}{2}$ and $H_q a^{-q} \Gamma_0 = 1/(2\theta)$. Hence, on every realization,

$$\sup_{B(2\bar{R})} |P - \mathfrak{t}| \leq \frac{1}{2} + \frac{E_0}{2\theta}. \quad (\text{K.3})$$

On G this is at most one. On $B(R)$, Lemma B.2(a),(c) give $|f - \mathfrak{t}| \leq \varrho_0$ and $\|\nabla^2(f - \mathfrak{t})\|_{\text{op}} \leq \varrho_2$, and Lemma J.5(a) with $\rho = \bar{R}$ gives $\|\nabla^2(P - \mathfrak{t})\|_{\text{op}} \leq q^4/\bar{R}^2$ on G . This proves the bounds on G .

Procedure 5 Polynomial pilot and batched residual correction

Require: $T, d, \beta, L_\beta, R, \sigma$ (also α in the strongly convex case); the pilot, guard level U , starting point and schedule $(n_j, B_j, h_j, \mu_{j,k}, \omega_{j,k})$ of Lemma K.2 or Lemma K.3

- 1: In one round acquire all Q_{pil} pilot values and fit P
 - 2: Select $x_P \in \arg \min_{B(R)} P$; set $z^{(0)} = x_P$ (Lemma K.2) or $z^{(0)} = 0$ (Lemma K.3)
 - 3: **for** $j = 1, \dots, J$ **do**
 - 4: $z_1 \leftarrow z^{(j-1)}$
 - 5: **for** $k = 1, \dots, n_j$ **do**
 - 6: $\hat{v}_z \leftarrow \bar{g}_{h_j, B_j}(z_k) - \nabla P(z_k)$ from a fresh stencil block
 - 7: $y_k \leftarrow \arg \min_{x \in B(R)} \{P(x) + \langle \hat{v}_z, x \rangle + \|x - z_k\|^2 / (2\mu_{j,k})\}$
 - 8: $\hat{v}_y \leftarrow \bar{g}_{h_j, B_j}(y_k) - \nabla P(y_k)$ from a fresh stencil block
 - 9: $z_{k+1} \leftarrow \text{Proj}_{B(R)}(y_k - \mu_{j,k}(\hat{v}_y - \hat{v}_z))$
 - 10: $z^{(j)} \leftarrow \sum_{k=1}^{n_j} \omega_{j,k} y_k / \sum_{k=1}^{n_j} \omega_{j,k}$
 - 11: **return** $z^{(J)}$ if $P(z^{(J)}) \leq P(x_P) + 2U$, and x_P otherwise
-

By (J.6) with $2K$ nodes and $\sigma^2/\ell \leq \theta^2/16$, $\mathbb{E}[E_0 \mathbf{1}_{G^c}] \leq K\theta 2^{-r}$. By Lemma J.8 and (K.3), $f(\hat{x}) - f^* \leq 4\varrho_0 + 3 + E_0/\theta$ on every realization, which is at most $4UE_0/\theta$ on G^c . $\square$

K.2 A statistical rate in logarithmically many rounds

Procedure 5 runs a pilot and then J epochs of the corrected method (J.7). Epoch j makes n_j iterations with B_j frames per stencil, radius h_j , steps $\mu_{j,k}$ and averaging weights $\omega_{j,k}$, and starts from the output of the previous epoch. The algorithm uses $Q_{\text{pil}} + 4md \sum_{j=1}^J n_j B_j$ calls in $1 + 2 \sum_{j=1}^J n_j$ rounds: in each iteration the second center depends on the first stencil, while the B_j frames of one stencil are queried together. Polynomial minimization and the guard need no calls.

Theorem K.2 (Logarithmically many rounds). *Let $\beta > 2$, $\alpha, L_\beta, \sigma, R > 0$, and let $f \in \mathcal{F}_{\text{sc}}(\beta, L_\beta; \alpha, \infty, R)$ with centered noise at level σ . Run Procedure 5 with the pilot of Section K.1, where r is the smallest odd integer $r \geq 3$ with $4UK2^{-r} \leq \Phi_T$, with $Q_{\text{pil}} = 2Kr\ell$, the guard level U , the starting point $z^{(0)} = x_P$, and, for $k_0 = \max\{7, 16\Lambda/\alpha\}$ and $n = \lceil 2k_0 \rceil$,*

$$n_j = n, \quad B_j = 2^{j-1}, \quad h_j = \left(\frac{d^{1+\tau} \sigma^2}{L_\beta^2 n B_j} \right)^{1/(2\beta)}, \quad \mu_{j,k} = \frac{8}{\alpha(k + k_0)}, \quad \omega_{j,k} = k + k_0 - 1,$$

for the epochs $j = 1, \dots, J$, where

$$J = \left\lceil \log_2 \left(1 + \frac{T - Q_{\text{pil}}}{4mdn} \right) \right\rceil. \quad (\text{K.4})$$

If

$$Q_{\text{pil}} \leq T/2, \quad T - Q_{\text{pil}} \geq 4mdn, \quad 2U16^{-J} \leq \Phi_T, \quad (\text{K.5})$$

then the algorithm uses at most T calls in exactly $1 + 2nJ$ rounds, which is at most a constant multiple of $1 + \log T$, with the constant depending only on $\beta, \alpha, L_\beta, R$, and its output satisfies

$$\mathbb{E}[f(\hat{x}) - f^*] \leq C_\beta \Phi_T.$$

Proof. The epochs use $4mdn(2^J - 1) \leq T - Q_{\text{pil}}$ calls, and n depends only on $\beta, \alpha, L_\beta, R$. Write $p = p_\beta$ and $A_j = \mathbb{E}[(f(z^{(j)}) - f^*) \mathbf{1}_G]$. On G , $|f - P| \leq U$ on $B(R)$ (Lemma K.1), so $A_0 \leq 2U$.

Consider epoch j . Its starting point $z^{(j-1)} \in B(R)$ is measurable with respect to the earlier rounds, and all its steps are at most $8/(\alpha(k_0 + 1))$. Hence Lemma J.9, with B_j frames, gives (F.2) with $a = \alpha/4$, $c = 2$, $\mathcal{B} = 2b_{h_j}^2/\alpha$ and $\mathcal{V} = 8V_j$, where $V_j = b_{h_j}^2 + \varsigma_{h_j}^2/B_j$. Lemma F.1(a) with $N = n$ and the convexity of f give

$$A_j \leq \frac{\alpha k_0^2}{8n^2} \mathbb{E}[\|z^{(j-1)} - x^*\|^2 \mathbf{1}_G] + \frac{b_{h_j}^2}{\alpha} + \frac{64V_j}{\alpha n}. \quad (\text{K.6})$$

Strong convexity and $n \geq 2k_0$ bound the first term by $A_{j-1}/16$. For $Q_j = 4mdnB_j$, the radius h_j gives

$$L_\beta^2 d^{-\tau} h_j^{2\beta-2} = \frac{d\sigma^2}{nB_j h_j^2} = (4m)^p \alpha \Phi_{Q_j},$$

so $A_j \leq A_{j-1}/16 + C_\beta \Phi_{Q_j}$. Since $\Phi_{Q_{j-1}} = 2^p \Phi_{Q_j}$ and $2^p/16 \leq 1/8$, induction gives $A_J \leq 2U16^{-J} + C_\beta \Phi_{Q_J}$ with a new C_β . Maximality of J gives $Q_J > (T - Q_{\text{pil}})/4 \geq T/8$, so (K.5) yields $A_J \leq C_\beta \Phi_T$. On G the guard does not increase the value (Lemma J.8), and on G^c Lemma K.1 and the choice of r bound its contribution by Φ_T . $\square$

Size of the pilot. For fixed $\beta, \alpha, L_\beta, \sigma, R$ and sufficiently large K , the quantities a, θ and ℓ in (K.1) are bounded above and below by positive multiples of $K^{-1/(2\gamma)}$, $K^{-(q/\gamma+1)/2}$ and $K^{q/\gamma+1}$, respectively; these multiples depend only on the fixed parameters. Since $\Phi_T \geq cT^{-p_\beta}$, there is also such a constant C for which $r \leq C(1 + \log(2+d+T))$. Consequently,

$$Q_{\text{pil}} \leq CK^{2+q/\gamma}(1 + \log(2+d+T)).$$

For integer β , the right-hand side is at most another such constant times $d^{\beta^2-1}(1 + \log(2+d+T))$, and the onset conditions (K.5) hold once T exceeds a polynomial in d , up to logarithmic factors. This startup is much larger than those of Lemmas J.11 and J.17. The leading constant C_β depends only on β ; the startup and the number of rounds also depend on $\alpha, L_\beta, \sigma, R$.

K.3 A polynomial-pilot method for convex objectives

Let $f \in C^q(\mathbb{R}^d)$ be convex with $[f]_\beta \leq L_\beta$ and a global minimizer $x^* \in B(R)$; strong convexity is not assumed. We use the pilot of Section J.3 on $B(2R)$ with tolerance 1, $\ell = \max\{1, \lceil 16\sigma^2 \rceil\}$ and an odd $r \geq 3$, so that $Q_{\text{pil}} = 2Kr\ell$. Let $U_1 = \Upsilon_1$ and Λ_1 be the numbers (J.4) with $\theta = 1$, and

$$A_{\text{pil}} = 2U_1 + 2\varrho_0 + 2\Gamma_0\varrho_0 + 2\Gamma_0.$$

Theorem K.3 (Polynomial-pilot upper bound for convex objectives). *Under the assumptions just stated, let $N \geq 1$ and*

$$h = \left(\frac{d^{1+\tau}\sigma^2}{L_\beta^2 N} \right)^{1/(2\beta)}, \quad V_h = b_h^2 + s_h^2, \quad \mu = \min \left\{ \frac{1}{2\Lambda_1}, \frac{R}{\sqrt{8NV_h}} \right\},$$

with b_h, s_h from (J.2). Run Procedure 5 with this pilot, the guard level U_1 , the starting point $z^{(0)} = 0$ and one epoch ($J = 1$) with $n_1 = N$, $B_1 = 1$, $h_1 = h$, $\mu_{1,k} = \mu$ and $\omega_{1,k} = 1$. The output satisfies

$$\mathbb{E}[f(\hat{x}) - f^*] \leq C_\beta \left\{ \frac{\Lambda_1 R^2}{N} + RL_\beta^{1/\beta} \sigma^{(\beta-1)/\beta} d^{(\beta-1-\tau)/(2\beta)} N^{-(\beta-1)/(2\beta)} \right\} + A_{\text{pil}} K 2^{-r}. \quad (\text{K.7})$$

The algorithm uses $Q_{\text{pil}} + 4mdN$ calls in $1 + 2N$ rounds.

Proof. Let $G = \{E_0 \leq 1\}$, $\Delta_k = \mathbb{E}[\|z_k - x^*\|^2 \mathbf{1}_G]$ and $F_k = \mathbb{E}[(f(y_k) - f^*) \mathbf{1}_G]$. On G the residual Hessian is at most Λ_1 on $B(R)$ (Lemma J.7) and $\mu\Lambda_1 \leq 1/2$, so Lemma J.9 with $\alpha = 0$ gives $\Delta_{k+1} \leq \Delta_k - 2\mu F_k + 4\mu R b_h + 8\mu^2 V_h$. Summing over $k \leq N$, and using $\Delta_1 \leq R^2$ and convexity, we obtain for the average $\bar{y} = z^{(1)}$ of the y_k

$$\mathbb{E}[(f(\bar{y}) - f^*) \mathbf{1}_G] \leq \frac{R^2}{2\mu N} + 2R b_h + 4\mu V_h \leq \frac{\Lambda_1 R^2}{N} + 2\sqrt{2}R\sqrt{V_h/N} + 2R b_h.$$

The radius gives $X := L_\beta^2 d^{-\tau} h^{2\beta-2} = d\sigma^2/(Nh^2)$, hence $b_h^2 = C_b^2 X$, $V_h/N \leq (C_b^2 + C_f + C_\sigma)X$ and $\sqrt{X} = L_\beta^{1/\beta} \sigma^{(\beta-1)/\beta} d^{(\beta-1-\tau)/(2\beta)} N^{-(\beta-1)/(2\beta)}$. On G the guard does not increase the objective. On G^c , Lemma J.8 and (J.5) bound its gap by $A_{\text{pil}} E_0$, and (J.6) gives $\mathbb{E}[E_0 \mathbf{1}_{G^c}] \leq K 2^{-r}$. $\square$

With the whole budget, let

$$S_T = RL_\beta^{1/\beta} \sigma^{(\beta-1)/\beta} d^{1-(2+\tau)/(2\beta)} T^{-(\beta-1)/(2\beta)},$$

let r be the smallest odd integer $r \geq 3$ with $A_{\text{pil}} K 2^{-r} \leq S_T$, and $N = \lfloor (T - Q_{\text{pil}})/(4md) \rfloor$. If $Q_{\text{pil}} \leq T/2$, $T \geq 16md$ and $\Lambda_1 R^2 d/T \leq S_T$, then $T/4 \leq 4mdN \leq T - Q_{\text{pil}}$, and (K.7) gives $\mathbb{E}[f(\hat{x}) - f^*] \leq C_\beta S_T$ with at most T calls.

Ellipsoidal norms. Let Σ be a known invertible matrix and $\|x\|_\Sigma = \|\Sigma x\|$. Lemmas K.2 and K.3 apply to $\tilde{f}(z) = f(\Sigma^{-1}z)$, with each query z made at $\Sigma^{-1}z$. Function values and their errors are unchanged, so these results hold, with the same calls, rounds and noise levels, when strong convexity, the localization of x^* , the Hölder seminorms ($[f]_{\beta, \text{op}, \Sigma} = [\tilde{f}]_{\beta, \text{op}}$ and $[f]_{\beta, \text{HS}, \Sigma} = [\tilde{f}]_{\beta, \text{HS}}$) and the Lipschitz constant of η are measured in $\|\cdot\|_\Sigma$ and its dual norm. No factor depending on the condition number of Σ appears.

L CONVEX OBJECTIVES: UPPER BOUNDS

This appendix proves the upper bounds of Section 6 for minimization over a compact convex set Q , where the minimizer may lie on the boundary with a nonzero gradient; the class $\mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$ is the case $Q = B(R)$ (Lemma L.1). The gradient-scale bound, Lemma L.6, rests on projected methods with a bounded bias (Sections L.1 and L.2). The value-scale bound, Lemma L.9, rests on descent on the ball average of f (Section L.3). Section L.4 simplifies their constants. Section L.5 proves the upper bound of Lemma L.10, whose lower bound is proved in Section M, and Section L.6 gives the achievable level of Lipschitz noise.

Setting. For a nonempty closed convex set $K \subset \mathbb{R}^d$, Proj_K is the Euclidean projection onto K , and for $r \geq 0$

$$K_r = K + rB(1) = \{x \in \mathbb{R}^d : \text{dist}(x, K) \leq r\};$$

this set is closed and convex, and $K_r + sB(1) = K_{r+s}$ for $r, s \geq 0$. The random vector W is uniform on $B(1)$ and U is uniform on $\mathbb{S}^{d-1}$. For a function g defined near $x + hB(1)$ we write $g_h(x) = \mathbb{E}g(x + hW)$. With $\gamma = \beta - q$ as in Section 2, we put $c_\gamma = 1/((1 + \gamma)(2 + \gamma)) \leq 1/2$. Rounds and noise obey Definition 1.

Definition L.1 (Constrained problems). Let $Q \subset \mathbb{R}^d$ be nonempty, closed and convex with diameter $D = \sup_{x, y \in Q} \|x - y\| \in (0, \infty)$; thus Q is compact. For a closed convex set $K \subset \mathbb{R}^d$, let $\mathcal{F}_{\text{cvx}}^K(\beta, L_\beta; L)$ be the set of real functions f that are defined and q times continuously differentiable on an open set containing K and satisfy

$$0 \preceq \nabla^2 f(x) \preceq LI \quad (x \in K), \quad \|D^q f(x) - D^q f(y)\| \leq L_\beta \|x - y\|^\gamma \quad (x, y \in K),$$

where the norm of $D^q f(x) - D^q f(y)$ is $\|\cdot\|_{\text{op}}$ ($\tau = 0$) or $\|\cdot\|_{\text{HS}}$ ($\tau = \tau_\beta$). A *constrained problem* consists of Q , a starting point $x_0 \in Q$, a number $R \in (0, D]$ and an objective f with a minimizer $u \in \arg \min_Q f$ such that $\|x_0 - u\| \leq R$. Each result below states the smoothness of f near Q that it uses. A method knows Q through Proj_Q , and it knows x_0, R, D and the parameters of the class. All its queries lie in Q_h , where h is the radius of its stencil, and it outputs a point $\hat{x} \in Q$; its error is $\mathbb{E}[f(\hat{x}) - \min_Q f]$.

By integration along segments, the Hessian inequalities imply that f is convex with an L -Lipschitz gradient on every convex set on which they hold. If the projection onto $Q \cap B(x_0, R)$ is available, Q may be replaced by this intersection and D by its diameter, which is at most $\min\{D, 2R\}$.

Known parameters. The methods of this appendix know $\text{Proj}_Q, x_0, \beta, L, R, D, \sigma, d$, the target accuracy ε and the constants of Lemma D.3 (explicit for $m = 1$). They know L_β , except for the method of Lemma L.9(b). They do not use M, δ, f or η ; the numbers $\bar{M}$ below are design values. We assume $L > 0$ and $L_\beta > 0$: an affine objective satisfies the assumptions with every $L > 0$, and if $L_\beta = 0$, every statement holds with any $L'_\beta > 0$ in place of L_β . The case $\sigma = 0$ is included; the terms containing σ then vanish.

The estimate. The next lemma is Lemma D.3 for an objective that is smooth only near the set that contains the centers.

Lemma L.2 (The estimate near a set). *Let $K \subset \mathbb{R}^d$ be nonempty, closed and convex, let $h > 0$, and put $\vartheta = 2$ if the Hölder bound is measured in the Hilbert–Schmidt norm and q is even, and $\vartheta = 1$ otherwise. Let f be q times continuously differentiable on an open set containing $K_{\vartheta h}$, with $\|D^q f(x) - D^q f(y)\| \leq L_\beta \|x - y\|^\gamma$ for all $x, y \in K_{\vartheta h}$, and let the noise be admissible at levels (M, σ, δ) . Consider round k of Procedure 1 with $B \geq 1$ frames, the deterministic radius h and an $\mathcal{F}_{k-1}$ -measurable center z with values in K . Then all queries of the round lie in K_h , and the conclusions (i)–(iv) of Lemma D.3 hold. In particular*

$$\bar{g} = \nabla f(z) + b(z) + \zeta + \Delta, \tag{L.1}$$

where $b(z)$ is $\mathcal{F}_{k-1}$ -measurable, $\zeta + \Delta$ is $\mathcal{F}_k$ -measurable, $\mathbb{E}[\zeta \mid \mathcal{F}_{k-1}] = 0$,

$$\|b(z)\| \leq C_b L_\beta d^{-\tau/2} h^{\beta-1} + A_1 M, \quad \|\Delta\| \leq \frac{\sqrt{d} S_1 \delta}{h} \text{ almost surely,}$$

and

$$\mathbb{E}[\|\zeta\|^2 \mid \mathcal{F}_{k-1}] \leq \frac{1}{B} \left(C_f L_\beta^2 d^{-\tau} h^{2\beta-2} + C_\eta M^2 + C_\sigma \frac{d\sigma^2}{h^2} \right). \quad (\text{L.2})$$

The bound (L.2) also holds for $d = 1$.

Proof. All queries lie in K_h . The Taylor, Pizzetti and Poincaré arguments in Section C use only the ball $z + hB(1)$, except for the Hilbert–Schmidt bias with even q : its additional difference of radius $h' = hd^{-1/4}$ uses $z + (h + h')B(1) \subseteq K_{2h}$. So the stated neighborhood contains every point used by those proofs. The surrogate gradient is continuous on K by dominated convergence on compact subsets, so its value at the predictable center is measurable. The noise part of the proof of Lemma D.3 is unchanged, and its stronger $d = 1$ variance bound implies (L.2). $\square$

For $m = 1$, that is, for $\beta \in (2, 3]$, the stencil is the central difference: $t_1 = a_1 = 1$, and $\mathcal{S}_h f = f_h$, $\mathcal{S}_h \eta = \eta_h$ in (C.5). Hence

$$\nabla f(z) + b(z) = \nabla f_h(z) + \nabla \eta_h(z), \quad \|\nabla \eta_h(z)\| \leq M, \quad (\text{L.3})$$

and $S_1 = S_2 = A_1 = 1$, $C_\eta = 4$ and $C_\sigma = (1 + 2^{-1/2})^2$. In the operator geometry also $C_b = c_1 = 1$ and $C_f = 4c_1^2 = 4$ (Section D and (C.3)).

L.1 Projected methods with a bounded bias

In this subsection Q is as in Lemma L.1, and φ is differentiable on an open set containing Q , convex on Q , with $\|\nabla \varphi(x) - \nabla \varphi(y)\| \leq L \|x - y\|$ for $x, y \in Q$. Let $(\mathcal{F}_k)_{k \geq 0}$ be a filtration; in the applications it is the filtration of Definition 1, and step k is round k . In step k an $\mathcal{F}_{k-1}$ -measurable point $z \in Q$, the *center*, is submitted to an oracle, which returns an $\mathcal{F}_k$ -measurable vector

$$g_k = \nabla \varphi(z) + b_k + \zeta_k, \quad \|b_k\| \leq b \text{ a.s.}, \quad \mathbb{E}[\zeta_k \mid \mathcal{F}_{k-1}] = 0, \quad \mathbb{E}[\|\zeta_k\|^2 \mid \mathcal{F}_{k-1}] \leq v_B^2. \quad (\text{L.4})$$

In the applications b_k contains the adversarial part of the estimate, which may depend on ζ_k ; so b_k and ζ_k are not assumed to be $\mathcal{F}_k$ -measurable. The comparator $u \in Q$ is an arbitrary fixed point. In the applications it minimizes f on Q , possibly at a boundary point with $\nabla f(u) \neq 0$, and in Lemma L.9 it need not minimize $\varphi = f_h$. For $x, y \in Q$ the segment $[x, y]$ lies in Q , so integration of $\nabla \varphi$ along it and convexity give

$$\varphi(y) \leq \varphi(x) + \langle \nabla \varphi(x), y - x \rangle + \frac{L}{2} \|y - x\|^2, \quad \varphi(y) \geq \varphi(x) + \langle \nabla \varphi(x), y - x \rangle. \quad (\text{L.5})$$

Lemma L.3 (Projected stochastic gradient). *Let $z_1 = x_0 \in Q$ and $0 < \nu \leq 1/(2L)$. For $k = 1, \dots, N$ submit the center z_k , and let $z_{k+1} = \text{Proj}_Q(z_k - \nu g_k)$. Let $\bar{z}_N = \frac{1}{N} \sum_{k=2}^{N+1} z_k$. For every $u \in Q$,*

$$\mathbb{E}[\varphi(\bar{z}_N) - \varphi(u)] \leq \frac{\|x_0 - u\|^2}{2\nu N} + \nu v_B^2 + Db. \quad (\text{L.6})$$

In particular, let $\|x_0 - u\| \leq R$, $\bar{v}_B \geq v_B$ and $\nu = \min\{1/(2L), R/(\bar{v}_B \sqrt{2N})\}$, with $\nu = 1/(2L)$ if $\bar{v}_B = 0$. Then

$$\mathbb{E}[\varphi(\bar{z}_N) - \varphi(u)] \leq \frac{LR^2}{N} + \frac{\sqrt{2} R \bar{v}_B}{\sqrt{N}} + Db. \quad (\text{L.7})$$

Proof. This is the standard analysis of projected stochastic gradient descent for smooth objectives, with the bias kept. By induction, $z_k \in Q$ is $\mathcal{F}_{k-1}$ -measurable. Fix k . The projection inequality $\langle z_k - \nu g_k - z_{k+1}, u - z_{k+1} \rangle \leq 0$ gives

$$\nu \langle g_k, z_{k+1} - u \rangle \leq \frac{1}{2} (\|z_k - u\|^2 - \|z_{k+1} - u\|^2 - \|z_{k+1} - z_k\|^2). \quad (\text{L.8})$$

By (L.5) for the pairs (z_k, z_{k+1}) and (z_k, u) , $\varphi(z_{k+1}) - \varphi(u) \leq \langle \nabla \varphi(z_k), z_{k+1} - u \rangle + \frac{L}{2} \|z_{k+1} - z_k\|^2$. Write $\nabla \varphi(z_k) = g_k - b_k - \zeta_k$, bound $\langle g_k, z_{k+1} - u \rangle$ by (L.8), $-\langle b_k, z_{k+1} - u \rangle$ by bD , and $-\langle \zeta_k, z_{k+1} - z_k \rangle$ by $\nu \|\zeta_k\|^2 + \|z_{k+1} - z_k\|^2/(4\nu)$. Since $\nu \leq 1/(2L)$, the terms with $\|z_{k+1} - z_k\|^2$ have a nonpositive total coefficient, and

$$\varphi(z_{k+1}) - \varphi(u) \leq \frac{\|z_k - u\|^2 - \|z_{k+1} - u\|^2}{2\nu} + \nu \|\zeta_k\|^2 + bD - \langle \zeta_k, z_k - u \rangle. \quad (\text{L.9})$$

The last term has mean zero, because $z_k - u$ is bounded and $\mathcal{F}_{k-1}$ -measurable. Summing over $k = 1, \dots, N$, taking expectations and using the convexity of φ on Q proves (L.6). For the tuned step, $1/\nu \leq 2L + \bar{v}_B\sqrt{2N}/R$ gives $R^2/(2\nu N) \leq LR^2/N + R\bar{v}_B/\sqrt{2N}$, and $\nu\bar{v}_B^2 \leq R\bar{v}_B/\sqrt{2N}$. $\square$

The accelerated method below is a variant of the AC-SA method of Lan (2012) for convex objectives, run with a biased oracle.

Lemma L.4 (Accelerated projected method). *Let $\Gamma_0 = 0$, $\Gamma_k = k(k+1)/2$ and $\rho_k = k/\Gamma_k = 2/(k+1)$ for $k \geq 1$. Let $v_0 = x_0 \in Q$ and $0 < \nu_0 \leq 1/(4L)$. For $k = 1, \dots, N$ submit the center z_k , where*

$$z_k = (1 - \rho_k)x_{k-1} + \rho_k v_{k-1}, \quad v_k = \text{Proj}_Q(v_{k-1} - \nu_0 k g_k), \quad x_k = (1 - \rho_k)x_{k-1} + \rho_k v_k.$$

Then for every $u \in Q$,

$$\mathbb{E}[\varphi(x_N) - \varphi(u)] \leq \frac{\|x_0 - u\|^2}{\nu_0 N(N+1)} + \frac{(2N+1)\nu_0 v_B^2}{3} + Db. \quad (\text{L.10})$$

In particular, let $\|x_0 - u\| \leq R$, $\bar{v}_B \geq v_B$ and $\nu_0 = \min\{1/(4L), R/(\bar{v}_B N^{3/2})\}$, with $\nu_0 = 1/(4L)$ if $\bar{v}_B = 0$. Then

$$\mathbb{E}[\varphi(x_N) - \varphi(u)] \leq \frac{4LR^2}{N^2} + \frac{2R\bar{v}_B}{\sqrt{N}} + Db. \quad (\text{L.11})$$

Proof. All iterates lie in Q , and z_k, v_{k-1}, x_{k-1} are $\mathcal{F}_{k-1}$ -measurable. Write $\nu_k = \nu_0 k$. Smoothness, convexity, and $x_k - z_k = \rho_k(v_k - v_{k-1})$ give

$$\varphi(x_k) - \varphi(u) \leq (1 - \rho_k)(\varphi(x_{k-1}) - \varphi(u)) + \rho_k \langle \nabla \varphi(z_k), v_k - u \rangle + \frac{1}{2} L \rho_k^2 \|v_k - v_{k-1}\|^2.$$

Split the oracle error as in the proof of Lemma L.3: apply the projection inequality to $\langle g_k, v_k - u \rangle$, bound the bias term by bD , and use $-\langle \zeta_k, v_k - v_{k-1} \rangle \leq \nu_k \|\zeta_k\|^2 + \|v_k - v_{k-1}\|^2/(4\nu_k)$. The remaining coefficient of $\|v_k - v_{k-1}\|^2$ is nonpositive because $2L\rho_k\nu_k \leq 1$. Multiplication by Γ_k yields

$$\begin{aligned} \Gamma_k(\varphi(x_k) - \varphi(u)) &\leq \Gamma_{k-1}(\varphi(x_{k-1}) - \varphi(u)) + \frac{\|v_{k-1} - u\|^2 - \|v_k - u\|^2}{2\nu_0} \\ &\quad + \nu_0 k^2 \|\zeta_k\|^2 + kbD - k \langle \zeta_k, v_{k-1} - u \rangle. \end{aligned}$$

The last term has mean zero. Summing, dropping the final nonpositive distance, and using $\sum k^2 = N(N+1)(2N+1)/6$ proves (L.10). Finally $1/\nu_0 \leq 4L + \bar{v}_B N^{3/2}/R$, $(2N+1)/3 \leq N$, and $N\nu_0 v_B^2 \leq R\bar{v}_B/\sqrt{N}$ give (L.11). $\square$

Remark L.5 (The bias does not accumulate). In Lemma L.4 the quadratic upper model uses the exact gradient $\nabla \varphi(z_k)$, and the bias enters only through $\rho_k \langle b_k, u - v_k \rangle \leq \rho_k bD$, because v_k is projected onto the bounded set Q . So the accelerated method pays the same bias term Db as the plain method; for an unbounded feasible set the argument gives no bound. Treated as an inexact oracle in the sense of Devolder et al. (2014), a gradient error of norm b on a set of diameter D has accuracy of order bD , and the fast gradient method accumulates it into a term of order NDb . A bound of the same form as (L.10), with noise weighted by k/Γ_N and multiplied by the diameter, is Proposition 3.5 of Cohen et al. (2018).

L.2 The stencil as a biased gradient

The next corollary uses only Lemma L.2 and the lemmas above. Its accelerated form is Procedure 6, which with $m = 1$ is also the method of Lemma L.9(d).

Corollary L.6 (Direct projected methods). *Let Q , x_0 , R and D be as in Lemma L.1, let $\varepsilon > 0$, and put*

$$h = (4C_b)^{-\frac{1}{\beta-1}} \left(\frac{\varepsilon d^{\tau/2}}{L_\beta D} \right)^{\frac{1}{\beta-1}}, \quad \bar{M} = \frac{\varepsilon}{8A_1 D}, \quad \bar{v}_h^2 = C_f L_\beta^2 d^{-\tau} h^{2\beta-2} + C_\eta \bar{M}^2 + C_\sigma \frac{d\sigma^2}{h^2}.$$

Let ϑ be as in Lemma L.2, and let f be q times continuously differentiable on an open set containing $Q_{\vartheta h}$, with

$$0 \leq \nabla^2 f(x) \preceq LI \quad (x \in Q), \quad \|D^q f(x) - D^q f(y)\| \leq L_\beta \|x - y\|^\gamma \quad (x, y \in Q_{\vartheta h}),$$

and with a minimizer $u \in \arg \min_Q f$ such that $\|x_0 - u\| \leq R$; for instance, $f \in \mathcal{F}_{\text{cvx}}^{Q_{\vartheta h}}(\beta, L_\beta; L)$ with such a minimizer. Let the noise be admissible at levels (M, σ, δ) . Consider the following two methods; in both, the oracle of round k is Procedure 1 at the center z_k , with radius h and B frames.

Procedure 6 Projected acceleration with batched scalar stencils

Require: Q , Proj_Q , $x_0 \in Q$; h, m, N, B, ν_0 from Lemma L.6(b), or $m = 1$ and h, N, B, ν_0 from Lemma L.9(d)

```

1:  $v_0 \leftarrow x_0$ 
2: for  $k = 1, \dots, N$  do
3:    $\rho_k \leftarrow 2/(k+1)$ ;  $z_k \leftarrow (1 - \rho_k)x_{k-1} + \rho_k v_{k-1}$ 
4:   Acquire  $\bar{g}_k$  at  $z_k$  by Procedure 1, radius  $h$ ,  $B$  frames
5:    $v_k \leftarrow \text{Proj}_Q(v_{k-1} - \nu_0 k \bar{g}_k)$ 
6:    $x_k \leftarrow (1 - \rho_k)x_{k-1} + \rho_k v_k$ 
7: return  $x_N \in Q$ 
    
```

(a) Run Lemma L.3 with $\varphi = f$ and $z_1 = x_0$, output $\bar{z}_N$, and take

$$N = \left\lceil \frac{8LR^2}{\varepsilon} \right\rceil, \quad T_0 = \frac{256mdR^2\bar{v}_h^2}{9\varepsilon^2}, \quad B = \left\lceil \frac{T_0}{2mdN} \right\rceil, \quad \nu = \min\left\{ \frac{1}{2L}, \frac{R\sqrt{B}}{\bar{v}_h\sqrt{2N}} \right\}.$$

(b) Run Lemma L.4 with $\varphi = f$ and $v_0 = x_0$, output x_N , and take

$$N = \left\lceil \sqrt{32LR^2/\varepsilon} \right\rceil, \quad T_0 = \frac{512mdR^2\bar{v}_h^2}{9\varepsilon^2}, \quad B = \left\lceil \frac{T_0}{2mdN} \right\rceil, \quad \nu_0 = \min\left\{ \frac{1}{4L}, \frac{R\sqrt{B}}{\bar{v}_h N^{3/2}} \right\}.$$

Suppose $M \leq \bar{M}$ and $\delta \leq \varepsilon h / (8S_1 D \sqrt{d})$. Then both methods query only points of Q_h , both outputs $\hat{x}$ lie in Q and satisfy $\mathbb{E}[f(\hat{x}) - \min_Q f] \leq \varepsilon$, and the number of calls is $T = 2mdNB \leq T_0 + 2mdN$, where, explicitly,

$$T_0 = \theta \left[\frac{256mC_\sigma(4C_b)^{\frac{2}{\beta-1}}}{9} \frac{\sigma^2}{\varepsilon^2} d^{2-\frac{\tau}{\beta-1}} R^2 \left(\frac{L_\beta D}{\varepsilon} \right)^{\frac{2}{\beta-1}} + md \frac{R^2}{D^2} \left(\frac{16C_f}{9C_b^2} + \frac{4C_\eta}{9A_1^2} \right) \right], \quad (\text{L.12})$$

with $\theta = 1$ in (a) and $\theta = 2$ in (b).

Proof. Let z_k be the center of round k ; it is $\mathcal{F}_{k-1}$ -measurable and lies in Q . By Lemma L.2 with $K = Q$, the queries lie in Q_h , and the returned vector is $g_k = \nabla f(z_k) + b_k + \zeta_k$ with $b_k = b(z_k) + \Delta_k$, $\mathbb{E}[\zeta_k | \mathcal{F}_{k-1}] = 0$ and, since $M \leq \bar{M}$, $\mathbb{E}[\|\zeta_k\|^2 | \mathcal{F}_{k-1}] \leq \bar{v}_h^2/B$. The choice of h gives $C_b L_\beta d^{-\tau/2} h^{\beta-1} = \varepsilon/(4D)$, so

$$\|b_k\| \leq b := \frac{\varepsilon}{4D} + A_1 M + \frac{\sqrt{d} S_1 \delta}{h} \leq \frac{\varepsilon}{4D} + \frac{\varepsilon}{8D} + \frac{\varepsilon}{8D} = \frac{\varepsilon}{2D}.$$

Thus the setting of Section L.1 holds for $\varphi = f$, which is convex with an L -Lipschitz gradient on Q , with $v_B^2 = \bar{v}_h^2/B$ and $Db \leq \varepsilon/2$.

In case (a), (L.7) with $\bar{v}_B = \bar{v}_h/\sqrt{B}$ bounds the error by $LR^2/N + \sqrt{2R\bar{v}_h}/\sqrt{NB} + Db$. Here $LR^2/N \leq \varepsilon/8$, and $NB \geq T_0/(2md) = 128R^2\bar{v}_h^2/(9\varepsilon^2)$ gives $\sqrt{2R\bar{v}_h}/\sqrt{NB} \leq 3\varepsilon/8$. In case (b), (L.11) bounds the error by $4LR^2/N^2 + 2R\bar{v}_h/\sqrt{NB} + Db$, where $4LR^2/N^2 \leq \varepsilon/8$ and $NB \geq 256R^2\bar{v}_h^2/(9\varepsilon^2)$ gives $2R\bar{v}_h/\sqrt{NB} \leq 3\varepsilon/8$. In both cases the sum is at most ε , and $f(u) = \min_Q f$.

Finally, $B \leq T_0/(2mdN) + 1$ gives $T \leq T_0 + 2mdN$, and inserting

$$L_\beta^2 d^{-\tau} h^{2\beta-2} = \frac{\varepsilon^2}{16C_b^2 D^2}, \quad \bar{M}^2 = \frac{\varepsilon^2}{64A_1^2 D^2}, \quad \frac{1}{h^2} = (4C_b)^{\frac{2}{\beta-1}} d^{-\frac{\tau}{\beta-1}} \left(\frac{L_\beta D}{\varepsilon} \right)^{\frac{2}{\beta-1}}$$

into T_0 gives (L.12). □

L.3 Value analysis of the ball average

The method of this subsection is projected stochastic gradient descent driven by central differences, analyzed as descent on the ball average f_h . It uses central differences for every β : for $\beta > 3$ the stencil of Lemma D.3 has $m \geq 2$, and its surrogate $\mathcal{S}_h f$ has weights of both signs and need not be convex.

Lemma L.7 (Values of the ball average). *Let Q and D be as in Lemma L.1, and let $h > 0$.*

- (a) Let $\gamma \in (0, 1]$ and $\beta = 2 + \gamma$, and let f be twice continuously differentiable on an open set containing Q_h , with $\|\nabla^2 f(x) - \nabla^2 f(y)\|_{\text{op}} \leq L_\beta \|x - y\|^\gamma$ for all $x, y \in Q_h$. Then for $x \in Q$,

$$f_h(x) = f(x) + \frac{h^2}{2(d+2)} \text{tr} \nabla^2 f(x) + r_h(x), \quad |r_h(x)| \leq \frac{c_\gamma d}{d+\beta} L_\beta h^\beta \leq c_\gamma L_\beta h^\beta, \quad (\text{L.13})$$

and $|\text{tr} \nabla^2 f(x) - \text{tr} \nabla^2 f(y)| \leq d L_\beta \|x - y\|^\gamma$ for all $x, y \in Q_h$.

- (b) Let f be differentiable on an open set containing Q_h and convex on Q_h , with $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$ for $x, y \in Q_h$. Then f_h is defined and continuously differentiable on an open set containing Q , with $\nabla f_h(x) = \mathbb{E} \nabla f(x + hW)$ for $x \in Q$. It is convex on Q , $\|\nabla f_h(x) - \nabla f_h(y)\| \leq L \|x - y\|$ for $x, y \in Q$, and

$$0 \leq f_h(x) - f(x) \leq \frac{d}{2(d+2)} L h^2 \quad (x \in Q).$$

If moreover f is twice continuously differentiable on an open set containing Q_h , then $0 \leq \text{tr} \nabla^2 f(x) \leq dL$ for $x \in Q$.

- (c) Under the assumptions of (b), for all $x, y \in Q$,

$$f(x) - f(y) \leq f_h(x) - f_h(y) + \frac{Lh^2}{2}. \quad (\text{L.14})$$

Under the assumptions of (a) and (b), for all $x, y \in Q$,

$$f(x) - f(y) \leq f_h(x) - f_h(y) + S(h), \quad S(h) := \min \left\{ \frac{Lh^2}{2}, \frac{L_\beta D^\gamma h^2}{2} + 2c_\gamma L_\beta h^\beta \right\}. \quad (\text{L.15})$$

In particular, under (a) and (b), for $x, y \in Q$,

$$|\text{tr} \nabla^2 f(x) - \text{tr} \nabla^2 f(y)| \leq d \min\{L, L_\beta \|x - y\|^\gamma\} \leq d \min\{L, L_\beta D^\gamma\}.$$

Proof. (a) For $x \in Q$ and $\|v\| \leq h$, Taylor's formula with integral remainder and the Hölder condition give

$$|f(x+v) - f(x) - \langle \nabla f(x), v \rangle - \frac{1}{2} \nabla^2 f(x)[v, v]| \leq L_\beta \|v\|^{2+\gamma} \int_0^1 (1-t) t^\gamma dt = c_\gamma L_\beta \|v\|^\beta.$$

Averaging over $v = hW$, with $\mathbb{E}W = 0$, $\mathbb{E}WW^\top = I/(d+2)$ and $\mathbb{E}\|W\|^\beta = d/(d+\beta)$, proves the identity. The trace bound follows from $|\text{tr} A| \leq d \|A\|_{\text{op}}$.

(b) By Lemma B.4, whose proof is local, f_h is continuously differentiable near Q and $\nabla f_h(x) = \mathbb{E} \nabla f(x + hW)$ for $x \in Q$. Since $x + hW \in Q_h$ for $x \in Q$, averaging preserves convexity and the Lipschitz constant of the gradient on Q . Averaging $0 \leq f(x+v) - f(x) - \langle \nabla f(x), v \rangle \leq L \|v\|^2/2$ over $v = hW$, with $\mathbb{E}\|W\|^2 = d/(d+2)$, gives the value bounds. When the Hessian exists, its eigenvalues lie in $[0, L]$ on Q , which proves the trace bounds.

(c) The difference $(f(x) - f(y)) - (f_h(x) - f_h(y))$ equals $(f_h - f)(y) - (f_h - f)(x)$. Part (b) bounds it by $Lh^2/2$, and part (a) by $h^2 d L_\beta D^\gamma / (2(d+2)) + 2c_\gamma L_\beta h^\beta$. The last claim combines the trace bounds of (a) and (b). $\square$

Lemma L.8 (Central differences for objectives with a Lipschitz gradient). *Let $K \subset \mathbb{R}^d$ be nonempty, closed and convex, let $h > 0$, let f be differentiable on an open set containing K_h with $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$ for $x, y \in K_h$, and let the noise be admissible at levels (M, σ, δ) . Consider a round of Procedure 1 with $m = 1$, an $\mathcal{F}_{k-1}$ -measurable center z with values in K , the deterministic radius h and B frames. Then the queries lie in K_h ,*

$$\bar{g} = \nabla f_h(z) + \nabla \eta_h(z) + \zeta + \Delta, \quad \|\nabla \eta_h(z)\| \leq M, \quad \|\Delta\| \leq \frac{\sqrt{d} \delta}{h} \text{ a.s.},$$

and $\mathbb{E}[\zeta \mid \mathcal{F}_{k-1}] = 0$. For $d \geq 2$,

$$\mathbb{E}[\|\zeta\|^2 \mid \mathcal{F}_{k-1}] \leq \frac{1}{B} \left(\frac{d}{d-1} (Lh + M)^2 + C_\sigma \frac{d\sigma^2}{h^2} \right), \quad C_\sigma = (1 + 2^{-1/2})^2; \quad (\text{L.16})$$

for $d = 1$ the term with $Lh + M$ is absent. Part (iv) of Lemma D.3 holds with $\nabla f(z) + b(z)$ replaced by $\nabla f_h(z) + \nabla \eta_h(z)$ and with $S_1 = 1$. If moreover $\beta \in (2, 3]$, f is twice continuously differentiable on an open set containing K_h and $\|\nabla^2 f(x) - \nabla^2 f(y)\|_{\text{op}} \leq L_\beta \|x - y\|^\gamma$ for $x, y \in K_h$, then ζ is the vector of Lemma L.2 with $m = 1$ and the operator norm, and (L.2) holds for it as well.

Proof. The mean, the stochastic-noise contribution, the adversarial bound and the adapted decomposition are those of Lemma D.3 with $m = 1$; they use only continuity and oddness on $z + hB(1)$. It remains to bound the frame variance without a bound on $\nabla f(z)$.

Put $e_f(w) = (f(z + hw) - f(z - hw))/(2h) - \langle \nabla f(z), w \rangle$. On the unit ball, $\nabla e_f(w) = (\nabla f(z + hw) + \nabla f(z - hw))/2 - \nabla f(z)$ has norm at most Lh , so e_f is odd and Lh -Lipschitz there. Composition with the odd, nonexpansive projection onto that ball extends it to an odd Lh -Lipschitz function on $\mathbb{R}^d$. The Lipschitz-noise term $e_\eta(w) = (\eta(z + hw) - \eta(z - hw))/(2h)$ is odd and M -Lipschitz, so $e_f + e_\eta$ has Lipschitz constant $Lh + M$. By (B.5), the centered second moment of one frame is at most $d(Lh + M)^2/(d - 1)$. Independent frames divide it by B ; for $d = 1$ it is zero. Adding the orthogonal noise term proves (L.16). For the final claim, both constructions define $\zeta = \bar{g} - \nabla f_h(z) - \nabla \eta_h(z) - \Delta$; apply Lemma L.2. $\square$

Theorem L.9 (Value analysis). *Let Q , x_0 , R and D be as in Lemma L.1, and let $h > 0$. Let f be differentiable on an open set containing Q_h and convex on Q_h , with $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for $x, y \in Q_h$, and let f have a minimizer $u \in \arg \min_Q f$ such that $\|x_0 - u\| \leq R$. Every $f \in \mathcal{F}_{\text{cvx}}^{Q_h}(\beta, L_\beta; L)$, $\beta > 2$, with such a minimizer has these properties. Let the noise be admissible at levels (M, σ, δ) . The method is projected stochastic gradient descent on Q driven by central differences:*

- $z_1 = x_0$;
- in round $k = 1, \dots, N$, run Procedure 1 with $m = 1$ at the center z_k , with radius h and B frames ($2dB$ calls), obtain $\bar{g}_k$, and set $z_{k+1} = \text{Proj}_Q(z_k - \nu \bar{g}_k)$;
- output $\bar{z}_N = \frac{1}{N} \sum_{k=2}^{N+1} z_k$.

All queries lie in Q_h , and $\bar{z}_N \in Q$.

(a) (Error bound.) For all $N, B \in \mathbb{N}$ and $0 < \nu \leq 1/(2L)$,

$$\mathbb{E}[f(\bar{z}_N) - \min_Q f] \leq \frac{R^2}{2\nu N} + \frac{\nu v_h^2}{B} + D\left(M + \frac{\sqrt{d}\delta}{h}\right) + S(h), \quad (\text{L.17})$$

where

$$v_h^2 = 4L^2h^2 + 4M^2 + C_\sigma d\sigma^2/h^2, \quad S(h) = Lh^2/2.$$

If moreover $\beta \in (2, 3]$ and f is twice continuously differentiable on an open set containing Q_h , with $\|\nabla^2 f(x) - \nabla^2 f(y)\|_{\text{op}} \leq L_\beta \|x - y\|^\gamma$ for $x, y \in Q_h$, where $\gamma = \beta - 2$, then (L.17) also holds with

$$v_h^2 = \min\{4L^2h^2, 4L_\beta^2h^{2\beta-2}\} + 4M^2 + C_\sigma d\sigma^2/h^2, \quad S(h) = \min\left\{\frac{Lh^2}{2}, \frac{L_\beta D^\gamma h^2}{2} + 2c_\gamma L_\beta h^\beta\right\}.$$

(b) (Curvature branch, every β .) Let $\varepsilon > 0$, and take

$$h = \sqrt{\frac{\varepsilon}{8L}}, \quad \bar{M} = \frac{\varepsilon}{8D}, \quad \bar{v}_h^2 = 4L^2h^2 + 4\bar{M}^2 + C_\sigma \frac{d\sigma^2}{h^2},$$

$$N = \left\lceil \frac{8LR^2}{\varepsilon} \right\rceil, \quad T_0 = \frac{256dR^2\bar{v}_h^2}{9\varepsilon^2}, \quad B = \left\lceil \frac{T_0}{2dN} \right\rceil, \quad \nu = \min\left\{\frac{1}{2L}, \frac{R\sqrt{B}}{\bar{v}_h\sqrt{2N}}\right\}.$$

If $M \leq \bar{M}$ and $\delta \leq \varepsilon h/(8D\sqrt{d})$, then $\mathbb{E}[f(\bar{z}_N) - \min_Q f] \leq \varepsilon$, with N rounds and

$$T = 2dNB \leq \frac{2048C_\sigma}{9} \frac{d^2\sigma^2LR^2}{\varepsilon^3} + \frac{272}{9} \frac{dLR^2}{\varepsilon} + \left(2 + \frac{16R^2}{9D^2}\right)d \quad (\text{L.18})$$

calls. This method uses $L, R, D, \sigma, d, \varepsilon$, but neither L_β nor β .

(c) (Hölder branch, $\beta \in (2, 3]$.) Suppose moreover that f satisfies the Hölder condition of (a) on Q_h and that $0 < \varepsilon \leq 8L_\beta D^\beta$. Take

$$h = \sqrt{\frac{\varepsilon}{8L_\beta D^\gamma}}, \quad \bar{M} = \frac{\varepsilon}{8D}, \quad \bar{v}_h^2 = 4L_\beta^2h^{2\beta-2} + 4\bar{M}^2 + C_\sigma \frac{d\sigma^2}{h^2},$$

and define N, T_0, B, ν as in (b), with this $\bar{v}_h$. Then $h \leq D$, and the conclusion of (b) holds with

$$T \leq \frac{2048 C_\sigma}{9} \frac{d^2 \sigma^2}{\varepsilon^2} \cdot \frac{L_\beta D^\gamma R^2}{\varepsilon} + \frac{128}{9} \left(\frac{h}{D} \right)^{2\gamma} \frac{d L_\beta D^\gamma R^2}{\varepsilon} + \frac{16 d L R^2}{\varepsilon} + \left(2 + \frac{16 R^2}{9 D^2} \right) d. \quad (\text{L.19})$$

(d) (Acceleration.) In (b) and (c), replace the update by the method of Lemma L.4 with $\varphi = f_h$ and $v_0 = x_0$; it queries at the centers z_k and outputs $x_N \in Q$. Keep h and $\bar{v}_h$, and take

$$N = \lceil \sqrt{32 L R^2 / \varepsilon} \rceil, \quad T_0 = \frac{512 d R^2 \bar{v}_h^2}{9 \varepsilon^2}, \quad B = \left\lceil \frac{T_0}{2 d N} \right\rceil, \quad \nu_0 = \min \left\{ \frac{1}{4 L}, \frac{R \sqrt{B}}{\bar{v}_h N^{3/2}} \right\}.$$

Under the same conditions, $\mathbb{E}[f(x_N) - \min_Q f] \leq \varepsilon$ with $T = 2 d N B \leq T_0 + 2 d N$ calls. In the curvature branch this means

$$T \leq \frac{4096 C_\sigma}{9} \frac{d^2 \sigma^2 L R^2}{\varepsilon^3} + \frac{256}{9} \frac{d L R^2}{\varepsilon} + \frac{32 R^2}{9 D^2} d + 2 d \lceil \sqrt{32 L R^2 / \varepsilon} \rceil,$$

and in the Hölder branch

$$T \leq \frac{4096 C_\sigma}{9} \frac{d^2 \sigma^2}{\varepsilon^2} \cdot \frac{L_\beta D^\gamma R^2}{\varepsilon} + \frac{256}{9} \left(\frac{h}{D} \right)^{2\gamma} \frac{d L_\beta D^\gamma R^2}{\varepsilon} + \frac{32 R^2}{9 D^2} d + 2 d \lceil \sqrt{32 L R^2 / \varepsilon} \rceil.$$

Proof. (a) By Lemma L.8 with $K = Q$, in round k

$$\bar{g}_k = \nabla f_h(z_k) + b_k + \zeta_k, \quad b_k = \nabla \eta_h(z_k) + \Delta_k, \quad \|b_k\| \leq M + \frac{\sqrt{d} \delta}{h},$$

where $\bar{g}_k$ is $\mathcal{F}_k$ -measurable and $\mathbb{E}[\zeta_k | \mathcal{F}_{k-1}] = 0$. Since $\frac{d}{d-1}(Lh + M)^2 \leq 2(Lh + M)^2 \leq 4L^2 h^2 + 4M^2$ for $d \geq 2$, (L.16) gives $\mathbb{E}[\|\zeta_k\|^2 | \mathcal{F}_{k-1}] \leq v_h^2 / B$. In the Hölder case, (L.2) with $m = 1$, $\tau = 0$ and $C_f = C_\eta = 4$ applies to the same ζ_k by the last claim of Lemma L.8, so the smaller of the two bounds holds. By Lemma L.7(b), $\varphi = f_h$ satisfies the setting of Section L.1, and Lemma L.3 with the comparator u gives

$$\mathbb{E}[f_h(\bar{z}_N) - f_h(u)] \leq \frac{R^2}{2 \nu N} + \frac{\nu v_h^2}{B} + D \left(M + \frac{\sqrt{d} \delta}{h} \right).$$

Both $\bar{z}_N$ and u lie in Q , so (L.14) or (L.15) gives $f(\bar{z}_N) - f(u) \leq f_h(\bar{z}_N) - f_h(u) + S(h)$, and $f(u) = \min_Q f$. This proves (L.17).

(b) Here $S(h) = Lh^2/2 = \varepsilon/16$ and $v_h^2 \leq \bar{v}_h^2$. As in (L.7), the choice of ν gives $R^2/(2 \nu N) + \nu v_h^2/B \leq L R^2/N + \sqrt{2 R \bar{v}_h}/\sqrt{N B}$. In this bound $L R^2/N \leq \varepsilon/8$, and $N B \geq T_0/(2d) = 128 R^2 \bar{v}_h^2/(9 \varepsilon^2)$ gives $\sqrt{2 R \bar{v}_h}/\sqrt{N B} \leq 3 \varepsilon/8$. The conditions on M and δ give $D(M + \sqrt{d} \delta/h) \leq \varepsilon/4$, so the error is at most $13 \varepsilon/16$. For the calls, $B \leq T_0/(2 d N) + 1$ and $N \leq 8 L R^2/\varepsilon + 1$ give $T \leq T_0 + 16 d L R^2/\varepsilon + 2d$. With $h^2 = \varepsilon/(8L)$ and $\bar{M} = \varepsilon/(8D)$, the parts $4L^2 h^2 = L \varepsilon/2$, $4 \bar{M}^2 = \varepsilon^2/(16 D^2)$ and $C_\sigma d \sigma^2/h^2 = 8 C_\sigma d \sigma^2 L/\varepsilon$ of $\bar{v}_h^2$, multiplied by $256 d R^2/(9 \varepsilon^2)$, give (L.18).

(c) Now $L_\beta D^\gamma h^2/2 = \varepsilon/16$ and $(h/D)^2 = \varepsilon/(8 L_\beta D^\beta) \leq 1$. Since $\beta/2 = 1 + \gamma/2$,

$$L_\beta h^\beta = \frac{\varepsilon}{8} \left(\frac{\varepsilon}{8 L_\beta D^\beta} \right)^{\gamma/2} \leq \frac{\varepsilon}{8},$$

and $2c_\gamma \leq 1$ gives $S(h) \leq 3 \varepsilon/16$. The rest of the argument of (b) applies, and the error is at most $15 \varepsilon/16$. For the calls, insert

$$L_\beta^2 h^{2\beta-2} = \frac{\varepsilon}{8} L_\beta D^\gamma \left(\frac{h}{D} \right)^{2\gamma}, \quad C_\sigma \frac{d \sigma^2}{h^2} = \frac{8 C_\sigma d \sigma^2 L_\beta D^\gamma}{\varepsilon}$$

and $4 \bar{M}^2 = \varepsilon^2/(16 D^2)$ into T_0 ; this gives (L.19).

(d) Replace Lemma L.3 by Lemma L.4 for $\varphi = f_h$. By (L.11) with $\bar{v}_B = \bar{v}_h/\sqrt{B}$, the optimization error is at most $4 L R^2/N^2 + 2 R \bar{v}_h/\sqrt{N B} \leq \varepsilon/8 + 3 \varepsilon/8$, since $N B \geq 256 R^2 \bar{v}_h^2/(9 \varepsilon^2)$. The other terms are as in (b) and (c), and the transfer inequalities apply because $x_N \in Q$. The contributions to T_0 double, and $2 d N = 2 d \lceil \sqrt{32 L R^2 / \varepsilon} \rceil$. $\square$

L.4 Explicit call constants on a Euclidean ball

Proof of the upper bounds of Section 6. The gradient-scale bound follows from Lemma L.6 and the value-scale bound from Lemma L.9, applied with $Q = B(R)$, $x_0 = 0$ and $D = 2R$. We simplify the constants for a general Q , using $R \leq D$ and $\lceil y \rceil \leq y + 1$, and write $x = LR^2/\varepsilon$.

Gradient scale. Lemma L.6 uses $\lceil 8x \rceil$ rounds, or $\lceil \sqrt{32x} \rceil$ rounds in the accelerated form; in both forms $2mdN \leq 16md(1+x)$, because $\sqrt{32x} \leq 4 + 2x$. Since $R^2(L_\beta D/\varepsilon)^{2/(\beta-1)} = (L_\beta R^{\beta-1} D/\varepsilon)^{2/(\beta-1)}$, the bound (L.12) with $\theta \leq 2$ gives

$$T \leq T_0 + 2mdN \leq C_\beta \frac{\sigma^2}{\varepsilon^2} d^{2-\frac{2}{\beta-1}} \left(\frac{L_\beta R^{\beta-1} D}{\varepsilon} \right)^{\frac{2}{\beta-1}} + C_\beta md(1+x)$$

with

$$C_\beta = \max \left\{ \frac{512 m C_\sigma (4C_b)^{\frac{2}{\beta-1}}}{9}, \frac{32C_f}{9C_b^2} + \frac{8C_\eta}{9A_1^2} + 16 \right\},$$

which depends only on β . For $D = 2R$, $L_\beta R^{\beta-1} D = 2L_\beta R^\beta$, and the radius and the noise conditions of Lemma L.6 take the form stated in Section 6.

Value scale. Let $V = L$, or, for $\beta \in (2, 3]$ and the operator norm, $V = \min\{L, L_\beta D^{\beta-2}\}$. If $V = L$ we use Lemma L.9(b). If $V = L_\beta D^{\beta-2} < L$ we use Lemma L.9(c), which requires $\varepsilon \leq 8L_\beta D^\beta$. Part (d) of that theorem gives the accelerated forms. In all cases the radius is $h = \sqrt{\varepsilon/(8V)}$ and the conditions on the noise are $M \leq \varepsilon/(8D)$ and $\delta \leq \varepsilon h/(8D\sqrt{d})$. In the Hölder branch $h \leq D$ and $V < L$, so the second term of (L.19) is at most $\frac{128}{9}dx$. Hence, in both branches,

$$T \leq \frac{2048 C_\sigma}{9} \frac{d^2 \sigma^2}{\varepsilon^2} \frac{VR^2}{\varepsilon} + \frac{272}{9} dx + \frac{34}{9} d,$$

and for the accelerated form

$$T \leq \frac{4096 C_\sigma}{9} \frac{d^2 \sigma^2}{\varepsilon^2} \frac{VR^2}{\varepsilon} + \frac{256}{9} dx + \frac{32}{9} d + 2d(\sqrt{32x} + 1).$$

Both are at most

$$\frac{2^{12} C_\sigma}{9} \frac{d^2 \sigma^2}{\varepsilon^2} \frac{VR^2}{\varepsilon} + 32d(1+x);$$

for the second use $8\sqrt{2}\sqrt{x} \leq 32x/9 + 9$. □

L.5 Matching bounds on the unconstrained class

In this subsection and the next, $Q = B(R)$, $x_0 = 0$ and $D = 2R$, so that the problems of Lemma L.1 are those of the class $\mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$ and $\min_Q f = f^*$.

Lower bounds. Theorem 5(a), proved in Section M, gives constants $c_\beta, c'_\beta > 0$, depending only on β , with the following property in the operator geometry. Let $d \geq 256$, let the noise be independent $N(0, \sigma^2)$, which is admissible at levels $(0, \sigma, 0)$, let $L \in (0, \infty]$ and $0 < \varepsilon \leq c'_\beta \min\{L_\beta R^\beta, LR^2\}$. Then every algorithm with $\mathbb{E}[f(\hat{x}) - f^*] \leq \varepsilon$ for all $f \in \mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$ makes

$$T \geq c_\beta \frac{\sigma^2}{\varepsilon^2} \min \left\{ d^{5/2}, d^2 \left(\frac{L_\beta R^\beta}{\varepsilon} \right)^{2/\beta}, \frac{d^2 LR^2}{\varepsilon} \right\} \quad (\text{L.20})$$

calls. With $L_\varepsilon = C_{\text{cap}} L_\beta^{2/\beta} \varepsilon^{1-2/\beta}$ as in its proof, the last term is the smallest one if and only if $L \leq L_\varepsilon$ and $LR^2/\varepsilon \leq C_{\text{cap}} \sqrt{d}$.

Corollary L.10 (Curvature matched to the hard instances). *Consider the operator geometry, let $c_\beta, c'_\beta, C_{\text{cap}}$ be as in (L.20), let $d \geq 256$ and $L, \sigma > 0$, and for $M \leq \varepsilon/(16R)$ and $\delta \leq \varepsilon^{3/2}/(16R\sqrt{8dL})$ let $T^*(\varepsilon; L)$ be the smallest number of calls with which some algorithm achieves $\mathbb{E}[f(\hat{x}) - f^*] \leq \varepsilon$ for all $f \in \mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$ and all noise admissible at levels (M, σ, δ) .*

(a) If

$$\max\left\{\frac{LR^2}{C_{\text{cap}}\sqrt{d}}, \left(\frac{L}{C_{\text{cap}}L_\beta^{2/\beta}}\right)^{\frac{\beta}{\beta-2}}\right\} \leq \varepsilon \leq \min\{c'_\beta L_\beta R^\beta, c'_\beta LR^2, \sigma\sqrt{d}\}, \quad (\text{L.21})$$

then

$$\frac{c_\beta}{C_{\text{cap}}} \frac{d^2 \sigma^2 LR^2}{\varepsilon^3} \leq T^*(\varepsilon; L) \leq \frac{2048 C_\sigma + 294}{9} \frac{d^2 \sigma^2 LR^2}{\varepsilon^3},$$

with $C_\sigma = (1 + 2^{-1/2})^2$. The upper bound is attained by Lemma L.9(b) with $\lceil 8LR^2/\varepsilon \rceil$ rounds, and with twice the constant by its accelerated form (d) with $\lceil \sqrt{32LR^2/\varepsilon} \rceil$ rounds; neither uses β or L_β .

(b) If $L \geq L_\varepsilon$ and $L_\beta R^\beta d^{-\beta/4} \leq \varepsilon \leq c'_\beta L_\beta R^\beta$, then

$$c_\beta \frac{d^2 \sigma^2}{\varepsilon^2} \left(\frac{L_\beta R^\beta}{\varepsilon}\right)^{2/\beta} \leq T^*(\varepsilon; L) \leq \frac{2^{12} C_\sigma C_{\text{cap}}}{9} \frac{V}{L_\varepsilon} \frac{d^2 \sigma^2}{\varepsilon^2} \left(\frac{L_\beta R^\beta}{\varepsilon}\right)^{2/\beta} + 32d \left(1 + \frac{LR^2}{\varepsilon}\right),$$

where $V = L$, or $V = \min\{L, L_\beta(2R)^{\beta-2}\}$ if $2 < \beta \leq 3$ and $\varepsilon \leq 8L_\beta(2R)^\beta$.

The window (L.21) is nonempty if and only if its left end is at most its right end; this requires $LR^2 \leq c L_\beta R^\beta$ and $d \geq c''$ for numbers c, c'' depending only on β .

Proof. (a) The lower end of the window makes the last term of (L.20) the smallest one, which gives the lower bound, since Gaussian noise is admissible at every level. For the upper bound apply Lemma L.9(b) on $Q = B(R)$, $D = 2R$, with $h = \sqrt{\varepsilon/(8L)}$; its admissible levels are those defining $T^*(\varepsilon; L)$, and (L.18) gives $T \leq \frac{2048 C_\sigma}{9} \frac{d^2 \sigma^2 LR^2}{\varepsilon^3} + \frac{272}{9} \frac{dLR^2}{\varepsilon} + \frac{22}{9}d$. Since $\varepsilon \leq \sigma\sqrt{d}$ and $LR^2/\varepsilon \geq 1/c'_\beta \geq 1$, the last two terms are at most $\frac{294}{9}d^2 \sigma^2 LR^2/\varepsilon^3$. For acceleration use $\lceil \sqrt{32x} \rceil \leq 16x + 3/2$ with $x = LR^2/\varepsilon$ in part (d) of that theorem.

(b) Since $L \geq L_\varepsilon$, the proof of Theorem 5(a) applies with $\ell = L_\beta$ whenever $\varepsilon \leq c'_\beta L_\beta R^\beta$, and the lower end of the range makes the second term of (L.20) the smallest one. The upper bound is the value-scale bound of Section L.4 with $d^2 \sigma^2 VR^2/\varepsilon^3 = (V/L_\varepsilon) C_{\text{cap}}(d^2 \sigma^2/\varepsilon^2)(L_\beta R^\beta/\varepsilon)^{2/\beta}$. $\square$

L.6 Admissible level of Lipschitz noise

Corollary L.11 (Admissible level of Lipschitz noise). *Let $\beta > 2$, in either geometry, let $\varepsilon > 0$ and $\sigma \geq 0$, and let $M_{\text{max}}(\varepsilon, T)$ be the admissible level of Section 2 on $\mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$. Then $M_{\text{max}}(\varepsilon, T) \geq \varepsilon/(16R)$ for every*

$$T \geq \frac{2048 C_\sigma}{9} \frac{d^2 \sigma^2 LR^2}{\varepsilon^3} + \frac{272}{9} \frac{dLR^2}{\varepsilon} + \frac{22}{9}d.$$

If $\varepsilon < LR^2/2$, then also $M_{\text{max}}(\varepsilon, T) \leq 2\varepsilon/R$ for every T (Lemma E.7). The constants $1/16$ and 2 do not depend on d, T, σ, β or the geometry.

Proof. Apply Lemma L.9(b) with $Q = B(R)$, $x_0 = 0$, $D = 2R$ and $\delta = 0$. It applies to every $f \in \mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$, for every β and every $\varepsilon > 0$, with $\bar{M} = \varepsilon/(8D) = \varepsilon/(16R)$, and for $D = 2R$ the right-hand side of (L.18) is the stated number of calls. The upper bound is the first consequence listed in Lemma E.7, with $p = 2$. $\square$

M THE SMOOTH CONVEX LOWER BOUND

This appendix proves Theorem 5(a). It uses the hidden-map construction of Rajaraman and Han (2026), whose hard functions are convex and Lipschitz but not differentiable; theorem and lemma numbers of that paper refer to its version 3. Section M.1 smooths the family: the Euclidean norm in the hidden tube is replaced by $\sqrt{1 + \|\cdot\|^2} - 1$, and the hinge that joins the function to a conic branch outside the unit ball (Rajaraman and Han, 2026, Section 4.2) is replaced by a hinge of class C^{q+2} . Section M.2 carries the information argument of Rajaraman and Han (2026) over to this family with explicit constants, and Section M.3 rescales the family to the convex classes.

M.1 A smooth hidden-map family

Throughout this appendix $d = 2d_0$, $\beta > 2$, $q = \lceil \beta \rceil - 1$ and $\gamma = \beta - q$. We write $B_2^n(R)$ for the closed Euclidean ball of radius R in $\mathbb{R}^n$ and $B_2^n = B_2^n(1)$. Norms of derivative tensors are multilinear operator norms. The symbols W , u , π , M_j , S_k , E_k , K_β^{tube} and U_j are local to this appendix: W is a matrix, and the constants are unrelated to the Hessian bound L and to the stencil constants S_1, S_2 .

The tube. For $0 < \vartheta \leq 1/2$, $r \geq 16\vartheta$, a $d_0 \times d_0$ matrix W , a vector $u \in \mathbb{R}^{d_0}$ and $z = (x, y)$ with $x, y \in \mathbb{R}^{d_0}$, set

$$\begin{aligned} B_W(x, y) &= Wx/r - y/(8\vartheta), & a_W(x, y) &= \sqrt{1 + \|B_W(x, y)\|^2} - 1, \\ b_u(x, y) &= \frac{1}{2} \|x\|^2 - \langle x, u \rangle, & f_{W,u}(x, y) &= \vartheta \{\log(e^{a_W(x, y)} + e^{b_u(x, y)}) - \log 2\}. \end{aligned} \quad (\text{M.1})$$

With $\|B_W\|$ in place of a_W and $\varepsilon = \vartheta$, this is the hard function of [Rajaraman and Han \(2026, Section 2\)](#). The bounded parameter set is $E_b = \{\|W\|_{\text{op}} \leq 4, \|u\| \leq 3/8\}$.

Lemma M.1 (Geometry of the smooth tube). *The function $f = f_{W,u}$ is smooth and convex on $\mathbb{R}^{2d_0}$, and $f(0) = 0$. It has the global minimizer $z_* = (u, (8\vartheta/r)Wu)$, and for every $z = (x, y)$,*

$$f(z) - f(z_*) \geq \frac{\vartheta}{2(1 + \exp(\|u\|^2/2))} \|x - u\|^2.$$

On E_b , $\|z_*\| \leq \rho = 3\sqrt{5}/8$ and f is 1-Lipschitz on the unit ball.

Proof. The proof follows that of [Rajaraman and Han \(2026, Lemmas A.1 and A.2\)](#) for the tube $\|B_W\|$. Both arguments of the log-sum-exp are convex, and log-sum-exp is convex and coordinatewise increasing; the minima 0 of a_W and $- \|u\|^2/2$ of b_u are both attained at z_* . Since $a_W \geq 0$, $b_u(z) + \|u\|^2/2 = \|x - u\|^2/2$, and $h(t) = \log(1 + e^t)$ satisfies $h'(t) \geq (1 + e^{\|u\|^2/2})^{-1}$ for $t \geq -\|u\|^2/2$, integrating h' gives the error bound. On E_b , ∇f is a convex combination of $\vartheta \nabla a_W(z) = \vartheta B_W^\top B_W(z)/(1 + \|B_W(z)\|^2)^{1/2}$, where $B_W^\top$ is the adjoint of the linear map B_W , and of $\vartheta(x - u, 0)$. Their norms are at most $\vartheta \|B_W\|_{\text{op}} \leq (16\vartheta^2/r^2 + 1/64)^{1/2} \leq \sqrt{5}/8$ and, on the unit ball, $\vartheta(1 + 3/8) \leq 11/16$. Finally $\|(8\vartheta/r)Wu\| \leq 2\|u\|$, so $\|z_*\| \leq \sqrt{5}\|u\| \leq \rho$. $\square$

A smooth hinge. Let $n = q + 1$ and $\lambda_0 = 1/64$, and define

$$\varphi_n(t) = \frac{(2n+1)!}{(n!)^2} t^n (1-t)^n \mathbf{1}_{[0,1]}(t), \quad H_n(t) = \int_{-\infty}^t \varphi_n(v) dv, \quad \Psi(s) = \lambda_0 \int_{-\infty}^{(s+\lambda_0)/\lambda_0} H_n(t) dt.$$

This hinge replaces the C^2 hinge of [Rajaraman and Han \(2026, Appendix C\)](#). It satisfies $\Psi \in C^{n+1} = C^{q+2}$, $0 \leq \Psi' \leq 1$, $\Psi'' \geq 0$ and $\Psi(s) = 0$ for $s \leq -\lambda_0$. Symmetry of φ_n gives $\Psi(s) = s + \lambda_0/2$ for $s \geq 0$, and convexity gives $\Psi(s) \geq s + \lambda_0/2$ for every s . Put

$$V_n = \max_t \varphi_n(t) = \frac{(2n+1)!}{(n!)^2 4^n}, \quad \vartheta_\beta = \min\{1/2, 1/V_n\}.$$

Since $\Psi''(s) = \varphi_n((s+\lambda_0)/\lambda_0)/\lambda_0$, we have $\lambda_0 \vartheta \|\Psi''\|_\infty \leq 1$ whenever $\vartheta \leq \vartheta_\beta$. Put $U_1 = 1$ and, for $2 \leq j \leq q+1$, $U_j = \lambda_0^{1-j} \left\| \varphi_n^{(j-2)} \right\|_{\infty, [0,1]}$; these are maxima of fixed polynomials, and $|\Psi^{(j)}| \leq U_j$ for $1 \leq j \leq q+1$.

Global extension. Set $\Gamma(z) = \lambda_0 + (\|z\| - 1)/2$ and define

$$g_{W,u}(z) = \begin{cases} \Gamma(z) + \Psi(\lambda_0 f_{W,u}(z) - \Gamma(z) - 2\lambda_0), & \|z\| < 1, \\ \Gamma(z), & \|z\| \geq 1. \end{cases} \quad (\text{M.2})$$

For arbitrary (W, u) , the same formula defines the mean $\mu_{W,u}$ of the auxiliary experiment in [Section M.2](#).

Lemma M.2 (Smooth extension and global decoding). *For $(W, u) \in E_b$, g is convex, globally $1/2$ -Lipschitz, and C^{q+1} on $\mathbb{R}^{2d_0}$. It equals Γ whenever $\|z\| \geq 30/31$, and z_* is a global minimizer. Let $\Pi(z) = z/(1 \vee \|z\|)$ be the radial projection and $\Pi^1(z) \in \mathbb{R}^{d_0}$ its x -component. Then*

$$\begin{aligned} g(z) - g(z_*) &\geq \lambda_0 \{f(\Pi(z)) - f(z_*)\} \\ &\geq \frac{\lambda_0 \vartheta}{2(1 + e^{9/128})} \|\Pi^1(z) - u\|^2. \end{aligned} \quad (\text{M.3})$$

Proof. Apart from the smoothness, these statements are [Rajaraman and Han \(2026, Lemma 4.3\)](#) together with its proof in [Rajaraman and Han \(2026, Appendix C.1\)](#), where the radius $30/31$, the Lipschitz constant $1/2$ and the first inequality in (M.3) are obtained for their tube and hinge. That proof uses only the following properties, which hold here by Lemma M.1 and the properties of Ψ listed above: f is convex and 1-Lipschitz on the unit ball, $f(0) = 0$, and z_* minimizes f with $\|z_*\| \leq \rho < 29/33$; Ψ is convex, $0 \leq \Psi' \leq 1$, $\Psi = 0$ on $(-\infty, -\lambda_0]$, $\Psi(s) = s + \lambda_0/2$ on $[0, \infty)$ and $\Psi(s) \geq s + \lambda_0/2$. In that proof g is written as $\Gamma + \Psi(\lambda_0 \tilde{f} - \Gamma - 2\lambda_0)$ with the convex 1-Lipschitz extension $\tilde{f}(z) = \inf_{v \in B_2^{2d_0}} \{f(v) + \|z - v\|\}$, which equals f on the unit ball. Since $\tilde{f}(z) \leq \|z\|$, the hinge argument is at most $(29 - 31\|z\|)/64$, hence at most $-\lambda_0$ for $\|z\| \geq 30/31$, where the hinge term therefore vanishes; so this function agrees with (M.2). The second inequality in (M.3) follows from Lemma M.1 and $\|u\|^2/2 \leq 9/128$.

It remains to prove smoothness. For $\|z\| \leq 1/4$, $f(z) \geq -\|z\|$, so the hinge argument is at least $(29 - 33\|z\|)/64 > 0$ and $g = \lambda_0 f - 3\lambda_0/2$; the nonsmooth norm in Γ cancels. On $0 < \|z\| < 1$ all ingredients are C^{q+1} , and $g = \Gamma$ on the collar $30/31 \leq \|z\| \leq 1$ and outside it. Hence $g \in C^{q+1}(\mathbb{R}^{2d_0})$. $\square$

Derivative constants. For $k \geq 1$, let $(1/2)_j$ denote the falling factorial and set

$$P_k = \sum_{a+2b=k} \frac{k! 2^a}{a! b!} |(1/2)_{a+b}|, \quad M_j = 2^{j-1} (j-1)!.$$

Let $\mathfrak{P}_k$ be the set of partitions of $\{1, \dots, k\}$, and define

$$\begin{aligned} C_k &= P_k + \sum_{\pi \in \mathfrak{P}_k} M_{|\pi|} \prod_{A \in \pi} (P_{|A|} + 2), \\ G_k &= \frac{1}{2} P_k 4^{k-1}, \quad S_k = \lambda_0 C_k + G_k, \\ E_k &= G_k + \lambda_0 C_k + \sum_{\pi \in \mathfrak{P}_k} U_{|\pi|} \prod_{A \in \pi} S_{|A|} \quad (1 \leq k \leq q+1), \\ K_\beta^{\text{tube}} &= (2E_q)^{1-\gamma} E_{q+1}^\gamma. \end{aligned}$$

Lemma M.3 (Global smoothness and Hessian cap). *For $0 < \vartheta \leq \vartheta_\beta$, $r \geq 16\vartheta$, and $(W, u) \in E_b$,*

$$\sup_z \|D^k g(z)\|_{\text{op}} \leq E_k \vartheta^{1-k} \quad (1 \leq k \leq q+1), \quad [g]_{\beta, \text{op}} \leq K_\beta^{\text{tube}} \vartheta^{1-\beta}, \quad 0 \preceq \nabla^2 g \preceq E_2 \vartheta^{-1} I.$$

The constants depend only on β .

Proof. We bound the derivatives of f on the unit ball, then those of g on the regions $\|z\| \leq 1/4$, $1/4 \leq \|z\| \leq 1$ and $\|z\| \geq 1$, and interpolate.

Step 1: derivatives of f . Faà di Bruno's formula for $h(v) = \sqrt{1 + \|v\|^2}$ has only blocks of size one or two. There are $k!/(a!b!2^b)$ partitions with a singletons and b pairs, and the sum of their absolute contributions is at most the corresponding summand of P_k times $(1 + \|v\|^2)^{(1-k)/2}$. Thus $\|D^k h(v)\|_{\text{op}} \leq P_k$, and, as $\|B_W\|_{\text{op}} \leq \vartheta^{-1}$, $\|D^k a_W\|_{\text{op}} \leq P_k \vartheta^{-k}$. On the unit ball the nonzero derivatives of b_u have norms at most 2. Write

$$f = \vartheta a_W + \vartheta h_0(b_u - a_W) - \vartheta \log 2, \quad h_0(t) = \log(1 + e^t).$$

Here $h'_0 = p$ with $p(t) = e^t/(1 + e^t) \in [0, 1]$, and each further derivative is obtained by applying $p(1-p) d/dp$ to the preceding polynomial in p . By induction, $h_0^{(j)}$ is a polynomial in p whose absolute coefficients sum to at most M_j . Faà di Bruno's formula for the composition and $\vartheta \leq 1$ give $\|D^k f\|_{\text{op}} \leq C_k \vartheta^{1-k}$ on the unit ball.

Step 2: derivatives of g . The same computation for $\|z\|$ shows $\|D^k \Gamma(z)\|_{\text{op}} \leq G_k$ for $\|z\| \geq 1/4$. On the annulus $1/4 \leq \|z\| \leq 1$ the derivatives of the hinge argument are at most $S_k \vartheta^{1-k}$. In Faà di Bruno's formula for Ψ composed with the hinge argument, the term of a partition π with $j = |\pi|$ blocks is at most $U_j \prod_{A \in \pi} S_{|A|}$ times $\vartheta^{j-k} \leq \vartheta^{1-k}$. This proves the bound $E_k \vartheta^{1-k}$ on the annulus. The same bound holds in the inner ball, where $g = \lambda_0 f - 3\lambda_0/2$, and outside the unit ball, where $g = \Gamma$.

Step 3: Hölder and Hessian bounds. With $t = \|z - z'\|$, the uniform bounds on $D^q g$ and $D^{q+1} g$ give

$$\|D^q g(z) - D^q g(z')\|_{\text{op}} \leq \min\{2E_q \vartheta^{1-q}, E_{q+1} \vartheta^{-q} t\} \leq K_\beta^{\text{tube}} \vartheta^{1-\beta} t^\gamma.$$

Convexity and the case $k = 2$ give the Hessian inequalities. $\square$

M.2 Information lower bound for the smooth tube

The next theorem is the estimation lower bound of [Rajaraman and Han \(2026, Theorem 4.2\)](#) for the smooth family, with explicit constants and with the amplitude and the noise level kept separate.

Theorem M.4 (Smooth-tube information lower bound). *Let $d_0 \geq 128$, $A, \sigma > 0$, $0 < \vartheta \leq \vartheta_\beta$, and $r \geq \max\{16\vartheta, d_0^{-1/4}\}$. Let π be the prior under which $\text{vec}(W) \sim N(0, I_{d_0^2}/d_0)$ and $u \sim N(0, I_{d_0}/(16d_0))$ are independent, and let $\pi_b = \pi(\cdot \mid E_b)$. The unknown function is $F_{W,u} = Ag_{W,u}$, and a query at $z_t \in \mathbb{R}^{2d_0}$ returns $F_{W,u}(z_t) + \sigma Z_t$, where $Z_1, Z_2, \dots$ are independent standard Gaussian variables, independent of (W, u) and of the seed of the algorithm. Queries and the output $\hat{z}$ are unrestricted, and the algorithm may know $A, \sigma, \vartheta, r, \beta$ and π . Put $\nu = A\lambda_0\vartheta/\sigma$. If*

$$T < \frac{1}{65536\nu^2} \min\{d_0^{5/2}, d_0^2/r^2\}, \quad (\text{M.4})$$

then every adaptive randomized algorithm with at most T queries has Bayes error $\mathbb{E}[F_{W,u}(\hat{z}) - \min F_{W,u}]$ under π_b greater than $A\lambda_0\vartheta/4096$. In particular, its expected error is greater than $A\lambda_0\vartheta/4096$ for some $(W, u) \in E_b$.

Proof. We follow the proof of the estimation bound of [Rajaraman and Han \(2026, Theorem 4.2\)](#) in Appendix C.3 there. It combines the reduction of their Lemma 2.7 with their Lemma C.1, which is the version of their Lemmas 2.1–2.3 for the auxiliary mean. Dividing the observations by σ gives their model with the mean multiplied by A/σ , so that ν plays the role of their $\lambda\varepsilon$. Apart from the decoding inequality (M.3), these proofs use the tube and the hinge only through the parameter derivatives of the mean (Step 2) and an anti-concentration bound for the weight of the second branch (Step 5). We verify both for the smooth family and keep track of the constants.

Write $\theta = (W, u)$. Let $\mathcal{H}_t$ be generated by the seed of the algorithm and the first t query–response pairs, so that z_t is $\mathcal{H}_{t-1}$ -measurable. For a density ρ let $J(\rho) = \int (\nabla \log \rho)(\nabla \log \rho)^\top \rho$. In the *auxiliary experiment* $\theta \sim \pi$, and a query at z_t returns $A\mu_\theta(z_t) + \sigma Z_t$; put $h_\theta = A\mu_\theta/\sigma$.

Step 1: reduction to the auxiliary experiment. For $g \sim N(0, I_{d_0})$ and $a > 1$ the Chernoff bound gives $\mathbb{P}(\|g\|^2 \geq ad_0) \leq \exp\{-d_0(a - 1 - \log a)/2\}$, which is at most $e^{-d_0/2}$ for $a = 4$ and at most $e^{-d_0/8}$ for $a = 9/4$. For a $d_0 \times d_0$ matrix G with independent standard Gaussian entries and a $1/4$ -net $\mathcal{N}$ of the unit sphere with $|\mathcal{N}| \leq 9^{d_0}$, we have $\|G\|_{\text{op}} \leq (4/3) \max_{v \in \mathcal{N}} \|Gv\|$; the case $a = 9$ and a union bound give $\mathbb{P}(\|G\|_{\text{op}} > 4\sqrt{d_0}) \leq e^{-d_0(4-3\log 3)} \leq e^{-d_0/2}$. For $W = d_0^{-1/2}G$ and $u = g/(4\sqrt{d_0})$ this yields

$$\mathbb{P}(E_b^c) \leq e^{-d_0/2} + e^{-d_0/8} \leq 2e^{-d_0/8}, \quad \mathbb{P}(\|u\| > 1/2) \leq e^{-d_0/2}.$$

Let $\hat{u} = \Pi^1(\hat{z})$, so that $\|\hat{u}\| \leq 1$. Since $\mathbb{E}\|u\|^4 = (1 + 2/d_0)/256$ and $d_0 \geq 128$, the Cauchy–Schwarz inequality gives

$$\mathbb{E}[\|\hat{u} - u\|^2 \mathbf{1}_{E_b^c}] \leq 2\mathbb{P}(E_b^c) + 2(\mathbb{E}\|u\|^4 \mathbb{P}(E_b^c))^{1/2} \leq 5e^{-d_0/16} \leq 1/512. \quad (\text{M.5})$$

On E_b , $\mu_\theta = g_\theta$ at every point of $\mathbb{R}^{2d_0}$, so, as in [Rajaraman and Han \(2026, Lemma 2.7\)](#), the law of the whole trajectory of the auxiliary experiment conditional on E_b is that of the actual experiment with prior π_b . With (M.5) this gives $\mathbb{E}_{\pi, \mu} \|\hat{u} - u\|^2 \leq \mathbb{E}_{\pi_b, g} \|\hat{u} - u\|^2 + 1/512$, where the subscripts indicate the prior and the mean function. On E_b , (M.3) and $e^{9/128} < 3$ give $F_\theta(z) - \min F_\theta \geq (A\lambda_0\vartheta/8) \|\Pi^1(z) - u\|^2$. Hence Bayes error at most $A\lambda_0\vartheta/4096$ under π_b would imply $\mathbb{E}_{\pi_b, g} \|\hat{u} - u\|^2 \leq 1/512$ and $\mathbb{E}_{\pi, \mu} \|\hat{u} - u\|^2 \leq 1/256$. It remains to show that under (M.4) no estimator has $\mathbb{E}\|\hat{u} - u\|^2 \leq 1/256$ in the auxiliary experiment; all expectations below refer to this experiment.

Step 2: parameter derivatives. At a query with $\|z_t\| \geq 1$, $\mu_\theta(z_t) = \Gamma(z_t)$ does not depend on θ . For $\|z_t\| < 1$ put

$$p_t = \frac{e^{b_u(z_t)}}{e^{a_W(z_t)} + e^{b_u(z_t)}}, \quad s_t = \lambda_0 f_\theta(z_t) - \Gamma(z_t) - 2\lambda_0.$$

As in [Rajaraman and Han \(2026, Appendix C.2\)](#), with $\varepsilon = \vartheta$ and $\lambda = \lambda_0$,

$$\begin{aligned}\nabla_u \mu_\theta(z_t) &= -\lambda_0 \vartheta \Psi'(s_t) p_t x_t, \\ \nabla_u^2 \mu_\theta(z_t) &= [\lambda_0 \vartheta \Psi'(s_t) p_t (1 - p_t) + \lambda_0^2 \vartheta^2 \Psi''(s_t) p_t^2] x_t x_t^\top, \\ \|\nabla_W \mu_\theta(z_t)\|_F &\leq \lambda_0 \vartheta / r;\end{aligned}\tag{M.6}$$

the last bound holds because $\partial f_\theta / \partial a_W \leq \vartheta$ and $\nabla_W a_W(z_t) = B_W(z_t) x_t^\top / (r(1 + \|B_W(z_t)\|^2)^{1/2})$ has Frobenius norm at most $\|x_t\| / r \leq 1/r$. Put $p_t = 0$ if $\|z_t\| \geq 1$, and

$$I_0 = 0, \quad I_T = \nu^2 \sum_{t=1}^T p_t^2 x_t x_t^\top, \quad \bar{I}_T = \mathbb{E}[I_T \mid \mathcal{H}_T].$$

Since $0 \leq \Psi' \leq 1$, $\lambda_0 \vartheta \|\Psi''\|_\infty \leq 1$ and $\|x_t\| \leq 1$, the bracket in (M.6) lies in $[0, \lambda_0 \vartheta p_t]$, and $B_t = \nabla_u^2 h_\theta(z_t)$ satisfies

$$(\nabla_u h_\theta)(\nabla_u h_\theta)^\top \preceq I_t - I_{t-1}, \quad B_t^2 \preceq I_t - I_{t-1}.\tag{M.7}$$

These bounds, the Gaussian prior and the at most linear growth of μ_θ in θ justify the differentiations under the integral sign and the integrations by parts below, first with compact cutoffs.

Step 3: posterior spread. Let $\rho_{T,u}$ be the marginal posterior density of u . Given (M.7), the proof of [Rajaraman and Han \(2026, Lemma 3.1\)](#) with $\eta = 1$ and $\delta = 1/4$ applies verbatim. It gives

$$J(\rho_{T,u}) \preceq 16d_0 I_{d_0} + \bar{I}_T - \bar{C}_T, \quad C_T = \sum_{t=1}^T Z_t B_t, \quad \bar{C}_T = \mathbb{E}[C_T \mid \mathcal{H}_T],\tag{M.8}$$

$$J(\rho_{T,u}) \preceq 17d_0 I_{d_0} + \frac{3}{2} \bar{I}_T \quad \text{with probability at least } 3/4.\tag{M.9}$$

Here $16d_0 I_{d_0}$ is the prior precision of u . For the second bound, that proof applies the matrix supermartingale inequality ([Tropp, 2011, Lemma 2.1](#)) to $X_t = -Z_t B_t$, using $\mathbb{E}[e^{-Z_t B_t} \mid \theta, \mathcal{H}_{t-1}, z_t] = e^{B_t^2/2}$ and the second bound in (M.7), and obtains $-\bar{C}_T \preceq \bar{I}_T/2 + \log(4d_0) I_{d_0}$ with probability at least $3/4$; then $\log(4d_0) \leq d_0$ gives (M.9).

Now suppose that $\mathbb{E} \|\hat{u} - u\|^2 \leq 1/256$. The posterior mean minimizes the conditional risk, so Markov's inequality gives $\text{tr Cov}(u \mid \mathcal{H}_T) \leq 1/64$ with probability at least $3/4$. Let S_T be the intersection of this event with the event in (M.9); then $\mathbb{P}(S_T) \geq 1/2$. The Cramér–Rao inequality $\text{Cov}(u \mid \mathcal{H}_T) \succeq J(\rho_{T,u})^{-1}$ (integration by parts gives $\mathbb{E}[(u - \mathbb{E}u)(\nabla \log \rho)^\top] = -I$, and a Schur complement concludes) shows that on S_T the trace of the inverse of $17d_0 I_{d_0} + \frac{3}{2} \bar{I}_T$ is at most $1/64$. If fewer than $d_0/2$ eigenvalues of $\bar{I}_T$ were at least d_0 , this trace would exceed $(d_0/2)/((37/2)d_0) = 1/37 > 1/64$. Thus on S_T at least $d_0/2$ eigenvalues of $\bar{I}_T$ are at least d_0 .

Step 4: information about W . Let $\rho_{t,W}$ be the marginal posterior density of W and $K_t = d_0^{-2} \text{Tr}_{\text{out}} J(\rho_{t,W})$, where Tr_{out} sums the d_0 diagonal blocks indexed by the output rows of W ([Rajaraman and Han, 2026, Section 2](#)). The proof of [Rajaraman and Han \(2026, Lemma 2.2\(1\)–\(2\)\)](#) in Section 3.2 there applies verbatim with $\|\nabla_W h_\theta\|_F \leq \nu/r$ from (M.6): $K_0 = I_{d_0}$, $K_t \preceq \mathbb{E}[K_{t+1} \mid \mathcal{H}_t]$, and

$$\mathbb{E} \text{tr } K_T \leq d_0 + \frac{\nu^2 T}{r^2 d_0^2}.\tag{M.10}$$

Step 5: anti-concentration. [Rajaraman and Han \(2026, Lemma 3.2\)](#) bounds small-ball probabilities by Fisher information with an unspecified constant; we use an explicit version. Let $Z \in \mathbb{R}^{d_0}$, $d_0 \geq 3$, have a differentiable density with scalar Fisher information $J_{\text{tr}} < \infty$. Integration by parts against $V_\delta(w) = (w - z)/(\|w - z\|^2 + \delta)$, whose divergence is at least $(d_0 - 2)/(\|w - z\|^2 + \delta)$, and the Cauchy–Schwarz inequality give $(d_0 - 2)\mathbb{E}(\|Z - z\|^2 + \delta)^{-1} \leq (J_{\text{tr}} \mathbb{E}(\|Z - z\|^2 + \delta)^{-1})^{1/2}$. Letting $\delta \downarrow 0$ gives Hardy's inequality $\mathbb{E} \|Z - z\|^{-2} \leq J_{\text{tr}}/(d_0 - 2)^2$, and Markov's inequality gives

$$\mathbb{P}(\|Z - z\| \leq s) \leq \frac{s^2 J_{\text{tr}}}{(d_0 - 2)^2} \leq \frac{9s^2 J_{\text{tr}}}{d_0^2}.\tag{M.11}$$

Fix a history and an in-ball query with $x_t \neq 0$ (the case $x_t = 0$ is trivial), and let $Z = W x_t$ with W distributed according to $\rho_{t-1,W}$. The argument in [Rajaraman and Han \(2026, Section 3.2\)](#) relates the scores of W and Z and gives

$$\frac{\text{tr } J(Z)}{d_0^2} x_t x_t^\top \preceq K_{t-1}, \quad \frac{\text{tr } J(Z)}{d_0^2} x_t^\top K_{t-1}^{-1} x_t \leq 1.\tag{M.12}$$

The matrix K_{t-1} is positive definite, since a positive differentiable density whose Fisher matrix has a null direction is constant along lines in that direction and hence not integrable.

Let $E = \{\|u\| \leq 1/2\}$ and $H_t = x_t^\top K_{t-1}^{-1} x_t$. On E , $b_u(z_t) \leq 1$, so $p_t^2 \mathbf{1}_E \leq \exp[-2(a_W(z_t) - 1)_+]$, which depends on θ only through W . This is the step that changes for the smooth tube: for $s \geq 0$,

$$\{a_W(z_t) \leq s\} = \{\|Wx_t - ry_t/(8\vartheta)\| \leq r\sqrt{s^2 + 2s}\},$$

whereas for the tube $\|B_W\|$ the radius is rs . Using $e^{-2(a-1)_+} = \int_1^\infty 2e^{-2(s-1)} \mathbf{1}_{\{a \leq s\}} ds$, (M.11) and (M.12), we obtain

$$H_t \mathbb{E}[p_t^2 \mathbf{1}_E \mid \mathcal{H}_{t-1}, z_t] \leq 9r^2 \int_1^\infty 2e^{-2(s-1)}(s^2 + 2s) ds = \frac{99}{2}r^2 \leq 50r^2. \quad (\text{M.13})$$

Step 6: projections and the contradiction. Let $\Lambda = 16(1 + \nu^2 T/(r^2 d_0^3))$. The construction of [Rajaraman and Han \(2026, Section 3.3\)](#) gives $\mathcal{H}_t$ -measurable orthogonal projections $I_{d_0} = Q_0 \succeq Q_1 \succeq \dots \succeq Q_T$ with $Q_t K_t Q_t \preceq \Lambda Q_t$: Q_t is obtained from Q_{t-1} by removing the eigendirections of $Q_{t-1} K_t Q_{t-1}$ with eigenvalues larger than Λ . The submartingale property and (M.10) give, as there,

$$\mathbb{E} \text{tr}(I_{d_0} - Q_T) \leq d_0/16. \quad (\text{M.14})$$

Moreover $K_t^{-1} \succeq Q_t/\Lambda$: the bound $\|K_t^{1/2} Q_t\|_{\text{op}}^2 \leq \Lambda$ gives $K_t^{1/2} Q_t K_t^{1/2} \preceq \Lambda I$, and conjugation by $K_t^{-1/2}$ gives the claim. Hence $x_t^\top Q_T x_t \leq x_t^\top Q_{t-1} x_t \leq \Lambda H_t$, and (M.13) gives $\mathbb{E}[\text{tr}(I_T Q_T) \mathbf{1}_E] \leq 50\Lambda \nu^2 T r^2$. Since $\text{tr} I_T \leq \nu^2 T$, $\mathbb{P}(E^c) \leq e^{-d_0/2} \leq d_0^{-1/2} \leq r^2$ and $\mathbb{E} \text{tr}(\bar{I}_T Q_T) = \mathbb{E} \text{tr}(I_T Q_T)$, we get

$$\mathbb{E} \text{tr}(\bar{I}_T Q_T) \leq 801 \left(\nu^2 T r^2 + \frac{\nu^4 T^2}{d_0^3} \right). \quad (\text{M.15})$$

On the other hand, by (M.14) and Markov's inequality, $\text{rank} Q_T \geq 3d_0/4$ with probability at least $3/4$. On the intersection of this event with S_T , which has probability at least $1/4$, the span of the eigenvectors of $\bar{I}_T$ with eigenvalues at least d_0 meets $\text{range} Q_T$ in dimension at least $d_0/4$, so $\text{tr}(\bar{I}_T Q_T) \geq d_0^2/4$ and

$$\mathbb{E} \text{tr}(\bar{I}_T Q_T) \geq d_0^2/16. \quad (\text{M.16})$$

By (M.4), $\nu^2 T r^2 < 2^{-16} d_0^2$ and $\nu^4 T^2/d_0^3 < 2^{-32} d_0^2$, so the right side of (M.15) is less than $801(2^{-16} + 2^{-32})d_0^2 < d_0^2/16$, which contradicts (M.16). Hence no estimator has $\mathbb{E} \|\hat{u} - u\|^2 \leq 1/256$ in the auxiliary experiment, and Step 1 completes the proof. $\square$

M.3 Scaling to the convex classes

Let $\mathcal{C}_{\beta, L_\beta, R}$ denote the convex functions $F : \mathbb{R}^d \rightarrow \mathbb{R}$ of class C^q with $[F]_{\beta, \text{op}} \leq L_\beta$ and a global minimizer in $B_2^d(R)$. A query at $X_t \in \mathbb{R}^d$ returns $F(X_t) + \xi_t$, where the ξ_t are independent $N(0, \sigma^2)$ variables, independent of the seed of the algorithm. Queries and the output are unrestricted. The algorithm may know $\beta, L_\beta, R, \sigma, T$, the hard family and its construction parameters, and, on the subclass with $\nabla^2 F \preceq LI$, also L .

Lemma M.5 (A quantitative hard family). *Let $\beta > 2$, $L_\beta, R, \sigma > 0$, and $d = 2d_0$ with $d_0 \geq 128$, and let $K_\beta^{\text{tube}}, E_2, \vartheta_\beta$ be the constants of Section M.1. For every accuracy scale*

$$0 < \varepsilon_0 \leq \frac{L_\beta R^\beta \vartheta_\beta^\beta}{K_\beta^{\text{tube}}}, \quad \vartheta = \left(\frac{K_\beta^{\text{tube}} \varepsilon_0}{L_\beta R^\beta} \right)^{1/\beta}, \quad r = \max\{16\vartheta, d_0^{-1/4}\},$$

every adaptive randomized algorithm using an integer budget

$$T < \frac{\sigma^2}{16\varepsilon_0^2} \min \left\{ d_0^{5/2}, \frac{d_0^2}{256\vartheta^2} \right\} \quad (\text{M.17})$$

has expected error greater than $\varepsilon_0/262144$ on some $F \in \mathcal{C}_{\beta, L_\beta, R}$. The hard functions also satisfy

$$0 \preceq \nabla^2 F \preceq \frac{E_2 \varepsilon_0}{R^2 \vartheta^2} I = \frac{E_2}{K_\beta^{\text{tube}}} L_\beta R^{\beta-2} \vartheta^{\beta-2} I. \quad (\text{M.18})$$

The same assertion therefore holds on the subclass with a known Hessian cap L whenever the right-hand side of (M.18) is at most L . Odd dimensions are covered by $d_0 = \lfloor d/2 \rfloor$ and one unused coordinate.

Proof. Set $A = \varepsilon_0/\vartheta$ and $F_\theta(x) = Ag_\theta(x/R)$ for $\theta \in E_b$. By the condition on ε_0 , $\vartheta \leq \vartheta_\beta$.

The class. By Lemma M.2, F_θ is convex and has a global minimizer in $B_2^d(R)$. By Lemma M.3,

$$[F_\theta]_{\beta, \text{op}} \leq AR^{-\beta} K_\beta^{\text{tube}} \vartheta^{1-\beta} = K_\beta^{\text{tube}} \varepsilon_0 R^{-\beta} \vartheta^{-\beta} = L_\beta,$$

and $\nabla^2 F_\theta \preceq AR^{-2} E_2 \vartheta^{-1} I$, which is (M.18).

The budget. A query of F_θ at X is a query of Ag_θ at X/R , and errors are preserved. Lemma M.4 excludes Bayes error at most $A\lambda_0\vartheta/4096 = \varepsilon_0/262144$ whenever

$$T < 2^{-16} \left(\frac{\sigma}{A\lambda_0\vartheta} \right)^2 \min\{d_0^{5/2}, d_0^2/r^2\}.$$

Since $A\vartheta = \varepsilon_0$, $\lambda_0 = 1/64$ and $r^2 = \max\{256\vartheta^2, d_0^{-1/2}\}$, this is (M.17). Bayes error greater than $\varepsilon_0/262144$ implies the same lower bound for the worst case.

Odd dimensions. Extending F_θ to $\mathbb{R}^{2d_0+1}$ by a coordinate on which it does not depend preserves the class, the errors and, after projection of the queries onto the first $2d_0$ coordinates, the Gaussian experiment. $\square$

Proof of Theorem 5(a). Let $C_{\text{cap}} = 2^{18(1-2/\beta)} E_2(K_\beta^{\text{tube}})^{-2/\beta}$ and $c'_0 = \vartheta_\beta^\beta / (2^{18} K_\beta^{\text{tube}})$. For $L \in (0, \infty]$ put

$$\ell = \min\{L_\beta, (L/C_{\text{cap}})^{\beta/2} \varepsilon^{-(\beta-2)/2}\},$$

with $\ell = L_\beta$ if $L = \infty$. Apply Lemma M.5 with ℓ in place of L_β , $\varepsilon_0 = 2^{18}\varepsilon$ and $d_0 = \lfloor d/2 \rfloor \geq 128$; the condition on ε_0 is $\varepsilon \leq c'_0 \ell R^\beta$. The hard functions have $[F]_{\beta, \text{op}} \leq \ell \leq L_\beta$, and by the definition of ϑ the cap in (M.18) equals

$$\frac{E_2 \varepsilon_0}{R^2 \vartheta^2} = E_2(K_\beta^{\text{tube}})^{-2/\beta} \ell^{2/\beta} \varepsilon_0^{1-2/\beta} = C_{\text{cap}} \ell^{2/\beta} \varepsilon^{1-2/\beta} \leq L,$$

so they belong to $\mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$. An ε -accurate method therefore needs at least the right-hand side of (M.17) calls; with $d_0 \geq d/3$ and $\vartheta^2 = (2^{18} K_\beta^{\text{tube}} \varepsilon / (\ell R^\beta))^{2/\beta}$ this is at least $c_\beta(\sigma^2/\varepsilon^2) \min\{d^{5/2}, d^2(\ell R^\beta/\varepsilon)^{2/\beta}\}$, and

$$\left(\frac{\ell R^\beta}{\varepsilon} \right)^{2/\beta} = \min\left\{ \left(\frac{L_\beta R^\beta}{\varepsilon} \right)^{2/\beta}, \frac{LR^2}{C_{\text{cap}} \varepsilon} \right\}.$$

The condition $\varepsilon \leq c'_0 \ell R^\beta$ holds if $\varepsilon \leq c'_0 L_\beta R^\beta$ and $\varepsilon \leq (c'_0)^{2/\beta} LR^2/C_{\text{cap}}$. The theorem follows with $c'_\beta = \min\{c'_0, (c'_0)^{2/\beta}/C_{\text{cap}}\}$ and c_β divided by $\max\{1, C_{\text{cap}}\}$. For $L \geq L_\varepsilon := C_{\text{cap}} L_\beta^{2/\beta} \varepsilon^{1-2/\beta}$ one has $\ell = L_\beta$, and the third term of the minimum is not the smallest. $\square$

M.4 The separable bounds for convex functions

Akhavan et al. (2020, 2024) state their bounds for strongly convex functions. This subsection derives from them the lower and the upper bound on $\mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$ in their row of Table 2 and compares the lower bound with Theorem 5(a). Let $A = L_\beta R^\beta/\varepsilon$, $S = \sigma^2/\varepsilon^2$ and $K = LR^2/\varepsilon$. The lower bound rescales the construction in the proof of Akhavan et al. (2024, Theorem 22) (see also Akhavan et al., 2020, Theorem 6.1), whose minimizers lie in the unit ball. Below the seminorm is the operator seminorm, the errors are independent $N(0, \sigma^2)$ with $\sigma > 0$, queries may be adaptive and unrestricted, and all constants depend only on β and on the fixed function ϕ .

The cited family. Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be the bounded smooth function of the cited proof, with $\phi' = 1$ on $[-1/4, 1/4]$ and $\phi' = 0$ outside $[-1, 1]$. For $h, a, k > 0$ and $\omega \in \{-1, 1\}^d$, set

$$f_\omega(x) = \frac{k}{2} \|x\|^2 + a \sum_{i=1}^d \omega_i \phi(x_i/h).$$

Its operator Hölder seminorm is at most $C_\beta a h^{-\beta}$. If $k \geq C_\beta a/h^2$, the Hessian lies between $kI/2$ and $2kI$, $(x_\omega^*)_i = -a\omega_i/(kh)$, and $\|x_\omega^*\| = \sqrt{d}a/(kh)$. The bounded differences between adjacent functions and the testing argument of the cited proof give, whenever $T \leq c_\beta \sigma^2/a^2$,

$$\sup_{\omega} \mathbb{E}_{\omega}[f_{\omega}(\hat{x}) - f_{\omega}(x_{\omega}^*)] \geq c_\beta d a^2/(k h^2), \quad (\text{M.19})$$

where $\mathbb{E}_{\omega}$ denotes expectation under f_{ω} .

Lower bound. Let $d \geq 1$, $\beta > 2$, $L_\beta, R > 0$ and $0 < \varepsilon \leq c_\beta L_\beta R^\beta$. Put $h_1 = (\varepsilon/(dL_\beta))^{1/\beta}$ and $h_2 = (\varepsilon/(\sqrt{d}RL_\beta))^{1/(\beta-1)}$, and choose

$$h = C_\beta \max\{h_1, h_2\}, \quad a = c_\beta L_\beta h^\beta, \quad k = C_\beta L_\beta \max\{h^{\beta-2}, \sqrt{d}h^{\beta-1}/R\}.$$

The functions have seminorm at most L_β and minimizers in $B(R)$, and the right side of (M.19) is at least a constant times $\min\{dL_\beta h^\beta, \sqrt{d}RL_\beta h^{\beta-1}\}$, which exceeds ε for a sufficiently large constant in h . Thus every ε -accurate method requires

$$T \geq c_\beta \frac{\sigma^2}{\varepsilon^2} \min \left\{ d^2, d^{\beta/(\beta-1)} \left(\frac{L_\beta R^\beta}{\varepsilon} \right)^{2/(\beta-1)} \right\}. \quad (\text{M.20})$$

Since $h_1, h_2 \leq (\varepsilon/L_\beta)^{1/\beta}$, $\sqrt{d}L_\beta h_1^{\beta-1}/R = L_\beta^{1/\beta} \varepsilon^{1-1/\beta} d^{1/\beta-1/2}/R$, $\sqrt{d}L_\beta h_2^{\beta-1}/R = \varepsilon/R^2$ and $\varepsilon \leq L_\beta R^\beta$, both terms of k are at most $C_\beta L_\beta^{2/\beta} \varepsilon^{1-2/\beta}$. Hence these functions belong to $\mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$ whenever $L \geq C_\beta L_\beta^{2/\beta} \varepsilon^{1-2/\beta}$. The two terms of the minimum in (M.20) are equal at $\varepsilon = L_\beta R^\beta d^{1-\beta/2}$, and the term d^2 is the smaller one for smaller ε .

Upper bound by regularization. For $f \in \mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$ put $f_\mu = f + \mu \|\cdot\|^2/2$ with $\mu = \varepsilon/R^2$. The minimizer of f_μ lies in $B(R)$, its seminorm is at most L_β , and $\mu I \preceq \nabla^2 f_\mu \preceq (L + \mu)I$; let $A \geq 1$. [Akhavan et al. \(2024, Theorem 18 and Table 1\)](#), applied to f_μ at accuracy $\varepsilon/2$, gives error at most ε for f with

$$T \leq C_\beta \{d^2 S A^{2/(\beta-1)} + d(1+K)^2\}$$

calls. The third term of that theorem requires a budget of order Q , where

$$Q^{\beta+1} = d^{\beta+2}(1+K)^{2\beta} S A^{-2} \leq [d(1+K)^2]^\beta [d^2 S A^{2/(\beta-1)}],$$

so it is absorbed by the two displayed terms. The leading statistical term alone suffices when it dominates $d(1+K)^2$. The bound of [Akhavan et al. \(2024, Corollary 19\)](#) is not used here, because its constant depends on L^2/σ^2 .

Size of the improvement. Let $A = L_\beta R^\beta/\varepsilon$ with $1 \leq A \leq d^{\beta/4}$. On this range the lower bound of Theorem 5(a) is $S d^2 A^{2/\beta}$ up to a constant factor, and its ratio to (M.20) is at least a constant times $d^{(\beta-2)/\beta}$. Indeed, for $A \geq d^{(\beta-2)/2}$ the prior minimum is d^2 and the ratio is $A^{2/\beta}$; for $A \leq d^{(\beta-2)/2}$ the prior minimum is $d^{\beta/(\beta-1)} A^{2/(\beta-1)}$ and the ratio is $d^{(\beta-2)/(\beta-1)} A^{-2/(\beta(\beta-1))}$, which decreases in A and is $d^{(\beta-2)/\beta}$ at $A = d^{(\beta-2)/2}$ and $d^{(2\beta-5)/(2\beta-2)} \geq d^{(\beta-2)/\beta}$ at $A = d^{\beta/4}$ when $\beta > 4$. For $A > d^{\beta/4}$ the new lower bound is $c_\beta S d^{5/2}$ and (M.20) is at most $c_\beta S d^2$, so the ratio is at least of order $d^{1/2}$. Both bounds hold on the same class $\mathcal{F}_{\text{cvx}}(\beta, L_\beta; L, R)$ with $L \geq C_\beta L_\beta^{2/\beta} \varepsilon^{1-2/\beta}$.

N NON-EUCLIDEAN GEOMETRY, CONSTRAINTS, AND AN ℓ_1 UPPER BOUND

This appendix proves the results of Section 7. Sections N.1 and N.2 transfer the Euclidean lower bounds of Sections E and M through image balls to Hölder regularity measured in an arbitrary norm, to minimizers localized in an arbitrary set and to constrained problems. Section N.3 evaluates these bounds on ℓ_p balls. Section N.4 proves an upper bound on the ℓ_1 ball that is linear in the dimension up to logarithmic factors for conditionally sub-Gaussian centered noise without an oblivious component; in a range of accuracies the lower bound is linear in the dimension as well.

Notation. Throughout, $\beta > 2$, $q = \lceil \beta \rceil - 1 \geq 2$ and $\gamma = \beta - q \in (0, 1]$. In this appendix $\|\cdot\|$ is a norm on $\mathbb{R}^d$ with closed unit ball $\mathcal{B}$ (written $\|\cdot\|_{\mathcal{B}}$ in Section 7), $\|u\|_* = \sup_{\|v\| \leq 1} \langle u, v \rangle$ is its dual norm, and the Euclidean norm is written $\|\cdot\|_2$. We write $\|\cdot\|_p$ for the ℓ_p norm, $B_p^n(R) = \{y \in \mathbb{R}^n : \|y\|_p \leq R\}$ and $B_p^n = B_p^n(1)$. For a j -linear form A on $\mathbb{R}^d$ and for $f \in C^q(\mathbb{R}^d)$ let

$$\|A\|_{\|\cdot\|} = \sup\{|A[u_1, \dots, u_j]| : \|u_1\| \leq 1, \dots, \|u_j\| \leq 1\}, \quad [f]_{\beta, \|\cdot\|} = \sup_{x \neq y} \frac{\|D^q f(x) - D^q f(y)\|_{\|\cdot\|}}{\|x - y\|^\gamma},$$

so that $[f]_{\beta, \|\cdot\|} = [f]_{\beta, \mathcal{B}}$ in the notation of Section 7 and $[f]_{\beta, \|\cdot\|_2} = [f]_{\beta, \text{op}}$; we write $[f]_{\beta, p} = [f]_{\beta, \|\cdot\|_p}$. For a nonempty set $K \subseteq \mathbb{R}^d$ let $\mathcal{C}(\beta, L_\beta; \|\cdot\|, K)$ be the class of convex $f \in C^q(\mathbb{R}^d)$ with $[f]_{\beta, \|\cdot\|} \leq L_\beta$ and $\arg \min f \cap K \neq \emptyset$. For a linear map $P : \mathbb{R}^d \rightarrow \mathbb{R}^n$ and a symmetric $d \times d$ matrix H let

$$\|P\|_{\mathcal{B} \rightarrow 2} = \max_{x \in \mathcal{B}} \|Px\|_2, \quad \|H\|_{\|\cdot\| \rightarrow \|\cdot\|_*} = \sup\{u^\top H v : \|u\| \leq 1, \|v\| \leq 1\},$$

and let $\|P\|_{2 \rightarrow 2}$ be the spectral norm of P . If $\|u\|_2 \leq \kappa_0 \|u\|$ for all u , then $\|A\|_{\|\cdot\|} \leq \kappa_0^q \|A\|_{\text{op}}$ for every q -linear form A and $\|x - y\|_2^\gamma \leq \kappa_0^\gamma \|x - y\|^\gamma$, hence

$$[f]_{\beta, \|\cdot\|} \leq \kappa_0^\beta [f]_{\beta, \text{op}}; \quad (\text{N.1})$$

for $\|\cdot\|_p$ with $1 \leq p \leq 2$ one may take $\kappa_0 = 1$.

Definition N.1 (Image radius). Let $C \subseteq \mathbb{R}^d$ be nonempty and $1 \leq n \leq d$. Image balls (P, z, r) of C with $P : \mathbb{R}^d \rightarrow \mathbb{R}^n$ are defined in Definition 2 with $\mathcal{K} = C$; their first condition is $\|P\|_{\mathcal{B} \rightarrow 2} \leq 1$. The *image radius* of C is

$$\mathcal{R}_n(\mathcal{B}, C) = \sup\{r > 0 : (P, z, r) \text{ is an image ball of } C \text{ for some } P, z\}, \quad \sup \emptyset = 0.$$

This is the quantity of Dorn et al. (2026), who write the supremum as a maximum. For compact C with $\mathcal{R}_n(\mathcal{B}, C) > 0$ the supremum is attained, because P and z stay bounded along a maximizing sequence and a limit of image balls of a compact set, with radii bounded away from 0, is again an image ball. Thus positive image radius is precisely the geometric condition used by the transfer theorem. For centrally symmetric convex C the center may be taken to be 0 (Dorn et al., 2026, Lemma J.1).

Oracle. We use the round model and the noise of Definition 1; unless stated otherwise, η is Lipschitz in the Euclidean norm. The lower bounds use independent $N(0, \sigma^2)$ noise and allow adaptation after every call, so they hold for every number of rounds.

N.1 Pull-back of hard functions

Lemma N.2 (Pull-back). *Let $K \subseteq \mathbb{R}^d$ be nonempty, let $1 \leq n \leq d$, and let (P, z, r) be an image ball of K . Let $g : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex and q times continuously differentiable, with $[g]_{\beta, \text{op}} \leq L_\beta$ and a minimizer $y^* \in z + rB_2^n$, and put $f = g \circ P$.*

(a) *f is convex and q times continuously differentiable, $D^j f(x)[u_1, \dots, u_j] = D^j g(Px)[Pu_1, \dots, Pu_j]$ for $1 \leq j \leq q$, and $[f]_{\beta, \|\cdot\|} \leq \|P\|_{\mathcal{B} \rightarrow 2}^\beta L_\beta \leq L_\beta$.*

(b) *$\nabla f(x) = P^\top \nabla g(Px)$ and $\nabla^2 f(x) = P^\top \nabla^2 g(Px)P$. Hence, for all $x \in \mathbb{R}^d$,*

$$\|\nabla f(x)\|_* \leq \|\nabla g(Px)\|_2, \quad \|\nabla^2 f(x)\|_{\|\cdot\| \rightarrow \|\cdot\|_*} \leq \|\nabla^2 g(Px)\|_{\text{op}}.$$

If $0 \preceq \nabla^2 g \preceq LI$ on $\mathbb{R}^n$, then $0 \leq u^\top \nabla^2 f(x)u \leq L\|u\|^2$ and $0 \preceq \nabla^2 f(x) \preceq L\|P\|_{2 \rightarrow 2}^2 I$ for all $x, u \in \mathbb{R}^d$.

(c) *There is $x^* \in K$ with $Px^* = y^*$; it minimizes f , $\min f = \min g$, and $f(x) - \min f = g(Px) - \min g$ for all x . In particular $f \in \mathcal{C}(\beta, L_\beta; \|\cdot\|, K)$.*

(d) *Let $\mathbf{A}$ be an algorithm on $\mathbb{R}^d$ with T calls in N rounds, and let $\mathbf{A}_P$ be the algorithm on $\mathbb{R}^n$ that uses the same random elements, runs a copy of $\mathbf{A}$, answers each query x of the copy by querying Px , and outputs $P\hat{x}$; it has T calls in N rounds. For every noise admissible at levels (M', σ, δ) in the run of $\mathbf{A}_P$ on g there is a noise*

admissible at levels $(M' \|P\|_{2 \rightarrow 2}, \sigma, \delta)$ in the run of A on f under which the random elements, the queries of A , the observed values and the error components coincide pathwise with those of the copy; independent $N(0, \sigma^2)$ noise corresponds to independent $N(0, \sigma^2)$ noise, and $\mathbb{E}[g(P\hat{x}) - \min g] = \mathbb{E}[f(\hat{x}) - \min f]$. Consequently, if A is ε -accurate at levels (M, σ, δ) on a class that contains $g \circ P$ for all g in a class $\mathcal{E}$ of functions as above, then A_P is ε -accurate on $\mathcal{E}$ at levels $(M/\|P\|_{2 \rightarrow 2}, \sigma, \delta)$, with the same numbers of calls and rounds.

Proof. For exact function values this is the image-ball reduction of [Dorn et al. \(2026\)](#); we check the regularity of order β and the noise model. The chain rule gives $D^j f(x) = D^j g(Px) \circ P^{\otimes j}$, so

$$\|D^q f(x) - D^q f(y)\|_{\|\cdot\|} \leq \|P\|_{\mathcal{B} \rightarrow 2}^q \|D^q g(Px) - D^q g(Py)\|_{\text{op}} \leq L_\beta \|P\|_{\mathcal{B} \rightarrow 2}^{q+\gamma} \|x - y\|^\gamma.$$

The norm bounds in (b) follow from $\|Pu\|_2 \leq \|u\|$ and $\|Pu\|_2 \leq \|P\|_{2 \rightarrow 2} \|u\|_2$, and (c) from $y^* \in z + rB_2^n \subseteq P(K)$.

For (d), let the run of A on f use the same random elements and the same variables $\xi_t^c, \xi_t^o, \delta_t$ as the run of A_P on g , and put $\eta = \eta' \circ P$, where η' is the Lipschitz component of the latter. Since $f + \eta = (g + \eta') \circ P$, induction over the rounds shows that the two runs observe the same values. The queries of a round are functions of the random elements and of the values of the earlier rounds, so both runs generate the same σ -algebras $\mathcal{F}_{k-1}$ and $\mathcal{E}_t$. Hence (C), (O) and (A) carry over, η is fixed before the run, and $|\eta(x) - \eta(y)| \leq M' \|P(x - y)\|_2 \leq M' \|P\|_{2 \rightarrow 2} \|x - y\|_2$. Finally $f(\hat{x}) - \min f = g(P\hat{x}) - \min g$ by (c). If the Lipschitz component on $\mathbb{R}^d$ is measured in $\|\cdot\|$ instead, the same computation gives the level M' . $\square$

N.2 Transfer of one-point lower bounds

Euclidean lower bounds. Write K_H, E_H and s_H for the constants K_β, E_2 and $\vartheta_\beta \leq 1/2$ of Lemma [M.5](#) (not those of Lemma [E.1](#)), and put $C_H = E_H s_H^{\beta-2} / K_H$. For $n \geq 1, z \in \mathbb{R}^n$ and $r > 0$ let $\mathcal{E}_n(z, r)$ be the class of convex $g \in C^q(\mathbb{R}^n)$ with $[g]_{\beta, \text{op}} \leq L_\beta$ and a global minimizer in $z + rB_2^n$, and let $\mathcal{E}_n^{\text{cap}}(z, r)$ be its subclass defined by $0 \leq \nabla^2 g \leq C_H L_\beta r^{\beta-2} I$. The translation $g \mapsto g(\cdot + z)$ maps these classes onto $\mathcal{E}_n(0, r)$ and $\mathcal{E}_n^{\text{cap}}(0, r)$ and preserves optimization errors. We use two bounds for independent $N(0, \sigma^2)$ noise and adaptive randomized algorithms with unrestricted queries and outputs. This noise is independent of the random elements of the algorithm, so both bounds hold in the model of Definition [1](#).

(H1) For $n \geq 256$ and $0 < \varepsilon \leq c'_H L_\beta r^\beta$, every algorithm with $\sup_{g \in \mathcal{E}_n^{\text{cap}}(0, r)} \mathbb{E}[g(\hat{y}) - \min g] \leq \varepsilon$ uses

$$T \geq c_H \frac{\sigma^2}{\varepsilon^2} \min\{n^{5/2}, n^2 (L_\beta r^\beta / \varepsilon)^{2/\beta}\}$$

calls, even when the Hessian cap and all parameters are known. This is Lemma [M.5](#) with R replaced by r , with $\varepsilon_0 = 2^{18}\varepsilon$ and with the half-dimension $m' = \lfloor n/2 \rfloor \geq 128$ (denoted m there), so that $m' \geq \frac{255}{512}n$; it gives

$$c'_H = \frac{s_H^\beta}{2^{18} K_H}, \quad c_H = 2^{-40} \left(\frac{255}{512} \right)^{5/2} \min\{1, 2^{-8} (2^{18} K_H)^{-2/\beta}\}.$$

The scale parameter ϑ of these hard functions is at most s_H , so [\(M.18\)](#) bounds their Hessians by $C_H L_\beta r^{\beta-2}$.

(V) For $n \geq 1$ and $0 < \varepsilon \leq c'_V L_\beta r^\beta$, every algorithm with $\sup_{g \in \mathcal{E}_n(0, r)} \mathbb{E}[g(\hat{y}) - \min g] \leq \varepsilon$ uses $T \geq c_V n^2 \sigma^2 / \varepsilon^2$ calls. This is the operator case of Lemma [E.6](#), whose proof applies to every $n \geq 1$ and gives $c_V = (3/(256\bar{K}_2))^2$ and $c'_V = 3(2\bar{K}_2)^\beta / (128\bar{K}_2 K_\beta)$, where $\bar{K}_2 = \max\{1, K_2\}$ and K_2, K_β are the constants of Lemma [E.1](#).

Theorem N.3 (Geometry transfer of one-point lower bounds). *Let $K \subseteq \mathbb{R}^d$ be nonempty, let $1 \leq n \leq d$, and let (P, z, r) be an image ball of K . Consider independent $N(0, \sigma^2)$ noise and an algorithm with T calls, in any number of rounds, such that $\mathbb{E}[f(\hat{x}) - \min f] \leq \varepsilon$ for every f in a class $\mathcal{F}$ of functions on $\mathbb{R}^d$.*

(a) *If $\mathcal{F} \supseteq \{g \circ P : g \in \mathcal{E}_n^{\text{cap}}(z, r)\}$, for instance if $\mathcal{F} = \mathcal{C}(\beta, L_\beta; \|\cdot\|, K)$, and if $n \geq 256$ and $0 < \varepsilon \leq c'_H L_\beta r^\beta$, then*

$$T \geq c_H \frac{\sigma^2}{\varepsilon^2} \min\left\{n^{5/2}, n^2 \left(\frac{L_\beta r^\beta}{\varepsilon}\right)^{2/\beta}\right\}.$$

(b) *If $\mathcal{F} \supseteq \{g \circ P : g \in \mathcal{E}_n(z, r)\}$, for instance if $\mathcal{F} = \mathcal{C}(\beta, L_\beta; \|\cdot\|, K)$, and if $0 < \varepsilon \leq c'_V L_\beta r^\beta$, then $T \geq c_V n^2 \sigma^2 / \varepsilon^2$.*

Every $f = g \circ P$ with $g \in \mathcal{E}_n^{\text{cap}}(z, r)$ satisfies, for all $x, u \in \mathbb{R}^d$,

$$0 \leq u^\top \nabla^2 f(x) u \leq C_H L_\beta r^{\beta-2} \|u\|^2, \quad \|\nabla f(x)\|_* \leq C_H L_\beta r^{\beta-2} (\|Px - z\|_2 + r), \quad (\text{N.2})$$

so (a) holds also on the subclass of $\mathcal{C}(\beta, L_\beta; \|\cdot\|, K)$ defined by (N.2), even when these bounds, P , z , r and all class parameters are known. If in addition $\|P\|_{2 \rightarrow 2} \leq 1$, these functions satisfy the Euclidean cap $0 \preceq \nabla^2 f \preceq C_H L_\beta r^{\beta-2} I$, and (a) holds on the corresponding subclass. If K is compact, (a) and (b) hold with $r = \mathcal{R}_n(\mathcal{B}, K)$; for general K they hold for every $r < \mathcal{R}_n(\mathcal{B}, K)$.

Proof. Let g_0 range over $\mathcal{E}_n^{\text{cap}}(0, r)$ for (a) and over $\mathcal{E}_n(0, r)$ for (b), and put $g = g_0(\cdot - z)$. By Lemma N.2(c), $g \circ P \in \mathcal{C}(\beta, L_\beta; \|\cdot\|, K)$. The algorithm A_P of Lemma N.2(d), with its queries shifted to $Px - z$ and its output to $P\hat{x} - z$, has on g_0 the error of the given algorithm on $g \circ P$, with the same calls, rounds and Gaussian noise. Bounds (H1) and (V) therefore give (a) and (b). If $g \in \mathcal{E}_n^{\text{cap}}(z, r)$ has the minimizer y^* , then, with $L = C_H L_\beta r^{\beta-2}$, $\|\nabla g(y)\|_2 \leq L \|y - y^*\|_2 \leq L(\|y - z\|_2 + r)$; together with Lemma N.2(b) this gives (N.2) and the Euclidean cap. The last assertion follows from the attainment of the image radius for compact K and from the definition of a supremum otherwise. $\square$

Corollary 1 follows from parts (a) and (b), with $c_\beta = \min\{c_H, c_V\}$ and $c'_\beta = \min\{c'_H, c'_V\}$, and from the following corollary.

Corollary N.4 (Restricted queries and functional constraints). *Let K and n be as in Lemma N.3, let $\phi_1, \dots, \phi_J : \mathbb{R}^d \rightarrow \mathbb{R}$ ($J \geq 0$) be fixed functions, let $X_\phi = \{x \in K : \phi_i(x) \leq 0, 1 \leq i \leq J\}$, and let (P, z, r) be an image ball of X_ϕ . Let $X_q, X_o \subseteq \mathbb{R}^d$ be arbitrary sets, for instance $X_q = K + \rho B_2^d$ with $\rho \geq 0$ and $X_o = X_\phi$. Consider an algorithm that evaluates the objective only at points of X_q , outputs $\hat{x} \in X_o$, and may in addition evaluate the functions ϕ_i at arbitrary points without these evaluations being counted. Suppose that the objective values carry independent $N(0, \sigma^2)$ noise, that the constraint values are returned exactly or with independent noise of a fixed law, and that*

$$\mathbb{E} \left[f(\hat{x}) - \min_{X_\phi} f \right] \leq \varepsilon \quad \text{for every } f \in \mathcal{C}(\beta, L_\beta; \|\cdot\|, X_\phi),$$

possibly together with further requirements on $\hat{x}$, such as bounds on the constraint violation. Then the number T of objective evaluations satisfies (a) and (b) of Lemma N.3.

Proof. Every $f \in \mathcal{C}(\beta, L_\beta; \|\cdot\|, X_\phi)$ has a global minimizer in X_ϕ , so $\min_{X_\phi} f = \min f$. Let A' run the given algorithm, forward its objective queries, answer its constraint queries by computing the ϕ_i (adding, if the constraint oracle is noisy, noise generated from additional independent random elements appended to Θ_k), and output $\hat{x}$. Then A' is an algorithm of the model with T calls and unrestricted queries and outputs, and $\mathbb{E}[f(\hat{x}) - \min f] \leq \varepsilon$ for every $f \in \mathcal{C}(\beta, L_\beta; \|\cdot\|, X_\phi)$. Apply Lemma N.3 with K replaced by X_ϕ . $\square$

The corollary applies to every class of constrained problems, for instance with the functional constraints of [Abronin et al. \(2026\)](#), that contains all objectives in $\mathcal{C}(\beta, L_\beta; \|\cdot\|, X_\phi)$ for one fixed admissible tuple $(\phi_1, \dots, \phi_J)$; the algorithm need not know the ϕ_i . For exact values, [Dorn et al. \(2026, Theorems 3.2 and 3.3\)](#) treat a known regularizer, whose values are free, in a similar way.

N.3 Lower bounds on ℓ_p balls

Lemma N.5 (Radius bounds). *Let $1 \leq p \leq 2$, $R > 0$ and $1 \leq n \leq d$, and let $P_n : \mathbb{R}^d \rightarrow \mathbb{R}^n$, $P_n x = (x_1, \dots, x_n)$, be the coordinate projection.*

- (a) *($P_n, 0, Rn^{1/2-1/p}$) is an image ball of $B_p^d(R)$, both for the norm $\|\cdot\|_p$ and for the Euclidean norm, and $\|P_n\|_{2 \rightarrow 2} = 1$. Hence $\mathcal{R}_n(B_p^d, B_p^d(R)) \geq Rn^{1/2-1/p}$ and $\mathcal{R}_n(B_2^d, B_p^d(R)) \geq Rn^{1/2-1/p}$.*
- (b) *Let $p = 1$, and let $k \leq n$ be a positive integer with $b = \lceil n/k \rceil$ satisfying $b5^k \leq d$. Then there is a linear $P : \mathbb{R}^d \rightarrow \mathbb{R}^n$ such that $(P, 0, 7R/(8\sqrt{b}))$ is an image ball of $B_1^d(R)$ for the norm $\|\cdot\|_1$. Hence $\mathcal{R}_n(B_1^d, B_1^d(R)) \geq 7R/(8\sqrt{b}) \geq \frac{7}{8}R\sqrt{k/(2n)}$. The map P constructed in the proof has $\|P\|_{2 \rightarrow 2} \geq (1 + 49/64)^{1/2} > 1$.*

Proof. (a) This is the coordinate-projection case of [Dorn et al. \(2026, Proposition I.1\)](#): P_n is a contraction for both norms, and $P_n(B_p^d(R)) = B_p^n(R) \supseteq Rn^{1/2-1/p}B_2^n$ by Hölder's inequality.

(b) Partition $\{1, \dots, n\}$ into b blocks $I_1, \dots, I_b$ of sizes at most k . A subset of the unit sphere of $\mathbb{R}^{I_\ell}$ with pairwise Euclidean distances larger than $1/2$ has at most 5^k points, because the balls of radius $1/4$ around its points are disjoint and lie in the ball of radius $5/4$. Let $\mathcal{N}_\ell$ be a maximal such subset. By maximality, every unit vector $u \in \mathbb{R}^{I_\ell}$ has some $v \in \mathcal{N}_\ell$ with $\|u - v\|_2 \leq 1/2$, and then $\langle u, v \rangle = 1 - \|u - v\|_2^2/2 \geq 7/8$. Hence the support function of the absolutely convex hull A_ℓ of $\mathcal{N}_\ell$ is at least $7/8$ on the unit sphere, and $A_\ell \supseteq \frac{7}{8}B_2^{I_\ell}$ by the separation theorem.

Regard the at most $b5^k \leq d$ vectors of all $\mathcal{N}_\ell$ as vectors of $\mathbb{R}^n$ supported on I_ℓ , and let P map $e_1, e_2, \dots$ to these vectors and the remaining coordinate vectors to 0. Since B_1^d is the absolutely convex hull of the $\pm e_i$, $\|P\|_{B_1^d \rightarrow 2}$ is the largest column norm, which is 1, and $P(B_1^d(R)) = RA$, where A is the absolutely convex hull of $\bigcup_\ell \mathcal{N}_\ell$. Let $\|y\|_2 \leq 7/(8\sqrt{b})$, and let y_ℓ be the restriction of y to I_ℓ . Then $y_\ell \in \frac{8}{7}\|y_\ell\|_2 A_\ell \subseteq \frac{8}{7}\|y_\ell\|_2 A$, and since A is convex and contains 0, $y = \sum_\ell y_\ell \in \frac{8}{7}(\sum_\ell \|y_\ell\|_2)A \subseteq A$, because $\sum_\ell \|y_\ell\|_2 \leq \sqrt{b}\|y\|_2 \leq 7/8$. Hence $\frac{7R}{8\sqrt{b}}B_2^n \subseteq P(B_1^d(R))$, and $b \leq n/k + 1 \leq 2n/k$ gives the last bound on the image radius.

Finally, let $v \in \mathcal{N}_1$ and choose $v' \in \mathcal{N}_1$ with $\langle -v, v' \rangle \geq 7/8$. The vector $P^\top v$ has the entries $\langle v, v \rangle = 1$ and $\langle v, v' \rangle \leq -7/8$, so $\|P\|_{2 \rightarrow 2} = \|P^\top\|_{2 \rightarrow 2} \geq (1 + 49/64)^{1/2}$. $\square$

Part (b) with $b = 1$, that is, $k = n \leq \log_5 d$, gives the radius $7R/8$, independent of n . [Dorn et al. \(2026, Proposition I.3\)](#) also give images of radius of order R of every ℓ_p ball with $1 \leq p \leq 2$, of dimension of order $\log d$ for $p = 1$.

For $1 \leq p \leq 2$ put

$$p^* = \frac{p}{p-1} \in [2, \infty] \text{ (} p^* = \infty \text{ for } p = 1\text{)}, \quad \theta_p = \frac{1}{p} - \frac{1}{2} \in \left[0, \frac{1}{2}\right],$$

and $\mathcal{C}_p(R) = \mathcal{C}(\beta, L_\beta; \|\cdot\|_p, B_p^d(R))$. Then $1 + 2/p^* = 3 - 2/p = 2 - 2\theta_p$.

Corollary N.6 (Lower bounds for the ℓ_p geometry). *Let $1 \leq p \leq 2$. Under independent $N(0, \sigma^2)$ noise, every algorithm with $\sup_{f \in \mathcal{C}_p(R)} \mathbb{E}[f(\hat{x}) - \min f] \leq \varepsilon$ uses T calls, where:*

(a) for every integer n with $256 \leq n \leq d$ and $\varepsilon \leq c'_H L_\beta R^\beta n^{-\beta\theta_p}$,

$$T \geq c_H \frac{\sigma^2}{\varepsilon^2} \min\left\{n^{5/2}, n^{1+2/p^*} \left(\frac{L_\beta R^\beta}{\varepsilon}\right)^{2/\beta}\right\};$$

(b) for every integer n with $1 \leq n \leq d$ and $\varepsilon \leq c'_V L_\beta R^\beta n^{-\beta\theta_p}$, $T \geq c_V n^2 \sigma^2 / \varepsilon^2$.

(c) Let $\varepsilon \leq c'_H L_\beta R^\beta$ and put

$$n_\varepsilon = \begin{cases} (c'_H L_\beta R^\beta / \varepsilon)^{1/(\beta\theta_p)}, & p < 2, \\ +\infty, & p = 2, \end{cases} \quad n_\times = (L_\beta R^\beta / \varepsilon)^{4/(\beta(4\theta_p+1))}, \quad n_* = \min\{d, \lfloor n_\varepsilon \rfloor\}.$$

If $n_* \geq 256$, then

$$T \geq c_H \frac{\sigma^2}{\varepsilon^2} \min\{n_*^{5/2}, n_*^{1+2/p^*} (L_\beta R^\beta / \varepsilon)^{2/\beta}\},$$

and the first term is the smaller one exactly when $n_* \leq n_\times$. In particular, provided $n_* \geq 256$:

(R1) if $d \leq \min\{n_\varepsilon, n_\times\}$, then $T \geq c_H d^{5/2} \sigma^2 / \varepsilon^2$;

(R2) if $n_\times \leq d \leq n_\varepsilon$, then $T \geq c_H d^{1+2/p^*} (L_\beta R^\beta / \varepsilon)^{2/\beta} \sigma^2 / \varepsilon^2$;

(R3) if $p < 2$, $d > n_\varepsilon$ and $\lfloor n_\varepsilon \rfloor \geq n_\times$, then, independently of d ,

$$T \geq \frac{c_H}{4} \left(\frac{c'_H L_\beta R^\beta}{\varepsilon}\right)^{\frac{2(3p-2)}{\beta(2-p)}} \left(\frac{L_\beta R^\beta}{\varepsilon}\right)^{2/\beta} \frac{\sigma^2}{\varepsilon^2}.$$

Procedure 7 Entropic descent with coordinate stencils on the ℓ_1 ball

Require: $\beta, L_\beta, R, G_\infty, \sigma, d, \varepsilon; h > 0, N \geq 1, \varsigma > 0; t_j, a_j$ from (C.1)

```

1:  $w_1 \leftarrow (2d)^{-1} \mathbf{1} \in \Delta_{2d}$ 
2: for  $k = 1, \dots, N$  do
3:    $x_k \leftarrow R(w_k^+ - w_k^-)$ ; draw independent uniform signs  $s_{k,i}$ 
4:   In one round obtain  $Y_{k,i,j}^\pm$  at  $x_k \pm ht_j s_{k,i} e_i$ 
5:    $g_{k,i} \leftarrow \frac{s_{k,i}}{2h} \sum_{j=1}^m a_j (Y_{k,i,j}^+ - Y_{k,i,j}^-)$  for  $i = 1, \dots, d$ 
6:    $G_k \leftarrow R(g_k, -g_k)$ 
7:    $w_{k+1,\ell} \leftarrow w_{k,\ell} e^{-\varsigma G_{k,\ell}} / \sum_{r=1}^{2d} w_{k,r} e^{-\varsigma G_{k,r}}$  for  $\ell = 1, \dots, 2d$ 
8: return  $\bar{x}_N = N^{-1} \sum_{k=1}^N x_k$ 
    
```

(d) For $p = 1$: $n_\varepsilon = (c'_H L_\beta R^\beta / \varepsilon)^{2/\beta}$ and $n_\times = (L_\beta R^\beta / \varepsilon)^{4/(3\beta)}$. If $\min\{d, \lfloor n_\varepsilon \rfloor\} \geq \max\{256, n_\times\}$, then

$$T \geq \frac{c_H}{2} \min\left\{d, \left(\frac{c'_H L_\beta R^\beta}{\varepsilon}\right)^{2/\beta}\right\} \left(\frac{L_\beta R^\beta}{\varepsilon}\right)^{2/\beta} \frac{\sigma^2}{\varepsilon^2}.$$

In particular, if $d \geq 256$ and $L_\beta R^\beta d^{-3\beta/4} \leq \varepsilon \leq c'_H L_\beta R^\beta d^{-\beta/2}$, then $T \geq c_H d (L_\beta R^\beta / \varepsilon)^{2/\beta} \sigma^2 / \varepsilon^2$, which is linear in d .

(e) For $p = 1$ and integers $b, k \geq 1$ with $b5^k \leq d$, $bk \geq 256$ and $\varepsilon \leq c'_H (7/8)^\beta L_\beta R^\beta b^{-\beta/2}$,

$$T \geq c_H \frac{\sigma^2}{\varepsilon^2} \min\left\{(bk)^{5/2}, \frac{49}{64} b k^2 \left(\frac{L_\beta R^\beta}{\varepsilon}\right)^{2/\beta}\right\}.$$

The bounds (a), (c) and (d) hold already on the subclass of $\mathcal{C}_p(R)$ defined by the Euclidean Hessian cap and the gradient bound

$$0 \leq \nabla^2 f \preceq C_H L_\beta r^{\beta-2} I, \quad \|\nabla f(x)\|_{p^*} \leq C_H L_\beta r^{\beta-2} (\|P_n x\|_2 + r),$$

where $r = Rn^{-\theta_p}$ and n is the dimension used. The bound (e) holds on the subclass defined by (N.2) for the map of Lemma N.5(b).

Proof. Parts (a), (b) and (e) follow from Lemma N.3 with the image balls of Lemma N.5. For the coordinate projection, $r = Rn^{-\theta_p}$ and $n^2 (L_\beta r^\beta / \varepsilon)^{2/\beta} = n^{1+2/p^*} (L_\beta R^\beta / \varepsilon)^{2/\beta}$; for (e), $n = bk$ and $r = 7R/(8\sqrt{b})$ give $n^2 r^2 = \frac{49}{64} b k^2 R^2$. In (c), the condition on ε in (a) means $n \leq n_\varepsilon$, and both terms of the minimum increase with n , so the best choice is $n = n_*$; the first term is the smaller one exactly when $n \leq n_\times$. This gives (R1)–(R3), where (R3) uses $\lfloor n_\varepsilon \rfloor \geq n_\varepsilon/2$ and $1 + 2/p^* \leq 2$. Part (d) is (c) with $\theta_1 = 1/2$ and $1 + 2/p^* = 1$, since $n_* \geq \min\{d, n_\varepsilon\}/2$, and $n_* = d$ under its last condition. The subclass assertions follow from (N.2) and $\|P_n\|_{2 \rightarrow 2} = 1$. $\square$

Part (b) also covers dimensions below 256. If $d \geq 256$ and $\varepsilon \leq \min\{c'_H, c'_V\} L_\beta R^\beta d^{-\beta\theta_p}$, then part (a) with $n = d$ is at least c_H/c_V times part (b) with $n = d$, because $d^{5/2} \geq d^2$ and $d^{1+2/p^*} (L_\beta R^\beta / \varepsilon)^{2/\beta} = d^2 (d^{-\beta\theta_p} L_\beta R^\beta / \varepsilon)^{2/\beta} \geq (c'_H)^{-2/\beta} d^2 \geq d^2$. Part (e) improves the regime (R3) at $p = 1$: with $b = \lfloor \frac{49}{64} n_\varepsilon \rfloor$ and $k = \lfloor \log_5(d/b) \rfloor$, its second term exceeds the bound of (R3) by a factor of order $k^2 \asymp \log^2(d/n_\varepsilon)$.

N.4 An upper bound on the ℓ_1 ball

We use the stencil of Section C: $m = \lceil q/2 \rceil$, nodes $t_j = j/m$, weights a_j solving (C.1), and the operator $D_{h,u}$ of (C.4), here with $u = e_i$. Its constants are S_1, S_2 and c_0 of (C.3), A_1 of (C.2) and $C_\sigma = (S_1 + (S_2/2)^{1/2})^2$ of Lemma D.3 (see Table 4); they depend only on β . For $2 < \beta \leq 3$ they are given in the row $m = 1$ of Table 4, and $c_0 = 1/2$.

Let $\mathcal{C}_1(R, G_\infty)$ be the class of convex $f \in C^q(\mathbb{R}^d)$ with $[f]_{\beta,1} \leq L_\beta$, $\arg \min f \cap B_1^d(R) \neq \emptyset$, and $\|\nabla f(x)\|_\infty \leq G_\infty$ for all $x \in B_1^d(R)$. Let $\Delta_{2d} = \{w \in [0, 1]^{2d} : \sum_\ell w_\ell = 1\}$, and for $w \in \mathbb{R}^{2d}$ let $w^+, w^- \in \mathbb{R}^d$ be its first and last d coordinates. Procedure 7 runs entropic mirror descent (Nemirovski and Yudin, 1983; Beck and Teboulle, 2003) on Δ_{2d} , mapped onto $B_1^d(R)$ by $w \mapsto R(w^+ - w^-)$, with coordinate stencils whose signs are drawn afresh in every round. Its parameters h, N, ς are fixed before the run; Lemma N.9(b) chooses them for a target accuracy ε .

Since x_k is a function of $g_1, \dots, g_{k-1}$, it is $\mathcal{F}_{k-1}$ -measurable. The queries of round k are functions of x_k and of the signs $s_{k,i}$, which belong to Θ_k , and their order does not depend on Θ_k . Thus the algorithm belongs to the model, with N rounds and $T = 2mdN$ calls, and g_k is the estimate (N.3) of the next lemma at the center $x = x_k$, with $B = 1$ and $s_i^{(1)} = s_{k,i}$. We write b_k and Ξ_k for the vectors b and Ξ of the lemma for this estimate.

Lemma N.7 (Coordinate stencils with random signs). *Let $f \in C^q(\mathbb{R}^d)$ be such that for every $z \in \mathbb{R}^d$ and every i the function $u \mapsto f(z + ue_i)$ has a γ -Hölder q -th derivative with constant L_β ; this holds if $[f]_{\beta,p} \leq L_\beta$ for some $p \in [1, \infty]$, because $\|e_i\|_p = 1$. Let $h > 0$ and an integer $B \geq 1$ be fixed, let k be a round, and let x be an $\mathcal{F}_{k-1}$ -measurable random point of $\mathbb{R}^d$. In round k , draw independent signs $s_i^{(\ell)}$ with $\mathbb{P}(s_i^{(\ell)} = 1) = \mathbb{P}(s_i^{(\ell)} = -1) = 1/2$, $1 \leq i \leq d$, $1 \leq \ell \leq B$, as part of Θ_k . Query, in an order that does not depend on Θ_k , the $2mdB$ points $X_{i,j,\ell}^\pm = x \pm ht_j s_i^{(\ell)} e_i$ ($1 \leq i \leq d$, $1 \leq j \leq m$, $1 \leq \ell \leq B$), observe the values $Y_{i,j,\ell}^\pm$, and set*

$$g_i = \frac{1}{B} \sum_{\ell=1}^B \frac{s_i^{(\ell)}}{2h} \sum_{j=1}^m a_j (Y_{i,j,\ell}^+ - Y_{i,j,\ell}^-), \quad 1 \leq i \leq d. \quad (\text{N.3})$$

Let the noise satisfy (C), (O) and (A), and let its Lipschitz component η satisfy

$$|\eta(y + ue_i) - \eta(y - ue_i)| \leq 2M |u| \quad (y \in \mathbb{R}^d, u \in \mathbb{R}, 1 \leq i \leq d); \quad (\text{N.4})$$

this holds if η is M -Lipschitz for $\|\cdot\|_p$ for some $p \in [1, \infty]$, in particular under (L). Let $\xi_{i,j,\ell}^\pm$ be the random error $\xi_t^c + \xi_t^o$ at the query $X_{i,j,\ell}^\pm$. Then g is $\mathcal{F}_k$ -measurable and $g = \nabla f(x) + b + \Xi$, where

$$\Xi_i = \frac{1}{B} \sum_{\ell=1}^B \frac{s_i^{(\ell)}}{2h} \sum_{j=1}^m a_j (\xi_{i,j,\ell}^+ - \xi_{i,j,\ell}^-),$$

and:

(a) $\|b\|_\infty \leq \bar{b}_h := c_0 L_\beta h^{\beta-1} + A_1 M + S_1 \delta / h$ almost surely;

(b) for every i , $\mathbb{E}[\Xi_i | \mathcal{F}_{k-1}] = 0$ and $\mathbb{E}[\Xi_i^2 | \mathcal{F}_{k-1}] \leq C_\sigma \sigma^2 / (Bh^2)$; hence $\mathbb{E}[\Xi | \mathcal{F}_{k-1}] = 0$ and

$$\mathbb{E}[\|\Xi\|_\infty^2 | \mathcal{F}_{k-1}] \leq \mathbb{E}[\|\Xi\|_2^2 | \mathcal{F}_{k-1}] \leq C_\sigma d \sigma^2 / (Bh^2);$$

(c) if the noise satisfies (SG) and $\xi^o \equiv 0$, then

$$\mathbb{E}[\|\Xi\|_\infty^2 | \mathcal{F}_{k-1} \vee \sigma(\Theta_k)] \leq S_2(1 + \log(2d))\sigma^2 / (Bh^2) \leq (1 + 1/\log 2) S_2 \log(2d)\sigma^2 / (Bh^2),$$

and the same bound holds for $\mathbb{E}[\|\Xi\|_\infty^2 | \mathcal{F}_{k-1}]$.

Proof. Decomposition. The queries are functions of x and Θ_k , and g is a function of Θ_k and of the values observed in round k , so g is $\mathcal{F}_k$ -measurable. For $s \in \{-1, 1\}$ and every $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$,

$$\frac{s}{2h} \sum_j a_j (\varphi(x + ht_j s e_i) - \varphi(x - ht_j s e_i)) = D_{h,e_i} \varphi(x),$$

since for $s = -1$ both the factor s and the difference change sign. Inserting the values, with adversarial errors $\delta_{i,j,\ell}^\pm$, gives $g_i = \partial_i f(x) + b_i + \Xi_i$ with

$$b_i = D_{h,e_i} f(x) - \partial_i f(x) + D_{h,e_i} \eta(x) + \frac{1}{B} \sum_{\ell=1}^B \frac{s_i^{(\ell)}}{2h} \sum_j a_j (\delta_{i,j,\ell}^+ - \delta_{i,j,\ell}^-).$$

Part (a). Lemma C.1 in dimension one, applied to $u \mapsto f(x + ue_i)$, gives $|D_{h,e_i} f(x) - \partial_i f(x)| \leq c_0 L_\beta h^{\beta-1}$. By (N.4), $|D_{h,e_i} \eta(x)| \leq \frac{1}{2h} \sum_j |a_j| 2Mht_j = A_1 M$, and by (A) the last term of b_i is at most $S_1 \delta / h$ in absolute value. These bounds hold for every realization of the signs.

Part (b). This is the noise part of the proof of Lemma D.3, with random signs in place of random frames. Split $\Xi_i = \Xi_i^c + \Xi_i^o$ into its centered and oblivious parts. The query t at $X_{i,j,\ell}^\pm$ enters Ξ_i with the Θ_k -measurable coefficient $w_t = \pm s_i^{(\ell)} a_j / (2hB)$, and the $2mB$ queries of coordinate i have $\sum_t w_t^2 = S_2 / (2Bh^2)$. By (C), as there, the terms $w_t \xi_t^c$ are centered and pairwise orthogonal, so $\mathbb{E}[\Xi_i^c | \mathcal{F}_{k-1}] = 0$ and $\mathbb{N}(\Xi_i^c)^2 \leq S_2 \sigma^2 / (2Bh^2)$, with $\mathbb{N}$ as in Section D.1. For the oblivious part, $\Xi_i^o = B^{-1} \sum_\ell s_i^{(\ell)} o_i^{(\ell)}$ with $o_i^{(\ell)} = \frac{1}{2h} \sum_j a_j (\xi_{i,j,\ell}^{o,+} - \xi_{i,j,\ell}^{o,-})$. By (O) and the argument of Lemma D.1(a), the signs are independent of $\mathcal{F}_{k-1}$ and of the oblivious errors of the round, and they have mean zero and are independent of each other; since $\mathbb{N}(o_i^{(\ell)}) \leq S_1 \sigma / h$, this gives $\mathbb{E}[\Xi_i^o | \mathcal{F}_{k-1}] = 0$ and $\mathbb{N}(\Xi_i^o)^2 \leq S_1^2 \sigma^2 / (Bh^2)$. Minkowski's inequality gives (b) for Ξ_i , and summing over i and $\|\cdot\|_\infty \leq \|\cdot\|_2$ give the rest of (b).

Part (c). Let $\mathcal{G} = \mathcal{F}_{k-1} \vee \sigma(\Theta_k)$. Now $\Xi_i = \sum_t w_t \xi_t^c$, summed over the $2mB$ queries t of coordinate i in their order, which does not depend on Θ_k . Each w_t takes two values, w_t and $\sum_{t' < t} w_{t'} \xi_{t'}^c$ are $\mathcal{E}_t$ -measurable, and $\mathcal{E}_t \supseteq \mathcal{G}$. Applying (SG) to these queries from the last to the first gives $\mathbb{E}[e^{\theta \Xi_i} | \mathcal{G}] \leq e^{\theta^2 v / 2}$ for $\theta \in \mathbb{R}$, where $v = \sigma^2 \sum_t w_t^2 = S_2 \sigma^2 / (2Bh^2)$. The Chernoff bound and a union bound over i give $\mathbb{P}(\|\Xi\|_\infty > u | \mathcal{G}) \leq 2de^{-u^2 / (2v)}$, so, with $w_0 = 2v \log(2d)$,

$$\mathbb{E}[\|\Xi\|_\infty^2 | \mathcal{G}] \leq w_0 + \int_{w_0}^\infty 2de^{-w / (2v)} dw = 2v(1 + \log(2d)).$$

This is the first bound of (c). The second follows from $\log 2 \leq \log(2d)$, and conditioning on $\mathcal{F}_{k-1}$ gives the last claim. $\square$

The random signs are needed for the oblivious part of (b): without them ($s_i^{(\ell)} \equiv 1$), the oblivious errors $\pm \sigma \operatorname{sgn}(a_j)$ at the points $x \pm ht_j e_i$ give $\Xi_i = S_1 \sigma / h$ for every i . Without (SG), the factor d in (b) cannot be removed. Let $B = 1$, fix j with $a_j \neq 0$, let the centered errors at the queries $X_{i,j,1}^\pm$, $1 \leq i \leq d$, be independent of each other and of everything else, equal to $\pm \sqrt{d} \sigma$ with probability $1/(2d)$ each and to 0 otherwise, and let all other errors vanish. Then (C) holds, and $\|\Xi\|_\infty = |a_j| \sqrt{d} \sigma / (2h)$ unless all these errors vanish, so $\mathbb{E}[\|\Xi\|_\infty^2] \geq (1 - e^{-1}) a_j^2 d \sigma^2 / (4h^2)$.

Lemma N.8 (Entropic mirror descent on the ℓ_1 ball). *Let $R, \varsigma > 0$, let $g_1, g_2, \dots \in \mathbb{R}^d$ be arbitrary, and define w_k, x_k and G_k by Procedure 7. Then $x_k \in B_1^d(R)$ for all k , and for every $x^* \in B_1^d(R)$ and $N \geq 1$,*

$$\sum_{k=1}^N \langle g_k, x_k - x^* \rangle \leq \frac{\log(2d)}{\varsigma} + \frac{\varsigma R^2}{2} \sum_{k=1}^N \|g_k\|_\infty^2.$$

Proof. The update of w_k is the prox-mapping, applied to ςG_k , of the entropy on Δ_{2d} , which is 1-strongly convex with respect to $\|\cdot\|_1$ (Nemirovski et al., 2009, Section 2.3); the associated prox-function is $V(w, u) = \text{KL}(u \| w)$, the Kullback–Leibler divergence. By Nemirovski et al. (2009, Lemma 2.1) and $\|G_k\|_\infty = R \|g_k\|_\infty$, every $u \in \Delta_{2d}$ satisfies

$$\text{KL}(u \| w_{k+1}) \leq \text{KL}(u \| w_k) - \varsigma \langle G_k, w_k - u \rangle + \frac{\varsigma^2 R^2}{2} \|g_k\|_\infty^2.$$

Summing over $k \leq N$ and using $\text{KL}(u \| w_1) \leq \log(2d)$ bounds $\sum_k \langle G_k, w_k - u \rangle$ by the right-hand side of the claim. Moreover $\|x_k\|_1 \leq R \sum_\ell w_{k,\ell} = R$. Given $x^* \in B_1^d(R)$, the vector u with $u_i^\pm = (\pm x_i^*)_+ / R + (1 - \|x^*\|_1 / R) / (2d)$ lies in Δ_{2d} and satisfies $R(u^+ - u^-) = x^*$, so that

$$\langle G_k, w_k - u \rangle = \langle g_k, x_k - x^* \rangle, \quad (\text{N.5})$$

and the claim follows. $\square$

Theorem N.9 (Upper bound for the ℓ_1 geometry). *Let $\beta > 2$, $L_\beta, R, G_\infty > 0$, let $\bar{b}_h$ be as in Lemma N.7, and let $\bar{x}_N$ be the output of Procedure 7. Consider noises admissible at levels (M, σ, δ) , and put $V_h = C_\sigma d \sigma^2 / h^2$ in general and $V_h = S_2(1 + \log(2d)) \sigma^2 / h^2$ for those that satisfy (SG) and $\xi^o \equiv 0$.*

(a) *For all $h, \varsigma > 0$, $N \geq 1$, $f \in \mathcal{C}_1(R, G_\infty)$ and such noises,*

$$\mathbb{E}[f(\bar{x}_N) - f^*] \leq \frac{\log(2d)}{\varsigma N} + \varsigma R^2 ((G_\infty + \bar{b}_h)^2 + V_h) + 2R\bar{b}_h.$$

If $\bar{G}^2 \geq 2(G_\infty + \bar{b}_h)^2 + 2V_h$ and $\varsigma = (2 \log(2d) / (R^2 \bar{G}^2 N))^{1/2}$, then $\mathbb{E}[f(\bar{x}_N) - f^*] \leq R \sqrt{2 \log(2d) \bar{G}^2 / N} + 2R\bar{b}_h$.

(b) Let $0 < \varepsilon < G_\infty R$ and put

$$h_{1,\varepsilon} = \left(\frac{\varepsilon}{8c_0 L_\beta R} \right)^{\frac{1}{\beta-1}}, \quad \bar{G}_\varepsilon^2 = 2 \left(G_\infty + \frac{\varepsilon}{4R} \right)^2 + 2V_{h_{1,\varepsilon}},$$

$$N = \left\lceil \frac{8R^2 \log(2d) \bar{G}_\varepsilon^2}{\varepsilon^2} \right\rceil, \quad \varsigma = \left(\frac{2 \log(2d)}{R^2 \bar{G}_\varepsilon^2 N} \right)^{1/2}.$$

If $M \leq \varepsilon/(16A_1 R)$ and $\delta \leq \varepsilon h_{1,\varepsilon}/(16S_1 R)$, then $\mathbb{E}[f(\bar{x}_N) - f^*] \leq \varepsilon$ for every $f \in \mathcal{C}_1(R, G_\infty)$ and every such noise. The algorithm uses N rounds of $2md$ calls, all queries lie in $B_1^d(R + h_{1,\varepsilon})$, and $T = 2mdN$ satisfies, under (SG) and $\xi^\circ \equiv 0$,

$$T \leq 72m d \log(2d) \frac{G_\infty^2 R^2}{\varepsilon^2} + 79m S_2(8c_0)^{\frac{2}{\beta-1}} d \log^2(2d) \frac{\sigma^2}{\varepsilon^2} \left(\frac{L_\beta R^\beta}{\varepsilon} \right)^{\frac{2}{\beta-1}},$$

and in general

$$T \leq 72m d \log(2d) \frac{G_\infty^2 R^2}{\varepsilon^2} + 32m C_\sigma(8c_0)^{\frac{2}{\beta-1}} d^2 \log(2d) \frac{\sigma^2}{\varepsilon^2} \left(\frac{L_\beta R^\beta}{\varepsilon} \right)^{\frac{2}{\beta-1}}.$$

The parameters depend only on $\beta, L_\beta, R, G_\infty, \sigma, d, \varepsilon$. If $\varepsilon \geq G_\infty R$, the output $\hat{x} = 0$ satisfies $f(0) - f^* \leq \varepsilon$ without any call.

(c) Dimension dependence. Let $d \geq 256$, $\sigma > 0$, $G_\infty \geq G_d := 2C_H L_\beta R^{\beta-1} d^{-(\beta-2)/2}$, and

$$L_\beta R^\beta d^{-3\beta/4} \leq \varepsilon \leq c'_H L_\beta R^\beta d^{-\beta/2}.$$

For independent $N(0, \sigma^2)$ noise, let $T^*(\varepsilon)$ be the minimax number of calls on $\mathcal{C}_1(R, G_\infty)$. Then

$$T^*(\varepsilon) \geq c_H d (L_\beta R^\beta / \varepsilon)^{2/\beta} \sigma^2 / \varepsilon^2,$$

and the sub-Gaussian upper bound in (b) also bounds $T^*(\varepsilon)$. Here $\varepsilon < G_d R \leq G_\infty R$; the window is nonempty exactly when $d \geq (c'_H)^{-4/\beta}$.

Proof. Part (a). Let $x^* \in \arg \min f \cap B_1^d(R)$. Convexity, Lemma N.8 and the decomposition $g_k = \nabla f(x_k) + b_k + \Xi_k$ of Lemma N.7 give

$$N(f(\bar{x}_N) - f^*) \leq \sum_{k=1}^N \langle \nabla f(x_k), x_k - x^* \rangle \leq \frac{\log(2d)}{\varsigma} + \frac{\varsigma R^2}{2} \sum_{k=1}^N \|g_k\|_\infty^2 + 2RN\bar{b}_h - \sum_{k=1}^N \langle \Xi_k, x_k - x^* \rangle,$$

since $\|x_k - x^*\|_1 \leq 2R$. The last sum has expectation zero by Lemma N.7(b), because $x_k - x^*$ is bounded and $\mathcal{F}_{k-1}$ -measurable, and $\mathbb{E}\|g_k\|_\infty^2 \leq 2(G_\infty + \bar{b}_h)^2 + 2V_h$, because $\|\nabla f(x_k)\|_\infty \leq G_\infty$. This proves the first bound of (a); the second follows by inserting ς .

Part (b). The choices of $h_{1,\varepsilon}$, M and δ give $\bar{b}_{h_{1,\varepsilon}} \leq \varepsilon/(8R) + \varepsilon/(16R) + \varepsilon/(16R) = \varepsilon/(4R)$, and the choice of N makes both terms of the tuned bound in (a) at most $\varepsilon/2$. Each query differs from a point of $B_1^d(R)$ in one coordinate by at most $h_{1,\varepsilon}$. Since $2(G_\infty + \varepsilon/(4R))^2 \leq 4G_\infty^2 + \varepsilon^2/(4R^2)$,

$$N \leq 1 + 2 \log(2d) + 32 \log(2d) \frac{G_\infty^2 R^2}{\varepsilon^2} + 16 \log(2d) \frac{R^2 V_{h_{1,\varepsilon}}}{\varepsilon^2}.$$

As $\varepsilon < G_\infty R$ and $1 \leq \log(2d)/\log 2$, multiplying by $2md$ bounds the first three terms by $72md \log(2d) G_\infty^2 R^2 / \varepsilon^2$. The two call bounds follow by inserting V_h , using $R^2 h_{1,\varepsilon}^{-2} = (8c_0)^{2/(\beta-1)} (L_\beta R^\beta / \varepsilon)^{2/(\beta-1)}$ and, in the sub-Gaussian case, $1 + \log(2d) \leq (1 + 1/\log 2) \log(2d)$ and $32(1 + 1/\log 2) < 79$. Finally, if $\varepsilon \geq G_\infty R$, convexity gives $f(0) - f^* \leq \|\nabla f(0)\|_\infty \|x^*\|_1 \leq G_\infty R$.

Part (c). The window is the condition of the last assertion of Lemma N.6(d). By the subclass assertion of Lemma N.6 with $p = 1$ and $n = d$, that lower bound holds on the subclass of $\mathcal{C}_p(R)$, $p = 1$, defined by $\|\nabla f(x)\|_\infty \leq C_H L_\beta r^{\beta-2} (\|x\|_2 + r)$ (and a Euclidean Hessian cap), where $r = R/\sqrt{d}$. For $x \in B_1^d(R)$ the right-hand side is at most $2C_H L_\beta r^{\beta-2} R = G_d$, so this subclass is contained in $\mathcal{C}_1(R, G_d) \subseteq \mathcal{C}_1(R, G_\infty)$. Since the lower bound is positive, no algorithm without calls is ε -accurate on $\mathcal{C}_1(R, G_d)$, whereas the output 0 has error at most $G_d R$ there (last assertion of (b) with G_d in place of G_∞); hence $\varepsilon < G_d R$. The upper bound is part (b) for Gaussian noise, which satisfies (SG) with $\xi^\circ \equiv 0$. The window is nonempty exactly when $d^{\beta/4} \geq 1/c'_H$. $\square$

Remark N.10 (Local smoothness suffices). The proof of Lemma N.9(a), (b) uses the Hölder condition only in Lemma N.7(a), through the functions $u \mapsto f(x_k + ue_i)$ on $[-h, h]$ with $x_k \in B_1^d(R)$; all queries lie in $B_1^d(R + h)$. Hence (a) and (b) hold for every f that is defined, convex and continuously differentiable on an open set containing $B_1^d(R + h)$, has a minimizer in $B_1^d(R)$ and $\|\nabla f\|_\infty \leq G_\infty$ on $B_1^d(R)$, and for which every function $u \mapsto f(x + ue_i)$, $x \in B_1^d(R)$, has a γ -Hölder q -th derivative with constant L_β on $[-h, h]$. Experiment 6 of Section P uses this form: for $f(x) = \sum_j \varphi(x_j - c_j)$ with $\varphi(t) = e^t - 1 - t$ the global seminorm is infinite, while the local constant is the largest value of φ''' at the coordinates that occur.

Remark N.11 (Comparison of the bounds). For fixed $\beta, L_\beta, R, \sigma, \varepsilon$ and $G_\infty \geq G_d$, both bounds of Lemma N.9(c) are linear in d up to the factor $\log^2(2d)$ on the whole range $(L_\beta R^\beta / \varepsilon)^{4/(3\beta)} \leq d \leq (c'_H L_\beta R^\beta / \varepsilon)^{2/\beta}$ of the window. Their ratio is at most

$$C_\beta \log^2(2d) (L_\beta R^\beta / \varepsilon)^{2/(\beta-1)-2/\beta} + C_\beta \log(2d) G_\infty^2 R^2 (\varepsilon / (L_\beta R^\beta))^{2/\beta} / \sigma^2,$$

so the remaining gap concerns the power of $1/\varepsilon$, which is $2 + 2/\beta$ in the lower bound and $2 + 2/(\beta - 1)$ in the upper bound. For larger d the lower bound of Lemma N.6(d) no longer grows with d , up to the logarithmic gain of Lemma N.6(e), while the upper bound still grows linearly. For noise that is only admissible at levels (M, σ, δ) , the upper bound grows like $d^2 \log(2d)$, and by the example after Lemma N.7 the second moment of $\|\Xi\|_\infty$ can indeed be of order $d\sigma^2/h^2$. Since $s_H \leq 1/2$, the threshold of Lemma N.9(c) satisfies $(c'_H)^{-4/\beta} \geq 2^{4+72/\beta} K_H^{4/\beta}$ (for instance $2^{28} K_H^{4/3}$ at $\beta = 3$); with the constants of Lemma M.5, part (c) is therefore a statement about large dimensions.

Upper bounds for $p > 2$ are open. For $p = 2$ and fixed L_β, R, σ , the known convex upper bound is $O(d^2 \varepsilon^{-2\beta/(\beta-1)})$ (Bach and Perchet, 2016; Gasnikov et al., 2024a), and Lemmas L.6 and L.9 give a sharper comparison when a Hessian bound is available.

O FUNCTIONAL CONSTRAINTS

This appendix treats convex problems with functional constraints whose values are observed with noise. Section O.1 introduces the problem, the multipliers and the oracle. Section O.2 shows that no method controls both the objective error and the violation if the values of the constraints carry an unknown constant error. Section O.3 analyzes stochastic Mirror-Prox with biased operator estimates, and Section O.4 builds these estimates from Procedure 1 and calibrated values of the constraints. Lower bounds for problems whose optimal value is not changed by the constraints are given by Lemma N.4.

O.1 Problem, multipliers and oracle

Problem. Let $Q \subset \mathbb{R}^d$ be nonempty, closed and convex with diameter $D \in (0, \infty)$, and let $g_0 = f, g_1, \dots, g_{n_c}$ ($n_c \geq 1$) be real functions that are differentiable on an open set containing Q and convex on Q . We consider

$$\text{minimize } f(x) \text{ over } x \in Q \text{ subject to } g_i(x) \leq 0, \quad i = 1, \dots, n_c, \quad (\text{O.1})$$

and write $g = (g_1, \dots, g_{n_c})$. We assume that the feasible set $\mathcal{X} = \{x \in Q : g(x) \leq 0\}$ is nonempty. It is compact, so $f^* = \min_{\mathcal{X}} f$ is attained at some $x^* \in \mathcal{X}$. The *violation* of $x \in Q$ is $v(x) = \max\{0, g_1(x), \dots, g_{n_c}(x)\}$. A method outputs $\hat{x} \in Q$; we bound $\mathbb{E}[f(\hat{x}) - f^*]_+$ and $\mathbb{E}v(\hat{x})$, where $[a]_+ = \max\{a, 0\}$. The objective error $f(\hat{x}) - f^*$ may be negative at infeasible points.

Multipliers. We assume a known number $\bar{y} \geq 0$ and a vector $y^* \in \mathbb{R}^{n_c}$ with

$$y^* \geq 0, \quad \|y^*\|_1 \leq \bar{y}, \quad \inf_{x \in Q} (f(x) + \langle y^*, g(x) \rangle) = f^*. \quad (\text{O.2})$$

Such a vector is a Lagrange multiplier of (O.1); Lemma O.3 discusses its existence and its size. Fix $\rho > \bar{y}$; the method below uses $\rho = \bar{y} + 1$. Put

$$Y_\rho = \{y \in \mathbb{R}^{n_c} : y \geq 0, \|y\|_1 \leq \rho\}, \quad K = Q \times Y_\rho \subset \mathbb{R}^{d+n_c}, \quad D_K = (D^2 + 2\rho^2)^{1/2}.$$

We use the Euclidean norm on $\mathbb{R}^{d+n_c}$ and write $w = (x, y)$ for its points; the Euclidean projection onto K is $\text{Proj}_K(x, y) = (\text{Proj}_Q x, \text{Proj}_{Y_\rho} y)$. For $y, y' \in Y_\rho$ we have $\langle y, y' \rangle \geq 0$, hence $\|y - y'\|^2 \leq \|y\|_1^2 + \|y'\|_1^2 \leq 2\rho^2$, and D_K bounds the diameter of K .

Lagrangian, operator and gap. For $(x, y) \in K$ let

$$\Psi(x, y) = f(x) + \langle y, g(x) \rangle, \quad H(x, y) = \left(\nabla f(x) + \sum_{i=1}^{n_c} y_i \nabla g_i(x), -g(x) \right).$$

For each $y \in Y_\rho$ the function $\Psi(\cdot, y)$ is convex on Q , and $\Psi(x, \cdot)$ is affine with gradient $g(x)$. We assume that H is M_H -Lipschitz on K for some $M_H > 0$, that is, $\|H(w) - H(w')\| \leq M_H \|w - w'\|$ for $w, w' \in K$; Lemma [O.2](#) gives a sufficient condition. The *gap* of $w = (x, y) \in K$ is

$$\text{gap}(x, y) = \max_{y' \in Y_\rho} \Psi(x, y') - \min_{x' \in Q} \Psi(x', y) = \max_{(x', y') \in K} (\Psi(x, y') - \Psi(x', y)) \geq 0.$$

Lemma O.1 (The gap controls the objective error and the violation). *Assume [\(O.2\)](#). For every $(x, y) \in K$,*

$$f(x) - f^* + \rho v(x) \leq \text{gap}(x, y) \quad \text{and} \quad f(x) - f^* \geq -\bar{y} v(x).$$

Consequently $[f(x) - f^*]_+ \leq \text{gap}(x, y)$ and $v(x) \leq \text{gap}(x, y)/(\rho - \bar{y})$.

Proof. The support function of Y_ρ gives $\max_{y' \in Y_\rho} \langle y', g(x) \rangle = \rho v(x)$, and feasibility of x^* gives $\min_{x' \in Q} \Psi(x', y) \leq \Psi(x^*, y) \leq f^*$. This proves the first inequality. The multiplier condition gives $f(x) - f^* \geq -\langle y^*, g(x) \rangle \geq -\bar{y} v(x)$. The two conclusions follow from these inequalities and $\text{gap} \geq 0$. $\square$

Lemma O.2 (A Lipschitz constant of the operator). *Let $f, g_1, \dots, g_{n_c}$ be twice continuously differentiable on an open set containing Q , with $0 \preceq \nabla^2 g_i(x) \preceq LI$ for $x \in Q$ and $i = 0, \dots, n_c$, and $\|\nabla g_i(x)\| \leq G_g$ for $x \in Q$ and $i = 1, \dots, n_c$. Then H is M_H -Lipschitz on K with $M_H = (1 + \rho)L + 2\sqrt{n_c} G_g$.*

Proof. For $w = (x, y)$ and $w' = (x', y')$ in K , the Hessian bounds, integration along segments and $\|y - y'\|_1 \leq \sqrt{n_c} \|y - y'\|$ show that the two blocks of $H(w) - H(w')$ have norms at most $(1 + \rho)L \|x - x'\| + \sqrt{n_c} G_g \|y - y'\|$ and $\sqrt{n_c} G_g \|x - x'\|$. Add these bounds and use $\|x - x'\|, \|y - y'\| \leq \|w - w'\|$. $\square$

Remark O.3 (The size of the multipliers). *Slater points.* Let $x_o \in Q$ and $\theta > 0$ with $g_i(x_o) \leq -\theta$ for all i . Every y^* as in [\(O.2\)](#) satisfies $\|y^*\|_1 \leq (f(x_o) - f^*)/\theta$, because $f^* \leq f(x_o) + \langle y^*, g(x_o) \rangle \leq f(x_o) - \theta \|y^*\|_1$. If $\|\nabla f\| \leq G_0$ on Q , then $f(x_o) - f^* \leq G_0 D$, so $\bar{y} = G_0 D/\theta$ is admissible. If x_o lies in the relative interior of Q , a vector y^* as in [\(O.2\)](#) exists ([Rockafellar, 1970](#), Theorem 28.2).

Smoothness does not bound the multipliers. On $Q = [0, 1] \subset \mathbb{R}$, for $c > 0$, the problem of minimizing $f(x) = -cx$ subject to $x - \frac{1}{2} \leq 0$ has $f^* = -c/2$. For $y \geq 0$ the infimum over $[0, 1]$ of $f(x) + y(x - \frac{1}{2}) = (y - c)x - y/2$ is $-y/2$ if $y \geq c$ and $y/2 - c$ if $y < c$, so the unique multiplier is $y^* = c$. All derivatives of order at least two of f and of the constraint vanish, so the Hessian bound and the Hölder constants are zero for every c . Thus $\bar{y}$ is a parameter of the problem that is not controlled by the smoothness parameters.

Oracle. The method works in rounds as in Section [2](#). Here $\mathcal{F}_{k-1}$ is generated by all queries and all returned values of rounds $1, \dots, k-1$ and by $\Theta_1, \dots, \Theta_{k-1}$. The queries of a round are ordered in a way that does not depend on Θ_k . For a query t of round k , the σ -algebra $\mathcal{E}_t$ is generated by $\mathcal{F}_{k-1}$, Θ_k , the queries of round k and all error components of the queries of round k that precede t .

Definition O.4 (Value calls and level calls). A query is of one of two kinds.

- (i) A *value call* at X_t returns the $n_c + 1$ values

$$Y_{i,t} = g_i(X_t) + \eta_i(X_t) + \xi_{i,t}^c + \xi_{i,t}^o + \delta_{i,t}, \quad i = 0, \dots, n_c,$$

where, for each i , η_i is fixed before the run and the four components satisfy [\(L\)](#)–[\(A\)](#) at levels (M, σ, δ) with the σ -algebras above; in [\(O\)](#), the vector of the errors $\xi_{i,t}^o$ of the value calls of round k is independent of Θ_k given $\mathcal{F}_{k-1}$.

- (ii) A *level call* at X_t returns, for a level $\sigma_g \geq 0$, the n_c values

$$Z_{i,t} = g_i(X_t) + \xi_{i,t}^c + \delta_{i,t}, \quad i = 1, \dots, n_c,$$

with $\mathbb{E}[\xi_{i,t}^c \mid \mathcal{E}_t] = 0$, $\mathbb{E}[(\xi_{i,t}^c)^2 \mid \mathcal{E}_t] \leq \sigma_g^2$ and $|\delta_{i,t}| \leq \delta$.

The components of one call may depend on each other. We count the calls of each kind and the returned scalar values.

Thus the values that enter differences may carry all four kinds of error of Definition 1, while the levels $g_i(x)$ are observed with centered noise and a bounded adversarial error only. A level call is a value call whose constraint components have neither a Lipschitz nor an oblivious part, with the objective component discarded; an oracle whose constraint components satisfy these restrictions provides both kinds of calls, with $\sigma_g = \sigma$. Lemma O.5 shows that some restriction of this kind on the levels is necessary.

O.2 The levels of the constraints must be calibrated

Proposition O.5 (A constant error in the constraint values). *Let $d = n_c = 1$, $Q = [-1, 1]$, $f(x) = -x$ and $0 < a < 1$. For $\theta \in \{-a, a\}$ let $g_\theta(x) = x - \theta$, and let f_θ^* and $v_\theta(x) = [x - \theta]_+$ be the optimal value and the violation of the problem (O.1) with the constraint g_θ .*

- (a) *Both problems satisfy the assumptions of Section O.1 with $\bar{y} = 1$: the minimizer is $x_\theta^* = \theta$, $f_\theta^* = -\theta$, the vector $y^* = 1$ satisfies (O.2), and $x_o = -1$ satisfies $g_\theta(x_o) \leq -(1 - a)$. All derivatives of order at least two of f and g_θ vanish.*
- (b) *Let the values of the objective be exact, and let the values of the constraint carry the fixed error $\eta_\theta \equiv \theta$, which is a constant and hence M -Lipschitz for every $M \geq 0$; all other error components vanish. Then every returned value of the constraint at a query X equals X , whatever θ . Consequently, for every method, whatever its number of calls and rounds and wherever it queries, the output $\hat{x} \in Q$ has the same law under both problems, and*

$$\max_{\theta \in \{-a, a\}} \max \left\{ \mathbb{E}_\theta [f(\hat{x}) - f_\theta^*]_+, \mathbb{E}_\theta v_\theta(\hat{x}) \right\} \geq a.$$

The output $\hat{x} = 0$ attains equality.

- (c) *The same conclusion holds if the constant θ is, instead of η_θ , an oblivious error $\xi_t^\circ = \theta$ of every constraint value, which is admissible at levels $(0, \sigma, 0)$ for $\sigma \geq a$, or an adversarial error $\delta_t = \theta$, which is admissible at levels $(0, 0, \delta)$ for $\delta \geq a$.*

Proof. (a) The feasible set is $[-1, \theta]$ and f is decreasing, so $x_\theta^* = \theta$ and $f_\theta^* = -\theta$. For every x , $f(x) + g_\theta(x) = -\theta = f_\theta^*$, so $y^* = 1$ satisfies (O.2) with $\bar{y} = 1$. Finally $g_\theta(-1) = -1 - \theta \leq -1 + a$, and f, g_θ are affine.

(b) The returned value of the constraint at X is $g_\theta(X) + \theta = X$, and that of the objective is $-X$. So under both problems every returned value is the same function of the query. By induction over the rounds, the queries, the returned values and the output are the same measurable functions of the random elements $\Theta_1, \Theta_2, \dots$ of the method under both problems, and their laws coincide. For every $x \in Q$,

$$[f(x) - f_a^*]_+ + v_{-a}(x) = [a - x]_+ + [x + a]_+ \geq (a - x) + (x + a) = 2a.$$

Taking expectations under the common law of $\hat{x}$ gives $\mathbb{E}_a[f(\hat{x}) - f_a^*]_+ + \mathbb{E}_{-a}v_{-a}(\hat{x}) \geq 2a$, so one of the two terms is at least a . For $\hat{x} = 0$ we have $[f(0) - f_\theta^*]_+ = [\theta]_+$ and $v_\theta(0) = [-\theta]_+$, so the maximum over θ and over the two errors equals a .

(c) In both cases the returned value of the constraint at X is again X , and the argument of (b) applies. A constant error is independent of the randomness of the method and has second moment a^2 . $\square$

Remark O.6 (Relation to [Abbronin et al. \(2026\)](#)). [Abbronin et al. \(2026\)](#) study (O.1) with $f(x) = \mathbb{E}F(x, \xi)$ and $g(x) = \mathbb{E}G(x, \xi)$ for an oracle that returns $(F(x, \xi), G(x, \xi))$ at two points with a common sample ξ . Their gradient estimates use differences at a common sample, which cancel every error that does not depend on the point, and their Lipschitz assumption on the samples (their Assumption 2.2) concerns only such differences. The levels enter the dual updates; for them they assume $\mathbb{E}\|G(x, \xi) - g(x)\|_\infty^2 \leq \sigma_g^2$ (their Assumption 2.3), which they call a genuine additional assumption (their Remark 2.7(ii)). Lemma O.5 shows that an assumption of this kind cannot be dropped. Our level calls play the role of their Assumption 2.3, with an additional adversarial error; our bound on each component implies theirs with $n_c \sigma_g^2$ in place of σ_g^2 .

Their Theorem 9.3 shows that the term of order σ_g^2/ε^2 in the number of level calls of Lemma O.9 cannot be avoided: its instances $G_i(x, \xi) = x - \xi_i$ are realized by our level calls, while under the constant Lipschitz errors $\eta_i \equiv \mathbb{E}\xi_i$ the value calls carry no information on the levels.

O.3 Mirror–Prox with biased operator estimates

The method is the stochastic Mirror–Prox method of Juditsky et al. (2011) in the Euclidean setup, that is, a stochastic extragradient method with projections onto K . Each iteration takes two rounds, because the second estimate is requested at a point that depends on the first one.

Theorem O.7 (Stochastic Mirror–Prox with bounded bias). *Assume the setting of Section O.1, including (O.2). Let $(\mathcal{F}_k)_{k \geq 0}$ be a filtration. In round k the method submits an $\mathcal{F}_{k-1}$ -measurable point $w \in K$ and receives an $\mathcal{F}_k$ -measurable estimate*

$$\hat{H} = H(w) + \mu_k + \zeta_k,$$

where ζ_k is $\mathcal{F}_k$ -measurable, $\mathbb{E}[\zeta_k | \mathcal{F}_{k-1}] = 0$, $\mathbb{E}[\|\zeta_k\|^2 | \mathcal{F}_{k-1}] \leq V_K$, and $\|\mu_k\| \leq b_K$ almost surely. Let $w_1 \in K$ be deterministic and $0 < \nu \leq 1/(2M_H)$. For $t = 1, \dots, N$:

- in round $2t - 1$, obtain the estimate $\hat{H}_t$ at w_t and put $w_{t+1/2} = \text{Proj}_K(w_t - \nu \hat{H}_t)$;
- in round $2t$, obtain the estimate $\hat{H}_{t+1/2}$ at $w_{t+1/2}$ and put $w_{t+1} = \text{Proj}_K(w_t - \nu \hat{H}_{t+1/2})$.

Let $\hat{w}_N = (\hat{x}_N, \hat{y}_N) = \frac{1}{N} \sum_{t=1}^N w_{t+1/2}$. Then

$$\mathbb{E} \text{gap}(\hat{w}_N) \leq \frac{D_K^2}{2\nu N} + 2\nu V_K + D_K \sqrt{\frac{V_K}{N}} + 3D_K b_K. \quad (\text{O.3})$$

With $\nu = \min\{1/(2M_H), D_K/(2\sqrt{NV_K})\}$, and $\nu = 1/(2M_H)$ if $V_K = 0$,

$$\mathbb{E} \text{gap}(\hat{w}_N) \leq \text{Err}_N := \frac{M_H D_K^2}{N} + 3D_K \sqrt{\frac{V_K}{N}} + 3D_K b_K. \quad (\text{O.4})$$

Consequently $\hat{x}_N \in Q$, $\mathbb{E}[f(\hat{x}_N) - f^*]_+ \leq \text{Err}_N$ and $\mathbb{E}v(\hat{x}_N) \leq \text{Err}_N/(\rho - \bar{y})$. The method uses $2N$ rounds.

Proof. A bound of this form is Juditsky et al. (2011, Theorem 1), whose oracle assumption (Assumption I there) also allows a bias; we give the short argument because here the errors are martingale differences with a history-dependent bias and because its constants are used below. Fix $u \in K$ and put $r = \|w_t - w_{t+1/2}\|$ and $r' = \|w_{t+1/2} - w_{t+1}\|$. Adding the projection inequalities for $w_{t+1/2}$, tested at w_{t+1} , and for w_{t+1} , tested at u , bounds $\nu \langle \hat{H}_{t+1/2}, w_{t+1/2} - u \rangle$ by $(\|w_t - u\|^2 - \|w_{t+1} - u\|^2 - r^2 - r'^2)/2$ plus $\nu \langle \hat{H}_{t+1/2} - \hat{H}_t, w_{t+1/2} - w_{t+1} \rangle$. In this term the operator difference contributes at most $\nu M_H r r' \leq (r^2 + r'^2)/4$, the two biases at most $2\nu b_K D_K$, and the noise at most $\nu^2 \|\zeta_{t+1/2} - \zeta_t\|^2 + r'^2/4$. Replacing $\hat{H}_{t+1/2}$ by $H(w_{t+1/2})$ on the left costs at most $\nu b_K D_K$ and a martingale term, so

$$\langle H(w_{t+1/2}), w_{t+1/2} - u \rangle \leq \frac{\|w_t - u\|^2 - \|w_{t+1} - u\|^2}{2\nu} + \nu \|\zeta_{t+1/2} - \zeta_t\|^2 + 3b_K D_K - \langle \zeta_{t+1/2}, w_{t+1/2} - u \rangle. \quad (\text{O.5})$$

For $u = (x', y')$, convexity in x and affinity in y give $\langle H(x, y), (x, y) - u \rangle \geq \Psi(x, y') - \Psi(x', y)$. Sum (O.5), use Jensen's inequality, write $w_{t+1/2} - u = (w_{t+1/2} - u_0) - (u - u_0)$ for a fixed $u_0 \in K$, and take the supremum over u :

$$N \text{gap}(\hat{w}_N) \leq D_K^2/(2\nu) + \nu \sum_t \|\zeta_{t+1/2} - \zeta_t\|^2 + 3Nb_K D_K - \sum_t \langle \zeta_{t+1/2}, w_{t+1/2} - u_0 \rangle + D_K \left\| \sum_t \zeta_{t+1/2} \right\|.$$

The inner products have mean zero. Martingale differences are orthogonal, so $\mathbb{E} \|\zeta_{t+1/2} - \zeta_t\|^2 \leq 2V_K$ and $\mathbb{E} \|\sum_t \zeta_{t+1/2}\| \leq \sqrt{NV_K}$, which proves (O.3). For the tuned step, $1/\nu \leq 2M_H + 2\sqrt{NV_K}/D_K$ and $2\nu V_K \leq D_K \sqrt{V_K/N}$, which proves (O.4). Lemma O.1 gives the last two bounds. $\square$

O.4 Operator estimates from the stencil

Fix a round k , a radius $h > 0$, integers $B, B_L \geq 1$ and an $\mathcal{F}_{k-1}$ -measurable point $w = (x, y) \in K$. The operator estimate $\hat{H}(w) = (\hat{H}^x, \hat{H}^y)$ at w consists of two parts.

- (i) *Value calls.* Run Procedure 1 at the center x with radius h and B frames, and apply its formula to each of the $n_c + 1$ components of the returned values; this gives vectors $\bar{g}_0, \dots, \bar{g}_{n_c} \in \mathbb{R}^d$. Put $\hat{H}^x = \bar{g}_0 + \sum_{i=1}^{n_c} y_i \bar{g}_i$. Since the formula is linear in the values, $\hat{H}^x$ is Procedure 1 applied to the values of the Lagrangian $\Psi(\cdot, y)$ at the frozen multiplier y . This part uses $2mdB$ value calls and $(n_c + 1)2mdB$ scalar values.
- (ii) *Level calls.* Request B_L level calls at x , with returned vectors $Z_{\cdot, \ell} \in \mathbb{R}^{n_c}$, $\ell = 1, \dots, B_L$, and put $\hat{H}^y = -\frac{1}{B_L} \sum_{\ell=1}^{B_L} Z_{\cdot, \ell}$. This part uses B_L level calls and $n_c B_L$ scalar values.

All queries of the estimate are determined at the start of the round, so the estimate takes one round; the value calls precede the level calls.

Lemma O.8 (Bias and variance of the operator estimate). *Let ϑ be as in Lemma L.2, and let $g_0 = f, g_1, \dots, g_{n_c}$ be q times continuously differentiable on an open set containing $Q_{\vartheta h}$, with $\|D^q g_i(x) - D^q g_i(x')\| \leq L_\beta \|x - x'\|^\gamma$ for $x, x' \in Q_{\vartheta h}$ and $i = 0, \dots, n_c$, in the chosen norm of Section 2 (with the corresponding exponent τ). Let the calls satisfy Lemma O.4. Then the value calls lie in Q_h , the level calls are made at $x \in Q$, and $\hat{H}(w) = H(w) + \mu + \zeta$, where μ is $\mathcal{F}_{k-1}$ -measurable, ζ is $\mathcal{F}_k$ -measurable, $\mathbb{E}[\zeta \mid \mathcal{F}_{k-1}] = 0$,*

$$\|\mu\| \leq b_K := (1 + \rho) \left(C_b L_\beta d^{-\tau/2} h^{\beta-1} + A_1 M + \frac{\sqrt{d} S_1 \delta}{h} \right) + \sqrt{n_c} \delta, \quad (\text{O.6})$$

and, with $\Omega_h = C_f L_\beta^2 d^{-\tau} h^{2\beta-2} + C_\eta M^2 + C_\sigma d \sigma^2 / h^2$, where $C_b, C_f, C_\eta = 4A_1^2$ and C_σ are the constants of Lemma D.3,

$$\mathbb{E}[\|\zeta\|^2 \mid \mathcal{F}_{k-1}] \leq V_K := 2(1 + \rho)^2 \left(\frac{\Omega_h}{B} + \frac{d S_1^2 \delta^2}{h^2} \right) + 2n_c \left(\frac{\sigma_g^2}{B_L} + \delta^2 \right). \quad (\text{O.7})$$

Proof. Write $\mathbf{N}(X) = (\mathbb{E}[\|X\|^2 \mid \mathcal{F}_{k-1}])^{1/2}$.

Primal block. Fix $i \in \{0, \dots, n_c\}$. The i -th components of the value calls form an oracle for g_i as in Definition 1 at levels (M, σ, δ) , with the σ -algebras of Lemma O.4. The proofs of Lemma D.3 and Lemma L.2 use the σ -algebras only through the $\mathcal{F}_{k-1}$ -measurability of the center and the radius, the independence of Θ_k and $\mathcal{F}_{k-1}$, the condition (C) with σ -algebras $\mathcal{E}_t$ that contain Θ_k and the errors of the earlier queries of the round, the condition (O), and the fact that the order of the queries does not depend on Θ_k . These properties hold here; the larger σ -algebras and the level calls, which follow the value calls, do not affect them. Hence Lemma L.2, with its convex set (called K there) equal to Q , and part (iv) of Lemma D.3 give $\bar{g}_i = \nabla g_i(x) + \mu_i + \tilde{\zeta}_i$, where $\mu_i = b_i(x) + \mathbb{E}[\Delta_i \mid \mathcal{F}_{k-1}]$ is $\mathcal{F}_{k-1}$ -measurable with $\|\mu_i\| \leq C_b L_\beta d^{-\tau/2} h^{\beta-1} + A_1 M + \sqrt{d} S_1 \delta / h$, and $\tilde{\zeta}_i = \zeta_i + \Delta_i - \mathbb{E}[\Delta_i \mid \mathcal{F}_{k-1}]$ is $\mathcal{F}_k$ -measurable with $\mathbb{E}[\tilde{\zeta}_i \mid \mathcal{F}_{k-1}] = 0$ and $\mathbf{N}(\tilde{\zeta}_i) \leq (\Omega_h / B)^{1/2} + \sqrt{d} S_1 \delta / h$; here $b_i(x)$, ζ_i and Δ_i are the terms $b(z)$, ζ and Δ of Lemma D.3 for g_i . With $y_0 = 1$,

$$\hat{H}^x = \nabla f(x) + \sum_{i \geq 1} y_i \nabla g_i(x) + \mu^x + \zeta^x, \quad \mu^x = \sum_{i=0}^{n_c} y_i \mu_i, \quad \zeta^x = \sum_{i=0}^{n_c} y_i \tilde{\zeta}_i.$$

Since y is $\mathcal{F}_{k-1}$ -measurable, $y \geq 0$ and $\sum_{i=0}^{n_c} y_i \leq 1 + \rho$, we have $\|\mu^x\| \leq (1 + \rho) \max_i \|\mu_i\|$ and $\mathbb{E}[\zeta^x \mid \mathcal{F}_{k-1}] = 0$, and the conditional Minkowski inequality gives $\mathbf{N}(\zeta^x)^2 \leq 2(1 + \rho)^2 (\Omega_h / B + d S_1^2 \delta^2 / h^2)$.

Dual block. Let $\bar{\xi}$ and Δ^L be the averages over the B_L level calls of the vectors $(\xi_{i, \ell}^c)_i$ and $(\delta_{i, \ell})_i$ of their errors. Then $\hat{H}^y = -g(x) + \mu^y + \zeta^y$ with $\mu^y = -\mathbb{E}[\Delta^L \mid \mathcal{F}_{k-1}]$ and $\zeta^y = -(\bar{\xi} + \Delta^L - \mathbb{E}[\Delta^L \mid \mathcal{F}_{k-1}])$, so that $\|\mu^y\| \leq \sqrt{n_c} \delta$, ζ^y is $\mathcal{F}_k$ -measurable and $\mathbb{E}[\zeta^y \mid \mathcal{F}_{k-1}] = 0$. For $\ell < \ell'$ the error $\xi_{i, \ell}^c$ is measurable with respect to the σ -algebra $\mathcal{E}_t$ of the ℓ' -th level call, so the cross terms vanish and $\mathbf{N}(\bar{\xi})^2 \leq n_c \sigma_g^2 / B_L$. Moreover $\mathbf{N}(\Delta^L - \mathbb{E}[\Delta^L \mid \mathcal{F}_{k-1}])^2 \leq n_c \delta^2$, and the conditional Minkowski inequality gives $\mathbf{N}(\zeta^y)^2 \leq 2n_c (\sigma_g^2 / B_L + \delta^2)$.

Both blocks. Put $\mu = (\mu^x, \mu^y)$ and $\zeta = (\zeta^x, \zeta^y)$. Then $\|\mu\| \leq \|\mu^x\| + \|\mu^y\| \leq b_K$ and $\mathbb{E}[\|\zeta\|^2 \mid \mathcal{F}_{k-1}] = \mathbf{N}(\zeta^x)^2 + \mathbf{N}(\zeta^y)^2 \leq V_K$; no independence between the blocks is needed. The value calls are the queries of Procedure 1 at $x \in Q$ and lie in Q_h . $\square$

Procedure 8 Calibrated zeroth-order Mirror-Prox

Require: $x_0 \in Q$, Proj_K ; value and level calls as in Lemma O.4; h, B, B_L, N, ν as in Lemma O.9

- 1: $w_1 \leftarrow (x_0, 0)$
 - 2: **for** $t = 1, \dots, N$ **do**
 - 3: $a_t \leftarrow \hat{H}(w_t)$, the operator estimate of Section O.4 (one round)
 - 4: $w_{t+1/2} \leftarrow \text{Proj}_K(w_t - \nu a_t)$
 - 5: $b_t \leftarrow \hat{H}(w_{t+1/2})$ (one round)
 - 6: $w_{t+1} \leftarrow \text{Proj}_K(w_t - \nu b_t)$
 - 7: **return** the x -component of $N^{-1} \sum_{t=1}^N w_{t+1/2}$
-

Corollary O.9 (Mirror-Prox with the stencil). *Assume the setting of Section O.1 with $\rho = \bar{y} + 1$, so that $D_K = (D^2 + 2\rho^2)^{1/2}$, and let H be M_H -Lipschitz on K for some $M_H > 0$, for instance with M_H of Lemma O.2 when it is positive. The method uses the quantities below, which depend on $\sigma, \sigma_g, \bar{y}, D, L_\beta, M_H, n_c, d$ and ε , but not on M and δ ; upper bounds on $\sigma, \sigma_g, \bar{y}, D, L_\beta, M_H$ and n_c suffice. Let $\varepsilon > 0$ and put*

$$\begin{aligned}
 h &= \left(\frac{\varepsilon d^{\tau/2}}{24(1+\rho)C_b L_\beta D_K} \right)^{\frac{1}{\beta-1}}, & \bar{M} &= \frac{\varepsilon}{24(1+\rho)A_1 D_K}, & \bar{\delta} &= \min \left\{ \frac{\varepsilon h}{42(1+\rho)S_1 D_K \sqrt{d}}, \frac{\varepsilon}{42\sqrt{n_c} D_K} \right\}, \\
 \bar{\Omega} &= C_f L_\beta^2 d^{-\tau} h^{2\beta-2} + C_\eta \bar{M}^2 + C_\sigma \frac{d\sigma^2}{h^2}, & N &= \left\lceil \frac{4M_H D_K^2}{\varepsilon} \right\rceil, \\
 B &= \max \left\{ 1, \left\lceil \frac{864(1+\rho)^2 D_K^2 \bar{\Omega}}{N \varepsilon^2} \right\rceil \right\}, & B_L &= \max \left\{ 1, \left\lceil \frac{864 n_c \sigma_g^2 D_K^2}{N \varepsilon^2} \right\rceil \right\}, \\
 \bar{V} &= \frac{2(1+\rho)^2 \bar{\Omega}}{B} + \frac{2n_c \sigma_g^2}{B_L} + \frac{\varepsilon^2}{432 D_K^2}, & \nu &= \min \left\{ \frac{1}{2M_H}, \frac{D_K}{2\sqrt{NV}} \right\}.
 \end{aligned}$$

Let $g_0 = f, g_1, \dots, g_{n_c}$ satisfy the assumptions of Lemma O.8 on $Q_{\vartheta h}$, and let the calls satisfy Lemma O.4 with $M \leq \bar{M}$ and $\delta \leq \bar{\delta}$. Run Procedure 8, that is, the method of Lemma O.7 from $w_1 = (x_0, 0)$, for some $x_0 \in Q$, with the step ν and the operator estimates of Lemma O.8 with radius h , B frames and B_L level calls. Then the output $\hat{x} = \hat{x}_N \in Q$ satisfies

$$\mathbb{E}[f(\hat{x}) - f^*]_+ \leq \varepsilon \quad \text{and} \quad \mathbb{E}v(\hat{x}) \leq \varepsilon$$

after $2N$ rounds, all value calls lie in Q_h , and the numbers T_V of value calls and T_L of level calls satisfy

$$\begin{aligned}
 T_V &\leq 3456 m C_\sigma (24 C_b)^{\frac{2}{\beta-1}} \frac{\sigma^2}{\varepsilon^2} d^{2-\frac{\tau}{\beta-1}} \left(\frac{S_K}{\varepsilon} \right)^{\frac{2}{\beta-1}} + 6md \left(\frac{C_f}{C_b^2} + \frac{C_\eta}{A_1^2} \right) + 4mdN, & S_K &= (1+\rho)^\beta L_\beta D_K^\beta, \\
 T_L &\leq \frac{1728 n_c \sigma_g^2 D_K^2}{\varepsilon^2} + 2N, & N &\leq \frac{4M_H D_K^2}{\varepsilon} + 1.
 \end{aligned}$$

Each value call returns $n_c + 1$ scalar values and each level call returns n_c .

Proof. By Lemma O.8, every operator estimate of the run has the form required in Lemma O.7, with b_K of (O.6) and V_K of (O.7). Since $M \leq \bar{M}$, we have $\Omega_h \leq \bar{\Omega}$. Since $\delta \leq \bar{\delta}$, the terms of V_K with δ are at most $4\varepsilon^2/(1764D_K^2) \leq \varepsilon^2/(432D_K^2)$. Hence $V_K \leq \bar{V}$, and Lemma O.7 applies with $\bar{V}$ in place of V_K and the step ν :

$$\mathbb{E} \text{gap}(\hat{w}_N) \leq M_H D_K^2 / N + 3D_K \sqrt{\bar{V}/N} + 3D_K b_K.$$

The choice of N gives $M_H D_K^2 / N \leq \varepsilon/4$. The choices of B and B_L give $2(1+\rho)^2 \bar{\Omega}/B \leq N\varepsilon^2/(432D_K^2)$ and $2n_c \sigma_g^2 / B_L \leq N\varepsilon^2/(432D_K^2)$, so $\bar{V} \leq N\varepsilon^2/(144D_K^2)$ and $3D_K \sqrt{\bar{V}/N} \leq \varepsilon/4$. The choices of $h, \bar{M}$ and $\bar{\delta}$ bound the three parts of b_K by $\varepsilon/(24D_K)$, $\varepsilon/(24D_K)$ and $\varepsilon/(21D_K)$, so $3D_K b_K \leq \varepsilon/8 + \varepsilon/8 + \varepsilon/7 < \varepsilon/2$. Thus $\mathbb{E} \text{gap}(\hat{w}_N) \leq \varepsilon$, and Lemma O.1 with $\rho - \bar{y} = 1$ gives the two claims.

Each of the $2N$ rounds uses $2mdB$ value calls and B_L level calls. Since $NB \leq N + 864(1+\rho)^2 D_K^2 \bar{\Omega}/\varepsilon^2$ and $NB_L \leq N + 864n_c \sigma_g^2 D_K^2/\varepsilon^2$, we get $T_L \leq 2N + 1728n_c \sigma_g^2 D_K^2/\varepsilon^2$ and $T_V \leq 4mdN + 3456md(1+\rho)^2 D_K^2 \bar{\Omega}/\varepsilon^2$. By the choice of h and $\bar{M}$, the first two terms of $\bar{\Omega}$ equal $C_f \varepsilon^2/(576(1+\rho)^2 C_b^2 D_K^2)$ and $C_\eta \varepsilon^2/(576(1+\rho)^2 A_1^2 D_K^2)$; they contribute $6md(C_f/C_b^2 + C_\eta/A_1^2)$ to T_V . The term $C_\sigma d\sigma^2/h^2$ contributes the first term of the bound on T_V , since $h^{-2} = (24C_b(1+\rho)L_\beta D_K/\varepsilon)^{2/(\beta-1)} d^{-\tau/(\beta-1)}$. $\square$

P DETAILS OF THE EXPERIMENTS

Code, randomness and reporting. The supplementary script `experiments/run_all.py`, which uses NumPy (Harris et al., 2020), produces Experiments 1–7 and the robustness checks reported with Experiments 3 and 5. All random streams are derived from `SeedSequence(20260927)`. Instances, tuning runs, reported runs, and methods use separate streams, and tuning does not reuse evaluation data. The reported run used Python 3.13.16 and NumPy 2.5.3 on two cores of a 2.80 GHz Intel Xeon processor. It took about 46 minutes, including 32 minutes for Experiment 5; `experiments/run_log.txt` records the time of each run.

Vertical uncertainty bars are centered on the plotted markers and represent one standard error over independent runs. When a standard error is smaller than its marker, the bar is covered by the marker. Paired ratios in Experiment 3 and its robustness check use the delta method. Calls required for a target accuracy are interpolated between adjacent points of a doubling grid in log–log coordinates. Their standard errors use 400 fixed bootstrap resamples; the grid is extended until every resample crosses the target. Slopes are least-squares fits in log–log coordinates. Experiments 5 and 6 obtain their slope errors from bootstrap resamples across independent dimensions, while the budget check in Experiment 5 uses the delta method for the mean errors. Haar frames are generated by QR factorization of a standard Gaussian matrix, with the correction $Q \leftarrow Q \text{diag}(\text{sgn diag } R)$. Each oracle call evaluates the objective at its stated query and adds the noise assigned to that call. Figures 2 and 3 display all seven experiments together with the robustness checks for Experiments 3 and 5. Panels (a)–(d) of Figure 1 are Experiments 3, 4, 5, and 7.

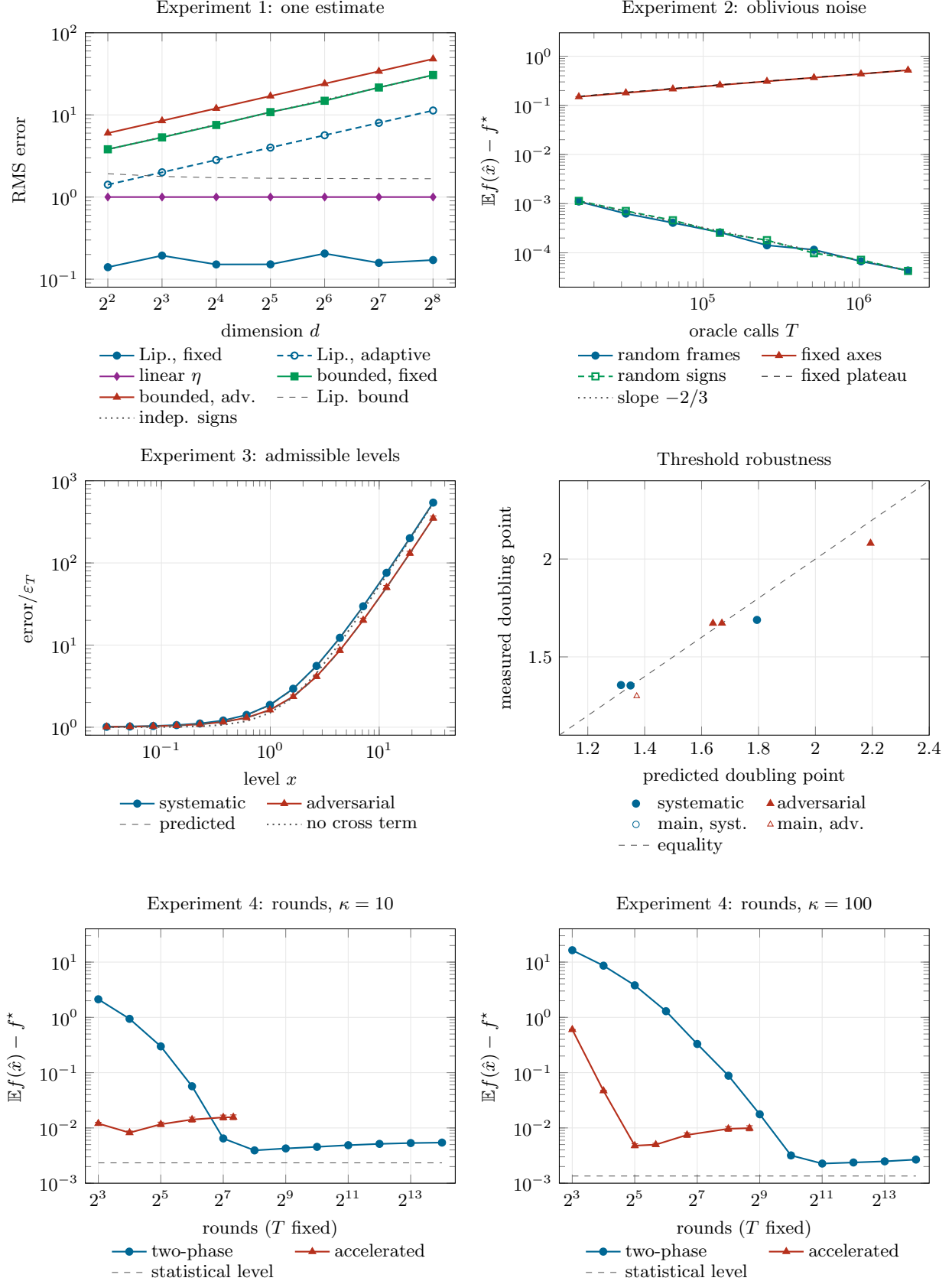


Figure 2: Experiments 1–4 and the robustness check for Experiment 3.

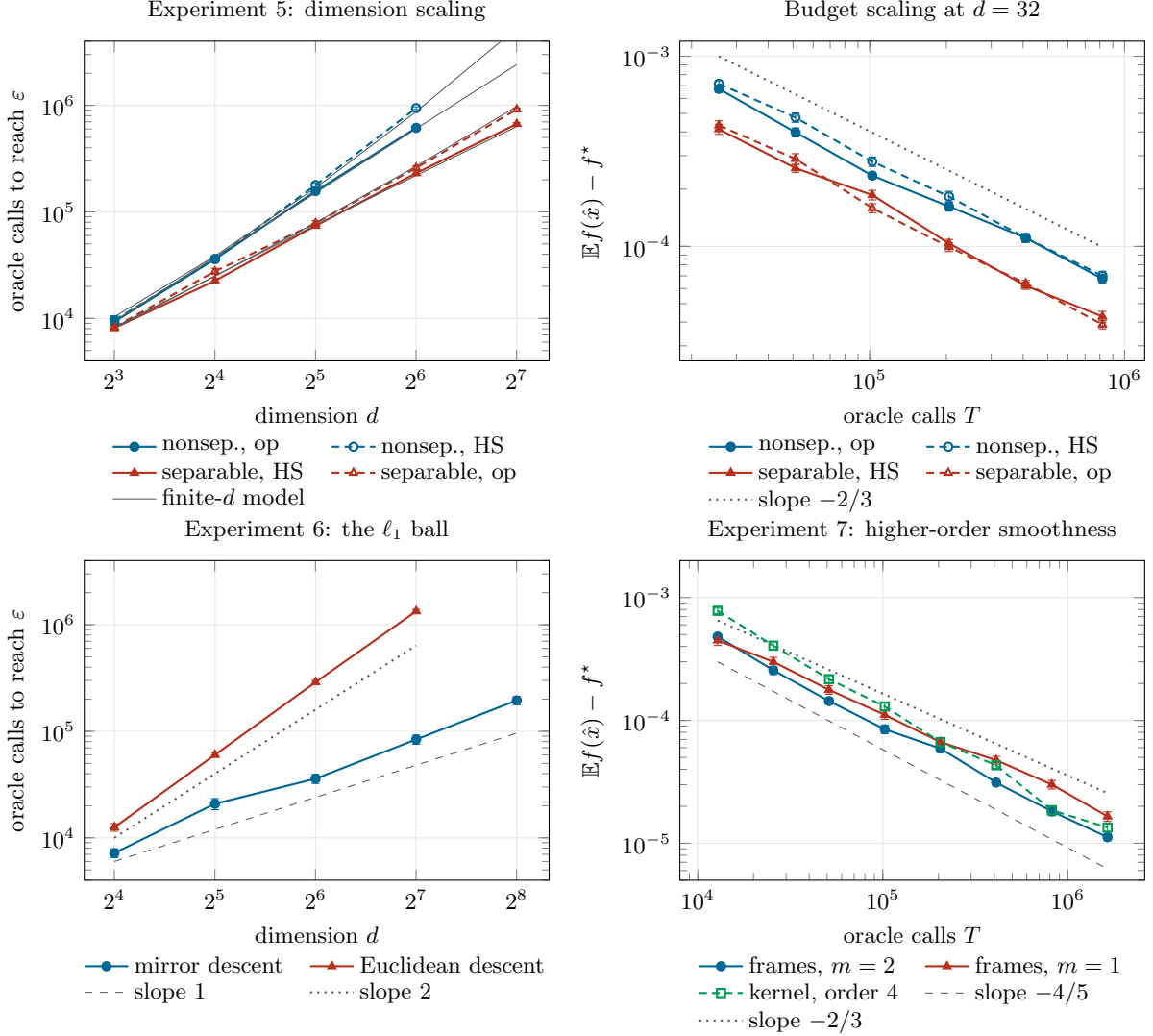


Figure 3: Experiments 5–7 and the robustness check for Experiment 5.

Experiment 1: one estimate under systematic and bounded errors. The objective is $f(x) = \frac{1}{2} \|x\|^2$, the center is $z = e_1/2$, the stencil has $m = 2$ and $h = 0.1$, oracle values are exact, and $d \in \{4, 8, \dots, 256\}$; the same 400 frames per d serve all error models. The Lipschitz error is $\eta(x) = M \sum_{k=1}^{50} w_k |\langle v_k, x \rangle - c_k|$ with $M = 1$, independent uniform unit vectors v_k , weights (w_k) from the flat Dirichlet distribution and offsets $c_k = \langle v_k, z \rangle + (h/\sqrt{d})\zeta_k$, $\zeta_k \sim U[-1, 1]$, so that every kink passes close to z ; η is drawn once per d , before the frames. This one function does not probe the supremum over all Lipschitz errors; as a check of the order M , the linear error $\eta(x) = M \langle v, x \rangle$ shifts every frame by exactly Mv , and the measured error has norm 1.000000 for every d . The bounded error fixed in advance is $\delta \operatorname{sgn} \sin(1000 \langle u, x \rangle)$ with $\delta = Mh$ and a fresh uniform unit vector u for each frame, also drawn before the frames; its values depend on the frame only through the query points, so it is a bounded error as in (A), not oblivious noise. Independent random signs of size δ are not simulated; they serve as the reference value $M\sqrt{S_2 d}/2$. The adversarial error sees the frame and adds $\pm \delta \operatorname{sgn}(a_j) \operatorname{sgn} \langle U_i, w \rangle$ to the value at $z \pm ht_j U_i$ for a fixed unit vector w ; the error of the estimate then has norm exactly $S_1 M\sqrt{d} = 3\sqrt{d}$ for every frame. The root-mean-square error under the Lipschitz error stays between 0.14 and 0.21 for all d , almost all of it a deterministic shift (the norm of the mean error is at least 96% of the root-mean-square error), and below the bound $A_1 M\sqrt{d/(d-1)}$ of Lemma D.2(d), which decreases from 1.92 to 1.67. The error under the bounded error fixed in advance agrees within 2.3% (at most 1.6 standard errors) with the reference $M\sqrt{S_2 d}/2 \approx 1.90\sqrt{d}$. A Lipschitz error chosen after the frame as in Lemma D.5 gives an error of norm exactly $M\sqrt{d/2}$.

Shared objective and method for Experiments 2, 3, and 5. Here $d = 20$, except in Experiment 5 and the robustness checks described with Experiments 3 and 5, and $f(x) = \frac{1}{2} \|x\|^2 + c g(x)$ with $g(x) = \langle w, x \rangle \|x\|^2 e^{-\|x\|^2}$, $c = 1/4$ and $w = d^{-1/2}(1, \dots, 1)$, so that $x^* = 0$, $f^* = 0$ and $\nabla^2 f(0) = I$. The eigenvalues of $\nabla^2 g$ have absolute value at most 1.69123 on $\mathbb{R}^d$, attained at $x = \pm 1.1495w$; this value is computed numerically, on a grid of step 0.005 after reduction to the plane spanned by w and x and with five local refinements, and it is not a certified bound. The methods use the bound 1.7, that is, $\alpha = 1 - 1.7c = 0.575$, $L = 1 + 1.7c = 1.425$ and $\kappa = 2.48$; so $\alpha = 1$ in the normalizations of Experiment 3 and its robustness check is the curvature at x^* , not the constant of a class. Since g is odd and $UU^\top = I$, every frame with $m = 1$ at the minimizer equals $b = ch^2 e^{-h^2} w$, the gradient of the ball average of f at 0, and so does every coordinate stencil; this bias does not depend on d . The noise is $N(0, \sigma^2)$ with $\sigma = 0.03$. All methods use $m = 1$, the radius $h = \sqrt{2} T^{-1/6}$ (except in Experiment 5 and its budget-scaling check), and variant (a) of Procedure 2 with $N = 200$ rounds ($k_0 = 16\kappa = 39.7$) and $R = 2$, from a random start x_0 with $\|x_0\| = 1$ drawn once per experiment. A round consists of B frames, that is, $2dB$ calls. Without noise and with $h = 10^{-4}$, this schedule ends at $f(\hat{x}) - f^* \leq 3.4 \cdot 10^{-18}$, so the optimization error is negligible. The odd term of f decays, so a very large radius would see only the quadratic, whose minimizer is x^* ; in the reported runs $h \leq 0.31$ for f , so the radii stay near the analysis.

Experiment 2: oblivious noise. The budgets are $T = 1.6 \cdot 10^4 \cdot 2^k$, $k = 0, \dots, 7$, with 20 runs each. In addition to the $N(0, \sigma^2)$ noise, every query $z + ht_j u$ of a round receives $+b$ and every query $z - ht_j u$ receives $-b$, with $b = \sigma$; this oblivious error depends only on the position of the query in the round. It is a crafted failure mode of stencils along fixed directions, not a model of all oblivious noise. The methods are (i) random frames, (ii) the stencil along $e_1, \dots, e_d$ with the same query order, and (iii) coordinate stencils with signs drawn in every round (Lemma N.7). For (ii) the oblivious error shifts every coordinate of the estimate by b/h , and the predicted plateau is $f(x_b) - f^*$, where x_b minimizes $f(x) + (b/h) \sum_i x_i$ (computed by Newton’s method). In (iii) the random sign of coordinate i multiplies both the difference and the direction, so the oblivious shift $\pm b/h$ enters with a random sign and has mean zero; in (i) the random frame plays this role. Method (ii) stays at 0.97 to 1.00 times its plateau, which grows from 0.15 to 0.52 because b/h grows as h decreases. For (i) and (iii) the oblivious error adds a term of mean zero and covariance $(b/h)^2 I$ to every estimate. The ratio of their errors lies between 0.85 and 1.27 (standard errors 5 to 10% of each error), and their fitted log-log slopes are -0.66 and -0.68 over all budgets and -0.59 and -0.67 over $T \geq 2.56 \cdot 10^5$, against the rate $-2/3$.

Experiment 3: admissible levels. Here $T = 1.024 \cdot 10^6$ ($B = 128$), $h = \sqrt{2} T^{-1/6} = 0.141$ and 20 runs. A unit vector v is drawn before the runs. The systematic error is $\eta(x) = M \langle v, x \rangle$. The adversarial error sees the frames and adds $\pm \delta \operatorname{sgn}(a_j) \operatorname{sgn} \langle U_i, v \rangle$ to the value at $z \pm ht_j U_i$, which adds $(\delta S_1/h) \sum_i \operatorname{sgn} \langle U_i, v \rangle U_i$ to the frame, a vector with mean $a_\delta v$, $a_\delta = (\delta S_1/h) d \mathbb{E} |u_1|$ for u uniform on the sphere ($d \mathbb{E} |u_1| = 3.61$ for $d = 20$). The 15 log-spaced values of M and of δ were set from a pilot estimate of ε_T on a separate stream. All 31 settings share the frames and the Gaussian noise of every call, and the errors are reported as ratios to $\varepsilon_T = 3.03 \cdot 10^{-5}$, the error at $M = \delta = 0$. They are plotted against $x = M/\sqrt{\alpha \varepsilon_T}$ and $x = \delta \sqrt{d}/(h \sqrt{\alpha \varepsilon_T})$, from 0.031 to 31, with $\alpha = 1 = \nabla^2 f(0)$.

The predictions follow the mean of the estimate. Near $x^* = 0$ the estimate has mean $b + av$ up to higher-order terms, with $a = M$ for the systematic error, since $\nabla \eta = Mv$, and $a = a_\delta$ for the adversarial one. The method then tends to $-(b + av)$, where f exceeds its value at $-b$ by $a \langle b, v \rangle + a^2/2$ up to higher-order terms. We therefore predict the ratio $1 + (F(a) + a \langle b, v \rangle)/\varepsilon_T$, where $F(a) = f(x_a) - f^*$ and x_a minimizes $f(x) + a \langle v, x \rangle$ (computed by Newton’s method); for the systematic error, $F(M)$ is the error that remains with exact values of $f + \eta$. Here $\langle b, v \rangle = 0.346 \sqrt{\alpha \varepsilon_T}$, and the squared bias $\|b\|^2/2$ is $0.39 \varepsilon_T$. The ratio of the measured to the predicted error lies between 1.00 and 1.02 for both errors at all 15 levels, at most 0.53 standard errors away. The errors cross $2\varepsilon_T$ at $x \approx 1.07$ (systematic) and $x \approx 1.30$ (adversarial) by linear interpolation in log-log coordinates between the two grid points around the crossing (1.08 and 1.34 by a power law $1 + cx^p$ through them), while the predictions cross 2 at 1.11 and 1.37 (computed on a fine grid); without the cross term $a \langle b, v \rangle$ the first would cross 2 at 1.41. For $x \geq 6$ the errors grow with fitted slopes 1.96 and 1.94, and at the largest level the ratios are 542.7 and 351.8, against the predictions 540.9 and 350.0. The systematic and adversarial errors are tested separately; their joint effect is not.

Experiment 4: rounds against calls. Here $d = 20$, $m = 2$, the noise is independent $N(0, \sigma^2)$ with $\sigma = 0.1$, the budget is $T = 2md \cdot 2^{14} = 1,310,720$ calls and the radius is $h = 0.3 (d^2 \sigma^2 / T)^{1/8} = 0.0613$. The objective is $f(x) = \frac{1}{2} x^\top A x - b^\top x + 0.1 \log \cosh \langle w, x \rangle$ with A diagonal, eigenvalues log-spaced in $[1, \kappa]$, $\kappa \in \{10, 100\}$, a

unit vector w and a minimizer x^* of norm 2, drawn once per κ ; then $I \preceq \nabla^2 f \preceq (\kappa + 0.1)I$. Its third derivative is small, so this experiment concerns rounds, not the order of the stencil. The methods use $R = 4$, $\alpha = 1$ and $L = \kappa + 0.1$. Each of the 20 runs draws a pool of $n = 2^{14}$ frames and of the noise of their calls, which both methods use in order, for both values of κ and all numbers of rounds (common random numbers). Without noise, the estimate at 20 random centers in $B(R)$ differs from ∇f by at most $1.0 \cdot 10^{-9}$. For $N \in \{8, 16, \dots, 2^{14}\}$ rounds the batch has $B = 2^{14}/N$ frames. The two-phase method is variant (a) of Procedure 2. The accelerated method is variant (b) with $N_0 = \lceil 5\sqrt{L/\alpha} \rceil$ (16 and 51) rounds per stage and with its iterates projected onto $B(R)$. If $N < N_0$, it runs one stage of N rounds with B frames; otherwise it runs $J = \min\{\lfloor N/N_0 \rfloor, \lfloor \log_2(2^{14}/N_0 + 1) \rfloor\}$ stages, stage j with $\lfloor 2^{13+j}/(N_0(2^J - 1)) \rfloor$ frames per round, so that the batches double up to rounding. Rounds beyond the last stage are not used: the method uses at most 160 rounds for $\kappa = 10$ and 408 for $\kappa = 100$, and Figure 1(b) shows it against the rounds it uses, once per schedule. The rounding leaves at most 1.1% of the budget unused. The statistical level is $\bar{\sigma}^2 \text{tr}(\nabla^2 f(x^*)^{-1})/(2n)$ with $\bar{\sigma}^2 = \sigma^2 S_2/(2h^2)$, the asymptotic error of Polyak–Ruppert averaging (Polyak and Juditsky, 1992) of n frames with exact gradients and the same noise; it is $2.34 \cdot 10^{-3}$ for $\kappa = 10$ and $1.35 \cdot 10^{-3}$ for $\kappa = 100$. The experiment compares methods at one budget; it is not a measurement of the least number of rounds.

The two-phase method uses N rounds. Its error is 24 times the statistical level at $N = 64$, 2.8 times it at $N = 128$ and 1.7 to 2.3 times it for $N \geq 256$ when $\kappa = 10$, and 13 times it at $N = 512$, 2.4 times it at $N = 1024$ and 1.7 to 2.0 times it for $N \geq 2048$ when $\kappa = 100$. The accelerated method attains its smallest error with a single stage: 3.5 times the statistical level with 16 rounds for $\kappa = 10$ and with 32 rounds for $\kappa = 100$, where 16 rounds still give 35 times this level. With more rounds it runs more stages, its last stage has fewer frames per round, and its error rises to 5.0 to 6.7 times the statistical level with 32 to 160 rounds for $\kappa = 10$, and to 5.5 to 7.3 times it with 102 to 408 rounds for $\kappa = 100$. The number of rounds the two-phase method needs grows by a factor 8 from $\kappa = 10$ to $\kappa = 100$, that of the accelerated method by a factor 2.

Experiment 5: calls against the dimension. Besides the objective f above we use the separable objective $f_s(x) = \sum_j \psi(x_j)$, $\psi(t) = \frac{1}{2}t^2 + ct^3e^{-t^2}$, with the profile of f on the line through w along each axis. Its minimizer is 0, $\nabla^2 f_s(0) = I$, the eigenvalues of $\nabla^2 f_s$ lie in $[1 - 1.6912c, 1 + 1.6912c]$, so the methods use the same α and L , and $\|D^3 f_s(0)\|_{\text{op}} = \|D^3 f(0)\|_{\text{op}} = 6c$. Since $\nabla^2 f_s$ is diagonal, its Lipschitz constant is $\sup |\psi'''|$ in the operator and in the Hilbert–Schmidt norm, while $D^3 f(0)$ has Hilbert–Schmidt norm of order $c\sqrt{d}$. A frame at the minimizer has bias $ch^2e^{-h^2}w$ for f , and $\frac{3ch^2}{d+2}(1, \dots, 1)$, of norm $\frac{3ch^2\sqrt{d}}{d+2}$, for f_s to leading order in h . Both objectives are run with both radius rules $h = C(d^{2+\tau}\sigma^2/T)^{1/6}$ of (3), $\tau = 0$ (op) and $\tau = 1$ (HS), with variant (a) of Procedure 2, $N = 200$ rounds of B frames, $T = 2dNB$, the target $\varepsilon = 2 \cdot 10^{-4}$ and a uniform unit start drawn once per d . For the matched pairs (non-separable, op) and (separable, HS) the constant $C \in \{2^{k/4} : -4 \leq k \leq 4\}$ was tuned once, on a separate stream at $d = 16$ with 10 runs, budgets $B = 1, \dots, 32$ and common randomness across C , by minimizing the interpolated number of calls to reach ε ; the chosen values are $C = 1.68$ and $C = 1.19$, both inside the grid. The other rule of each objective uses the constant that gives the same radius at $d = 16$, so that the two rules differ only in how the radius scales with d . For each d the number of frames per round is halved or doubled until the mean error brackets ε , with 40 runs for $d \leq 16$ and 20 for $d \geq 32$; the non-separable objective is not run at $d = 128$. For f_s the radii reach 0.49 ($d = 128$), but a frame moves each coordinate by h times a coordinate of a uniform unit vector, and the next term of the bias is smaller than the leading one by a factor of order h^2/d .

With the matched rules the calls grow with the fitted slopes 2.03 ± 0.03 (non-separable) and 1.62 ± 0.03 (separable) over $8 \leq d \leq 64$, and 1.61 ± 0.02 for the separable objective up to $d = 128$; with the other rules the slopes are 2.21 ± 0.05 and 1.64 ± 0.03 over $8 \leq d \leq 64$, and 1.68 ± 0.02 for the separable objective up to $d = 128$. The ratio of the calls of the two objectives with the matched rules grows from 1.14 at $d = 8$ to 2.66 at $d = 64$. The rule of the other seminorm needs 1.03, 1.00, 1.14 and 1.53 times the calls of the matched rule for the non-separable objective ($d = 8, \dots, 64$), and 1.02, 1.23, 1.06, 1.12 and 1.37 times them for the separable one ($d = 8, \dots, 128$). At $d = 16$ the two rules give the same radius, so the ratio 1.23 there, 2.4 standard errors from 1, is a fluctuation of the runs; it indicates their variability.

To compare the slopes with a finite- d prediction we model the error after the method as $Ak_b(d)^2h^4 + Vd^2\sigma^2/(Th^2)$, with $k_b = c$ for f and $k_b = 3c\sqrt{d}/(d+2)$ for f_s , the leading terms of the squared bias and of the noise. With the radius rule and C optimal at $d = 16$, where the bias term is then half the noise term, the calls to reach ε are

proportional to

$$\left(\frac{1}{2}v(16)\frac{u(d)}{u(16)} + v(d)\right)^{3/2}, \quad u(d) = k_b(d)^2 d^{2(2+\tau)/3}, \quad v(d) = d^{(4-\tau)/3}, \quad (\text{P.1})$$

and we fit one constant per curve by least squares on the log scale. The model neglects higher-order terms of the bias, the variance of the frames without noise and the optimization error. For $8 \leq d \leq 64$ it has the slopes 2.00 and 2.13 for the non-separable objective with the op and the HS rule, and 1.59 and 1.67 for the separable one with the HS and the op rule (1.57 and 1.72 up to $d = 128$); its slopes as $d \rightarrow \infty$ are 2, 3, $3/2$ and 2. The measured slopes differ from those of the model by +0.03, +0.08, +0.03 and -0.03 (+0.03 and -0.04 up to $d = 128$). At all 18 points the residuals $\log(\text{measured}/\text{model})$ are at most 0.104 in absolute value, that is, within 11%, and at most 2.7 standard errors of the logarithm of the calls; `run.log.txt` lists them.

Experiment 6: the ℓ_1 ball. The objective is $f(x) = \sum_j \varphi(x_j - c_j)$ with $\varphi(t) = e^t - 1 - t$ and $c = \frac{1}{2}e_1$: it is convex, its minimizer is 1-sparse and lies in $B_1(1)$, and $f(0) - f^* = 0.107$. Since φ''' is unbounded on $\mathbb{R}$, f does not belong to the global class $\mathcal{C}_1(R, G_\infty)$ of Lemma N.9, and we use its local form, Lemma N.10. The target is $\varepsilon = 0.01$ and the noise is $N(0, \sigma^2)$ with $\sigma = 0.03$. Both methods use $m = 1$ and the radius of Lemma N.9(b), $h = (\varepsilon/(4L_\beta))^{1/2} = 0.0299$, with $L_\beta = e^{1+h}$. Every query of method (i) differs from a point of $B_1(1)$ in one coordinate by at most h , and every query of method (ii) lies in $B_1(1) + hB_2(1)$; in both cases all coordinates are at most $1 + h$ in absolute value. On this set $\nabla^2 f = \text{diag}(e^{x_j - c_j})$, so e^{1+h} bounds the Hölder constant of $\nabla^2 f$ in the ℓ_1 operator norm, and the local constant of Lemma N.10. The noise of one coordinate of either estimate has standard deviation $\sigma/(\sqrt{2}h) = 0.71$. Method (i) is entropic mirror descent with coordinate stencils, Procedure 7, with the step $\varsigma = \mu(2\log(2d)/(\bar{G}^2 N))^{1/2}$; here $\bar{G}^2 = 2(G_\infty + \varepsilon/4)^2 + 2(1 + \log(2d))\sigma^2/h^2$ is $\bar{G}_\varepsilon^2$ of Lemma N.9(b) for Gaussian noise, and $G_\infty = e - 1 = 1.72$ bounds $\|\nabla f\|_\infty$ on $B_1(1)$. Method (ii) is projected stochastic gradient descent onto $B_1(1)$ with the exact Euclidean projection, one frame per round and the step $\mu/(\bar{G}_2\sqrt{N})$, where $\bar{G}_2^2 = 2(G_2 + \varepsilon/4)^2 + d\sigma^2/h^2$ and $G_2 = 1.76$ bounds $\|\nabla f\|$ on $B_1(1)$. Both output the average of their centers and use $2d$ calls per round. The multiplier $\mu \in \{2^k : k = -2, \dots, 4\}$ was tuned in the same way for both methods, on a separate stream at $d = 32$ with 10 runs, budgets $N = 128, \dots, 8192$ and common randomness across μ , by minimizing the interpolated number of calls to reach ε ; the chosen values are $\mu = 8$ for (i) and $\mu = 2$ for (ii), both inside the grid. For each d the number of rounds is halved or doubled until the mean error brackets ε (20 runs; 10 and 6 for (ii) at $d = 64$ and 128). Method (i) needs $7.2 \cdot 10^3, 2.1 \cdot 10^4, 3.6 \cdot 10^4, 8.4 \cdot 10^4$ and $2.0 \cdot 10^5$ calls for $d = 16, 32, \dots, 256$, that is, 225 to 381 rounds; the fitted slope in d is 1.14 ± 0.06 for $16 \leq d \leq 128$ and 1.15 ± 0.04 up to 256. Method (ii) needs $1.3 \cdot 10^4, 6.0 \cdot 10^4, 2.9 \cdot 10^5$ and $1.34 \cdot 10^6$ calls for $d \leq 128$, that is, 393 to 5,230 rounds, and a fitted slope of 2.25 ± 0.04 . It was not run at $d = 256$: extrapolating its number of rounds gives about $1.2 \cdot 10^4$, and one literal 256×256 frame takes about 6 ms. The instance is sparse and favorable to the ℓ_1 geometry; it illustrates the upper bound and is not a minimax comparison.

Experiment 7: higher order. The objective is $f_7(x) = \frac{1}{2}\|x\|^2 + c\langle w, x \rangle (1 - e^{-\|x\|^2})$ with $c = 1/4$, $w = d^{-1/2}(1, \dots, 1)$ and $d = 16$. It is C^∞ with bounded derivatives of every order, so it belongs to the classes of every β ; $x^* = 0$, $f^* = 0$, $\nabla^2 f_7(0) = I$, and the eigenvalues of the Hessian of the odd term g_7 have absolute value at most 1.9518 (computed numerically as for g), so the methods use $\alpha = 1 - 1.96c = 0.51$ and $L = 1.49$. Unlike the odd term of f , that of f_7 does not decay: at a very large radius every estimate sees $x + cw$, whose root $-cw$ has error $2.7 \cdot 10^{-2}$, so wide tuning grids cannot favor huge radii. At the minimizer the bias is ch^2w for central differences, to leading order, and $O(h^4)$ for $m = 2$ and for the kernel estimate; noiseless means over many frames or directions at $h = 0.1, 0.2, 0.4$ confirm the orders $(2.49 \cdot 10^{-3}, 9.80 \cdot 10^{-3}, 3.70 \cdot 10^{-2}$ for $m = 1$ and $3.1 \cdot 10^{-6}, 4.9 \cdot 10^{-5}, 7.5 \cdot 10^{-4}$ for $m = 2$). Three estimates with $4dB$ values per round drive variant (a) of Procedure 2 with $N = 200$ rounds: (i) B frames with $m = 2$ and the radius rule of $\beta = 5$, $h = C(d^2\sigma^2/T)^{1/10}$; (ii) $2B$ frames with $m = 1$ and the rule of $\beta = 3$, $h = C(d^2\sigma^2/T)^{1/6}$; (iii) $2dB$ directions of the ℓ_2 -randomized kernel estimate of Akhavan et al. (2020, 2024), $\frac{d}{2h}(Y(x + hr\zeta) - Y(x - hr\zeta))K(r)\zeta$ with ζ uniform on the sphere, r uniform on $[-1, 1]$ and the kernel $K(r) = \frac{15r}{4}(5 - 7r^2)$ of order 4, with the rule of $\beta = 5$. The cited works use one such estimate per iteration of stochastic gradient descent; here all three estimates drive the same batched method, so that only the estimate differs. Each constant $C \in \{2^{k/2} : -4 \leq k \leq 8\}$ was tuned at $T = 2.05 \cdot 10^5$ on a separate stream with 10 runs and common randomness across C , by minimizing the mean error; the chosen values are 2.83, 1.41 and 2.83, all inside the grid. The budgets are $T = 1.28 \cdot 10^4 \cdot 2^k$, $k = 0, \dots, 7$, with 20 runs each. The fitted slopes in T are $-0.76, -0.67$ and -0.85 over all budgets and $-0.80, -0.67$ and -0.82 for $T \geq 2.05 \cdot 10^5$, against the rates $-4/5, -2/3$ and $-4/5$ of the analyses. The error of central differences is 0.93 to 1.65 times that of the frames with $m = 2$, growing with T , and the error of the kernel estimate is 1.02 to 1.62 times it.

Additional checks for Experiments 3 and 5. *Thresholds across dimensions and budgets.* We repeat Experiment 3 at $(d, T) = (10, 1.024 \cdot 10^6)$, $(40, 1.024 \cdot 10^6)$, and $(20, 2.56 \cdot 10^5)$, using nine log-spaced levels from $x = 0.25$ to 4.5 and independent streams for the instance, pilot, and reported runs. The measured systematic doubling points are 1.69, 1.36, and 1.35, compared with predictions 1.79, 1.32, and 1.35. The adversarial values are 2.08, 1.67, and 1.67, compared with 2.19, 1.64, and 1.67. Their variation comes from the instance-dependent cross term $a(b, v): \langle b, v \rangle / \sqrt{\alpha \varepsilon T} = -0.34, 0.10, \text{ and } 0.07$.

Budget scaling at fixed dimension. We also rerun the four configurations of Experiment 5 at $d = 32$, using its tuned constants, $N = 200$ rounds, $B = 2, 4, \dots, 64$ frames, and 20 independent runs per budget. Thus T ranges from $2.56 \cdot 10^4$ to $8.19 \cdot 10^5$ oracle calls and the radii range from 0.17 to 0.39. Because h is proportional to $T^{-1/6}$, the predicted error is $T^{-2/3}$. The four fitted slopes are -0.648 ± 0.017 , -0.677 ± 0.017 , -0.668 ± 0.020 , and -0.703 ± 0.018 ; for $T \geq 1.02 \cdot 10^5$ they are -0.60 ± 0.03 , -0.67 ± 0.04 , -0.71 ± 0.04 , and -0.67 ± 0.03 . The interpolated calls required for $\varepsilon = 2 \cdot 10^{-4}$ are $1.39 \cdot 10^5$, $1.77 \cdot 10^5$, $8.8 \cdot 10^4$, and $7.8 \cdot 10^4$, compared with $1.57 \cdot 10^5$, $1.78 \cdot 10^5$, $7.4 \cdot 10^4$, and $7.8 \cdot 10^4$ in the independent dimension-scaling runs of Experiment 5.

Q DISCUSSION AND OPEN QUESTIONS

Random noise costs calls but not rounds: the accelerated method keeps the $O(\sqrt{\kappa} \log(1/\varepsilon))$ rounds of Nesterov’s method with exact gradients, and only the number of calls per round grows with d and σ . Persistent errors are governed by their regularity rather than their size. A Lipschitz error fixed before the run is not averaged out: for $\eta(x) = M \langle v, x \rangle$ every estimate is shifted by Mv . But the shift is $O(M)$, without a factor $\sqrt{d}$, so the error costs $O(M^2/\alpha)$ in value and is admissible up to a Lipschitz constant of order $\sqrt{\alpha \varepsilon}$ in every dimension, however large its values are. An error known only to be bounded, such as a rounding error, is admissible only up to $\min\{\sqrt{d}, \sqrt{\log(4T)}\} \delta_\varepsilon$, even if it is a fixed function of the query, and only for $\varepsilon \leq \varepsilon_0$; the achievable level $c_\beta \delta_\varepsilon$ does not grow with the budget, since the radius cannot exceed the order of h_ε .

Limitations. The methods need β , L_β , σ , α and L ; adaptation to unknown parameters is not studied. The counts concern oracle calls and adaptive depth: a round assumes unlimited parallelism, a Haar frame costs $O(d^3)$ arithmetic operations, and the methods of Theorem 3 minimize polynomials at a cost we do not bound. Some constants are far from practical: the factor in c_N of Theorem 2(b) is below 10^{-45} for $\beta = 3$ and $N = 3$; the method of Theorem 3(b) starts with $O(d^{\beta^2-1} \log(d+T))$ calls for integer β , that is d^8 for $\beta = 3$ and d^{15} for $\beta = 4$; and with our constants the range of accuracies in Lemma L.10(a), and that of the ℓ_1 lower bound of Section 7, are nonempty only for d above large thresholds, which exceed 2^{28} for the latter at $\beta = 3$. Under unrestricted Hessian magnitude, our analysis of the initial polynomial fit uses centered noise. The upper bounds on constrained sets query the h -neighborhood of the set. The experiments check the predictions on particular instances, which are not worst cases: Hilbert–Schmidt smoothness appears only through the separable instance of Figure 1(c) and Experiment 5, the order β only through the comparison of the stencils with $m = 1$ and $m = 2$ in (d), and the methods of Lemma L.1 and Theorem 3 are not tested; minimax lower bounds cannot be confirmed by finitely many instances.

Open questions. *Rounds.* The lower bound $N_\beta(d)$ of Theorem 2(c) concerns the rate in the number of calls per round for fixed d , while for a fixed accuracy Theorem 1(c) needs $O(\sqrt{\kappa} (1 + \log(LR^2/\varepsilon)))$ rounds in every dimension. How the minimal number of rounds depends on κ and ε is open, since our nested hard functions have $\kappa \leq 3$. Under the condition of Lemma L.1, we do not know whether $O(\sqrt{\kappa_{\text{eff}}} \log(1/\varepsilon))$ rounds suffice. *Convex functions.* The minimax number of calls is known only in the ranges of Lemma L.10(a) and, for $L \geq L_\varepsilon$ and $L_\beta R^\beta d^{-\beta/4} \leq \varepsilon \leq c'_\beta L_\beta R^\beta$, up to the factor $C_\beta V/L_\varepsilon$ between the leading statistical terms (Lemma L.10(b)); the full upper bound adds $32d(1 + LR^2/\varepsilon)$ calls, and for smaller ε the gap grows. The admissible adversarial level is known only up to a factor polynomial in $1/\varepsilon$ (Table 3). *Unbounded Hessians.* Can the rate of Theorem 3 be reached with a smaller startup, under oblivious noise or with systematic and adversarial errors, or in polynomial time?

R FURTHER RELATED WORK

Oracle calls. Finite differences of noisy values go back to Kiefer and Wolfowitz (1952); higher smoothness is exploited by differences of higher order (Fabian, 1967; Dippon, 2003), optimal in dimension one (Chen, 1988),

and by randomized differences weighted by a kernel (Polyak and Tsybakov, 1990). Omitting factors free of d and ε , the number of calls has since fallen from $d^2\varepsilon^{-(\beta+1)/(\beta-1)}$ (Bach and Perchet, 2016) to $d^2\varepsilon^{-\beta/(\beta-1)}$ (Akhavan et al., 2024, Theorem 18 and Corollary 19), with intermediate steps by Akhavan et al. (2020) and Novitskii and Gasnikov (2022). For integer β the class of Akhavan et al. (2024) is defined through derivatives of order β , but their bias bound (Akhavan et al., 2024, Lemma 6) uses only the Taylor remainder of order β , which is bounded on our classes. Their noise, like that of Lobanov (2024), may be oblivious: independent of the random directions and not necessarily centered. The lower bound $d^{\beta/(\beta-1)}\varepsilon^{-\beta/(\beta-1)}$ of Akhavan et al. (2020, Theorem 6.1) and Akhavan et al. (2024, Theorem 22) comes from separable functions, for which it is attained (Akhavan and Tsybakov, 2025). For a Hilbert–Schmidt bound on the Lipschitz constant of the Hessian, $\beta = 3$, Yu et al. (2024) established matching bounds of order $d^{3/2}\varepsilon^{-3/2}$ for centered noise and unrestricted Hessian magnitude; their Theorems 3.1 and 3.2 are asymptotic in the number of calls with the other parameters fixed. With a bounded Hessian alone, d^2/ε^2 calls are necessary (Shamir, 2013, improving Jamieson et al., 2012), but the third derivatives of the hard functions are not bounded uniformly in ε . Wang et al. (2018) study local minimax rates for noisy optimization of smooth functions, a neighboring problem with a different notion of optimality. Akhavan et al. (2026) minimize a partial derivative $\partial_j^k g$ of an unknown function from noisy values of g , with the optimal rate $N^{-(\beta-1)/(\beta+k)}$ in the number N of queries; for $k = 0$ their bound is $d^{(2\beta-1)/\beta}N^{-(\beta-1)/\beta}$. Hessians that grow at most linearly with the gradient (Zhang et al., 2020) were treated for two-point feedback by Lobanov and Gasnikov (2025). Zeroth-order methods are also used for nonconvex problems such as black-box attacks on neural networks (Chen et al., 2017) and fine-tuning of language models (Malladi et al., 2023), which lie outside our assumptions.

Directions. Our estimate uses all axes of a uniformly random orthonormal basis. Finite differences along random orthogonal directions, which include spherical smoothing and coordinate descent as special cases, were analyzed by Kozak et al. (2023), and stochastic descent along $\ell \leq d$ structured orthogonal directions by Rando et al. (2024). Their oracles and classes differ from ours; we use stencils of high order and the noise of Definition 1, in which the frames make an oblivious error average out and a fixed Lipschitz error cost $O(M)$ (Proposition 1).

Rounds. Parallel first-order methods need many rounds in high dimension (Nemirovski, 1994; Diakonikolas and Guzmán, 2020), and zeroth-order methods for functions with a bounded Hessian need about $d \log \log n$ rounds of n calls to keep the sequential rate (Duchi et al., 2018). For highly smooth functions, batched kernel methods with bounds on the rounds exist for convex (Lobanov et al., 2024), Polyak–Łojasiewicz (Gasnikov et al., 2024b) and strongly convex (Lobanov, 2024, Theorem 2) objectives; the last one makes d^2 calls, up to logarithmic factors, in $O(\sqrt{\kappa} \log(1/\varepsilon))$ rounds for $\kappa = L/\alpha$. Compared with it, Theorem 1(a) has no logarithmic factor in the error bound, holds under Hilbert–Schmidt smoothness with the power d^s , and admits systematic and adversarial errors at the levels of Theorem 1(c), which are sharp up to the factors stated there.

Admissible errors. Systematic and adversarial errors have mostly been studied for less smooth functions and for two-point feedback, where two values share one random sample (Duchi et al., 2015; Shamir, 2017). There Gasnikov et al. (2022) bounded the calls, the rounds and the largest admissible adversarial error, which Pasechnyuk-Vilensky et al. (2026) determined up to logarithms, building on approximately convex optimization (Belloni et al., 2015; Singer and Vondrák, 2015; Li and Risteski, 2016). In Pasechnyuk-Vilensky et al. (2026) the errors are fixed functions on a Euclidean ball, queries and output lie in the ball, and the guarantee holds with high probability; their lower bound $\varepsilon^{3/2}/(R\sqrt{dL})$ on the admissible level (Table 3) is proved on the subclass with a Lipschitz constant of order $\sqrt{L}\varepsilon$ on the ball, and their upper bound for a range of accuracies and for d at least of the order of the logarithm of the number of queries. Lipschitz errors were treated by Gasnikov et al. (2016a) and by Dvinskikh et al. (2023), whose constant is of order $\varepsilon/(R\sqrt{d})$ for exact values and two-point feedback, and accelerated methods with bounded errors by Gorbunov et al. (2022). For highly smooth Polyak–Łojasiewicz functions, kernel methods tolerate bounded deterministic errors of order $\varepsilon^{\beta/(2(\beta-1))}d^{-\beta/(\beta-1)}$ (Gasnikov et al., 2024b, Theorems 1 and 3).

Convex functions and other geometries. For convex functions the optimal dependence on d and ε is open. Finite differences use $d^2\varepsilon^{-2\beta/(\beta-1)}$ calls (Bach and Perchet, 2016), as does the method of Akhavan et al. (2024) applied to a regularized objective (Section M.4). Smoothing with a bounded Hessian (Flaxman et al., 2005; Saha and Tewari, 2011) uses $d^2L\varepsilon^{-3}$ calls (Gasnikov et al., 2024a, Section 4.4), a bound stated with the second moment of the values in place of σ^2 ; our bounds depend on σ but not on the size of the values, since the stencil uses differences of values. Bandit methods use ε^{-2} calls up to logarithmic factors, removed for $d = 1$ only

recently (Carpentier et al., 2026), but with a larger power of d (Lattimore and György, 2023; Fokkema et al., 2024; Lattimore, 2024): the methods of Lattimore and György (2023); Fokkema et al. (2024) use $d^p \varepsilon^{-2}$ calls with $3 \leq p \leq 7/2$ for centered sub-Gaussian noise and functions that are Lipschitz on the domain, as ours are with constant $2LR$ on $B(R)$. The lower bound derived from the separable functions of Akhavan et al. (2020, 2024) (Section M.4) is smaller in d or in $1/\varepsilon$. For convex Lipschitz functions, smoothing uses $d^2 \varepsilon^{-4}$ calls (Flaxman et al., 2005), which Rajaraman and Han (2026) proved necessary for ε at least of order $d^{-1/4}$ by hiding curvature near a tube defined by an unknown linear map. For other norms, a linear map onto a Euclidean ball transfers lower bounds, as Dorn et al. (2026) showed for exact values. One-point methods with nearly linear dependence on d were known on the simplex for $\beta \leq 2$ and Euclidean constants (Gasnikov et al., 2017), and two-point ones adapted to ℓ_1 by Gasnikov et al. (2016b); Akhavan et al. (2022). For first-order methods, a mismatch between the norm of smoothness and the feasible set also changes the complexity: for a Euclidean-Lipschitz gradient on the ℓ_1 ball, LR^2/N^3 after N gradients is the minimax rate up to logarithms in high dimension, $d \geq N + 1$ for deterministic and $d \geq 32N + 1$ for randomized methods (Loginov et al., 2026).

S COMPARISON WITH CONCURRENT WORK

Zhang et al. (2026) study noisy zeroth-order minimization of strongly convex functions whose derivative of order q is Hölder continuous in the multilinear operator norm, from values with independent Gaussian noise and with adaptive queries anywhere in $\mathbb{R}^d$. In our notation their class is $\mathcal{F}_{\text{sc}}(\beta, 1; 1, 3, 1)$ with the operator seminorm and $\sigma = 1$, and their main theorem states that for every fixed $\beta > 2$ and all integers $d, T \geq 1$ the minimax error is of order $\min\{1, (d^2/T)^{(\beta-1)/\beta}\}$, so that $d^2 \varepsilon^{-\beta/(\beta-1)}$ calls are necessary and sufficient for small ε . Their Appendix E extends this to other fixed values of the parameters, with constants that depend on them. The two works were carried out independently. This appendix lists where they overlap and where they differ.

Overlap. For the operator seminorm, Lemma E.5 with $s = 2$ is their main theorem, and Theorem 1(b) contains their lower bound. Both lower bounds perturb a quadratic by a linear function with direction $\omega/\sqrt{d}$, $\omega \in \{-1, 1\}^d$, under one radial cutoff, and both bound the sum over the coordinates of the squared changes of a value under a sign flip uniformly in the query point (Lemma E.3, Step 3). The upper bound of this order was known (Akhavan et al., 2024); they also prove it with a spherical estimator at the radii $2^{-j}h$ combined with Lagrange weights and with projected stochastic gradient descent, one estimate per iteration.

Differences.

- *Dependence on the parameters.* In Theorem 1 the constants depend only on β : for $\delta = 0$, $0 \leq M \leq \alpha R$ and budgets above the two thresholds stated after Theorem 1, the minimax error is of order $\Phi_T + M^2/\alpha$ jointly in all parameters, and Theorem 1(b) holds for every budget with the explicit saturation level Φ_0 . Zhang et al. (2026, Section 5) do not claim a formula that is sharp jointly in these parameters.
- *Hilbert–Schmidt smoothness.* They do not consider the Hilbert–Schmidt seminorm, for which the dimension factor is $d^{3/2}$ for even q and $d^{2-(q-1)/(2(\beta-1))}$ for odd q (Theorem 1 and Lemma E.5); for even q this needs the ridge functions of Lemma E.4. They discuss the result of Yu et al. (2024) for a Frobenius-norm bound on the Lipschitz constant of the Hessian, $\beta = 3$, as a neighboring class.
- *Rounds.* Their method computes one estimate per iteration, so it uses as many sequential rounds as iterations. Theorem 1(a) attains the same number of calls in $O(\sqrt{\kappa} \log(1/\varepsilon))$ rounds, and Theorem 2 bounds the number of rounds from below.
- *Noise.* They consider independent Gaussian noise. Here the random noise may be biased if it is independent of the directions of the method, and Theorem 1(c) determines the admissible levels of Lipschitz systematic errors and, up to $C_\beta \min\{\sqrt{d}, \sqrt{\log(4T)}\}$, of adversarial errors.
- *Other classes.* Functions whose Hessian is not bounded from above (Lemma I.1 and Theorem 3), convex functions (Section 6) and other norms and constraints (Section 7) are not considered there.

Estimators. Both upper bounds cancel the lower-order terms of the Taylor expansion by combining several radii, as in Fabian (1967) and Polyak and Tsybakov (1990). Their estimator differentiates along one random direction at the radii $2^{-j}h$; ours uses the nodes j/m along all axes of a random orthonormal basis and averages such frames within a round (Section 4).