THE COMPLEXITY OF CONVEX OPTIMIZATION WITH MISMATCHED GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimal first-order methods on non-Euclidean domains such as the ℓ_1 ball $B_1^n(R) = \{x \in \mathbb{R}^n : \|x\|_1 \le R\}$ pair the prox-function with the norm in which smoothness is measured. When the gradient is L-Lipschitz in the Euclidean norm only, the accelerated method with a Euclidean prox-setup reduces the functional gap $f(x_N) - \min_{B_1^n(R)} f$ to $O(LR^2/N^2)$ after N first-order queries (an entropic ℓ_1 prox-setup replaces L by the $\ell_1 \to \ell_\infty$ constant $L_1 \le L$, at the cost of a factor $\log n$ and with the same exponent). The lower bound of Guzmán and Nemirovski is of order LR^2/N^3, and whether the upper bound can be brought down to that order is a question of A. S. Nemirovski. We show that for this mismatched problem the minimax value of the gap is of order LR^2/N^3 up to logarithmic factors, for deterministic and for randomized methods alike, and already in dimension proportional to the number of queries. The upper bound is attained by a Steiner-point level method: every supporting hyperplane seen so far is kept as a level cut, and the next query is made near the Steiner point of the resulting localization polytope. The analysis rests on a single geometric fact: along nested subsets of the ball the Steiner points travel a distance that is polylogarithmic in the dimension, in contrast to $\sqrt{n}$ for the Euclidean ball. All queries stay feasible, and a randomized selector keeps the internal work polynomial. The same geometry yields optimal rates for nonsmooth objectives, Hölder gradients, higher-order oracles and Lipschitz monotone operators, the last with a matching deterministic lower bound. For convex quadratics, a curvature-learning method attains the optimal rate on every ℓ_p ball with $1 \le p < 2$ and without logarithmic loss. Experiments confirm the predicted N^{-3} behaviour on objectives that are hard for Euclidean methods.

1 INTRODUCTION

Consider minimizing a convex function f over the ℓ_1 ball $B_1^n(R) = \{x \in \mathbb{R}^n : \|x\|_1 \le R\}$, or over the probability simplex, when all we know about the smoothness of f is a bound in the Euclidean norm,

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2.$$

This situation is common. The ℓ_1 ball and the simplex are the standard feasible sets for sparse and probabilistic models, while for a least-squares loss the natural smoothness constant is the largest eigenvalue of the Gram matrix, a Euclidean quantity.

First-order methods for constrained problems are built around a *prox-setup*: a prox-function that is strongly convex with respect to the same norm in which the gradient is Lipschitz (Nemirovsky & Yudin, 1983; Ben-Tal & Nemirovski, 2001; Beck & Teboulle, 2003; Gasnikov et al., 2026); methods for relatively smooth functions follow the same principle (Hanzely et al., 2021). With the Euclidean prox-function, accelerated projected gradient reduces the optimality gap to $O(LR^2/N^2)$ after N gradient queries. The entropy prox-function is adapted to the geometry of the ℓ_1 ball, but it needs the constant $L_1 \le L$ of ∇f as a map from $(\mathbb{R}^n, \|\cdot\|_1)$ to $(\mathbb{R}^n, \|\cdot\|_\infty)$ and gives $O(L_1 R^2 \log n/N^2)$ (Juditsky & Nemirovski, 2011); when L_1 is close to L, this is no better.

We compare rates by their *exponent*: a bound of order $N^{-\alpha}$, up to logarithmic factors, has exponent α. Both methods above have exponent two. The lower bound of Guzmán & Nemirovski (2015) for this class is of order LR^2/N^3 when $n > N$, so it has exponent three. Which exponent is correct was posed as an open problem by Guzmán (2015); Vorontsova et al. (2019) attribute the question to

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

A. S. Nemirovski. The gap does not come from a poor choice of constants: the lower bound holds already for functions whose smallest Euclidean and ℓ_1 smoothness constants coincide (Section D.2).

Two known approaches each capture half of what is needed. Prox methods accelerate, but their progress is measured in the norm of the prox-function, and in the Euclidean norm the ℓ_1 ball is as large as the Euclidean ball of the same radius. Cutting-plane methods, such as the method of centers of gravity or the volumetric method, use the shape of the feasible set: they keep every cut and query a center of the remaining localization set (Levin, 1965; Vaidya, 1996; Lee et al., 2015; Jiang et al., 2020). Their guarantees come from volume reduction and require order $n \log(1/\varepsilon)$ queries, which says nothing in the regime $N \le n$ where the lower bound lives. Bundle and level methods combine cuts with prox steps and are optimal when the prox-function matches the smoothness norm (Lemaréchal et al., 1995; Lan, 2015). For convex quadratics, Ouyang (2026) recently obtained the rate $O(LR^2/N^3)$ on the ℓ_1 ball by learning the Hessian from gradient differences; that argument does not extend beyond quadratics.

Our approach. We show that the exponent is three. The method is an accelerated level method that keeps all its cuts, but replaces the prox step by a move of its center to the *Steiner point* of the current localization polytope, the average of the maximizers of a random Gaussian linear function over the polytope (Shephard, 1968; Schneider, 2014); as in every accelerated method, the queries are convex combinations of this center and the current iterate. Bubeck et al. (2020) used this point for chasing nested convex bodies, because it moves little when the body shrinks. On the ℓ_1 ball this effect is strong: along any nested sequence of N subsets of $B_1^n(R)$, the Steiner points travel a total Euclidean distance of order $R\sqrt{\log N \log n}$, whereas inside a Euclidean ball the bound is of order $R\sqrt{n \log N}$. An accelerated level method needs exactly such a bound. Every query either permits an accelerated step, whose error term is the squared displacement of the center, or moves the center far; the short path limits both kinds of queries, and balancing them gives the error LR^2/N^3 up to logarithmic factors.

Contributions.

- *The smooth case.* For convex functions with Euclidean L-Lipschitz gradient on $B_1^n(R)$, the minimax error after N first-order queries is LR^2/N^3 up to the factor $\log(2N)\log(2n)$, for deterministic and for randomized methods, when n is at least proportional to N (Section 2). The upper bound is attained by an explicit method whose queries all lie in the feasible set; its randomized version solves polynomially many linear programs.

- *Other problem classes.* Keeping the geometric argument and changing the one-step inequality, we obtain methods with the optimal exponent for nonsmooth objectives (1 instead of $1/2$), Hölder continuous gradients and higher-order oracles, and with the exponent 2 instead of 1, optimal for deterministic methods, for monotone variational inequalities and convex–concave saddle-point problems. Restarts give the strongly convex and strongly monotone versions.

- *Lower bounds.* We extend the deterministic first-order lower bounds of Guzmán & Nemirovski (2015) to randomized methods and higher-order oracles, and prove lower bounds for quadratics, smooth strongly convex functions and Lipschitz variational inequalities. All of them match the rates above up to logarithmic factors, and they also hold for methods that take prox steps with any known prox-function, for instance the entropy.

- *Quadratics.* A curvature-learning method removes the logarithmic factors for convex quadratics and extends the result of Ouyang (2026) from the ℓ_1 ball to every ℓ_p ball with $1 \le p < 2$, with the optimal exponent $1 + 2/p$.

Table 1 compares these rates with those of standard methods, which measure all distances in the Euclidean norm, and with the lower bounds known before. Lower bounds proved for the Euclidean ball do not apply on $B_1^n(R)$: our upper bounds are below them. Experiments on instances that are hard for Euclidean methods confirm the short Steiner paths and show fewer oracle calls for a given accuracy, at a much higher cost per call (Section 6).

Table 1: Known bounds and ours on $B_1^n(R)$ with Euclidean constants; new results are in bold.

	Upper bound		Lower bound		
Class	known	ours	known	ours	Results
Convex minimization: error after N queries					
G-Lipschitz f	$O(\frac{GR}{\sqrt{N}})$	$\boldsymbol{\widetilde{O}(\frac{GR}{N})}$	$\Omega(\frac{GR}{N})^\dagger$	$\boldsymbol{\Omega(\frac{GR}{N})}$	D.1, D.6, E.1
ν-Hölder ∇f	$O(\frac{L_{1,\nu}R^{1+\nu}}{N^{(1+3\nu)/2}})$	$\boldsymbol{\widetilde{O}(\frac{L_{1,\nu}R^{1+\nu}}{N^{1+2\nu}})}$	$\Omega(\frac{L_{1,\nu}R^{1+\nu}}{N^{1+2\nu}})^\dagger$	$\boldsymbol{\Omega(\frac{L_{1,\nu}R^{1+\nu}}{N^{1+2\nu}})}$	D.1, D.8, E.3
L-Lipschitz ∇f	$O(\frac{LR^2}{N^2})$	$\boldsymbol{\widetilde{O}(\frac{LR^2}{N^3})}$	$\Omega(\frac{LR^2}{N^3})^\dagger$	$\boldsymbol{\Omega(\frac{LR^2}{N^3})}$	Thm. 1
Hölder $D^k f$	$O(\frac{L_{k,\nu}R^s}{N^{(3s-2)/2}})$	$\boldsymbol{\widetilde{O}(\frac{L_{k,\nu}R^s}{N^{2s-1}})}$	—	$\boldsymbol{\Omega(\frac{L_{k,\nu}R^s}{N^{2s-1}})}$	D.4, D.8, E.4
quadratic, $p=1$	$O(\frac{LR^2}{N^3})$	$O(\frac{LR^2}{N^3})$	—	$\boldsymbol{\Omega(\frac{LR^2}{N^3})^\dagger}$	H.2, H.9
quadratic, $p>1$	$O(\frac{LR^2}{N^2})$	$\boldsymbol{O(\frac{LR^2}{N^{1+2/p}})}$	—	$\boldsymbol{\Omega(\frac{LR^2}{N^{1+2/p}})^\dagger}$	H.2, H.9
Strongly convex minimization: queries for error ε					
G-Lipschitz f	$O(\frac{G^2}{\mu_1\varepsilon})$	$\boldsymbol{\widetilde{O}(\frac{G}{\sqrt{\mu_1\varepsilon}})}$	—	—	F.2
ν-Hölder ∇f	$O(\varrho^{\frac{1}{1+3\nu}})$	$\boldsymbol{\widetilde{O}(\varrho^{\frac{1}{2+4\nu}})}$	—	—	F.2
L-Lipschitz ∇f	$O(\kappa_1^{1/2}\log\frac{1}{\varepsilon})$	$\boldsymbol{\widetilde{O}(\kappa_1^{1/3}\log\frac{1}{\varepsilon})}$	—	$\boldsymbol{\widetilde{\Omega}(\kappa_1^{1/3}\log\frac{1}{\varepsilon})^{\dagger\ddagger}}$	F.2, F.9
Hölder $D^k f$	$O(\bar\kappa^{\frac{2}{3s-2}}+\log\log\frac{1}{\varepsilon})$	$\boldsymbol{\widetilde{O}(\bar\kappa^{\frac{1}{2s-1}}+\log\frac{1}{\varepsilon})}$	—	—	F.3
quadratic	$O(\kappa_p^{1/2}\log\frac{1}{\varepsilon})$	$\boldsymbol{O(\kappa_p^{p/(p+2)}\log\frac{1}{\varepsilon})}$	—	$\boldsymbol{\Omega(\kappa_p^{p/(p+2)})^\dagger}$	H.5, H.11
Monotone variational inequalities: weak gap after N calls					
L-Lipschitz F	$O(\frac{LR^2}{N})$	$\boldsymbol{\widetilde{O}(\frac{LR^2}{N^2})}$	—	$\boldsymbol{\Omega(\frac{LR^2}{N^2})^\dagger}$	G.2, G.7
ν-Hölder F	$O(\frac{L_{0,\nu}R^{1+\nu}}{N^{(1+\nu)/2}})$	$\boldsymbol{\widetilde{O}(\frac{L_{0,\nu}R^{1+\nu}}{N^{1+\nu}})}$	—	—	G.3
Strongly monotone variational inequalities: calls for $\|x-x^\star\|_1\le\varepsilon$					
L-Lipschitz F	$O(\kappa_1\log\frac{R}{\varepsilon})$	$\boldsymbol{\widetilde{O}(\kappa_1^{1/2}\log\frac{R}{\varepsilon})}$	—	—	G.2

$L_{k,\nu}$ is the ν-Hölder constant of $D^k f$ (of F for $k=0$) and $s=k+\nu$, with $\nu<1$ and $s>2$ in the strongly convex Hölder rows; μ_p is the modulus of strong convexity or monotonicity in ℓ_p, $\kappa_p=L/\mu_p$, $\bar\kappa=L_{k,\nu}R^{s-2}/\mu_1$, $\varrho=L_{1,\nu}^2\mu_1^{-1-\nu}\varepsilon^{\nu-1}$; the quadratic rows are on $B_p^n(R)$, $1\le p<2$. Known upper bounds are those of Euclidean methods and, for $p=1$, of Ouyang (2026); known lower bounds are those of Guzmán & Nemirovski (2015); Guzmán (2015). $\widetilde{O}$ and $\widetilde{\Omega}$ hide polylogarithmic factors; lower bounds need n at least proportional to N; †: deterministic methods only; ‡: if $\kappa_1\ge 2n\ge 2\kappa_1^{1/3}$; —: none known.

2 Setting and main result

For $1\le p\le\infty$, $\|\cdot\|_p$ is the ℓ_p norm on $\mathbb{R}^n$, $B_p^n(R)=\{x:\|x\|_p\le R\}$, and $e_1,\dots,e_n$ are the coordinate vectors. Unless stated otherwise, regularity constants are Euclidean: f is *L-smooth* if $\|\nabla f(x)-\nabla f(y)\|_2\le L\|x-y\|_2$ for all x,y. We write $f^\star=\min_K f$ for the feasible set K, and log is the natural logarithm.

Oracle model. A method for minimizing f over a known set K chooses query points $x_1,\dots,x_N\in\mathbb{R}^n$ one after another, each depending on the previous answers; at every query the *first-order oracle* returns $f(x_t)$ and $\nabla f(x_t)$. After N queries the method outputs a point $\widehat{x}\in K$. Only oracle calls are counted, while computations with known objects, such as linear programs over K, are free; this is the information-based complexity of Nemirovsky & Yudin (1983). A query is *feasible* if it lies in K. Let $\mathcal{F}_L$ be the class of convex L-smooth functions on $\mathbb{R}^n$. The minimax error of deterministic methods on the ℓ_1 ball is

$$\mathcal{E}_N^{\mathrm{det}}=\inf_{\text{methods}}\sup_{f\in\mathcal{F}_L}\Big[f(\widehat{x})-\min_{B_1^n(R)}f\Big],$$

and $\mathcal{E}_N^{\mathrm{ran}}$ is defined in the same way for randomized methods, with the error averaged over the internal randomness of the method.

Theorem 1 (Minimax rate for smooth functions)**.** *For every $N\ge 1$,*

$$\mathcal{E}_N^{\mathrm{det}}\ge\frac{LR^2}{16(N+1)^3}\quad(n\ge N+1),\qquad\mathcal{E}_N^{\mathrm{ran}}\ge\frac{LR^2}{2^{50}N^3}\quad(n\ge 32N+1),$$

and, in every dimension n,

$$\mathcal{E}_N^{\mathrm{ran}} \le \mathcal{E}_N^{\mathrm{det}} \le \frac{2^{35}\, LR^2 \log(2N)\log(2n)}{N^3}.$$

The upper bound is attained by a method whose queries all lie in $B_1^n(R)$.

Thus the answer to the question of Section 1 is N^{-3}: when n is at least proportional to N, the minimax error is LR^2/N^3 up to the factor $\log(2N)\log(2n)$, and randomization does not change the exponent. Some condition on the dimension is necessary, because in a fixed dimension cutting-plane methods converge linearly. The deterministic lower bound is the construction of Guzmán & Nemirovski (2015) with an explicit constant; the randomized lower bound and the upper bound are new. The upper bound holds, with the same proof, on every polytope contained in a translate of $B_1^n(R)$, in particular on the simplex. Its constant is large, so Theorem 1 is a complexity result rather than a practical iteration count.

Section 3 presents the method and proves the upper bound up to routine steps, Section 4 explains the lower bounds, and Section 5 shows how the same construction gives the other rows of Table 1.

3 The method: short Steiner paths and an accelerated level method

Throughout this section K is a nonempty polytope, given explicitly by linear inequalities, such that $K \subseteq c + B_1^n(R)$ for some known $c \in \mathbb{R}^n$ and $R > 0$. The main example is $K = B_1^n(R)$ with $c = 0$.

3.1 Steiner points move little

Let Z be a standard Gaussian vector in $\mathbb{R}^n$. For a nonempty compact convex set $C \subseteq \mathbb{R}^n$, the maximizer $x_C(Z)$ of the linear function $x \mapsto \langle Z, x\rangle$ over C is unique with probability one, and the *Steiner point* of C is its expectation,

$$\mathrm{st}(C) = \mathbb{E}\, x_C(Z) \ \in C. \tag{1}$$

This is the classical Steiner point of convex geometry written in Gaussian form (Shephard, 1968; Schneider, 2014). It belongs to C because it is an average of points of C, and it can be estimated by averaging the solutions of a few linear programs. The property we need is its stability: when C shrinks, the point $\mathrm{st}(C)$ moves little.

Lemma 3.1 (Short Steiner paths)**.** *Let $T \ge 1$ and let $C_0 \supseteq C_1 \supseteq \cdots \supseteq C_T$ be nonempty compact convex subsets of $c + B_1^n(R)$. Then*

$$\sum_{t=1}^{T} \|\mathrm{st}(C_t) - \mathrm{st}(C_{t-1})\|_2 \ \le\ 4R\sqrt{(1+\log T)(1+\log(2n))}. \tag{2}$$

Proof sketch. Let $h_C(z) = \max_{x\in C}\langle z, x\rangle$ be the support function of C. Gaussian integration by parts gives $\mathrm{st}(C) = \mathbb{E}[Z\, h_C(Z)]$. Let v_t be the unit vector in the direction of $\mathrm{st}(C_{t-1}) - \mathrm{st}(C_t)$ and $A(Z) = \max\{0, \langle v_1, Z\rangle, \dots, \langle v_T, Z\rangle\}$. The sets are nested, so every difference $h_{C_{t-1}} - h_{C_t}$ is nonnegative, and the differences telescope:

$$\sum_{t=1}^{T}\|\mathrm{st}(C_{t-1}) - \mathrm{st}(C_t)\|_2 = \mathbb{E}\sum_{t=1}^{T}\langle v_t, Z\rangle\,\big(h_{C_{t-1}}(Z) - h_{C_t}(Z)\big) \le \mathbb{E}\Big[A(Z)\big(h_{C_0}(Z) - h_{C_T}(Z)\big)\Big]. \tag{3}$$

Since $C_T \subseteq C_0 \subseteq c + B_1^n(R)$, we have $0 \le h_{C_0}(Z) - h_{C_T}(Z) \le 2R\|Z\|_\infty$, and the Cauchy–Schwarz inequality gives

$$\mathbb{E}\Big[A(Z)\big(h_{C_0}(Z) - h_{C_T}(Z)\big)\Big] \ \le\ 2R\,\big(\mathbb{E}A(Z)^2\big)^{1/2}\big(\mathbb{E}\|Z\|_\infty^2\big)^{1/2}. \tag{4}$$

Both factors are maxima of Gaussian variables with variance at most one: $\mathbb{E}A(Z)^2 \le 2(1+\log T)$ and $\mathbb{E}\|Z\|_\infty^2 \le 2(1+\log(2n))$. The full proof is in Section B. □

The dimension enters only through $\log n$, because an ℓ_1 ball has small Gaussian width: its support function is $R\|Z\|_\infty$, of order $R\sqrt{\log n}$. For a Euclidean ball the same argument gives $2R\sqrt{2(1+\log T)\,n}$, and the experiments of Section 6 show that movement of order $\sqrt{n}$ indeed occurs there. Neither logarithm in (2) can be dropped in general: along the faces $\mathrm{conv}\{Re_1, \dots, Re_j\}$, $j = n, n-1, \dots, 1$, of the simplex, the Steiner points travel a distance of order $R \log n$ (Proposition B.4). In the algorithms we use the slightly larger *path budget*

$$P_N = 4R\sqrt{\mathsf{b}_N \mathsf{c}_n}, \qquad \mathsf{b}_N = 1 + \lceil \log_2 N \rceil, \quad \mathsf{c}_n = 1 + \lceil \log_2(2n) \rceil, \tag{5}$$

which bounds the left-hand side of (2) for every chain of length $T \le N$.

3.2 An accelerated level method with Steiner centers

Fix a target level $\ell \in \mathbb{R}$. A query at x defines the *level cut* $\{u : f(x) + \langle \nabla f(x), u - x \rangle \le \ell\}$; by convexity it contains every point u with $f(u) \le \ell$. Intersecting K with the cuts produces nested localization polytopes $C_0 \supseteq C_1 \supseteq \cdots$, and if one of them is empty, then $\min_K f > \ell$. Accelerated level methods (Lan, 2015) combine such cuts with the three sequences x_t, y_t, z_t of an accelerated method and choose z_t by a prox step onto the current polytope. Algorithm 1 chooses instead a point z_t close to the Steiner point of C_t. This point is not selected to decrease any distance, so each iteration, or *trial*, checks how far the center moved: a short move leads to the accelerated update, while a long move is *rejected*, although its cut is kept. We call a point $z \in C$ with $\|z - \mathrm{st}(C)\|_2 \le \eta$ an *η-approximate Steiner point* of C.

Algorithm 1 SP-Level(ℓ, N): accelerated level method with Steiner centers

Input: polytope $K \subseteq c + B_1^n(R)$; first-order oracle of f; constant L; level ℓ; horizon N

$\eta_N \leftarrow P_N/(4N)$, $\delta_N \leftarrow 4P_N/N$ ▷ center accuracy and acceptance threshold
$C_0 \leftarrow K$, $A_0 \leftarrow 0$; $y_0 \leftarrow$ any point of K; $z_0 \leftarrow$ an η_N-approximate Steiner point of K
for $t = 1, \dots, N$ **do**
  $a_t \leftarrow \big(1 + \sqrt{1 + 4LA_{t-1}}\big)/(2L)$, $B_t \leftarrow A_{t-1} + a_t$ ▷ so that $La_t^2 = B_t$
  $x_t \leftarrow (A_{t-1}y_{t-1} + a_t z_{t-1})/B_t$; query $f(x_t)$ and $g_t = \nabla f(x_t)$
  $C_t \leftarrow C_{t-1} \cap \{u : f(x_t) + \langle g_t, u - x_t \rangle \le \ell\}$
  if $C_t = \varnothing$ **then return** "$\min_K f > \ell$"
  $z_t \leftarrow$ an η_N-approximate Steiner point of C_t
  if $\|z_t - z_{t-1}\|_2 \le \delta_N$ **then** ▷ accept
    $y_t \leftarrow (A_{t-1}y_{t-1} + a_t z_t)/B_t$, $A_t \leftarrow B_t$
  else ▷ reject
    $y_t \leftarrow y_{t-1}$, $A_t \leftarrow A_{t-1}$
return y_N

Proposition 3.2 (Level method)**.** *Let K be a nonempty polytope given by linear inequalities with $K \subseteq c + B_1^n(R)$, and let f be convex with L-Lipschitz Euclidean gradient on K. For every level $\ell \in \mathbb{R}$ and horizon $N \ge 1$, Algorithm 1 makes at most N first-order queries, all in K, and either certifies that $\min_K f > \ell$ or returns $y_N \in K$ with*

$$f(y_N) \;\le\; \ell + \frac{48LP_N^2}{N^3} \;=\; \ell + \frac{768\,LR^2\,\mathsf{b}_N\mathsf{c}_n}{N^3}. \tag{6}$$

Proof sketch. The points x_t, y_t and z_t are convex combinations of points of K, so all queries are feasible. If $\min_K f \le \ell$, a minimizer satisfies every level cut, so an empty C_t certifies $\min_K f > \ell$. Assume that all C_t are nonempty. Two facts drive the proof.

The energy inequality. For $A \ge 0$, $a > 0$, $B = A + a$, $x = (Ay + az)/B$ and $y^+ = (Ay + az^+)/B$, if z^+ lies in the level cut at x, then

$$B\,[f(y^+) - \ell] \;\le\; A\,[f(y) - \ell] + \frac{La^2}{2B}\|z^+ - z\|_2^2. \tag{7}$$

It follows from the descent lemma between x and y^+, after the linear term is split into the convexity inequality at x, evaluated at y with weight A, and the level cut at z^+ with weight a. The rule $La_t^2 = B_t$ makes the last coefficient equal to $\frac{1}{2}$.

The path bound. Let $d_t = \|z_t - z_{t-1}\|_2$. The polytopes C_t are nested, so Lemma 3.1 and the accuracy η_N of the centers give, for accepted and rejected trials together,

$$\sum_{t=1}^{N} d_t \;\le\; P_N + 2N\eta_N \;=\; \tfrac{3}{2}P_N. \tag{8}$$

A rejected trial has $d_t > \delta_N = 4P_N/N$, so fewer than $3N/8$ trials are rejected and more than $N/2$ are accepted. Summing (7) over the accepted trials, where $d_t \le \delta_N$, gives

$$A_N\,[f(y_N) - \ell] \;\le\; \frac{1}{2}\sum_{t\ \text{accepted}} d_t^2 \;\le\; \frac{\delta_N}{2}\sum_{t=1}^{N} d_t \;\le\; \frac{3P_N^2}{N},$$

and the weights of more than $N/2$ accepted steps satisfy $A_N \ge N^2/(16L)$. Dividing by A_N gives (6); Section C has the details. □

The mechanism behind the rate is visible in the last display. As in every accelerated method, the weight A_N grows like N^2/L. For a prox step the squared displacements would sum to order R^2; here the path bound makes them sum to $O(P_N^2/N)$, which gains one power of N.

3.3 Finding the level and computing the centers

The level method needs a target level, while $f^\star = \min_K f$ is unknown. One query at a point $x_0 \in K$ brackets $f^\star$: with $g_0 = \nabla f(x_0)$ and a vertex v minimizing $\langle g_0, u\rangle$ over $u \in K$, convexity and smoothness give

$$f(x_0) + \langle g_0, v - x_0\rangle \;\le\; f^\star \;\le\; f(v) \;\le\; f(x_0) + \langle g_0, v - x_0\rangle + \tfrac{L}{2}\|v - x_0\|_2^2. \tag{9}$$

The complete method, SP-Accel (Algorithm 2 in Section C), bisects this bracket. At the midpoint level it runs Algorithm 1 with the smallest horizon $N = 2^j$ whose guarantee (6) is at most a quarter of the current width. A certificate $\min_K f > \ell$ halves the bracket; otherwise the returned point shrinks it by the factor $3/4$. The widths decrease geometrically and the horizons are minimal powers of two, so the horizons are dominated by a geometric series: the total number of queries is at most $1 + 24N(\varepsilon)$, where $N(\varepsilon)$ is the horizon that the rule assigns to the width ε, which bounds every horizon used (Proposition C.2). The search for the level therefore costs only a constant factor, and choosing ε from the budget gives the upper bound of Theorem 1.

It remains to compute the centers. Computing the Steiner point of a known polytope costs no oracle calls, and a finite, though exponential-time, procedure approximates it to any accuracy (Proposition C.3). A practical estimator uses sampling: for independent Gaussian vectors $Z_1, \dots, Z_m$, the average of the maximizers of $\langle Z_i, u\rangle$ over $u \in C$ lies in C, is an unbiased estimate of $\mathrm{st}(C)$, and concentrates at the rate $R/\sqrt{m}$. With m polynomial in $N(\varepsilon)$ and in the logarithm of the inverse failure probability, all guarantees hold with high probability, and the method solves polynomially many linear programs (Proposition C.4).

4 Lower bounds

Deterministic methods. The hard functions of Guzmán & Nemirovski (2015) are maxima of signed coordinates, built by a resisting oracle. At the t-th query x_t, the oracle takes, among the coordinates it has not used yet, one with the largest $|x_{t,i}|$, sets $v_t = \mathrm{sign}(x_{t,i})\, e_i$, and answers according to

$$\varphi_t(x) = \max_{j \le t}\big\{\langle v_j, x\rangle - (j-1)\tau\big\}$$

with a small offset $\tau > 0$. Near every past query, each form revealed later lies at least τ below the current maximum, so every completion of φ_t gives the same answers, and one fixed function reproduces the whole interaction. After N queries one coordinate is still unused. Giving it the sign of the corresponding coordinate of the output $\widehat{x}$ defines v_{N+1} and the final function $\varphi = \varphi_{N+1}$ with $\varphi(\widehat{x}) \ge -N\tau$, while the feasible point $-R\sum_{j \le N+1} v_j/(N+1)$ has value $-R/(N+1)$. Smoothing φ by a Moreau envelope makes it L-smooth and gives the bound $LR^2/[16(N+1)^3]$ of Theorem 1. For order-k oracles we smooth by convolution with a compactly supported product kernel, whose derivative bounds do not depend on the dimension; this gives the exponent $2(k+\nu) - 1$. The bounds allow queries anywhere in $\mathbb{R}^n$ and any computation between queries, so they also apply to methods that take prox steps with a known prox-function, for example the entropy on the simplex.

Randomized methods. Against a randomized method, the hard function has to be chosen before the interaction, as a random function (Diakonikolas & Guzmán, 2020; Agarwal & Hazan, 2018). We use a smoothed maximum $\max_i(x_i - a_i)$ with independent exponential shifts a_i. Given the answers so far, the shifts that have not been revealed are independent and, by the memoryless property, exponential above known thresholds. Because the smoothing parameter is small compared with the mean shift, one query reveals at most two new coordinates on average, a constant fraction of the coordinates remains hidden after N queries when $n \geq 32N + 1$, and the output cannot locate the minimizer. This gives the randomized bound of Theorem 1 and its analogue for higher-order oracles. Section D contains the proofs.

5 The same geometry for other problem classes

The proof of Proposition 3.2 uses smoothness only through the energy inequality (7) and the geometry only through Lemma 3.1. Keeping the second and replacing the first gives methods for the other rows of Table 1, except the quadratic ones, which use a different mechanism; Sections E to H contain the statements and proofs.

Nonsmooth objectives. Let f be convex and G-Lipschitz, with a subgradient g_t returned at each query. The method SP-Cut queries an approximate Steiner point x_t of the current polytope and keeps only the points u with $\langle g_t, x_t - u\rangle \geq \varepsilon$. Such deep cuts may remove every minimizer; their role is to force the Steiner point to move. While the polytope is nonempty, the new Steiner point lies in it, hence at distance at least ε/G from x_t, and Lemma 3.1 allows only $O(GP_N/\varepsilon)$ such moves. Once the polytope is empty, linear-programming duality gives weights $\lambda_t \geq 0$, $\sum_t \lambda_t = 1$, with

$$\max_{u\in K} \sum_t \lambda_t \langle g_t, x_t - u\rangle < \varepsilon, \tag{10}$$

and by convexity the average $\sum_t \lambda_t x_t$ is ε-optimal. The certificate uses the stored subgradients only, and the error after N queries is $\widetilde{O}(GR/N)$.

Hölder gradients and higher-order oracles. If ∇f is ν-Hölder with constant $L_{1,\nu}$, the remainder $\frac{L}{2}\|y^+ - x\|_2^2$ of the descent lemma behind (7) becomes $L_{1,\nu}\|y^+ - x\|_2^{1+\nu}/(1+\nu)$, and the same counting gives the exponent $1 + 2\nu$. With an oracle of order k, each trial computes an approximate stationary point v over K of a regularized Taylor model of f (Nesterov & Polyak, 2006), queries f at v, and uses the supporting hyperplane of f at v, not that of the model, as the level cut; the exponent becomes $2(k + \nu) - 1$. The model may be nonconvex, since the stationarity of v is certified by one linear program. For a Hessian with Lipschitz constant $L_{2,1}$ this gives $\widetilde{O}(L_{2,1}R^3/N^5)$, compared with $N^{-7/2}$ for optimal Euclidean methods (Arjevani et al., 2019; Kovalev & Gasnikov, 2022).

Strong convexity. If f is also μ_1-strongly convex in ℓ_1, a point with error ε lies within ℓ_1 distance $\sqrt{2\varepsilon/\mu_1}$ of the minimizer. Restarting the method on the intersection of K with shrinking ℓ_1 balls, which are again polytopes inside translates of ℓ_1 balls, gives $\widetilde{O}(\kappa_1^{1/3}\log(1/\varepsilon))$ queries for smooth f, $\kappa_1 = L/\mu_1$, and the other counts of Table 1. There is a subtlety: a function that is L-smooth in ℓ_2 and μ_1-strongly convex in ℓ_1 always has $\kappa_1 \geq n$. The complexity is governed by the *slack* $\Lambda = \kappa_1 - n$, since one projected gradient step localizes the minimizer in an ℓ_1 ball whose radius depends on Λ rather than on κ_1. For every $\Lambda \geq 8$ we determine the deterministic complexity up to polylogarithmic factors: with $\bar{n} = \min\{n, \Lambda^{1/3}\}$, it is $\min\{\bar{n}, (LR^2/\varepsilon)^{1/3}\}$ plus $\bar{n}\log_+(LR^2/(\bar{n}^3\varepsilon))$, where $\log_+ = \max\{0, \log\}$ (Section F).

Variational inequalities and saddle-point problems. For a monotone operator F on K, the accuracy of a point $\bar{w}$ is its weak gap $\max_{u\in K}\langle F(u), \bar{w} - u\rangle$, and for a convex–concave function on a product of polytopes it is the saddle gap. The certificate (10) carries over with $F(w_t)$ and w_t in place of g_t and x_t and bounds both gaps. For an L-Lipschitz F, the method SP-VI makes an extragradient step $w_t = \mathrm{proj}_K(z_t - F(z_t)/(2L))$, with the Euclidean projection proj_K onto K, from the current approximate Steiner point z_t and cuts deeply with $F(w_t)$. A nonempty cut at depth ε forces the next center to move by at least $\frac{2}{3}\sqrt{\varepsilon/L}$, so with ε of order LP_N^2/N^2 the method stops within N trials. The weak gap after $2N$ operator calls is therefore $\widetilde{O}(LR^2/N^2)$, against $O(LR^2/N)$

for extragradient and Mirror-Prox (Korpelevich, 1976; Nemirovski, 2004). A skew-symmetric affine operator whose columns are revealed two per query shows that the exponent two is optimal for deterministic methods; the lower bound applies also to bilinear games on products of ℓ_1 balls and of simplices. The same scheme handles Hölder operators and Taylor models of higher order, and restarts give the strongly monotone version.

Quadratics. For a convex quadratic, gradient differences give Hessian–vector products, and Ouyang (2026) accelerates by learning a positive semidefinite lower model of the Hessian through symmetric rank-one (SR1) updates, combined with a Huber regularizer. Replacing the Huber function by a power–Huber function, quadratic near zero and equal to $|t|^p/p$ up to a constant elsewhere, extends the method, which we call CLA (curvature-learning acceleration), to every ℓ_p ball with $1 \le p < 2$. It gives the error $O(LR^2/N^{1+2/p})$ with a constant depending only on p and without logarithmic factors, which matches a lower bound (Section H). The mechanism needs the residual between the function and its learned model to remain a positive semidefinite quadratic, and this fails already in dimension one for nonquadratic functions.

6 Experiments

The experiments check the two ingredients of the method, short Steiner paths and the level method, and compare the methods with standard ones on instances that are hard for Euclidean methods. All instances live on the unit ℓ_1 ball (games: on a product of two such balls), their Euclidean constants are at most one, and errors are computed from known minimizers. Steiner points are estimated by averaging the solutions of m linear programs with random Gaussian objectives, solved with HiGHS; the m directions are shared by all centers of a run, which works better than fresh directions at these sample sizes. SP-Accel doubles its horizons adaptively, and SP-Cut and SP-VI set the cut depth from their certificate, because the theoretical parameters are far too conservative at these budgets. The baselines are accelerated projected gradient with the Euclidean and the entropy prox-function (AGD-E, AGD-Ent), projected gradient (PG), Frank–Wolfe variants, a Euclidean bundle-level method, projected subgradient (PSG), entropic mirror descent (MD-Ent) and extragradient (EG). Section I describes the instances, the implementations and a larger controlled study.

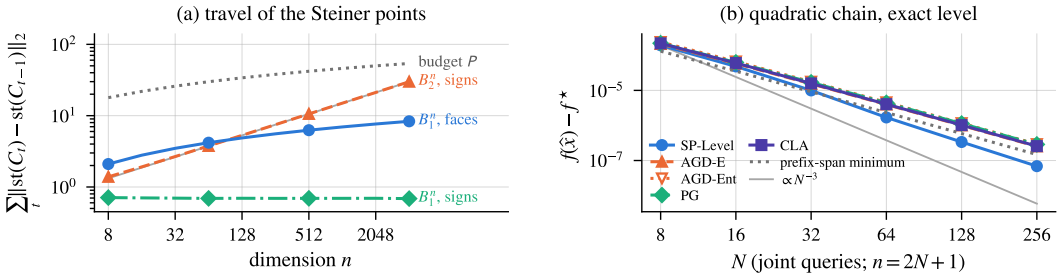

Figure 1: (a) Total movement of Steiner points along nested chains in the unit ℓ_1 and ℓ_2 balls. (b) Error on the rotated quadratic chain, $n = 2N + 1$.

Steiner paths. Figure 1(a) shows the quantity bounded in Lemma 3.1 for the chain $C_t = K \cap \{x : x_1 \le 0, \dots, x_t \le 0\}$, $t \le n$ ("signs"). In the Euclidean ball the Steiner points travel about $0.47\sqrt{n}$. In the ℓ_1 ball they travel about 0.69 for every n between 8 and 4096, a factor 43 less at $n = 4096$. The chain of simplex faces from Section 3.1 ("faces") travels 8.34 at $n = 4096$, in agreement with its $R \log n$ growth. The dotted curve is the path budget P_N at $N = n$.

The level method. Figure 1(b) uses Nesterov's worst-case quadratic chain, rotated so that the minimizer is 5-sparse, in dimension $n = 2N + 1$. On this family Euclidean and entropic methods remove only a small part of the initial gap; the dotted line is the best error attainable in the span of the first N chain coordinates. SP-Level, run at the exact level $\ell = f^\star$ with $m = 256$, has a fitted slope -2.3 over $N = 8, \dots, 256$, and at $N = 256$ its error is 0.27 times that of AGD-E. These budgets are far below the regime of Proposition 3.2, whose constant keeps the guarantee above the initial gap up to $N \approx 3 \cdot 10^7$.

The complete method. In a study with fixed dimension $n = 65$, 5-sparse minimizers and 64 queries, SP-Accel, which also searches for the level, leaves 0.23 of the initial gap on the quadratic chain and 0.11 on a nonquadratic one (medians over three instances). A feasible accelerated method leaves 0.71 and 0.74, and the bundle-level method 0.39 and 0.58. The advantage nearly disappears at $n = 129$ (0.85 against 0.87 on one instance), and with 64 directions on the budget-indexed chains of Fig. 1(b) the complete method does not improve on its initial point (Section I.4). It is also paid for in time: a run of SP-Accel takes 74–94 seconds, mostly in linear programs, against milliseconds for accelerated gradient.

Curvature learning. On a quadratic chain in dimension $n = 2N + 1$ with the 1-sparse minimizer $0.9e_1$, the error of CLA drops by a factor 8.7 between $N = 400$ and $N = 800$, a local slope of -3.1 that matches the predicted N^{-3}, while the error of AGD-E drops by 4.0; at $N = 800$ CLA is 3.6 times more accurate. The gain depends on the sparsity: at $n = 65$ and $N = 128$ (so $N > n$), CLA improves on AGD-E by a factor 6 for a 1-sparse minimizer but not for a 5-sparse or a dense one.

Nonsmooth problems and games. On ℓ_∞ regression with $n = 4N + 1$, the error of SP-Cut decays like $N^{-1.8}$, against $N^{-0.7}$ for PSG and $N^{-0.3}$ for MD-Ent, and its certificate stays within a factor 1.25 of the true error for $N \geq 64$. On bilinear games with $n = 2Q$, the saddle gap of SP-VI after $Q = 256$ operator calls is $8 \cdot 10^{-7}$, against $5.6 \cdot 10^{-3}$ for EG. The bundle-level method, which, unlike SP-Cut, uses function values, is more accurate than SP-Cut on the nonsmooth instances of the controlled study.

Overall, the path bound is visible already in small dimension, the level method and CLA decay faster than the Euclidean rate on the chains, and the deep-cut methods need far fewer calls than subgradient and extragradient methods. Every Steiner-point method is slower in wall-clock time by several orders of magnitude, because each center costs m linear programs; a faster estimator of Steiner points is the main practical open problem.

7 Conclusion

When the gradient is Lipschitz in the Euclidean norm and the feasible set is an ℓ_1 ball, the minimax error of first-order methods in dimension at least proportional to N is LR^2/N^3 up to logarithmic factors, not the rate LR^2/N^2 of methods whose prox-function matches the smoothness norm. The reason is geometric: Steiner points of nested subsets of the ℓ_1 ball move little, and an accelerated level method with Steiner centers turns this stability into acceleration. The same mechanism doubles the exponent for nonsmooth objectives and for monotone variational inequalities, and gives optimal exponents for Hölder and higher-order classes. Four questions remain open: whether the logarithmic factors are necessary; whether Steiner points can be approximated deterministically in polynomial time, or fast enough to make the methods competitive in running time; whether the exponent two for monotone operators is also optimal for randomized methods; and which lower bounds hold in the five rows of Table 1 that have none. Concurrent work is discussed in Section J.

Reproducibility statement

Each result states its assumptions, and complete proofs are in Sections B to H. Algorithms 1 to 9 specify our methods, and Section I gives the instances, parameters, seeds, software and hardware of all experiments.

AI use statement

Generative AI tools were used for language editing and to improve the clarity of the manuscript. We reviewed all resulting changes and take full responsibility for the final content.

References

Naman Agarwal and Elad Hazan. Lower bounds for higher-order convex optimization. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 774–792, 2018. URL https://proceedings.mlr.press/v75/agarwal18a.html.

Yossi Arjevani, Shai Shalev-Shwartz, and Ohad Shamir. On lower and upper bounds in smooth and strongly convex optimization. *Journal of Machine Learning Research*, 17(126):1–51, 2016. URL https://jmlr.org/papers/v17/15-106.html.

Yossi Arjevani, Ohad Shamir, and Ron Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, 178(1–2):327–360, 2019. doi: 10.1007/s10107-018-1293-1. URL https://arxiv.org/abs/1705.07260.

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.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542.

Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. SIAM, 2001. doi: 10.1137/1.9780898718829.

Sébastien Bubeck, Bo'az Klartag, Yin Tat Lee, Yuanzhi Li, and Mark Sellke. Chasing nested convex bodies nearly optimally. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1496–1508, 2020. doi: 10.1137/1.9781611975994.91. arXiv:1811.00999.

Brian Bullins and Kevin A. Lai. Higher-order methods for convex-concave min-max optimization and monotone variational inequalities. *SIAM Journal on Optimization*, 32(3):2208–2229, 2022. doi: 10.1137/21M1395764.

Juan Pablo Contreras, Cristóbal Guzmán, and David Martínez-Rubio. Non-Euclidean high-order smooth convex optimization extended abstract. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *PMLR*, pp. 1330, 2025. URL https://proceedings.mlr.press/v291/contreras25a.html. Extended abstract; full version arXiv:2411.08987.

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.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *International Conference on Machine Learning*, pp. 272–279, 2008. doi: 10.1145/1390156.1390191.

Pavel Dvurechensky, Alexander Gasnikov, Sergey Omelchenko, and Alexander Tiurin. Adaptive similar triangles method: a stable alternative to Sinkhorn's algorithm for regularized optimal transport, 2017. URL https://arxiv.org/abs/1706.07622.

Olivier Fercoq and Zheng Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *IMA Journal of Numerical Analysis*, 39(4):2069–2095, 2019. doi: 10.1093/imanum/drz007. URL https://arxiv.org/abs/1709.02300.

Ankit Garg, Robin Kothari, Praneeth Netrapalli, and Suhail Sherif. Near-optimal lower bounds for convex optimization for all orders of smoothness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 29874–29884, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/fa6c94460e902005a0b660266190c8ba-Abstract.html.

Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Evgeniya Vorontsova, Daniil Selikhanovych, and César A. Uribe. Optimal tensor methods in smooth convex and uniformly convex optimization. In *Conference on Learning Theory*, volume 99 of *PMLR*, pp. 1374–1391, 2019. URL https://proceedings.mlr.press/v99/gasnikov19a.html.

Alexander V. Gasnikov, Darina M. Dvinskikh, Yuriy V. Dorn, Dmitry I. Kamzolov, Yury V. Maximov, Vladislav V. Matyukhin, and Roland F. Hildebrand. How to look into the black box: the works of Yu. E. Nesterov (on the award of the 2026 Gauss prize). *Uspekhi Matematicheskikh Nauk*, 2026. In Russian. To appear.

Geovani N. Grapiglia and Yurii Nesterov. Tensor methods for minimizing convex functions with Hölder continuous higher-order derivatives. *SIAM Journal on Optimization*, 30(4):2750–2779, 2020. doi: 10.1137/19M1259432. URL https://arxiv.org/abs/1904.12559.

Branko Grünbaum. Partitions of mass-distributions and of convex bodies by hyperplanes. *Pacific Journal of Mathematics*, 10(4):1257–1261, 1960. doi: 10.2140/pjm.1960.10.1257.

Cristóbal Guzmán. Open problem: The oracle complexity of smooth convex optimization in non-standard settings. In *Conference on Learning Theory*, volume 40 of *PMLR*, pp. 1761–1763, 2015. URL https://proceedings.mlr.press/v40/Guzman15.html.

Cristóbal Guzmán and Arkadi Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1–14, 2015. doi: 10.1016/j.jco.2014.08.003. URL https://arxiv.org/abs/1307.5001.

Filip Hanzely, Peter Richtárik, and Lin Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79:405–440, 2021. doi: 10.1007/s10589-021-00273-8. URL https://arxiv.org/abs/1808.03045.

Haotian Jiang, Yin Tat Lee, Zhao Song, and Sam Chiu-wai Wong. An improved cutting plane method for convex optimization, convex-concave games, and its applications. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 944–953, 2020. doi: 10.1145/3357713.3384284.

Anatoli Juditsky and Arkadi Nemirovski. First-order methods for nonsmooth convex large-scale optimization, I: General purpose methods. In Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright (eds.), *Optimization for Machine Learning*, pp. 121–148. MIT Press, 2011. doi: 10.7551/mitpress/8996.003.0007.

Narendra Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4 (4):373–395, 1984. doi: 10.1007/BF02579150.

Leonid G. Khachiyan. A polynomial algorithm in linear programming. *Soviet Mathematics Doklady*, 20:191–194, 1979. URL https://www.mathnet.ru/eng/dan42319. Russian original: Doklady Akademii Nauk SSSR, 244(5):1093–1096.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12(4):747–756, 1976. URL https://cs.uwaterloo.ca/~y328yu/classics/extragrad.pdf.

Dmitry Kovalev and Alexander Gasnikov. The first optimal acceleration of high-order methods in smooth convex optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35339–35351, 2022. doi: 10.52202/068431-2561. URL https://arxiv.org/abs/2205.09647.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank–Wolfe optimization variants. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://papers.nips.cc/paper/2015/hash/c058f544c737782deacefa532d9add4c-Abstract.html.

Guanghui Lan. Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. *Mathematical Programming*, 149(1–2):1–45, 2015. doi: 10.1007/s10107-013-0737-x.

Yin Tat Lee, Aaron Sidford, and Sam Chiu-wai Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. In *IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1049–1065, 2015. doi: 10.1109/FOCS.2015.68.

Claude Lemaréchal, Arkadii Nemirovskii, and Yurii Nesterov. New variants of bundle methods. *Mathematical Programming*, 69:111–147, 1995. doi: 10.1007/BF01585555.

A. Yu. Levin. On an algorithm for the minimization of convex functions. *Soviet Mathematics Doklady*, 6:286–290, 1965. URL https://www.mathnet.ru/eng/dan30770. Russian original: Doklady Akademii Nauk SSSR, 160(6):1244–1247.

Tianyi Lin and Michael I. Jordan. Perseus: A simple and optimal high-order method for variational inequalities. *Mathematical Programming*, 209(1):609–650, 2025. doi: 10.1007/s10107-024-02075-2.

Colin McDiarmid. On the method of bounded differences. In Johannes Siemons (ed.), *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

Renato D. C. Monteiro and Benar F. Svaiter. Iteration-complexity of a Newton proximal extragradient method for monotone variational inequalities and inclusion problems. *SIAM Journal on Optimization*, 22(3):914–935, 2012. doi: 10.1137/11083085X.

Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004. doi: 10.1137/S1052623403425629.

Arkadi S. Nemirovsky and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, 1983. URL https://www2.isye.gatech.edu/~nemirovs/Nemirovskii_Yudin_1983.pdf. Translated from the Russian by E. R. Dawson.

Yurii Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1–2):381–404, 2015. doi: 10.1007/s10107-014-0790-0.

Yurii Nesterov and Boris T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006. doi: 10.1007/s10107-006-0706-8.

Yuyuan Ouyang. An $O(1/T^3)$ algorithm for minimizing convex quadratic functions over the L_1 ball, 2026. arXiv:2609.10314.

Rolf Schneider. *Convex Bodies: The Brunn–Minkowski Theory*, volume 151 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 2nd edition, 2014. doi: 10.1017/CBO9781139003858. URL https://assets.cambridge.org/97811076/01017/frontmatter/9781107601017_frontmatter.pdf.

G. C. Shephard. A uniqueness theorem for the Steiner point of a convex region. *Journal of the London Mathematical Society*, s1-43(1):439–444, 1968. doi: 10.1112/jlms/s1-43.1.439.

Fedor Stonyakin, Alexander Gasnikov, Pavel Dvurechensky, Alexander Titov, and Mohammad Alkousa. Generalized mirror prox algorithm for monotone variational inequalities: Universality and inexact oracle. *Journal of Optimization Theory and Applications*, 194(3):988–1013, 2022. doi: 10.1007/s10957-022-02062-7. URL https://arxiv.org/abs/1806.05140.

Pravin M. Vaidya. A new algorithm for minimizing convex functions over convex sets. *Mathematical Programming*, 73(3):291–341, 1996. doi: 10.1007/BF02592216.

Evgeniya Vorontsova, Alexander Gasnikov, and Eduard Gorbunov. Accelerated directional search with non-Euclidean prox-structure. *Automation and Remote Control*, 80:693–707, 2019. doi: 10.1134/S0005117919040076. URL https://arxiv.org/abs/1710.00162.

Haihan Zhang, Wendao Wu, Chenheng Zhang, Cong Fang, Haoxuan Li, and Zhouchen Lin. Matching higher-order oracle complexity for smooth monotone variational inequalities, 2026a. URL https://arxiv.org/abs/2609.13462. arXiv:2609.13462v1, September 11, 2026.

Xinliang Zhang, Lesi Chen, Linxuan Pan, Chengchang Liu, Junchi Yang, and Jingzhao Zhang. Optimal high-order methods for solving monotone variational inequalities, 2026b. URL https://arxiv.org/abs/2609.23557. arXiv:2609.23557v1, September 20, 2026.

David Martínez-Rubio and Cristóbal Guzmán. Stable movement for nondual Lipschitz convex optimization: Efficiency and nearly optimal oracle rates, 2026. URL https://arxiv.org/abs/2609.20701. arXiv:2609.20701v1, September 17, 2026.

David Martínez-Rubio, Brian Bullins, Cristóbal Guzmán, and Mathieu Molina. The first-order oracle complexity of Lipschitz convex optimization in nondual settings, 2026a. URL https://arxiv.org/abs/2609.20687. arXiv:2609.20687v1, September 17, 2026.

David Martínez-Rubio, Brian Bullins, Cristóbal Guzmán, and Mathieu Molina. Near-optimal acceleration for smooth ℓ_p/ℓ_q nondual convex first-order oracle optimization, 2026b. URL https://arxiv.org/abs/2609.21880. arXiv:2609.21880v1, September 18, 2026.

CONTENTS OF THE APPENDIX

A Notation and conventions **16**

B Steiner points and short paths **17**

C The accelerated level method **20**

C.1 Proof of the guarantee of the level method 20

C.2 Bracket search and the proof of Theorem 1 21

C.3 Computing approximate Steiner points 23

C.4 Inexact oracles and finite precision 24

C.5 Other domains and smoothness metrics 27

D Lower bounds **29**

D.1 Deterministic first-order and nonsmooth bounds 29

D.2 Equal smoothness constants 30

D.3 Higher-order oracles 31

D.4 The simplex and prox steps 33

D.5 Randomized methods 33

E Nonsmooth, Hölder and higher-order objectives **38**

E.1 Nonsmooth objectives: deep cuts and a certificate 38

E.2 Hölder continuous gradients 39

E.3 Oracles of higher order 41

E.4 Gradient oracles: proximal deep cuts 44

E.5 A center-of-gravity method for small dimension 47

F Strong convexity **48**

F.1 Norm compatibility 49

F.2 Restarts 49

F.3 The slack and localization by one gradient 51

F.4 The deterministic complexity for $\Lambda \geq 8$ 53

G Variational inequalities and saddle-point problems **55**

G.1 Lipschitz operators 55

G.2 Operator models with power-law error 57

G.3 A lower bound for monotone operators 59

H Convex quadratics **61**

H.1 Mechanism 61

H.2 The method and its rate 62

H.3 Restarts and the quadratic assumption 63
H.4 Proofs for curvature learning 64
H.5 Lower bounds for quadratics 68

I Experimental details **71**
I.1 Setting and further results 71
I.2 Implementation of the Steiner-point methods 73
I.3 Instances 73
I.4 The complete bracket search 74
I.5 The gradient-only method 74
I.6 Implementation of CLA and the quadratic families 75
I.7 Sensitivity to selector sampling 76
I.8 Controlled study at fixed dimension 77

J Concurrent work **80**

A NOTATION AND CONVENTIONS

Norms and sets. Throughout, $n \geq 1$ is the dimension, $e_1, \dots, e_n$ are the coordinate vectors, $\mathbf{1}$ is the all-ones vector, $\langle x, y \rangle$ is the Euclidean inner product, $1 \leq p \leq 2$, and q is the conjugate exponent, $1/p + 1/q = 1$ ($q = \infty$ for $p = 1$). We write $B_p^n(R) = \{x \in \mathbb{R}^n : \|x\|_p \leq R\}$, $B_p(c, R) = c + B_p^n(R)$, and $\Delta_n(R) = \{x \in \mathbb{R}^n : x \geq 0,\ \sum_i x_i = R\}$ for the simplex of radius R, with $\Delta_n = \Delta_n(1)$. The feasible set K is nonempty, convex and compact. For the Steiner-point methods it is a *known polytope*: a polytope given to the method by linear inequalities and contained in a translate $c + B_1^n(R)$ of an ℓ_1 ball; the main examples are $B_1^n(R)$ and $\Delta_n(R)$. Only the objective is accessed through the oracle. If x has at most d nonzero coordinates, then

$$\|x\|_2 \leq \|x\|_p \leq d^{1/p-1/2}\|x\|_2, \tag{11}$$

and we write $\beta_p = 2/p - 1 \in [0, 1]$, so that $\|x\|_p^2 \leq n^{\beta_p}\|x\|_2^2$ on $\mathbb{R}^n$. The support function of a nonempty compact convex set C is $h_C(z) = \max_{x \in C} \langle z, x \rangle$, and proj_C is the Euclidean projection onto C.

Regularity constants. Unless stated otherwise, L is a Euclidean smoothness constant, $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$. We write L_1 for the Lipschitz constant of ∇f from $(\mathbb{R}^n, \|\cdot\|_1)$ to $(\mathbb{R}^n, \|\cdot\|_\infty)$. These are upper bounds known to the method. For an integer $k \geq 1$ and $0 \leq \nu \leq 1$, $L_{k,\nu}$ is a Hölder constant of the kth derivative in the Euclidean operator norm,

$$\|D^k f(x) - D^k f(y)\| \leq L_{k,\nu}\|x - y\|_2^\nu, \qquad \|T\| = \sup_{\|h_1\|_2, \dots, \|h_k\|_2 \leq 1} |T[h_1, \dots, h_k]| \tag{12}$$

(for $\nu = 0$ the right-hand side is read as $L_{k,0}$ for $x \neq y$). Thus $L_{1,1} = L$. A nonsmooth function is G-Lipschitz if $|f(x) - f(y)| \leq G\|x - y\|_2$. A function is μ_p-strongly convex with respect to $\|\cdot\|_p$ on K if $f(y) \geq f(x) + \langle g, y - x \rangle + \frac{\mu_p}{2}\|y - x\|_p^2$ for all $x, y \in K$ and $g \in \partial f(x)$, and $\kappa_p = L/\mu_p$. For a differentiable convex h, $D_h(u, z) = h(u) - h(z) - \langle \nabla h(z), u - z \rangle$ is its Bregman divergence. We use the integer logarithmic factors

$$\mathsf{b}_N = 1 + \lceil \log_2 N \rceil, \qquad \mathsf{c}_n = 1 + \lceil \log_2(2n) \rceil, \tag{13}$$

which are computed by comparing powers of two and satisfy $\mathsf{b}_N \geq 1 + \log N$ and $\mathsf{c}_n \geq 1 + \log(2n)$. We also write $t_+ = \max\{0, t\}$, $\log_+ t = \max\{0, \log t\}$ and $\lceil t \rceil_+ = \max\{0, \lceil t \rceil\}$.

Oracles. At a query point x, the *first-order oracle* (also called the joint oracle) returns $(f(x), \nabla f(x))$, the *gradient oracle* returns $\nabla f(x)$ only, and the *oracle of order* k returns $f(x)$ and $D^j f(x)$ for $j \leq k$. A *subgradient oracle* returns one element of $\partial f(x)$, with or without $f(x)$ (each result says which), and the *operator oracle* of a variational inequality returns $F(x)$. Queries are *global* (anywhere in $\mathbb{R}^n$) unless they are called *feasible* (in K). All upper bounds use feasible queries, except the finite-difference products for quadratics in Section H, and all lower bounds allow global queries.

What is counted. A query is one call of the oracle at one point, and every query is counted, including the queries of trials that a level method rejects, the second query of each trial of SP-Taylor and SP-VI, finite-difference probes, and the queries that compute an initial bracket; for SP-Prox every gradient of the inner loops is counted. Computations with known functions and sets are not counted. These include the minimization of an explicitly given model over K, Euclidean projections, linear programs over known polytopes, centroids and approximate Steiner points of known polytopes, matrix computations with stored vectors, and comparisons of real numbers. This is the information-based complexity model of Nemirovsky & Yudin (1983). It does not measure arithmetic cost; Section C states which internal operations are finite or polynomial.

Minimax error. For a class $\mathcal{F}$ the minimax error of deterministic methods is $\mathcal{E}_N^{\mathrm{det}}(\mathcal{F}; K) = \inf_{\mathcal{A}} \sup_{f \in \mathcal{F}} [f(\widehat{x}_{\mathcal{A},f}) - \min_K f]$, where $\mathcal{A}$ ranges over deterministic methods with at most N queries and $\widehat{x}_{\mathcal{A},f} \in K$ is the output. In $\mathcal{E}_N^{\mathrm{ran}}(\mathcal{F}; K)$ the methods may be randomized and the error is replaced by its expectation over the internal randomness of the method; our randomized lower bounds exhibit, for every method, a fixed function on which the expected error is large. We write $f^\star = \min_K f$.

Dimension and mismatch. The deterministic lower bounds of Propositions D.1 and D.4 hold for $n \geq N+1$ (Corollary D.2: $n \geq N+2$; quadratics: $n \geq 3N+2$; Proposition G.7: $n \geq 4N+6$), and the randomized ones for $n \geq 32N+1$ (smooth and higher order) or $n \geq 2^{p+2}N$ (nonsmooth, $8N$ on the ℓ_1 ball); the upper bounds hold in every dimension. A convex quadratic on $\mathbb{R}^n$ is identified by $n+1$ gradient queries (Section H), so lower bounds for quadratics require $N < n$. For every function the smallest valid L_1 is at most the smallest valid L, which is at most n times the former; both extremes occur, for $f = \frac{1}{2}\|x\|_2^2$ and $f = \frac{1}{2}(\mathbf{1}^\top x)^2$. When $L \asymp nL_1$, the entropic rate $L_1 R^2 \log n / N^2$ is better than $LR^2 \log(2N) \log(2n)/N^3$ in the regime $N \leq n$; the new rate matters when L and L_1 are comparable, and by Corollary D.2 the lower bound persists when the two smallest constants coincide.

Table 2: Main symbols.

Symbol	Meaning
n, N	dimension; number of queries, or horizon of one call of a level method
$B_p^n(R)$, $B_p(c,R)$, $\Delta_n(R)$	ℓ_p ball of radius R centered at 0 or c; simplex of radius R
p, q, β_p	$1 \leq p \leq 2$, $1/p + 1/q = 1$, $\beta_p = 2/p - 1$
K; C, C_t	feasible set; localization polytopes
h_C, $\mathrm{st}(C)$, $\omega(C)$	support function, Steiner point, Gaussian width (Section B)
L, L_1	Euclidean and $\ell_1 \to \ell_\infty$ gradient Lipschitz constants
$L_{k,\nu}$, G	Hölder constant of $D^k f$; Lipschitz constant of f
s	$k + \nu$, the order of Hölder smoothness
μ_p, κ_p, Λ	strong convexity modulus in $\|\cdot\|_p$; L/μ_p; slack $\kappa_1 - n$ (Section F)
$\bar{\kappa}$	$L_{k,\nu} R^{s-2}/\mu_1$, condition number of order s (Section F.2)
b_N, c_n, P_N	$1 + \lceil \log_2 N \rceil$, $1 + \lceil \log_2(2n) \rceil$, path budget $4R\sqrt{\mathsf{b}_N \mathsf{c}_n}$
η_N, δ_N	center accuracy $P_N/(4N)$ and acceptance threshold $4P_N/N$ of SP-Level
ℓ; $\underline{f}$, $\overline{f}$, w	level; ends and width of the bracket for $f^\star$
φ_t, φ	prefix maximum and final hard function of the resisting oracle (Section 4)
A, a, B; x, y, z	weights ($B = A + a$) and sequences of a level method
F, T_z, r, M	operator; operator model at z, its error exponent and constant (Section G)
$\psi_{\tau,p}$, Ψ_I	power–Huber function and regularizer with promoted set I (Section H)

The algorithms are SP-Level (Algorithm 1), SP-Accel (Algorithm 2), SP-Cut (Algorithm 3), SP-Taylor (Algorithm 4), SP-Prox (Algorithm 5), CG-Cut (Algorithm 6), SP-Restart (restarts, Section F.2), SP-VI (Algorithm 7) and CLA (Algorithm 8).

B Steiner points and short paths

This section proves the geometric facts used by all Steiner-point methods: the basic properties of the Steiner point, the path bound of Lemma 3.1 in a general form, and its tightness.

Steiner point and Gaussian width. Let $Z \sim \mathcal{N}(0, I_n)$. For a nonempty compact convex $C \subseteq \mathbb{R}^n$ the Steiner point is

$$\mathrm{st}(C) = \mathbb{E}\big[Z\, h_C(Z)\big] = \mathbb{E}\big[\nabla h_C(Z)\big] \in C, \tag{14}$$

where the equality of the two expectations is Gaussian integration by parts and $\nabla h_C(Z)$ is the almost surely unique maximizer of $\langle Z, x\rangle$ over $x \in C$, so that (14) agrees with (1) (Lemma B.1). Since $\mathbb{E}\|Z\|_2^2 = n$ and h_C is positively homogeneous, $\mathbb{E}[Z h_C(Z)] = n\,\mathbb{E}[\theta\, h_C(\theta)]$ for θ uniform on the unit sphere, so (14) is the classical Steiner point (Schneider, 2014; Shephard, 1968); Bubeck et al. (2020) use this spherical form for nested convex body chasing. We measure the size of a set by its root-mean-square Gaussian width

$$\omega(C) = \big(\mathbb{E}\, h_C(Z)^2\big)^{1/2}. \tag{15}$$

It depends on the position of C; the bounds below use translates $K - c$. For instance, $h_{B_1^n(R)}(Z) = R\|Z\|_\infty$ and $h_{B_2^n(R)}(Z) = R\|Z\|_2$, so

$$\omega(B_1^n(R)) \leq R\sqrt{2(1 + \log(2n))}, \qquad \omega(B_2^n(R)) = R\sqrt{n}, \tag{16}$$

by Lemma B.2 below and $\mathbb{E}\|Z\|_2^2 = n$.

Lemma B.1 (Basic properties)**.** *Let $C \subseteq \mathbb{R}^n$ be nonempty, compact and convex. Then h_C is convex and Lipschitz continuous with constant $\max_{x\in C}\|x\|_2$, differentiable almost everywhere with $\nabla h_C(z) \in C$ at every point of differentiability, and*

$$\mathrm{st}(C) = \mathbb{E}[Zh_C(Z)] = \mathbb{E}[\nabla h_C(Z)] \in C.$$

If C is symmetric about c, then $\mathrm{st}(C) = c$; in particular $\mathrm{st}(B_1^n(R)) = 0$.

Proof. Let $\rho = \max_{x\in C}\|x\|_2$. As a maximum of linear functions, h_C is convex, and $|h_C(z) - h_C(z')| \le \max_{x\in C} |\langle z - z', x\rangle| \le \rho\|z - z'\|_2$. By Rademacher's theorem h_C is differentiable almost everywhere. Let z be a point of differentiability and let $x \in C$ be any maximizer of $\langle z, \cdot\rangle$ over C. For every v and every real t, $h_C(z + tv) \ge \langle z + tv, x\rangle = h_C(z) + t\langle v, x\rangle$; dividing by $t > 0$ and by $t < 0$ and letting $t \to 0$ gives $\langle \nabla h_C(z), v\rangle = \langle v, x\rangle$ for all v, so $\nabla h_C(z) = x \in C$ (in particular the maximizer is unique at such z).

Integration by parts. Fix i and all coordinates of z except $z_i = u$, and put $\varphi(u) = h_C(z)$. The function φ is ρ-Lipschitz, hence absolutely continuous on bounded intervals with $|\varphi'| \le \rho$ almost everywhere, and $|\varphi(u)| \le \rho\|z\|_2$. For $a > 0$,

$$\int_{-a}^{a} u\varphi(u)e^{-u^2/2}\,du = \left[-\varphi(u)e^{-u^2/2}\right]_{-a}^{a} + \int_{-a}^{a} \varphi'(u)e^{-u^2/2}\,du.$$

The boundary term tends to 0 as $a \to \infty$ because φ grows at most linearly, and both integrands are integrable on $\mathbb{R}$, so $\int u\varphi(u)e^{-u^2/2}du = \int \varphi'(u)e^{-u^2/2}du$. Since $\|zh_C(z)\|_2 \le \rho\|z\|_2^2$ is Gaussian-integrable, Fubini's theorem gives $\mathbb{E}[Z_i h_C(Z)] = \mathbb{E}[\partial_i h_C(Z)]$, where $\partial_i h_C$ is the partial derivative, which exists almost everywhere and coincides with the ith component of ∇h_C at the points of differentiability. Hence $\mathbb{E}[Zh_C(Z)] = \mathbb{E}[\nabla h_C(Z)]$.

Membership. The random vector $V = \nabla h_C(Z)$ is bounded and takes values in the closed convex set C almost surely. For every closed half-space $\{x : \langle a, x\rangle \le b\} \supseteq C$ we have $\langle a, \mathbb{E}V\rangle = \mathbb{E}\langle a, V\rangle \le b$, and C is the intersection of the closed half-spaces containing it, so $\mathbb{E}V \in C$.

Symmetry. Translation gives $h_C(z) = \langle c, z\rangle + h_{C-c}(z)$ and $\mathbb{E}[Z\langle c, Z\rangle] = c$, so $\mathrm{st}(C) = c + \mathrm{st}(C - c)$. If $C - c = -(C - c)$, then $h_{C-c}(-z) = h_{C-c}(z)$, and since $-Z$ has the law of Z, $\mathrm{st}(C - c) = \mathbb{E}[Zh_{C-c}(Z)] = \mathbb{E}[(-Z)h_{C-c}(-Z)] = -\mathrm{st}(C - c)$, so $\mathrm{st}(C - c) = 0$. □

Lemma B.2 (Maxima of Gaussian variables)**.** *Let $X_1, \dots, X_T$ be centered Gaussian variables, not necessarily independent, with variances at most one, and let $A = \max\{0, X_1, \dots, X_T\}$. Then $\mathbb{E}A^2 \le 2(1 + \log T)$. In particular $\mathbb{E}\|Z\|_\infty^2 \le 2(1 + \log(2n))$.*

Proof. Each X_t satisfies $\mathbb{P}(X_t > u) \le e^{-u^2/2}$ for $u \ge 0$, so the union bound gives $\mathbb{P}(A > u) \le \min\{1, Te^{-u^2/2}\}$. With $u_0 = \sqrt{2\log T}$,

$$\mathbb{E}A^2 = \int_0^\infty 2u\,\mathbb{P}(A > u)\,du \le u_0^2 + \int_{u_0}^\infty 2uTe^{-u^2/2}\,du = 2\log T + 2.$$

For the second claim apply the first to the $2n$ variables $\pm Z_i$, whose maximum is $\|Z\|_\infty$. □

The path bound holds for nested subsets of any compact convex set; the domain enters only through the width of a centered copy.

Lemma B.3 (Short Steiner paths, general form)**.** *Let $T \ge 1$, let $K \subseteq \mathbb{R}^n$ be nonempty, compact and convex, and let $C_0 \supseteq C_1 \supseteq \cdots \supseteq C_T$ be nonempty compact convex subsets of K. Then, for every $c \in \mathbb{R}^n$,*

$$\sum_{t=1}^{T} \|\mathrm{st}(C_t) - \mathrm{st}(C_{t-1})\|_2 \;\le\; 2\sqrt{2(1 + \log T)}\,\omega(K - c). \tag{17}$$

Proof. Let v_t be the unit vector along $\mathrm{st}(C_{t-1}) - \mathrm{st}(C_t)$, or $v_t = 0$ if the two points coincide, and put $A(Z) = \max\{0, \langle v_1, Z\rangle, \dots, \langle v_T, Z\rangle\}$.

Step 1: telescoping. By Lemma B.1 and linearity of expectation,

$$\sum_{t=1}^{T}\|\mathrm{st}(C_{t-1})-\mathrm{st}(C_t)\|_2=\sum_{t=1}^{T}\langle v_t,\mathrm{st}(C_{t-1})-\mathrm{st}(C_t)\rangle=\mathbb{E}\sum_{t=1}^{T}\langle v_t,Z\rangle\big(h_{C_{t-1}}(Z)-h_{C_t}(Z)\big).$$

The sets are nested, so every difference $h_{C_{t-1}}(Z)-h_{C_t}(Z)$ is nonnegative, and $\langle v_t,Z\rangle\le A(Z)$. Hence

$$\sum_{t=1}^{T}\|\mathrm{st}(C_{t-1})-\mathrm{st}(C_t)\|_2\le\mathbb{E}\Big[A(Z)\sum_{t=1}^{T}\big(h_{C_{t-1}}(Z)-h_{C_t}(Z)\big)\Big]=\mathbb{E}\big[A(Z)\big(h_{C_0}(Z)-h_{C_T}(Z)\big)\big]. \tag{18}$$

Step 2: the range of the support functions. Fix $y\in C_T$. Since $C_0\subseteq K$, we have $h_{C_0}(Z)\le\langle c,Z\rangle+h_{K-c}(Z)$ and $h_{C_T}(Z)\ge\langle y,Z\rangle\ge\langle c,Z\rangle-h_{K-c}(-Z)$. Therefore

$$0\le h_{C_0}(Z)-h_{C_T}(Z)\le h_{K-c}(Z)+h_{K-c}(-Z), \tag{19}$$

and since $-Z$ has the law of Z, the second moment of the right-hand side is at most $4\omega(K-c)^2$.

Step 3: Cauchy–Schwarz. By (18), (19) and the Cauchy–Schwarz inequality,

$$\sum_{t=1}^{T}\|\mathrm{st}(C_{t-1})-\mathrm{st}(C_t)\|_2\le\big(\mathbb{E}A(Z)^2\big)^{1/2}\cdot 2\omega(K-c).$$

Each $\langle v_t,Z\rangle$ is Gaussian with variance at most one, so $\mathbb{E}A(Z)^2\le 2(1+\log T)$ by Lemma B.2. □

Proof of Lemma 3.1. Apply Lemma B.3 with $K=c+B_1^n(R)$. Then $K-c=B_1^n(R)$, and $\omega(B_1^n(R))\le R\sqrt{2(1+\log(2n))}$ by (16). □

The Gaussian vector appears only in the proof: (2) is a deterministic bound for every fixed chain, and it requires neither full-dimensional sets nor central cuts. The lemma is a variant of the mean-width bound of Bubeck et al. (2020, Lemma 3.2 and Theorem 2.1); the Gaussian form gives explicit constants and a bound that is sharper for chains shorter than n. For the Euclidean ball, (17) gives $2\sqrt{2(1+\log T)}\,R\sqrt{n}$. The path budget $P_N=4R\sqrt{\mathsf{b}_N\mathsf{c}_n}$ of (5) dominates the right-hand side of (2) for every $T\le N$, because $\mathsf{b}_N\ge 1+\log N$ and $\mathsf{c}_n\ge 1+\log(2n)$. Its two logarithmic factors cannot be removed from a bound that depends on the domain only through its width, even for the simplex.

Proposition B.4 (Tightness of the path bound)**.** *Let $n\ge 2$, $R>0$ and $C_j=\mathrm{conv}\{Re_1,\dots,Re_j\}$, $1\le j\le n$. Then $C_n\supseteq C_{n-1}\supseteq\cdots\supseteq C_1$ are nonempty compact convex subsets of $\Delta_n(R)\subseteq B_1^n(R)$, $\mathrm{st}(C_j)=\frac{R}{j}\sum_{i\le j}e_i$, and*

$$R(H_n-1)\le\sum_{j=2}^{n}\|\mathrm{st}(C_j)-\mathrm{st}(C_{j-1})\|_2\le RH_{n-1},\qquad H_n=\sum_{j=1}^{n}\frac{1}{j},$$

while $\omega(\Delta_n(R))\le R\sqrt{2(1+\log(2n))}$. Moreover, for every rule that selects a point of the current set (deterministic or randomized), there is a nested chain of n faces of $\Delta_n(R)$, revealed one at a time, along which the selected points move by at least $R(H_n-1)$ in expectation.

Proof. For $Z\sim\mathcal{N}(0,I_n)$ the maximizer of $\langle Z,x\rangle$ over C_j is Re_i with i the index of the largest of $Z_1,\dots,Z_j$, almost surely unique, and by symmetry each index has probability $1/j$; Lemma B.1 gives the formula for $\mathrm{st}(C_j)$. Then $\|\mathrm{st}(C_j)-\mathrm{st}(C_{j-1})\|_2^2=R^2/j^2+(j-1)R^2(\frac{1}{j-1}-\frac{1}{j})^2=R^2/(j(j-1))$, and $1/j\le 1/\sqrt{j(j-1)}\le 1/(j-1)$ gives the two bounds. The width bound follows from $h_{\Delta_n(R)}(Z)=R\max_i Z_i$, $(\max_i Z_i)^2\le\|Z\|_\infty^2$ and $\mathbb{E}\|Z\|_\infty^2\le 2(1+\log(2n))$ (Lemma B.2). For the last claim, let the current set be the face $C_S=\mathrm{conv}\{Re_i:i\in S\}$, $|S|=j\ge 2$, and let x be the selected point, so $x_i\ge 0$ and $\sum_{i\in S}x_i=R$. Remove an index $i\in S$ chosen uniformly at random (independently of the selector); the next selected point y lies in $C_{S\setminus\{i\}}$, so $y_i=0$ and $\mathbb{E}[\|x-y\|_2\mid x]\ge\frac{1}{j}\sum_{i\in S}x_i=R/j$. Summing over $j=n,\dots,2$ gives $R(H_n-1)$; averaging over the random order of removal shows that some fixed order gives at least this expected movement (for a deterministic rule the choice $i\in\arg\max_{i\in S}x_i$ gives the bound directly). □

Thus at $T = n - 1$ both factors of $4R\sqrt{(1+\log T)(1+\log(2n))}$ are of the right order for the simplex, and no online rule does better on such chains. Whether these logarithms are necessary in the oracle complexity of smooth minimization is open.

C THE ACCELERATED LEVEL METHOD

This section proves Proposition 3.2, analyzes the bracket search and proves Theorem 1, and then turns to the computation of Steiner points: an exact finite procedure, a randomized estimator with polynomially many linear programs, and a version for inexact oracles in finite precision. The last subsection extends the method to other domains and smoothness metrics. Throughout, $K \subseteq c + B_1^n(R)$ is a known polytope, and P_N, b_N, c_n are as in (5).

C.1 PROOF OF THE GUARANTEE OF THE LEVEL METHOD

We first isolate the one-step inequality, because Sections E.2 and E.3 reuse it with other remainders.

Lemma C.1 (One accepted step)**.** *Let f be convex with L-Lipschitz Euclidean gradient on K, let $A \geq 0$, $a > 0$, $B = A + a$, let $y, z, z^+ \in K$, and set $x = (Ay + az)/B$ and $y^+ = (Ay + az^+)/B$. If $f(x) + \langle \nabla f(x), z^+ - x\rangle \leq \ell$, then (7) holds.*

Proof. Since $y^+ - x = (a/B)(z^+ - z)$, the descent inequality between x and y^+ gives $Bf(y^+) \leq Bf(x) + a\,\langle g, z^+ - z\rangle + \frac{La^2}{2B}\|z^+ - z\|_2^2$ with $g = \nabla f(x)$. By $Bx = Ay + az$ its first two terms equal $A[f(x) + \langle g, y - x\rangle] + a[f(x) + \langle g, z^+ - x\rangle] \leq Af(y) + a\ell$ (convexity at x and the level cut), and subtracting $B\ell = A\ell + a\ell$ gives (7), also when $A = 0$. □

The proof uses only the descent inequality between x and y^+, convexity at x and the level cut at z^+. Replacing $\frac{L}{2}\|y^+ - x\|_2^2$ by any remainder $\phi(\|y^+ - x\|_2)$ therefore replaces $\frac{La^2}{2B}\|z^+ - z\|_2^2$ by $B\,\phi(\frac{a}{B}\|z^+ - z\|_2)$.

Proposition 3.2 (Level method)**.** *Let K be a nonempty polytope given by linear inequalities with $K \subseteq c + B_1^n(R)$, and let f be convex with L-Lipschitz Euclidean gradient on K. For every level $\ell \in \mathbb{R}$ and horizon $N \geq 1$, Algorithm 1 makes at most N first-order queries, all in K, and either certifies that $\min_K f > \ell$ or returns $y_N \in K$ with*

$$f(y_N) \;\leq\; \ell + \frac{48LP_N^2}{N^3} \;=\; \ell + \frac{768\,LR^2\,\mathsf{b}_N\mathsf{c}_n}{N^3}. \tag{6}$$

Proof. All iterates are convex combinations of points of K, so every query is feasible. If $\min_K f \leq \ell$, a minimizer u_* satisfies $f(x_t) + \langle g_t, u_* - x_t\rangle \leq f(u_*) \leq \ell$ for every t, hence $u_* \in C_t$ and no localization set is empty; an empty C_t therefore certifies $\min_K f > \ell$.

Assume $C_1, \dots, C_N$ are nonempty. The selected centers satisfy $\|z_t - \mathrm{st}(C_t)\|_2 \leq \eta_N$ and the sets C_t are nested in K, so Lemma 3.1 and the triangle inequality give (8), with $2N\eta_N = \frac{1}{2}P_N$. Every rejected trial has $d_t > \delta_N = 4P_N/N$, so their number r obeys $4rP_N/N < \frac{3}{2}P_N$, that is $r < 3N/8$, and $k = N - r > 5N/8 > N/2$ trials are accepted. On an accepted trial the membership $z_t \in C_t$ is the hypothesis of Lemma C.1, and the choice $La_t^2 = B_t$ makes $La_t^2/(2B_t) = \frac{1}{2}$, so $A_t[f(y_t) - \ell] \leq A_{t-1}[f(y_{t-1}) - \ell] + \frac{1}{2}d_t^2$; on a rejected trial $A_t = A_{t-1}$ and $y_t = y_{t-1}$. Telescoping from $A_0 = 0$, bounding $d_t \leq \delta_N$ on the accepted trials and using (8) gives

$$A_N[f(y_N) - \ell] \;\leq\; \frac{1}{2}\sum_{t\ \text{accepted}} d_t^2 \;\leq\; \frac{\delta_N}{2}\sum_{t=1}^{N} d_t \;\leq\; \frac{3P_N^2}{N}. \tag{20}$$

Number the accepted updates $j = 1, \dots, k$. The recursion $La_j^2 = A_j$ with $A_j = A_{j-1} + a_j$ means $a_j = (1 + \sqrt{1 + 4LA_{j-1}})/(2L)$, so $\sqrt{LA_j} = \frac{1}{2}(1 + \sqrt{1 + 4LA_{j-1}}) \geq \sqrt{LA_{j-1}} + \frac{1}{2}$ and $A_N = A_k \geq k^2/(4L) \geq N^2/(16L)$. Dividing (20) by A_N proves (6). □

C.2 Bracket search and the proof of Theorem 1

Algorithm 2 needs no knowledge of $f^\star = \min_K f$. One query at $x_0 \in K$ gives the bracket $\underline{f} \le f^\star \le f(v) \le \overline{f}$ with

$$v \in \arg\min_{u\in K} \langle g_0, u\rangle, \qquad \underline{f} = f(x_0) + \langle g_0, v - x_0\rangle, \qquad \overline{f} = \underline{f} + \tfrac{L}{2}\|v - x_0\|_2^2, \tag{21}$$

where $g_0 = \nabla f(x_0)$; the first inequality is the tangent lower bound, and the last one follows from smoothness at x_0. Since $\|v - x_0\|_2 \le \|v - x_0\|_1 \le 2R$, the width is at most $2LR^2$ for a general $K \subseteq c + B_1^n(R)$ and at most $LR^2/2$ for $K = B_1^n(R)$ and $x_0 = 0$. No query at v is needed. Algorithm 2 bisects the bracket, calling Algorithm 1 at the midpoint with the smallest power-of-two horizon whose guaranteed error (6) is at most a quarter of the width:

$$N(w) = \min\Big\{N = 2^j : \frac{48LP_N^2}{N^3} \le \frac{w}{4}\Big\} = \min\Big\{N = 2^j : N^3 \ge \frac{3072LR^2}{w}\,\mathsf{b}_N\mathsf{c}_n\Big\}. \tag{22}$$

A negative certificate raises the lower end to ℓ (width $w/2$); a positive answer stores the output and lowers the upper end to $\ell + w/4$ (width $3w/4$).

Algorithm 2 SP-Accel: bracket search for the level

Input: known polytope $K \subseteq c + B_1^n(R)$; first-order oracle of an L-smooth convex f; accuracy $\varepsilon > 0$

query $(f(x_0), g_0)$ at a point $x_0 \in K$ ($x_0 = 0$ if $K = B_1^n(R)$); compute $v, \underline{f}, \overline{f}$ by (21)
while $w := \overline{f} - \underline{f} > \varepsilon$ **do**
  $\ell \leftarrow (\underline{f} + \overline{f})/2$; $N \leftarrow N(w)$ by (22)
  run Algorithm 1 with (ℓ, N), restarting from $C_0 = K$, $A_0 = 0$
  if it certifies $\min_K f > \ell$ **then** $\underline{f} \leftarrow \ell$
  else $v \leftarrow y_N$; $\overline{f} \leftarrow \ell + w/4$
return v ▷ $f(v) - \min_K f \le \overline{f} - \underline{f} \le \varepsilon$

Proposition C.2 (Bracket search)**.** *Let f be convex with L-Lipschitz Euclidean gradient on $K = B_1^n(R)$.*

(a) For every $0 < \varepsilon < LR^2/2$, Algorithm 2 returns $v \in K$ with $f(v) - \min_K f \le \varepsilon$ after at most

$$1 + 24\,N(\varepsilon) \;=\; O\Big(1 + \Big[\frac{LR^2}{\varepsilon}\,\log(2n)\,\log\Big(2 + \frac{LR^2}{\varepsilon}\log(2n)\Big)\Big]^{1/3}\Big) \tag{23}$$

feasible first-order queries; explicitly, $N(\varepsilon) < [32\Theta(3 + \log_2\Theta)]^{1/3}$ with $\Theta = 3072LR^2\mathsf{c}_n/\varepsilon$.

(b) For every budget $T \ge 25$ there is a choice of ε (explicit in the proof) for which the method makes at most T queries and returns v with

$$f(v) - \min_K f \;\le\; \min\Big\{\frac{LR^2}{2},\; \frac{3072\cdot 48^3\,LR^2\,\mathsf{b}_T\mathsf{c}_n}{(T-1)^3}\Big\} = O\Big(\frac{LR^2\log(2T)\log(2n)}{T^3}\Big). \tag{24}$$

For $T < 25$ the point v of (21) has error at most $LR^2/2$.

The same statements hold on every known polytope $K \subseteq c + B_1^n(R)$, in particular on $\Delta_n(R)$, with $LR^2/2$ replaced by $2LR^2$, an upper bound on the initial bracket width of (21); the horizon rule (22) and the count $1 + 24N(\varepsilon)$ are unchanged.

Proof. We first treat $K = B_1^n(R)$, $x_0 = 0$. Write w for the current width $\overline{f} - \underline{f}$ and $N(w)$ for the horizon (22), which exists because for $N = 2^j$ one has $\mathsf{b}_N = 1 + j$ and $2^{3j}/(1 + j) \to \infty$.

Invariants. By (21), $\underline{f} \le \min_K f \le f(v) \le \overline{f}$ holds initially with $w = \frac{L}{2}\|v\|_2^2 \le LR^2/2$. If $w \le \varepsilon$ the method returns v after one query. Suppose the invariant holds at the start of a phase with width $w > \varepsilon$ and level $\ell = (\underline{f} + \overline{f})/2$. If the level method returns the certificate, $\min_K f > \ell$, so $\underline{f} \leftarrow \ell$ keeps the invariant and the new width is $w/2$. Otherwise Proposition 3.2 and the choice of $N(w)$

give $f(y_N) \le \ell + 48LP_N^2/N^3 \le \ell + w/4$, so $v \leftarrow y_N$, $\overline{f} \leftarrow \ell + w/4$ keeps the invariant and the new width is $3w/4$. The loop terminates because each phase multiplies the width by at most $3/4$, and at termination $f(v) - \min_K f \le \overline{f} - \underline{f} \le \varepsilon$. Every query is feasible by Proposition 3.2.

Phase count. Let $w_0 > w_1 > \cdots > w_{m-1} > \varepsilon$ be the widths at the beginnings of the m phases and $N_j = N(w_j)$. Each phase multiplies the width by at most $3/4$, so $w_j \ge (4/3)^{m-1-j} w_{m-1}$, and $N(\cdot)$ is nonincreasing, so $N_j \le N_{m-1} \le N(\varepsilon)$. Every N_j exceeds 1: the condition in (22) at $N = 1$ reads $1 \ge 3072LR^2\mathsf{c}_n/w_j$, which fails because $\mathsf{c}_n \ge 2$ and $w_j \le LR^2/2$. Hence $N_j/2$ is a power of two that violates (22), and using $\mathsf{b}_{N_j/2} \le \mathsf{b}_{N_j} \le \mathsf{b}_{N_{m-1}}$,

$$\frac{N_j^3}{8} < \frac{3072LR^2}{w_j}\mathsf{b}_{N_{m-1}}\mathsf{c}_n, \qquad N_{m-1}^3 \ge \frac{3072LR^2}{w_{m-1}}\mathsf{b}_{N_{m-1}}\mathsf{c}_n.$$

Dividing, $N_j^3 < 8(w_{m-1}/w_j)N_{m-1}^3 \le 8(3/4)^{m-1-j}N_{m-1}^3$, i.e. $N_j < 2(3/4)^{(m-1-j)/3}N_{m-1}$. Summing the geometric series,

$$\sum_{j=0}^{m-1} N_j < \frac{2}{1-(3/4)^{1/3}} N_{m-1} < 24\, N(\varepsilon),$$

because $(11/12)^3 = 1331/1728 > 3/4$ gives $(3/4)^{1/3} < 11/12$. With the initial query this proves the count $1 + 24N(\varepsilon)$ in (a).

Explicit horizon. Let $\Theta = 3072LR^2\mathsf{c}_n/\varepsilon > 1$, $J = \lceil \log_2 \Theta \rceil$ and $q = \lceil \log_2(2^{J+1}(J+2))/3 \rceil$. Then $q \le (J + 1 + \log_2(J+2))/3 + 1$ and $\log_2(J+2) \le J+1$ give $q + 1 \le 2(J+2)$, hence $(2^q)^3 \ge 2^{J+1}(J+2) \ge \Theta(q+1)$, which is (22) for $N = 2^q$ and $w = \varepsilon$ (note $\mathsf{b}_{2^q} = q+1$). Thus $N(\varepsilon) \le 2^q$ and, since $2^{3q} < 8 \cdot 2^{J+1}(J+2)$ and $2^J < 2\Theta$,

$$N(\varepsilon)^3 < 32\Theta(J+2) \le 32\Theta(3 + \log_2 \Theta),$$

which gives the explicit bound and the order in (23).

Fixed horizon. Let $T \ge 25$, let Q be the largest power of two with $24Q \le T - 1$, and $\varepsilon_Q = 3072LR^2\mathsf{b}_Q\mathsf{c}_n/Q^3$. If $\varepsilon_Q \ge LR^2/2$, return the point v of (21), whose error is at most $LR^2/2$ after one query. Otherwise run the method with accuracy ε_Q: Q satisfies (22) for $w = \varepsilon_Q$ with equality, so $N(\varepsilon_Q) \le Q$, the count is at most $1 + 24Q \le T$, and the error is at most ε_Q. By maximality of Q, $48Q > T - 1$, and $\mathsf{b}_Q \le \mathsf{b}_T$, so $\varepsilon_Q \le 3072 \cdot 48^3 LR^2\mathsf{b}_T\mathsf{c}_n/(T-1)^3$, which is (24).

General polytopes. For $K \subseteq c + B_1^n(R)$ the level method is unchanged: Lemma 3.1 applies to chains in $c + B_1^n(R)$, and z_0 is an η_N-approximate Steiner point of K computed as in Section C.3. The bracket (21) has width $\frac{L}{2}\|v - x_0\|_2^2 \le \frac{L}{2}(2R)^2 = 2LR^2$, the condition $N_j > 1$ still holds because $w_j \le 2LR^2 < 6144LR^2$, and the rest of the argument is the same with $LR^2/2$ replaced by $2LR^2$. □

Proof of Theorem 1. The lower bound for deterministic methods is Proposition D.1(b) with $p = 1$ and $\nu = 1$, and the lower bound for randomized methods is Proposition D.7. For the upper bound, run Algorithm 2 with the deterministic selector of Proposition C.3 and let $T = N$ in Proposition C.2(b) if $N \ge 25$: the method is deterministic, makes at most N feasible queries, and its error is at most

$$\frac{3072 \cdot 48^3 LR^2\mathsf{b}_N\mathsf{c}_n}{(N-1)^3} \le \frac{3072 \cdot 48^3 \cdot 8 \cdot 4}{\log^2 2} \cdot \frac{LR^2 \log(2N)\log(2n)}{N^3} \le \frac{2^{35} LR^2 \log(2N)\log(2n)}{N^3},$$

since $\mathsf{b}_N \le 2\log(2N)/\log 2$, $\mathsf{c}_n \le 2\log(2n)/\log 2$ and $N - 1 \ge N/2$. For $N \le 24$ the point v of (21) has error at most $LR^2/2$ after one query, which is at most $2^{35}LR^2\log(2N)\log(2n)/N^3$ for every $n \ge 1$, because $\log(2N)\log(2n) \ge \log^2 2$ and $24^3/(2\log^2 2) < 14387 < 2^{35}$. Randomized methods include deterministic ones, so $\mathcal{E}_N^{\mathrm{ran}} \le \mathcal{E}_N^{\mathrm{det}}$. □

The level cuts use function values; Section E.4 shows that the exponent is unchanged when only ∇f is returned, with two additional logarithmic factors.

C.3 Computing approximate Steiner points

Every localization set is a known polytope $C = K \cap \{\text{stored half-spaces}\}$, and the methods need, for a prescribed $\eta > 0$, a point $x(C, \eta) \in C$ with $\|x(C, \eta) - \mathrm{st}(C)\|_2 \le \eta$. We give a deterministic and a randomized implementation.

Proposition C.3 (Deterministic finite selector)**.** *For a known nonempty polytope $C \subseteq c + B_1^n(R)$ and $\eta > 0$, a point $x(C, \eta)$ can be computed by finitely many linear programs over C, one Euclidean projection onto C, and rational arithmetic with certified approximations of e^{-r} and $\sqrt{2\pi}$. The Gaussian integral (14) is truncated to a cube $[-a, a]^n$ and approximated by a midpoint rule whose nodes require h_C at rational points, and the result is projected onto C. The number of nodes is finite but grows exponentially in n.*

Proof. Since $C \subseteq c + B_1^n(R)$, by translation we may assume $c = 0$, so $\|x\|_2 \le \|x\|_1 \le R$ on C and $|h_C(z)| \le R\|z\|_\infty$. The polytope C is given by finitely many known linear inequalities (the 2^n facets $\langle \sigma, x\rangle \le R$, $\sigma \in \{-1, 1\}^n$, of the ball, or the facets of K, together with the stored cuts). Enumerating the nondegenerate systems of n active constraints, solving the linear systems and checking the remaining constraints yields all vertices of C, decides whether C is empty, and evaluates h_C at any rational point; lower-dimensional C are included, since the constraints defining the affine hull are among those enumerated. The Euclidean projection $\mathrm{proj}_C(w)$ is computed by projecting w onto the affine hull of every face (enumerating linearly independent subsets of active constraints, the empty set included), keeping the feasible candidates and choosing the closest one. This is correct: if $\mathrm{proj}_C(w)$ lies in the relative interior of a face C', then $w - \mathrm{proj}_C(w)$ lies in the normal cone of C at $\mathrm{proj}_C(w)$, which is spanned by the active normals, so $\mathrm{proj}_C(w)$ is the projection of w onto the affine hull of C', a candidate in the list.

Truncation. Let $B = \|Z\|_\infty$. For $a \ge 1$, the tail bound $\mathbb{P}(B > u) \le 2ne^{-u^2/2}$ gives $\mathbb{E}[B^2 \mathbf{1}_{\{B>a\}}] = a^2\mathbb{P}(B > a) + \int_a^\infty 2u\mathbb{P}(B > u)du \le 2n(a^2 + 2)e^{-a^2/2}$, and since $\|zh_C(z)\|_2 \le R\|z\|_2\|z\|_\infty \le R\sqrt{n}\,\|z\|_\infty^2$,

$$\left\|\mathbb{E}[Zh_C(Z)\mathbf{1}_{\{B>a\}}]\right\|_2 \le 2Rn^{3/2}(a^2 + 2)e^{-a^2/2}.$$

Choose the smallest even integer $a \ge 2$ with $2Rn^{3/2}(a^2 + 2)2^{-a^2/2} \le \eta/4$ (a condition with integer powers only, which implies the bound with $e^{-a^2/2}$ since $e > 2$).

Quadrature. On the cube $[-a, a]^n$ the integrand $\mathcal{I}(z) = zh_C(z)\varphi_n(z)$, $\varphi_n(z) = (2\pi)^{-n/2}e^{-\|z\|_2^2/2}$, is Lipschitz with constant $J = R(2\sqrt{n}\,a + na^3)$, since $|h_C| \le Ra$, $\mathrm{Lip}(h_C) \le R$, $\|z\|_2 \le \sqrt{n}\,a$, $\varphi_n \le 1$ and $\|\nabla\varphi_n\|_2 \le \sqrt{n}\,a$ on the cube. Split the cube into m^n equal subcubes of side $h = 2a/m$ with the smallest integer m such that $J\sqrt{n}\,h(2a)^n \le \eta/8$; the midpoint rule then approximates the truncated integral with Euclidean error at most $\eta/8$, and at every node h_C is evaluated by a linear program. The Gaussian density at the finitely many rational nodes is needed only to absolute accuracy $\theta = \eta/(8(2a)^n R\sqrt{n}\,a^2)$, which changes the quadrature by at most $\eta/8$ because $\|zh_C(z)\|_2 \le R\sqrt{n}\,a^2$ on the cube. Certified rational approximations of e^{-r} for rational r and of $\sqrt{2\pi}$ to any positive accuracy are obtained from convergent series with explicit remainders. The resulting finite sum w satisfies $\|w - \mathrm{st}(C)\|_2 \le \eta/4 + \eta/8 + \eta/8 = \eta/2$.

Projection. Return $x(C, \eta) = \mathrm{proj}_C(w) \in C$. Since $\mathrm{st}(C) \in C$ (Lemma B.1) and the projection is nonexpansive, $\|\mathrm{proj}_C(w) - \mathrm{st}(C)\|_2 = \|\mathrm{proj}_C(w) - \mathrm{proj}_C(\mathrm{st}(C))\|_2 \le \eta/2 \le \eta$. □

Proposition C.4 (Randomized selector)**.** *Let $C \subseteq c + B_1^n(R)$ be a known nonempty polytope, let $Z_1, \dots, Z_m$ be independent standard Gaussian vectors, let $V_i \in \arg\max_{u \in C} \langle Z_i, u\rangle$ (a linear program) and $\bar{x} = \frac{1}{m}\sum_i V_i$. Then $\bar{x} \in C$ on every outcome, $\mathbb{E}V_i = \mathrm{st}(C)$, and for $m \ge 4R^2/\eta^2$, $\mathbb{P}\{\|\bar{x} - \mathrm{st}(C)\|_2 > \eta\} \le \exp(-m\eta^2/(8R^2))$. Consequently, let $\delta \in (0, 1)$ and let every selector call in Algorithm 2 use $m_N = \lceil 32N^2(1 + \log(\bar{S}/\delta))/(\mathsf{b}_N\mathsf{c}_n)\rceil$ directions, where N is the horizon of the current level call and $\bar{S} = 2 + 48N(\varepsilon)$ bounds the number of selector calls. Then with probability at least $1 - \delta$ all guarantees of Proposition C.2 hold, the number of oracle queries is bounded by (23) on every outcome, and the total number of linear programs is $O\big(N(\varepsilon) + N(\varepsilon)^3 \log(N(\varepsilon)/\delta)/(\mathsf{b}_{N(\varepsilon)}\mathsf{c}_n)\big)$, polynomial in n, $N(\varepsilon)$ and $\log(1/\delta)$.*

Proof. By translation assume $c = 0$, so $\|V_i\|_2 \le R$. Fix a measurable tie-breaking rule for the maximizer; by Lemma B.1 the maximizer is almost surely unique and equals $\nabla h_C(Z_i)$, so $\mathbb{E}V_i = \mathrm{st}(C)$, and $\bar{x} \in C$ on every outcome by convexity. The V_i are independent with $\mathbb{E}\|V_i - \mathrm{st}(C)\|_2^2 \le R^2$, so $\mathbb{E}\|\bar{x} - \mathrm{st}(C)\|_2^2 \le R^2/m$ and $\mathbb{E}\|\bar{x} - \mathrm{st}(C)\|_2 \le R/\sqrt{m} \le \eta/2$ when $m \ge 4R^2/\eta^2$. Changing one V_i changes $\bar{x}$ by at most $2R/m$ in norm, hence changes $X = \|\bar{x} - \mathrm{st}(C)\|_2$ by at most $2R/m$. The bounded-differences inequality of McDiarmid (1989) gives $\mathbb{P}\{X > \mathbb{E}X + t\} \le \exp(-2t^2/(m(2R/m)^2)) = \exp(-mt^2/(2R^2))$; with $t = \eta/2 \le \eta - \mathbb{E}X$ this is the stated bound $\exp(-m\eta^2/(8R^2))$.

Now consider Algorithm 2 with every selector call replaced by this average with $m_N = \lceil 32N^2(1 + \log(\bar{S}/\delta))/(\mathsf{b}_N\mathsf{c}_n)\rceil$ directions, where N is the horizon of the current level call and $\eta_N = P_N/(4N) = R\sqrt{\mathsf{b}_N\mathsf{c}_n}/N$. Then $m_N\eta_N^2/(8R^2) \ge 4(1 + \log(\bar{S}/\delta))$ and $m_N \ge 4R^2/\eta_N^2$, so conditionally on the past (which fixes the current polytope, while the new directions are independent of it) each selector call fails, i.e. produces a point farther than η_N from the Steiner point, with probability at most $e^{-4}(\delta/\bar{S})^4 \le \delta/\bar{S}$. The algorithm is defined on every outcome. After N nonempty trials the level method returns the stored y_N (also when $A_N = 0$), and the bracket is updated by the prescribed formulas, so the widths shrink by the factors $1/2$ or $3/4$ whether or not the upper end is valid. The phase-count argument of Section C.2 uses only the widths, so it bounds the number of oracle queries by $1 + 24N(\varepsilon)$ on every outcome. Each level call makes one selector call per trial and one for z_0, so the number of selector calls is less than twice the query bound $S = 1 + 24N(\varepsilon)$; with $\bar{S} = 2S$, the union bound over the selector calls in their order of occurrence gives probability at least $1 - \delta$ that all of them succeed. On this event every chain of selected points is an η_N-accurate approximation of the Steiner points of a nested chain, so (8) holds for it, and all statements of Propositions 3.2 and C.2 apply without change.

Number of linear programs. All horizons used are powers of two at most $N(\varepsilon)$, and $N \mapsto N^2/\mathsf{b}_N$ is increasing on powers of two (the ratio of consecutive values is $4(j+1)/(j+2) \ge 2$), so every m_N is at most $m^* = \lceil 32N(\varepsilon)^2(1 + \log(2S/\delta))/(\mathsf{b}_{N(\varepsilon)}\mathsf{c}_n)\rceil$. Counting one linear minimization for the bracket, at most S emptiness tests and, over the fewer than $2S$ selector calls, at most $2Sm^*$ support linear programs, the total is at most $1 + S(2m^* + 1) = O(N(\varepsilon) + N(\varepsilon)^3 \log(N(\varepsilon)/\delta)/(\mathsf{b}_{N(\varepsilon)}\mathsf{c}_n))$. Each linear program maximizes a linear function over $\{(x,u) : -u \le x \le u,\ \sum_i u_i \le R\}$ intersected with at most $N(\varepsilon)$ stored cuts. □

C.4 Inexact oracles and finite precision

Proposition C.4 assumes exact oracle answers and linear programs solved exactly over the reals, with real Gaussian directions. On $K = B_1^n(R)$ these assumptions can be removed. Let R, L, ε be rational and let a query at a rational $x \in K$ return rational $(\widetilde{f}_x, \widetilde{g}_x)$ with

$$|\widetilde{f}_x - f(x)| \le \alpha := \frac{\varepsilon}{256}, \qquad \|\widetilde{g}_x - \nabla f(x)\|_2 \le \beta := \frac{\varepsilon}{512R}, \tag{25}$$

where the errors may be biased and may depend on the whole past, including the random bits drawn so far but not later ones. The answers lie on fixed grids $\widetilde{f}_x \in 2^{-k_f}\mathbb{Z}$, $\widetilde{g}_x \in 2^{-k_g}\mathbb{Z}^n$ whose exponents k_f, k_g are part of the input; componentwise accuracy $\varepsilon/(512Rn)$ suffices for the second condition. The *rounded level method* is Algorithm 1 with three changes. The weights of the accepted steps are $a_j = j/(2L)$, $A_j = j(j+1)/(4L)$, where j counts accepted steps, so no square root is needed. The level cut at x is the outward-rounded half-space $\{u : \langle \widetilde{g}_x, u\rangle \le c_x\}$ with $c_x = \rho\lceil(\ell + \tau - \widetilde{f}_x + \langle \widetilde{g}_x, x\rangle)/\rho\rceil$, $\tau = \alpha + 2R\beta = \varepsilon/128$, $\rho = \varepsilon/64$; it contains the true level set $\{f \le \ell\} \cap K$. The acceptance test is $d_t^2 \le \delta_N^2$, a comparison of rationals. The guarantee becomes $f(y_N) \le \ell + 48LP_N^2/N^3 + \varepsilon/32$. The *rounded bracket search* queries $x_0 = 0$, takes a vertex $v \in \arg\min_{u \in K}\langle \widetilde{g}_0, u\rangle$, uses the bracket $\underline{f} = \widetilde{f}_0 + \langle \widetilde{g}_0, v\rangle - \tau_0$, $\overline{f} = \underline{f} + LR^2/2 + 2\tau_0$, $\tau_0 = \alpha + R\beta = 3\varepsilon/512$, and the horizon rule with $w/8$ in place of $w/4$ in (22), i.e. $N'(w) = \min\{N = 2^j : N^3 \ge 6144LR^2\mathsf{b}_N\mathsf{c}_n/w\}$.

Proposition C.5 (Inexact oracle and finite-bit implementation)**.** *Let $K = B_1^n(R)$ and let the oracle satisfy (25).*

(a) With any selector that returns rational points $z_t \in C_t$ with $\|z_t - \mathrm{st}(C_t)\|_2 \le \eta_N$ (Proposition C.3 is one), the rounded bracket search returns $v \in K$ with $f(v) - \min_K f \le \varepsilon$ after at most

$S = 1 + 24N'(\varepsilon)$ *queries,* $N'(\varepsilon) = O(1 + [\frac{LR^2}{\varepsilon} \log(2n) \log(2 + \frac{LR^2}{\varepsilon} \log(2n))]^{1/3})$*; the errors do not accumulate with the number of steps.*

(b) For a known nonempty rational polytope $C \subseteq c + B_1^n(R)$ *with rational* c*, rational* $\eta \in (0, R]$ *and rational* $\gamma \in (0, 1)$*, a rational point* $\widehat{s} \in C$ *with* $\mathbb{P}\{\|\widehat{s} - \mathrm{st}(C)\|_2 \le \eta\} \ge 1 - \gamma$ *can be computed from finitely many uniformly random bits with* $m = \lceil 32(R/\eta)^2(1 + \lceil \log_2(1/\gamma) \rceil) \rceil$ *rational linear programs over* C*, in a number of bit operations polynomial in* n*,* R/η*,* $\log(1/\gamma)$ *and the bit lengths of* R*,* η *and of the description of* C*; the vertices of* C *are not enumerated.*

(c) Consequently, for rational $\delta \in (0, 1)$*, the rounded bracket search with the selectors of (b) (accuracy* $\eta = R/N$ *and failure probability* $\delta/(2S)$ *per selector call) returns a point with error at most* ε *with probability at least* $1 - \delta$*, makes at most* S *queries on every outcome, solves at most* $66S^3(1 + \lceil \log_2(2S/\delta) \rceil)$ *rational linear programs plus at most* S *emptiness tests, and its total number of bit operations is polynomial in* n*,* S*,* $\log(1/\delta)$*,* k_f*,* k_g*, the bit lengths of* $R, L, \varepsilon, \delta$ *and the bit length of the first oracle answer.*

The dependence on the first answer is unavoidable, since the affine part of f may be arbitrarily large. The stored cut constants are rounded to the grid $\rho\mathbb{Z}$ and bounded in terms of the first answer, so the bit lengths of the localization polytopes stay polynomial. The proposition concerns a randomized implementation in the bit model, and its running time is polynomial in the number of queries $S = \widetilde{\Theta}((LR^2/\varepsilon)^{1/3})$, not in $\log(1/\varepsilon)$. A deterministic polynomial-time selector and a practically fast one remain open (Section 7).

Lemma C.6 (Rounded level method)**.** *Let* $K = B_1^n(R)$*, let the oracle satisfy* (25)*, and let* $\tau = \alpha + 2R\beta = \varepsilon/128$*,* $\rho = \varepsilon/64$*. The rounded level method with level* ℓ *and horizon* N *makes at most* N *queries in* K*, and either all its cuts have empty intersection with* K*, in which case* $\min_K f > \ell$*, or it returns* $y_N \in K$ *with* $f(y_N) \le \ell + 48LP_N^2/N^3 + \varepsilon/32$*, provided every selected point is within* η_N *of the Steiner point of its polytope.*

Proof. For $x, u \in K$, $\|u - x\|_2 \le 2R$ and the Cauchy–Schwarz inequality give

$$\left| \widetilde{f}_x + \langle \widetilde{g}_x, u - x \rangle - f(x) - \langle \nabla f(x), u - x \rangle \right| \le \tau. \tag{26}$$

The cut contains the level set. If $f(u) \le \ell$, then $f(x) + \langle \nabla f(x), u - x \rangle \le f(u) \le \ell$, so $\langle \widetilde{g}_x, u \rangle \le \ell + \tau - \widetilde{f}_x + \langle \widetilde{g}_x, x \rangle \le c_x$ by (26) and the upward rounding. Hence an empty intersection certifies $\min_K f > \ell$.

Effect of the rounding. If $\langle \widetilde{g}_x, z' \rangle \le c_x$ with $z' \in K$, then, since $c_x < \ell + \tau - \widetilde{f}_x + \langle \widetilde{g}_x, x \rangle + \rho$, (26) gives $f(x) + \langle \nabla f(x), z' - x \rangle \le \ell + 2\tau + \rho = \ell + \varepsilon/32$.

Accepted steps. Number the accepted steps $j = 1, 2, \dots$ and let $a_j = j/(2L)$, $A_j = j(j+1)/(4L)$, so that $A_j = A_{j-1} + a_j$ and $La_j^2/(2A_j) = j/(2(j+1)) \le \frac{1}{2}$. Lemma C.1 with the level $\ell + \varepsilon/32$ (which the selected point z' satisfies by the previous paragraph) gives, for the accepted step with displacement $d = \|z' - z\|_2$,

$$A_j[f(y_j) - \ell] \le A_{j-1}[f(y_{j-1}) - \ell] + \tfrac{1}{2}d^2 + a_j\varepsilon/32;$$

telescoping and $\sum_{i \le j} a_i = A_j$ give $f(y_j) - \ell \le \frac{1}{2A_j} \sum_{\text{accepted}} d^2 + \varepsilon/32$, without any factor depending on the number of steps. The path bound (8), the counting of rejections and $A_k \ge k^2/(4L) \ge N^2/(16L)$ are as in the proof of Proposition 3.2, and give the claim. □

We first prove part (b) of Proposition C.5.

Proof of Proposition C.5(b). Translate by the rational vector c so that $C \subseteq B_2(0, R)$. Write $C = \{\mathsf{T}y + c' : y \in \widetilde{C}\}$ with a bounded rational polytope $\widetilde{C} = \{y \in \mathbb{R}^d : \mathsf{M}y \le \mathsf{q}\}$ given by p inequalities and rational T, c' (for the localization polytopes of our methods, $d = 2n$, $\widetilde{C}$ is the lifted description with variables (x, u), T the projection to x, $c' = 0$, and $p \le 3n + 1 + S$). Every vertex of $\widetilde{C}$ is determined by d linearly independent active inequalities, so $\widetilde{C}$ has at most $p^d =: V$ vertices and C is the convex hull of at most V points, the vertex images. For $Z \sim \mathcal{N}(0, I_n)$ let $X(Z)$ be the maximizer of $\langle Z, x \rangle$ over C; it is almost surely a unique vertex image and $\mathbb{E}X(Z) = \mathrm{st}(C)$.

Rounded directions. Let $\widehat{Z}$ be a rational vector coupled with Z so that $\mathbb{P}\{\|Z-\widehat{Z}\|_2 > h\} \le \zeta$, and let $\widehat{X} \in C$ be any rational optimizer of $\left\langle \widehat{Z}, x \right\rangle$ over C (an LP over $\widetilde{C}$ with objective $\mathsf{T}^\top \widehat{Z}$). We claim $\mathbb{P}\{\widehat{X} \ne X(Z)\} \le \zeta + V^2 h$. Suppose that $\|Z-\widehat{Z}\|_2 \le h$ and $v := X(Z) \ne \widehat{X}$. The optimal face of $\left\langle \widehat{Z}, \cdot \right\rangle$ contains $\widehat{X}$ and is the convex hull of the vertex images it contains, all with the same value of $\left\langle \widehat{Z}, \cdot \right\rangle$; since $\widehat{X} \ne v$, this face contains a vertex image $u \ne v$ with $\left\langle \widehat{Z}, u-v \right\rangle \ge 0$, while $\langle Z, v-u\rangle > 0$ by uniqueness. Thus $0 < \langle Z, (v-u)/\|v-u\|_2\rangle \le \|Z-\widehat{Z}\|_2 \le h$. For a fixed pair of distinct vertex images the left-hand side is a standard Gaussian, which falls in an interval of length h with probability at most $h/\sqrt{2\pi} < h$; the union bound over at most V^2 pairs proves the claim. Consequently $\|\mathbb{E}\widehat{X} - \mathrm{st}(C)\|_2 \le \mathbb{E}\|\widehat{X} - X(Z)\|_2 \le 2R(\zeta + V^2 h)$, and the choice $\zeta = \eta/(8R)$, $h = \eta/(8RV^2)$ makes the bias at most $\eta/2$; note $\log(1/h) = O(\log(R/\eta) + d \log p)$.

Finite Gaussian generator. Let $J = \lceil \log_2(2n/\zeta) \rceil$, $t_0 = \min\{t \in \mathbb{N} : t^2 \ge 2J\}$, $h' = h/n$, $\vartheta = 2^{-(t_0^2+2)}$, and let $k_b \ge 1$ be an integer with $2^{-k_b} \le \vartheta h'$ (we may take ζ and h to be the largest powers of two below the values prescribed above, so that all these quantities have bit length polynomial in the bit lengths of R, η and $\log V$). Then $\mathbb{P}\{\|Z\|_\infty > t_0\} \le 2ne^{-t_0^2/2} \le 2n2^{-J} \le \zeta$, and the standard normal density is at least ϑ on $[-t_0, t_0]$ (because $\sqrt{2\pi} < 4$ and $e^{t_0^2/2} < 2^{t_0^2}$), so the truncated quantile $\Phi_{t_0}^{-1}(u) = \max\{-t_0, \min\{t_0, \Phi^{-1}(u)\}\}$ is $1/\vartheta$-Lipschitz on $[0,1]$. For each coordinate draw k_b uniform bits, forming an integer $I_i \in \{0, \dots, 2^{k_b}-1\}$, put $U_i = (I_i + \frac{1}{2})2^{-k_b}$ and compute a rational $\widehat{Z}_i \in [-t_0, t_0]$ with $|\widehat{Z}_i - \Phi_{t_0}^{-1}(U_i)| \le h'/2$. For the analysis couple I_i with an independent uniform $\Theta_i \in [0,1]$, so that $\widetilde{U}_i = (I_i + \Theta_i)2^{-k_b}$ is uniform and $Z_i = \Phi^{-1}(\widetilde{U}_i)$ is standard Gaussian; then $|\widehat{Z}_i - \Phi_{t_0}^{-1}(\widetilde{U}_i)| \le h'/2 + 2^{-k_b-1}/\vartheta \le h'$, and on the event $\|Z\|_\infty \le t_0$ we have $\Phi_{t_0}^{-1}(\widetilde{U}_i) = Z_i$ and $\|\widehat{Z} - Z\|_2 \le \sqrt{n}\, h' \le h$. This is the coupling required above. To compute $\widehat{Z}_i$ we bisect $[l_0, r_0] = [-t_0, t_0]$: at the midpoint t compute a rational $\widetilde{\varphi}$ with $|\widetilde{\varphi} - \Phi(t)| \le \theta := \vartheta h'/16$ and set $r_0 = t$ if $\widetilde{\varphi} - \theta > U_i$, $l_0 = t$ if $\widetilde{\varphi} + \theta < U_i$, and otherwise return t (then $|\Phi(t) - U_i| \le 2\theta$ and $|t - \Phi_{t_0}^{-1}(U_i)| \le 2\theta/\vartheta \le h'/8$); return the midpoint as soon as $r_0 - l_0 \le h'$. Monotonicity keeps $\Phi_{t_0}^{-1}(U_i) \in [l_0, r_0]$, so at most $\lceil \log_2(2t_0/h') \rceil + 1$ evaluations of Φ are needed, and each is a truncation of the everywhere convergent series $\Phi(t) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \sum_{j \ge 0} \frac{(-1)^j t^{2j+1}}{2^j j! (2j+1)}$ with an explicit remainder, together with a rational approximation of $1/\sqrt{2\pi}$ of certified accuracy; all of this costs a number of bit operations polynomial in t_0^2, $\log(1/\theta)$ and the bit length of t. Since $t_0^2 = O(1 + \log(n/\zeta))$ and $k_b = O(1 + \log(n/\zeta) + \log(n/h))$, the generator uses nk_b random bits and polynomial work per direction.

Averaging. Draw m independent directions, solve the m linear programs, and return $\widehat{s} = \frac{1}{m}\sum_{i \le m} \widehat{X}_i \in C$ (a convex combination of points of C; no projection is needed). Let $\bar{s} = \mathbb{E}\widehat{X}_1$. As $\mathbb{E}\|\widehat{X}_1 - \bar{s}\|_2^2 \le R^2$, $\mathbb{E}\|\widehat{s} - \bar{s}\|_2 \le R/\sqrt{m} \le \eta/4$ for $m \ge 16R^2/\eta^2$, and changing one sample changes $\|\widehat{s} - \bar{s}\|_2$ by at most $2R/m$, so the bounded-differences inequality (McDiarmid, 1989) gives $\mathbb{P}\{\|\widehat{s} - \bar{s}\|_2 > \eta/2\} \le \exp(-m\eta^2/(32R^2)) \le \gamma$ for the stated m. With the bias bound, $\|\widehat{s} - \mathrm{st}(C)\|_2 \le \eta$ with probability at least $1-\gamma$. Each linear program has rational data of polynomial bit length, is solved exactly in polynomial time (Khachiyan, 1979; Karmarkar, 1984), and its basic optimal solutions have polynomial bit length by Cramer's rule; the average of m of them keeps polynomial length. If the polytope depends on earlier oracle answers and random choices, we condition on that history. The polytope and its vertex images are then fixed, the new bits are independent of it (the oracle errors may depend on the past but not on bits drawn later), and the bounds hold conditionally with the same p, d, R. □

Proof of Proposition C.5(a) *and* (c). *(a)* Let $\tau_0 = \alpha + R\beta = 3\varepsilon/512$ and let v be a rational vertex $\pm Re_i$ minimizing $\langle \widetilde{g}_0, \cdot \rangle$ over K. For every $u \in K$, convexity and (25) with $\|u\|_2 \le R$ give $f(u) \ge f(0) + \langle \nabla f(0), u\rangle \ge \widetilde{f}_0 + \langle \widetilde{g}_0, u\rangle - \tau_0 \ge \widetilde{f}_0 + \langle \widetilde{g}_0, v\rangle - \tau_0 = \underline{f}$, so $\underline{f} \le \min_K f$; and smoothness at 0 gives $f(v) \le f(0) + \langle \nabla f(0), v\rangle + LR^2/2 \le \underline{f} + 2\tau_0 + LR^2/2 = \overline{f}$. If $w = \overline{f} - \underline{f} \le \varepsilon$, return v. Otherwise $\varepsilon < w = LR^2/2 + 3\varepsilon/256$, so $\varepsilon < 128LR^2/253$ and $w < LR^2$. In a phase of width $w > \varepsilon$ with level ℓ and horizon $N = N'(w)$, Lemma C.6 either certifies $\min_K f > \ell$ or returns y_N with $f(y_N) \le \ell + w/8 + \varepsilon/32 \le \ell + w/4$. Hence the bracket

updates of Algorithm 2 keep the invariant $\underline{f} \le \min_K f \le f(v) \le \overline{f}$ and multiply w by $\frac{1}{2}$ or $\frac{3}{4}$. Every horizon exceeds 1 because $w < LR^2$, and the geometric summation of Section C.2 applies without change to the rule N', giving at most $1 + 24N'(\varepsilon)$ queries; the explicit order of $N'(\varepsilon)$ is obtained as in Section C.2 with 6144 in place of 3072. All comparisons are between rational numbers.

(b) was proved above.

(c) The width argument in the proof of Proposition C.4, and with it the bound S on the number of queries, does not depend on the accuracy of the selected points and applies here. As there, one selector call is made per trial and one per level call for z_0, so fewer than $2S$ calls are made; with failure probability $\delta/(2S)$ each, all are η_N-accurate with probability at least $1 - \delta$, and then (a) applies. A call in a level call with horizon N uses the rational accuracy $\eta = R/N$, which is at most $\eta_N = R\sqrt{\mathsf{b}_N \mathsf{c}_n}/N$ and at most R, so its points are η_N-accurate. Since $(R/\eta)^2 = N^2 \le S^2$, such a call solves at most $32S^2(1 + \lceil \log_2(2S/\delta) \rceil) + 1$ linear programs; multiplying by $2S$ gives the stated total, and the emptiness tests, one per trial, add at most S.

Bit lengths. On K, $\|\nabla f(x)\|_2 \le \|\nabla f(0)\|_2 + LR$ and $|f(x)| \le |f(0)| + R\|\nabla f(0)\|_2 + LR^2/2$, all levels lie in the initial bracket, and $|\langle \widetilde{g}_x, x \rangle| \le R(\|\nabla f(0)\|_2 + LR + \beta)$. Since the answers lie on the grids $2^{-k_f}\mathbb{Z}$, $2^{-k_g}\mathbb{Z}^n$, the coefficients $\widetilde{g}_t$ of the stored cuts have bit length $O(n(k_g + \log(2 + \|\nabla f(0)\|_2 + LR + \beta)))$. The constants $c_t \in \rho\mathbb{Z}$ satisfy $|c_t| \le |\ell| + |\widetilde{f}_{x_t}| + |\langle \widetilde{g}_{x_t}, x_t \rangle| + \tau + \rho$, hence have bit length $O(\log(2 + |f(0)| + R\|\nabla f(0)\|_2 + LR^2) + \mathrm{bits}(\rho))$, *independently of the bit length of the query point* x_t; here $\mathrm{bits}(\rho)$ is the bit length of the rational $\rho = \varepsilon/64$. The lifted localization polytopes therefore have polynomial descriptions, their vertices and the selected averages have polynomial bit length, the accepted-step weights $a_j/A_j = 2/(j+1)$ are rational, and the iterates x, y, being convex combinations with these weights of selected points, grow in length only additively over at most S steps. Together with (b) this gives the polynomial bound on the total number of bit operations. □

C.5 Other domains and smoothness metrics

For a known polytope K with V vertices contained in $B_2(c, R)$ with $c \in K$, the union bound and the tail integration of Lemma B.2 give $\omega(K - c) \le R\sqrt{2(1 + \log V)}$, so Proposition C.7 below, with $H = I$ and $\bar{\omega} = R\sqrt{2(1 + \log V)}$, gives the bounds of Proposition C.2 with $1 + \log(2n)$ replaced by $1 + \log V$. For the ℓ_p ball with $1 < p < 2$ and $q = p/(p-1)$, $\omega(B_p^n(R)) = R(\mathbb{E}\|Z\|_q^2)^{1/2}$ is at most $R(n\,\mathbb{E}|Z_1|^q)^{1/q}$ and at most $Rn^{1/q}(\mathbb{E}\|Z\|_\infty^2)^{1/2}$, and $(\mathbb{E}|Z_1|^q)^{2/q} = \Theta(q)$; Proposition C.7 then gives the error $O(LR^2 n^{2-2/p} \min\{q, \log(2n)\} \log(2N)/N^3)$, uniformly in $1 < p < 2$. This bound matches the lower bound $LR^2/N^{1+2/p}$ of Proposition D.1 up to logarithms when $N + 1 \le n = O(N)$, but not when $n \gg N$; in that regime the general question for $1 < p < 2$ remains open (for quadratics it is settled in Section H). For the Euclidean ball the bound is $O(LR^2 n \log(2N)/N^3)$, which does not improve on LR^2/N^2 when $N \le n$, as the Euclidean lower bound requires.

More generally, neither Lemma 3.1 nor the level method is specific to the ℓ_1 ball or to the Euclidean metric. The domain enters through the Gaussian width of its image under the square root of the smoothness metric. Let $H \succ 0$ be a known symmetric matrix, $\|x\|_H = \sqrt{x^\top H x}$, $\|g\|_{H^{-1}} = \sqrt{g^\top H^{-1} g}$, $S = H^{1/2}$, and let K be a known nonempty compact convex set. For a nonempty compact convex $C \subseteq K$ define the *H-Steiner point* $\mathrm{st}_H(C) = S^{-1}\,\mathrm{st}(SC) \in C$. We assume that the stored sets $C = K \cap \{\text{half-spaces}\}$ admit an emptiness test and η-accurate feasible approximations of $\mathrm{st}_H(C)$ in the norm $\|\cdot\|_H$, and that linear functions can be minimized over K and the certificate program $\min_\lambda\{\sum_j \lambda_j \langle g_j, x_j \rangle + h_K(-\sum_j \lambda_j g_j)\}$ of Algorithm 3 (a convex program, a linear program for polytopes) can be solved. These are computations with known sets, which are not counted (Section A); for polytopes, Section C.3 provides them after the change of variables $x \mapsto Sx$, with R replaced by the Euclidean radius of $S(K - c)$.

Proposition C.7 (General known geometry). *Let $\bar{\omega} \ge \omega(S(K - c))$ be known for some $c \in \mathbb{R}^n$ (the best choice is $c = \mathrm{st}_H(K)$), let $P_N^H = 2\sqrt{2}\,\bar{\omega}\sqrt{\mathsf{b}_N}$, and run SP-Cut, SP-Level and SP-Accel with $\|\cdot\|_H$ in place of $\|\cdot\|_2$, st_H in place of $\mathrm{st}(\cdot)$, and P_N^H in place of P_N (cut depth $4GP_N^H/N$; $\eta_N = P_N^H/(4N)$, $\delta_N = 4P_N^H/N$; horizon rule $N_H(w) = \min\{N = 2^j : N^3 \ge 1536L\bar{\omega}^2\mathsf{b}_N/w\}$; initial bracket $[\underline{f}, \underline{f} + W_0]$ with $\underline{f}$ from (21) and $W_0 = \pi L\bar{\omega}^2$).*

(a) If f is convex on K and the subgradient oracle returns g_x with $\|g_x\|_{H^{-1}} \le G$, then **SP-Cut** *with horizon N returns $\widehat{x} \in K$ with $f(\widehat{x}) - \min_K f < 4GP_N^H/N = 8\sqrt{2}\,G\bar{\omega}\sqrt{\mathsf{b}_N}/N$ after at most N queries.*

(b) If f is convex and differentiable on a neighborhood of K with $f(v) \le f(u) + \langle \nabla f(u), v-u\rangle + \frac{L}{2}\|v-u\|_H^2$ for $u, v \in K$ (for instance if $\|\nabla f(v) - \nabla f(u)\|_{H^{-1}} \le L\|v-u\|_H$), then the level method either certifies $\min_K f > \ell$ or returns y_N with $f(y_N) \le \ell + 48L(P_N^H)^2/N^3 = \ell + 384L\bar{\omega}^2\mathsf{b}_N/N^3$, and the bracket search returns an ε-solution after at most $1 + 24N_H(\varepsilon) = O(1 + [\frac{L\bar{\omega}^2}{\varepsilon}\log(2 + \frac{L\bar{\omega}^2}{\varepsilon})]^{1/3})$ first-order queries; with a budget of $T \ge 25$ queries the error is $O(L\bar{\omega}^2\log(2T)/T^3)$.

For $K = x_0 + B_1^n(R)$ one has $\omega(S(K-x_0)) = R(\mathbb{E}\|H^{1/2}Z\|_\infty^2)^{1/2} \le R\sqrt{2(1+\log(2n))}\max_i\sqrt{H_{ii}}$, which recovers Proposition C.2 at $H = I$ and allows anisotropic smoothness. For an ellipsoid $K = x_0 + \{u : u^\top Q u \le R^2\}$, $\omega(S(K-x_0))^2 = R^2\,\mathrm{tr}(HQ^{-1})$, so the rate is $O(LR^2\,\mathrm{tr}(HQ^{-1})\log(2N)/N^3)$; for the Euclidean ball it is $O(LR^2 n\log(2N)/N^3)$. The proposition does not show that the width is the right parameter for every K.

Proof. Change of variables. Put $\widetilde{K} = S(K-c)$, $\widetilde{f}(v) = f(S^{-1}v + c)$, so that $\nabla\widetilde{f}(v) = S^{-1}\nabla f(S^{-1}v+c)$, $\|S(u-v)\|_2 = \|u-v\|_H$ and $\|S^{-1}g\|_2 = \|g\|_{H^{-1}}$. The hypotheses of (a) and (b) become, for $\widetilde{f}$ on $\widetilde{K}$ with the same G and L, the properties used in the proofs of Propositions 3.2, C.2 and E.1: the subgradient inequality with $\|g\|_2 \le G$, and the descent inequality on $\widetilde{K}$ (the proofs use smoothness only through Lemma C.1 and the bracket). A query of $\widetilde{f}$ at $v \in \widetilde{K}$ is one query of f at $S^{-1}v + c \in K$; the Steiner point is translation equivariant (Lemma B.1), so the image $S(\mathrm{st}_H(C) - c)$ of $\mathrm{st}_H(C)$ is the Euclidean Steiner point of the image $S(C-c)$, and $\|z - \mathrm{st}_H(C)\|_H = \|S(z-c) - \mathrm{st}(S(C-c))\|_2$. So it suffices to prove (a) and (b) for the Euclidean methods on a general compact convex $\widetilde{K}$ with $\omega(\widetilde{K}) \le \bar{\omega}$; the translation c never enters the method.

Path budget. For a nested chain $\widetilde{C}_0 \supseteq \cdots \supseteq \widetilde{C}_T$ of nonempty compact convex subsets of $\widetilde{K}$ with $T \le N$, Lemma B.3 (with $c = 0$) gives $\sum_t \|\mathrm{st}(\widetilde{C}_t) - \mathrm{st}(\widetilde{C}_{t-1})\|_2 \le 2\sqrt{2(1+\log T)}\,\bar{\omega} \le 2\sqrt{2}\,\bar{\omega}\sqrt{\mathsf{b}_N} = P_N^H$, and with selector accuracy $P_N^H/(4N)$ the selected path has length at most $\frac{3}{2}P_N^H$. This is the only property of P_N used in the proofs of Proposition E.1 and Proposition 3.2, so both apply without change to P_N^H; this gives (a) and the level bound in (b), since $48L(P_N^H)^2/N^3 = 384L\bar{\omega}^2\mathsf{b}_N/N^3$.

Diameter and bracket. Let $a, b \in \widetilde{K}$ with $\|a-b\|_2 = \mathrm{diam}_2(\widetilde{K}) =: d$. Then $h_{\widetilde{K}}(Z) \ge \max\{\langle Z, a\rangle, \langle Z, b\rangle\} = \frac{1}{2}\langle Z, a+b\rangle + \frac{1}{2}|\langle Z, a-b\rangle|$, so $\mathbb{E}h_{\widetilde{K}}(Z) \ge \frac{1}{2}\mathbb{E}|\langle Z, a-b\rangle| = d/\sqrt{2\pi}$, and by Jensen's inequality $\omega(\widetilde{K}) \ge \mathbb{E}h_{\widetilde{K}}(Z)$ (the mean is nonnegative because $\mathbb{E}h_{\widetilde{K}}(Z) \ge \mathbb{E}\langle Z, u\rangle = 0$ for any $u \in \widetilde{K}$). Hence $d^2 \le 2\pi\bar{\omega}^2$ (the diameter is translation invariant). One query at $x_0 \in K$ gives $\underline{f} \le \min_K f$ as in (21) and, by the descent inequality at x_0, $f(v) \le \underline{f} + \frac{L}{2}d^2 \le \underline{f} + W_0$ for the linear minimizer v. Thus $[\underline{f}, \underline{f} + W_0]$ is a valid bracket of width $W_0 = \pi L\bar{\omega}^2$ that needs no maximization over K. If $\varepsilon \ge W_0$ one query suffices.

Bracket search. The horizon rule $N_H(w)$ is the condition $48L(P_N^H)^2/N^3 \le w/4$, so each phase behaves as in Section C.2. Every horizon exceeds 1: at $N = 1$ the rule requires $w \ge 1536L\bar{\omega}^2$, while $w \le W_0 = \pi L\bar{\omega}^2$. The geometric summation of Section C.2 applies to any rule of the form $N(w) = \min\{N = 2^j : N^3 \ge c_0\mathsf{b}_N/w\}$ with a constant $c_0 > 0$ for which every horizon exceeds 1, and gives the count $1 + 24N_H(\varepsilon)$. The explicit inversion with the scalar $\Theta = 1536L\bar{\omega}^2/\varepsilon$ gives $N_H(\varepsilon)^3 < 32\Theta(3 + \log_2\Theta)$, hence the stated order. The fixed-budget statement follows with $\varepsilon_Q = 1536L\bar{\omega}^2\mathsf{b}_Q/Q^3$ for the largest power of two Q with $24Q \le T - 1$, whose error is at most $\min\{W_0, \varepsilon_Q\} = O(L\bar{\omega}^2\log(2T)/T^3)$. □

The identity $\omega(S(K-x_0))^2 = R^2\,\mathrm{tr}(HQ^{-1})$ for the ellipsoid $K = x_0 + \{u : u^\top Q u \le R^2\}$ follows from $h_{S(K-x_0)}(z) = R\sqrt{z^\top SQ^{-1}Sz}$ and $\mathbb{E}Z^\top AZ = \mathrm{tr}\,A$; for $K = x_0 + B_1^n(R)$, $h_{S(K-x_0)}(z) = R\|Sz\|_\infty$, where each $(SZ)_i$ is centered Gaussian with variance H_{ii}, so the union bound over the $2n$ tails and the tail integration of Lemma B.2 give $\mathbb{E}\|SZ\|_\infty^2 \le 2\max_i H_{ii}(1+\log(2n))$; and for $K = x_0 + B_2^n(R)$, $\omega(S(K-x_0))^2 = R^2\,\mathrm{tr}\,H$. (The root-mean-square width is not translation

invariant, since $\omega(C+c)^2 = \omega(C)^2 + 2\langle \mathrm{st}(C), c\rangle + \|c\|_2^2$; for this reason the proposition is stated with a center.)

The two width bounds at the beginning of this subsection are proved in the same way. If K is the convex hull of V points $v_1, \dots, v_V$ of $B_2(c, R)$ and $c \in K$, then $h_{K-c}(z) = \max_i \langle v_i - c, z\rangle \geq 0$ and each $\langle v_i - c, Z\rangle$ is Gaussian with variance at most R^2, so the union bound and the tail integration of Lemma B.2 give $\mathbb{E}h_{K-c}(Z)^2 \leq 2R^2(1+\log V)$. For $K = B_p^n(R)$, $h_K(Z) = R\|Z\|_q$ with $q \geq 2$, and Jensen's inequality for the concave function $t \mapsto t^{2/q}$ gives $\mathbb{E}\|Z\|_q^2 \leq (\mathbb{E}\|Z\|_q^q)^{2/q} = (n\mathbb{E}|Z_1|^q)^{2/q}$.

D Lower bounds

All lower bounds in this section allow queries anywhere in $\mathbb{R}^n$ and require only that the output is feasible. For deterministic methods, each proof builds, for a given method, one fixed function that reproduces the whole transcript (Sections D.1, D.3 and D.4); for randomized methods the hard function is drawn at random before the interaction (Section D.5).

D.1 Deterministic first-order and nonsmooth bounds

The deterministic bounds use the signed-coordinate resisting oracle of Guzmán & Nemirovski (2015). At query x_t the adversary reveals the signed coordinate $v_t = \mathrm{sign}(x_{t,i_t})e_{i_t}$, where i_t maximizes $|x_{t,i}|$ over the coordinates not yet revealed, and answers, for an offset $\tau > 0$ fixed in the proofs, with the prefix maximum

$$\varphi_t(x) = \max_{j \leq t}\big\{\langle v_j, x\rangle - (j-1)\tau\big\}.$$

Every future form $\langle v_j, x\rangle - (j-1)\tau$, $j > t$, is at least τ below the current one at x_t, and differences of forms have ℓ_1-norm at most two; hence all completions of φ_t coincide with φ_t on the ℓ_∞ ball of radius $\tau/2$ around x_t. After N queries one coordinate is still unrevealed; aligning its sign with the output $\widehat{x}$ gives $\varphi(\widehat{x}) \geq -N\tau$, while the comparator $x^\circ = -d\sum_j v_j$ with $d = R/(N+1)^{1/p}$ has $\|x^\circ\|_p = R$ and $\varphi(x^\circ) = -d$. Guzmán & Nemirovski (2015) smooth this function by an infimal convolution; for first-order oracles we use the Moreau envelope, which gives explicit constants.

Proposition D.1 (First-order and nonsmooth lower bounds)**.** *Let $N \geq 1$, $n \geq N+1$, $1 \leq p \leq 2$ and $R > 0$. For every deterministic method that makes at most N global queries and returns $\widehat{x} \in B_p^n(R)$:*

(a) with the subgradient oracle (values and subgradients), there is a convex G-Lipschitz function and a fixed subgradient selection such that $f(\widehat{x}) - \min_{B_p^n(R)} f \geq GR/(N+1)^{1/p}$;

(b) with the first-order oracle, for every $0 < \nu \leq 1$ there is a convex function with ν-Hölder gradient, $\|\nabla f(x) - \nabla f(y)\|_2 \leq L_{1,\nu}\|x-y\|_2^\nu$, such that

$$f(\widehat{x}) - \min_{B_p^n(R)} f \;\geq\; \frac{2^{\nu-2}}{8^\nu} \cdot \frac{L_{1,\nu}R^{1+\nu}}{(N+1)^{\nu+(1+\nu)/p}}. \tag{27}$$

At $\nu = 1$ the bound is $LR^2/[16(N+1)^{1+2/p}]$, that is, $LR^2/[16(N+1)^3]$ on the ℓ_1 ball.

Proof. Set $m = N+1$ and use only the first m coordinates. For the nonsmooth claim consider $f_s(x) = G\max_{i \leq m} s_i x_i$, with signs revealed adaptively. At a query x, let a be the largest value at x of the previously revealed signed forms ($-\infty$ initially), and let u be the largest $|x_i|$ among unused coordinates. If $a \geq u$, return the earliest-revealed maximizer. Otherwise reveal a coordinate attaining u, with its sign aligned with x_i, and return its form and slope. Each answer reveals at most one sign. Every still-unknown signed form at that query is at most u, so all future sign completions preserve the value and the validity of the returned subgradient.

After the output, align one unrevealed sign with the corresponding coordinate of $\widehat{x}$ and complete the other signs arbitrarily. Then $f_s(\widehat{x}) \geq 0$, whereas $x_i^\circ = -s_i R/m^{1/p}$ for $i \leq m$, zero otherwise, is feasible and has value $-GR/m^{1/p}$. Every slope has Euclidean norm G. To obtain one fixed final oracle, order the coordinates by time of revelation, with unused ones last, and always return the

earliest maximizing form in this order. Earlier maximizing forms precede later ties, and a newly revealed form strictly exceeds every earlier one, so this rule reproduces the full transcript, including repeated queries and ties. This proves part (a) without using the feasibility of the output. (An oracle that returns the whole subdifferential is not covered.)

For the Hölder claim put

$$d = R/m^{1/p}, \qquad \tau = d/(2m), \qquad \lambda = \tau/4, \qquad \varphi(x) = \max_{i\le m}\{\langle v_i, x\rangle - (i-1)\tau\}.$$

Now a fresh coordinate is revealed at *every* query x_t: i_t maximizes $|(x_t)_i|$ among the unused coordinates, and $v_t = \operatorname{sign}((x_t)_{i_t})e_{i_t}$, with sign $+1$ at zero. The oracle uses the prefix maximum through t. Every later vector v_j, $j > t$, satisfies $\langle v_j, x_t\rangle \le \langle v_t, x_t\rangle$, so the offset puts its affine form at least τ below the current one. After the output, align the remaining signed coordinate v_m with that output. The final φ and all prefixes are convex and Euclidean 1-Lipschitz.

Let $e_\lambda\varphi(x) = \min_z\{\varphi(z) + \|z-x\|_2^2/(2\lambda)\}$. The unique proximal point $P(x)$ satisfies $(x - P(x))/\lambda \in \partial\varphi(P(x))$, so $\|x - P(x)\|_2 \le \lambda$. Monotonicity of $\partial\varphi$ shows that $I - P$ is nonexpansive. Comparing the defining minima at x and at $x+h$ gives $\nabla e_\lambda\varphi(x) = (x - P(x))/\lambda$. Thus the envelope has gradient norm at most one and gradient Lipschitz constant at most $1/\lambda$. Moreover,

$$\varphi(x) - \lambda/2 \le e_\lambda\varphi(x) \le \varphi(x), \tag{28}$$

by minimizing $\varphi(x) - \|z-x\|_2 + \|z-x\|_2^2/(2\lambda)$ over the distance for the lower inequality, and using $z = x$ for the upper.

Take $b = 2^{\nu-1}L_{1,\nu}\lambda^\nu$ and $f = be_\lambda\varphi$. Its gradient difference is bounded by $\min\{2b, b\|x-y\|_2/\lambda\}$, and the inequality $\min\{u, v\} \le u^{1-\nu}v^\nu$ gives the required Hölder condition. It remains to check that the oracle stays consistent after smoothing. At query x_t each future affine form is at least τ below the current one. A difference of two forms is 2-Lipschitz, so no future form can exceed the current one within distance $\tau/2$ of x_t. The proximal points of the full and of the prefix envelope are within $\lambda = \tau/4$ of their arguments, so for arguments within $\tau/8$ of x_t they remain in the region where full and prefix maxima coincide. Evaluating both minima at each other's proximal points shows that the envelopes are equal in that neighborhood; in particular their values and gradients agree at the query.

At the final output the last affine form has value at least $-N\tau$, while $x^\circ = -d\sum_i v_i$ has p-norm exactly R and $\varphi(x^\circ) = -d$. Hence (28) yields

$$f(\widehat{x}) - \min_{B_p^n(R)} f \ge b(d - N\tau - \lambda/2) \ge bd/2 = 2^{\nu-2}L_{1,\nu}\frac{d^{1+\nu}}{(8m)^\nu}.$$

Substituting $d = R(N+1)^{-1/p}$ gives (27). The comparator need not minimize the envelope, and this hard function need not have an unconstrained minimizer. □

D.2 Equal smoothness constants

The following corollary shows that the gap between N^{-2} and N^{-3} does not come from functions whose Euclidean constant greatly overestimates their ℓ_1 constant.

Corollary D.2 (Equal smoothness constants)**.** *Let $N \ge 1$, $n \ge N+2$, $1 \le p \le 2$ and $R, L > 0$. For every deterministic first-order method with at most N global queries and output $\widehat{x} \in B_p^n(R)$ there is a convex C^1 function f with*

$$f(\widehat{x}) - \min_{B_p^n(R)} f \;\ge\; \frac{LR^2}{16(N+1)^{1+2/p}},$$

whose smallest gradient Lipschitz constants in ℓ_2 and in ℓ_1 are both equal to L, also when restricted to pairs of points in the interior of $B_p^n(R)$.

Proof. Put $m = N+1$ and write each point of $\mathbb{R}^n$ as $x = (u, z, w)$, where $u \in \mathbb{R}^m$, $z \in \mathbb{R}$, and $w \in \mathbb{R}^{n-m-1}$ (the last block is omitted when $n = m+1$). Given a function $f_0 : \mathbb{R}^m \to \mathbb{R}$, define

$$f(u, z, w) = f_0(u) + \frac{L}{2}z^2.$$

We simulate the prescribed algorithm with a first-order oracle for f_0. A query (u, z, w) uses one query of f_0 at u, and the algorithm receives

$$\left(f_0(u) + \frac{L}{2}z^2,\ (\nabla f_0(u), Lz, 0)\right).$$

The simulation reproduces all future query coordinates of the algorithm, including z and w; queries can be arbitrarily large and need not lie in any previous linear span. The simulated method returns the first block $\widehat{u}$ of the final output, a feasible point of $B_p^m(R)$, so Proposition D.1, at $\nu = 1$, supplies one fixed convex function f_0 with Euclidean gradient Lipschitz constant at most L and

$$f_0(\widehat{u}) - \min_{B_p^m(R)} f_0 \geq \frac{LR^2}{16(N+1)^{1+2/p}}.$$

The function f built from this f_0 reproduces all simulated answers. Projection onto the first block maps $B_p^n(R)$ into $B_p^m(R)$, and every point of the latter embeds by setting $z = w = 0$. Since the added quadratic is nonnegative, $\min_{B_p^n(R)} f = \min_{B_p^m(R)} f_0$, and the same observation at the output proves the displayed lower bound.

For any $x = (u, z, w)$ and $x' = (u', z', w')$, block orthogonality gives

$$\|\nabla f(x) - \nabla f(x')\|_2^2 \leq L^2\big(\|u - u'\|_2^2 + |z - z'|^2\big) \leq L^2\|x - x'\|_2^2.$$

Moreover, Euclidean smoothness of f_0 implies

$$\|\nabla f_0(u) - \nabla f_0(u')\|_\infty \leq L\|u - u'\|_1.$$

Consequently

$$\|\nabla f(x) - \nabla f(x')\|_\infty \leq L\max\{\|u - u'\|_1, |z - z'|\} \leq L\|x - x'\|_1.$$

Both constants are therefore at most L. For $x = 0$ and $x' = te_{m+1}$ with $0 < t < R$, both Lipschitz ratios are equal to L, and these points lie in the interior of the feasible ball. □

The extra coordinate fixes both smallest constants and carries no information about f_0.

D.3 Higher-order oracles

For oracles of order k we smooth the same maximum of signed coordinates by convolution with a compactly supported product kernel on the $m = N + 1$ active coordinates (Lemma D.3 below). Its support is an ℓ_∞ ball, so smoothing preserves the local agreement exactly, and since the score of the one-dimensional factor of the kernel has Fisher information 10, the directional derivatives of the product kernel are bounded in L^1 uniformly in the dimension. Repeated smoothing and offset maxima are standard in lower-bound constructions (Agarwal & Hazan, 2018; Contreras et al., 2025); the point of the kernel below is that its *Euclidean* derivative bounds do not depend on the dimension.

Lemma D.3 (Compact product smoothing)**.** *Let $k \geq 1$, $\rho > 0$, and let $\varphi : \mathbb{R}^m \to \mathbb{R}$ be globally Euclidean G-Lipschitz. Define*

$$\chi(t) = \begin{cases} \frac{15}{16}(1 - t^2)^2, & |t| \leq 1, \\ 0, & |t| > 1, \end{cases} \qquad \chi_h(z) = h^{-m}\prod_{i=1}^m \chi(z_i/h), \quad \mathcal{S}_{\rho,k}\varphi = \varphi * \chi_h^{*k}, \qquad h = \rho/k.$$

Here χ_h^{*k} denotes convolution of k copies of χ_h. Put $\theta = k\sqrt{10}/\rho$. Then $\mathcal{S}_{\rho,k}\varphi \in C^{k+1}$ and

$$\|D^j\mathcal{S}_{\rho,k}\varphi(x)\| \leq G\theta^{j-1} \qquad (1 \leq j \leq k+1), \tag{29}$$

$$\|D^k\mathcal{S}_{\rho,k}\varphi(x) - D^k\mathcal{S}_{\rho,k}\varphi(y)\| \leq 2^{1-\nu}G\theta^{k-1+\nu}\|x - y\|_2^\nu \qquad (0 < \nu \leq 1). \tag{30}$$

Convexity is preserved. If φ is convex and G_∞-Lipschitz in ℓ_∞, then

$$\varphi \leq \mathcal{S}_{\rho,k}\varphi \leq \varphi + G_\infty\rho. \tag{31}$$

If two functions agree on $B_\infty(x, a)$ with $a > \rho$, their smoothings agree on $B_\infty(x, a - \rho)$.

Proof. The density χ is even, nonnegative and C^1 on $\mathbb{R}$, with both its value and first derivative zero at ± 1. Direct integration gives

$$\int \chi = 1, \qquad \int t\chi(t)\,dt = 0, \qquad \int \chi' = 0, \qquad \int_{-1}^{1} \frac{\chi'(t)^2}{\chi(t)}\,dt = \int_{-1}^{1} 15t^2\,dt = 10.$$

The ratio is taken on $(-1, 1)$, where $\chi > 0$; the score need not be bounded near the endpoints. For independent Z_i with density χ, the scores $S_i = \chi'(Z_i)/\chi(Z_i)$ have mean zero and second moment ten. Thus, for every $u \in \mathbb{R}^m$, independence and Cauchy–Schwarz yield

$$\|D_u \chi_h\|_{L^1} = h^{-1}\mathbb{E}\left|\sum_i u_i S_i\right| \le \frac{\sqrt{10}}{h}\|u\|_2. \tag{32}$$

The total convolution noise has mean zero and infinity norm at most $kh = \rho$. This gives locality, convexity preservation, and (31) by Jensen's inequality and the infinity-norm Lipschitz bound. (The Euclidean norm of the noise may be as large as $\rho\sqrt{m}$, hence the constant G_∞ in (31).)

For the derivative bounds, put one weak derivative on φ and each remaining derivative on a different kernel factor. For $1 \le j \le k+1$,

$$D_{u_1} \cdots D_{u_j} \mathcal{S}_{\rho,k}\varphi = (D_{u_1}\varphi) * (D_{u_2}\chi_h) * \cdots * (D_{u_j}\chi_h) * \chi_h^{*(k-j+1)}.$$

At $j = 1$ all k factors are undifferentiated; at $j = k+1$ the last factor is omitted. Since $\|D_{u_1}\varphi\|_{L^\infty} \le G\|u_1\|_2$, Young's inequality and (32) bound this expression by $G\theta^{j-1}\prod_{i=1}^{j}\|u_i\|_2$. These distributional derivatives have continuous representatives: if J is the combined L^1 kernel and v tends to zero, their translation difference is bounded by $G\|u_1\|_2\|J(\cdot + v) - J\|_{L^1}$, which tends to zero. Integrating continuous weak partial derivatives along coordinate lines shows that they are classical derivatives, so $\mathcal{S}_{\rho,k}\varphi \in C^{k+1}$. Since the directions $u_1, \ldots, u_j$ are arbitrary, this is the tensor-norm bound (29), without polarization or coordinate-expansion loss. The case $j = k$ at x and at y, and the case $j = k+1$ integrated along the segment, give

$$\|D^k \mathcal{S}_{\rho,k}\varphi(x) - D^k \mathcal{S}_{\rho,k}\varphi(y)\| \le \min\{2G\theta^{k-1}, G\theta^k\|x - y\|_2\}.$$

Applying $\min\{a, b\} \le a^{1-\nu}b^\nu$ proves (30), with $\nu = 1$ read directly. □

Proposition D.4 (Lower bound for order-k methods)**.** *Let $k, N \ge 1$ be integers, $0 < \nu \le 1$, $s = k + \nu$, $1 \le p \le 2$, $R, L_{k,\nu} > 0$, and $n \ge N + 1$. For every deterministic method that makes at most N global order-k queries and returns $\widehat{x} \in B_p^n(R)$ there is a convex function $f \in C^{k+1}(\mathbb{R}^n)$ satisfying* (12) *such that*

$$f(\widehat{x}) - \min_{B_p^n(R)} f \;\ge\; \frac{2^{\nu-2}}{(8k\sqrt{10})^{s-1}} \cdot \frac{L_{k,\nu}R^s}{(N+1)^{s(1+1/p)-1}}. \tag{33}$$

Proof. If the method stops early, pad its transcript with repeated queries. Set $m = N + 1$ and work on the first m coordinates. Define

$$d = R/m^{1/p}, \qquad \tau = d/(2m), \qquad \rho = \tau/4, \qquad b = 2^{\nu-1}L_{k,\nu}\left(\frac{\rho}{k\sqrt{10}}\right)^{s-1}.$$

At query x_t, select an unused coordinate $i_t \le m$ maximizing $|(x_t)_{i_t}|$; use the smallest index to break ties. Choose $v_t = \operatorname{sign}((x_t)_{i_t})e_{i_t}$, with positive sign at zero, and set

$$\varphi_t(x) = \max_{1 \le j \le t}\{\langle v_j, x\rangle - (j-1)\tau\}.$$

The oracle returns the value and the derivatives through order k of $b\mathcal{S}_{\rho,k}\varphi_t$ at x_t, the smoothing always acting on the first m coordinates. A later vector v_j, $j > t$, uses a coordinate unused at time t, so $\langle v_j, x_t\rangle \le \langle v_t, x_t\rangle$. Its shifted form is therefore at least $(j-t)\tau \ge \tau$ below form t there. The difference of two signed coordinate slopes has ℓ_1 norm at most two. Consequently every future completion φ of the maximum agrees with φ_t throughout $B_\infty(x_t, \tau/2)$, and Lemma D.3 gives the identity

$$\mathcal{S}_{\rho,k}\varphi = \mathcal{S}_{\rho,k}\varphi_t \quad \text{on } B_\infty(x_t, \tau/4).$$

Hence the value and all requested derivatives at x_t are those of every completion, for queries of any size and position, repeated ones included.

After the output, align the remaining coordinate's sign with its coordinate in $\widehat{x}$ and call the resulting vector v_m. Let $\varphi(x) = \max_{j\le m}\{\langle v_j, x\rangle - (j-1)\tau\}$ and $f = b\mathcal{S}_{\rho,k}\varphi$. This one fixed function reproduces all answers. Its slopes have Euclidean and ℓ_1 norm one, so the smoothing lemma with $G = G_\infty = 1$ proves the claimed regularity and convexity. The aligned final sign gives $\varphi(\widehat{x}) \ge -N\tau$, while $x^\circ = -d\sum_{j=1}^m v_j$ has p-norm exactly R and $\varphi(x^\circ) = -d$. Jensen's inequality at the output and the upper bias bound at the comparator give

$$f(\widehat{x}) - \min_{B_p^n(R)} f \ge b(d - N\tau - \rho) = bd\left(\frac{1}{2} + \frac{3}{8m}\right) \ge bd/2.$$

Substitution proves (33). □

For $p = 2$ the exponent $s(1 + 1/p) - 1 = (3s - 2)/2$ is the optimal Euclidean exponent of methods of order k (Arjevani et al., 2019; Grapiglia & Nesterov, 2020; Kovalev & Gasnikov, 2022); for $p < 2$ it is larger by $s(1/p - 1/2)$, and for $p = 1$ it equals $2s - 1$, the exponent attained by SP-Taylor (Proposition E.4). For $k = 1$, smoothing by a Moreau envelope (Proposition D.1) gives better constants and also covers the nonsmooth case.

D.4 The simplex and prox steps

Proposition D.5 (Exact oracle reduction to a simplex)**.** *The linear map $A : \Delta_{2d} \to B_1^d(1)$, $A(w) = w_+ - w_-$ for $w = (w_+, w_-) \in \mathbb{R}^d \times \mathbb{R}^d$, is onto and has Euclidean operator norm $\sqrt{2}$. If $\bar{f} = f \circ A$, a query of $\bar{f}$ through order k is simulated with one query of f through order k and known linear transformations, and*

$$L_{k,\nu}(\bar{f}) \le 2^{(k+\nu)/2} L_{k,\nu}(f), \qquad G(\bar{f}) \le \sqrt{2}\, G(f).$$

The constrained minima coincide, and the error of w for $\bar{f}$ equals that of Aw for f.

Proof. For $x \in B_1^d(1)$ take its positive and negative parts and add half the remaining mass $1 - \|x\|_1$ to both entries of one signed pair. This gives a preimage in Δ_{2d}. Conversely $\|w_+ - w_-\|_1 \le \sum_i (w_{+,i} + w_{-,i}) = 1$. Since $AA^\top = 2I$, its operator norm is $\sqrt{2}$. The chain rule gives $D^k\bar{f}(w)[h_1, \dots, h_k] = D^k f(Aw)[Ah_1, \dots, Ah_k]$. The change of the base point contributes $\|A\|^\nu$, and the k directions contribute $\|A\|^k$. Surjectivity proves the remaining claims. □

Thus the deterministic lower bounds for $p = 1$ transfer to the simplex, with twice the dimension and the stated constant loss; a larger simplex is mapped onto $B_1^d(1)$ by sending the additional vertices to zero.

The lower bounds count oracle queries and allow any computation with known quantities between them. They therefore also apply to methods built on a non-Euclidean prox-setup: a prox step with a known prox-function, for instance the entropy on the simplex, whose inputs are chosen from the transcript (and, for a randomized method, from its random bits) is such a computation. In particular, Propositions D.4 and D.8 give the exponent $2s - 1$ on $B_1^n(R)$ also for these methods.

D.5 Randomized methods

The resisting oracle above adapts the function to the queries and therefore gives no bound for randomized methods. The bounds below draw one random function *before* the interaction. All three proofs follow the same scheme. For a *deterministic* method we bound the error averaged over the random function; to do so we pass to a more informative oracle, whose answers determine the true ones, and we keep track of the conditional law of the unrevealed parameters given the history. This law stays explicit: a uniformly random subset of the unchecked indices for a hidden support (by exchangeability), and independent exponential residuals above known thresholds for random shifts (by the product structure and the memoryless property). One query reveals only a few parameters, so with dimension linear in N much of the function remains unknown. For a *randomized* method we fix its random seed, including the randomization of the output, apply the deterministic bound and

average over the seed; exchanging the two expectations gives one fixed function whose expected error over the seed is at least the bound. Queries are global and return values together with (sub)gradients, and a method that stops early is padded with dummy queries.

Proposition D.6 (Randomized nonsmooth lower bound)**.** *Let $G, R > 0$, $N \geq 1$, $1 \leq p \leq 2$ and $n \geq 2^{p+2}N$. For every randomized method with at most N global queries of values and subgradients and output in $B_p^n(R)$ there is a set $S \subseteq \{1, \dots, n\}$ with $|S| = 2N$ such that $f_S(x) = G \max_{i \in S} x_i$, with the fixed subgradient selection $Ge_{\min \arg\max_{i \in S} x_i}$, satisfies*

$$\mathbb{E}\big[f_S(\widehat{x}) - \min_{B_p^n(R)} f_S\big] \;\geq\; \frac{GR}{2(2N)^{1/p}};$$

for $p = 1$ the bound is $GR/(4N)$, and it also holds on $\Delta_n(R)$ for $\widetilde{f}_S(z) = G \max_{i \in S}(-z_i)$ with the selection $-Ge_{\min \arg\max_{i \in S}(-z_i)}$.

Proposition D.7 (Randomized smooth lower bound)**.** *Let $L, R > 0$, $N \geq 1$ and $n \geq 32N + 1$. For every randomized method with at most N global first-order queries and output in K, where $K = B_1^n(R)$ or $K = \Delta_n(R)$, there is a convex $f \in C^1(\mathbb{R}^n)$ with L-Lipschitz gradient such that $\mathbb{E}[f(\widehat{x}) - \min_K f] \geq LR^2/(2^{50}N^3)$. If in addition $\mu_1 > 0$ and the slack $\Lambda = L/\mu_1 - n$ satisfies $\Lambda \geq 2^{51}N^3$, the function can be taken μ_1-strongly convex in ℓ_1 on $\mathbb{R}^n$ with the bound $LR^2/(2^{52}N^3)$.*

Proposition D.8 (Randomized higher-order lower bound)**.** *Let $k \geq 1$, $0 < \nu \leq 1$, $s = k + \nu$, $L_{k,\nu}, R > 0$, $N \geq 1$ and $n \geq 32N + 1$. For every randomized method with at most N global order-k queries and output in K, where $K = B_1^n(R)$ or $K = \Delta_n(R)$, there is a convex $f \in C^{k+1}(\mathbb{R}^n)$ satisfying (12) such that*

$$\mathbb{E}\big[f(\widehat{x}) - \min_K f\big] \;\geq\; c'_{k,\nu} \frac{L_{k,\nu} R^s}{N^{2s-1}},$$

where $c'_{k,\nu} > 0$ depends only on k and ν and is explicit in Section D.5.3.

The hard functions of Propositions D.7 and D.8 are smoothed coordinate maxima $\max_i(x_i - a_i)$ with independent exponential shifts a_i. The smoothing is a Moreau envelope for $k = 1$ and a compact product kernel for general k. The proofs use that a full order-k answer at x is determined by the shifts of the coordinates near the maximum (the active coordinates of the Moreau envelope, or those within 2ρ of the maximum for the kernel with smoothing parameter ρ), and that on average at most two of these are revealed for the first time at each query. The constants are not optimized. Randomized lower bounds of this type are classical (Diakonikolas & Guzmán, 2020; Agarwal & Hazan, 2018; Garg et al., 2021); the statements above are for Euclidean smoothness on an ℓ_1 domain, with global queries and dimension linear in N, matching the upper bounds of Sections C and E.

D.5.1 Hidden supports: proof of Proposition D.6

Proof of Proposition D.6. The family. Let $0 \leq N < m \leq n$, let S be uniform among the m-element subsets of $\{1, \dots, n\}$, and $f_S(x) = G \max_{i \in S} x_i$ on $\mathbb{R}^n$. Then $|f_S(x) - f_S(y)| \leq G\|x - y\|_\infty \leq G\|x - y\|_2$, every subgradient lies in $G \operatorname{conv}\{e_i : i \in S\}$ and has norm at most G. The oracle returns $(f_S(x), Ge_{i(x,S)})$ with $i(x, S) = \min \arg\max_{i \in S} x_i$, a fixed rule. The minimum over $K = B_p^n(R)$ is $-GRm^{-1/p}$: if $t = \max_{i \in S} x_i < 0$ then $R^p \geq \sum_{i \in S} |x_i|^p \geq m|t|^p$, and the value is attained at $x_i = -Rm^{-1/p}$ on S and 0 elsewhere.

A more informative oracle. Fix a deterministic method. The extended history consists of the elements of S found so far, the indices excluded so far, and the set U of unchecked indices. At a query x, order U by decreasing x_i (ties by increasing index) and test membership in S in this order up to and including the first positive answer; the tested indices leave U and their status is revealed. Each query thus reveals exactly one new element of S, even when a previously found element already determines the true answer. The true answer $(f_S(x), Ge_{i(x,S)})$ is determined by these data: the new positive index is the best among the previously unknown elements of S (with the same tie rule), and comparing it with the known elements gives $i(x, S)$. Hence a lower bound for this oracle also holds for the original one.

Conditional distribution. After any extended history, $S \cap U$ is uniform among the subsets of U of its size. Each membership test is chosen by the revealed data, and its answer only includes or excludes

the tested index, so all remaining admissible subsets stay equally likely; the query points are functions of the history and add no constraint. After N queries put $r = m - N > 0$ and $q = |U| \geq r$. We claim

$$\mathbb{E}\frac{1}{q} = \frac{m}{n(m-N)}. \tag{34}$$

Let r'/q' be the density of positive indices among the q' unchecked ones before an individual test. Before the Nth positive index is found, $r' \geq m-N+1 \geq 2$, so $q' \geq 2$; a test succeeds with probability r'/q', and the conditional expectation of the density after the test is $\frac{r'}{q'}\frac{r'-1}{q'-1} + (1-\frac{r'}{q'})\frac{r'}{q'-1} = \frac{r'}{q'}$. After the Nth positive index is found, we keep the density constant. The number of tests is at most $n-(m-N)$, so every branch can be padded to the same finite length, and repeated application of the identity shows that the expected final density equals the initial one, m/n. The final density is $(m-N)/q$, which proves (34).

Conditional risk. Let $\widehat{x} \in K$ be the output, a fixed point given the history. Since $S \cap U$ is nonempty and uniform,

$$\mathbb{E}[f_S(\widehat{x}) \mid \text{history}] \geq G\,\mathbb{E}\Big[\frac{1}{r}\sum_{i\in S\cap U}\widehat{x}_i \;\Big|\; \text{history}\Big] = \frac{G}{q}\sum_{i\in U}\widehat{x}_i \;\geq\; -GRq^{-1/p}, \tag{35}$$

where the last step is Hölder's inequality $\sum_{i\in U}|\widehat{x}_i| \leq q^{1-1/p}R$. For $1 \leq p \leq 2$ the function $t \mapsto t^{1/p}$ is concave, so Jensen's inequality and (34) give $\mathbb{E}[q^{-1/p}] \leq (m/(n(m-N)))^{1/p}$. Hence, for every deterministic method,

$$\mathbb{E}_S\big[f_S(\widehat{x}) - \min_K f_S\big] \geq GR\Big[m^{-1/p} - \Big(\frac{m}{n(m-N)}\Big)^{1/p}\Big]. \tag{36}$$

For a randomized method, (36) holds for every fixed seed, and exchanging the two expectations (S takes finitely many values and the errors are bounded on K) gives a fixed S whose expected error over the seed is at least the right-hand side.

Parameters. With $m = 2N$ and $n \geq 2^{p+2}N$, $(m/(n(m-N)))^{1/p} = (2/n)^{1/p} \leq \frac{1}{2}(2N)^{-1/p}$, so the bound is $GR/(2(2N)^{1/p})$; for $p = 1$, $n \geq 8N$ it is $GR/(4N)$.

Simplex. For $p = 1$ use $\widetilde{f}_S(z) = G\max_{i\in S}(-z_i)$ on $\Delta_n(R)$ with the oracle obtained from the one above at $-z$ (sign of the subgradient reversed). An output $z \in \Delta_n(R)$ gives $-z \in B_1^n(R)$, and the point $z_i = R/m$ on S, 0 elsewhere, lies in $\Delta_n(R)$ with value $-GR/m$, the minimum over the ball, so the argument applies without change. □

D.5.2 Exponential shifts and the Moreau envelope: proof of Proposition D.7

Proof of Proposition D.7. In this proof n denotes the number of *shifted* coordinates and $d = n+1$ the ambient dimension, which is the n of the statement; the hypothesis $n \geq 32N+1$ of Proposition D.7 thus reads $n \geq 32N$. Let $c_\star = 2^{17} = 131072$. We prove the bound $LR^2/(c_\star^2 n(n+1)^2)$ for $K = B_1^d(R)$; for $n = 32N$ it is at least $LR^2/(2^{50}N^3)$ because $32 \cdot 33^2 < 2^{16}$. Larger dimensions and the simplex are treated at the end.

The family. Let $\delta = R/d$, $a_0 = 0$, and let $a_1, \ldots, a_n$ be independent exponential with mean δ (density $\delta^{-1}e^{-a/\delta}$ on $[0,\infty)$). Put

$$\varphi_a(x) = \max_{0\leq i\leq n}(x_i - a_i), \qquad \lambda = \frac{\delta}{c_\star n},$$
$$h_a(x) = \inf_z\Big\{\varphi_a(z) + \frac{\|z-x\|_2^2}{2\lambda}\Big\}, \qquad f_a = L\lambda h_a.$$

The distribution is known to the method; only the realization a is unknown.

Smoothness. The function φ_a is convex and 1-Lipschitz in ℓ_2. Its conjugate equals $\langle a, p\rangle$ on Δ_d and $+\infty$ elsewhere, so $h_a(x) = \max_{p\in\Delta_d}\{\langle p, x-a\rangle - \frac{\lambda}{2}\|p\|_2^2\}$ has a unique maximizer $p(x)$, and the variational inequalities of the maximizers at x and y give $\lambda\|p(x)-p(y)\|_2^2 \leq \langle p(x)-p(y), x-y\rangle$. Hence $\nabla h_a = p$ is $1/\lambda$-Lipschitz, f_a is convex with L-Lipschitz gradient, and, since $\|p\|_2 \leq 1$ on Δ_d, $0 \leq \varphi_a - h_a \leq \lambda/2$.

Information. The maximizer is $p_i = (x_i - a_i - t)_+/\lambda$ with the unique threshold t such that $\sum_i (x_i - a_i - t)_+ = \lambda$, and $h_a(x) = t + \frac{\lambda}{2}\|p\|_2^2$. The value and gradient of f_a determine (p, t); on a coordinate with $p_i > 0$ they reveal $a_i = x_i - t - \lambda p_i$ exactly, and on a coordinate with $p_i = 0$ they reveal only $a_i \geq x_i - t$. Conversely, these data determine the answer, so the answer contains no further information on the unrevealed parameters. By induction on the queries, conditionally on any history the unrevealed parameters are independent with $a_i = b_i + E_i$, where b_i is the maximum of 0 and all past thresholds $x_i^{(j)} - t_j$, and E_i is exponential with mean δ. For the induction step, note that given the past the query point is fixed. On the event that the active set is J, the answer (t, p) is an affine injective function of the unrevealed active parameters (the revealed ones being fixed), and each inactive unrevealed parameter enters only through the constraint $a_i \geq x_i - t$. Hence, given the answer, the density of the inactive unrevealed parameters is proportional to $\prod_i \pi_i(a_i)\mathbf{1}\{a_i \geq x_i - t\}$, where π_i is the current conditional density of a_i; it keeps its product form, and the memoryless property gives the exponential law of the excess. (Equality of an unknown parameter with the threshold has probability zero.)

Number of revelations. Let $s_i = x_i - a_i$ and $M = \max_i s_i$; since $t \geq M - \lambda$, every active coordinate satisfies $s_i > M - \lambda$. Condition on the history, on the index attaining M and on the value of M. For any other unrevealed index the condition of not exceeding M truncates E_i from below; after this truncation, $\mathbb{P}\{s_i > M - \lambda \mid \cdot\} \leq 1 - e^{-\lambda/\delta}$ by the memoryless property (if the window below M is shorter than λ the probability is smaller). Counting the maximizer as one revelation, $\mathbb{E}[\text{new revelations} \mid \text{history}] \leq 1 + n(1 - e^{-\lambda/\delta}) \leq 1 + n\lambda/\delta \leq 2$.

Uncertain coordinates. After N queries let B be the set of unrevealed indices $i \geq 1$ with $b_i \leq \delta/4$. Since $b_i \leq a_i$, $|B| \geq \#\{i \geq 1 : a_i \leq \delta/4\} - \#\{\text{revealed}\}$, and with $1 - e^{-1/4} \geq 1/5$ and $n \geq 32N$,

$$\mathbb{E}|B| \geq n(1 - e^{-1/4}) - 2N \geq n/8. \tag{37}$$

The nonsmooth optimum. The equation $\sum_{i=0}^n (t_* - a_i)_+ = R$ has a unique positive root (as $a_0 = 0$), $\min_K \varphi_a = -t_*$, and the unique minimizer is $x_i^\star = -(t_* - a_i)_+$: the smallest ℓ_1-norm of a point with $\varphi_a \leq -t$ is $\sum_i (t - a_i)_+$ (each coordinate with $a_i < t$ must be at most $a_i - t$, the others are set to 0), and at $t = t_*$ this norm equals R. Since $a_i \geq 0$, $t_* \geq \delta$. For $x \in K$ let $r = \varphi_a(x) + t_* \geq 0$ and $w_i = (t_* - a_i)_+$; if $w_i > 0$ then $x_i + w_i \leq r$. Using $\sum_i w_i = R$ and $\|x\|_1 \leq R$,

$$\|x - x^\star\|_1 = \|x\|_1 + R - 2\sum_i \min\{(-x_i)_+, w_i\} \leq 2\sum_i (w_i - (-x_i)_+)_+ \leq 2dr, \tag{38}$$

because each term is 0 when $w_i = 0$, is $(w_i + x_i)_+ \leq r$ when $x_i < 0$, and is $w_i \leq w_i + x_i \leq r$ when $x_i \geq 0$. Next, change one a_i, $i \geq 1$, from u to v with $0 \leq u < v < \delta$; the roots satisfy $t_v \geq t_u \geq \delta$, so the coordinates 0 and i are active for both, and subtracting the two equations, $0 = 2(t_v - t_u) - (v - u) + \sum_{j \neq 0,i}[(t_v - a_j)_+ - (t_u - a_j)_+]$ with a nonnegative sum; hence $0 \leq t_v - t_u \leq (v - u)/2$ and $x_i^\star(v) - x_i^\star(u) \geq (v - u)/2$.

Conditional risk. Fix the history, an index $i \in B$, and then all other parameters; $a_i = b_i + E_i$ keeps its conditional law. The events $A_1 = \{E_i \leq \delta/8\}$ and $A_2 = \{3\delta/8 \leq E_i \leq \delta/2\}$ have probabilities at least $1/9$ and $e^{-3/8}(1 - e^{-1/8}) \geq \frac{5}{8} \cdot \frac{1}{9} > \frac{1}{16}$; on both, $a_i \leq 3\delta/4 < \delta$, and any two values from the two events differ by at least $\delta/4$, so the corresponding $x_i^\star$ differ by at least $\delta/8$. For two independent conditional copies $X_i^\star, \widetilde{X}_i^\star$ and any number z (in particular the output coordinate, fixed given the history), $\mathbb{E}|z - X_i^\star| \geq \frac{1}{2}\mathbb{E}|X_i^\star - \widetilde{X}_i^\star| \geq \mathbb{P}(A_1)\mathbb{P}(A_2)\delta/8 \geq \delta/2048$. Averaging over the other parameters and summing over the history-dependent set B, (37) gives $\mathbb{E}\|\widehat{x} - x^\star\|_1 \geq n\delta/16384$, and (38) with $n/d \geq 1/2$ gives $\mathbb{E}[\varphi_a(\widehat{x}) - \min_K \varphi_a] \geq n\delta/(32768d) \geq \delta/65536$. By $0 \leq \varphi_a - h_a \leq \lambda/2$ and $\min_K h_a \leq \min_K \varphi_a$, $\mathbb{E}[h_a(\widehat{x}) - \min_K h_a] \geq \delta/65536 - \lambda/2 \geq \delta/c_\star$, and multiplying by $L\lambda$ gives $LR^2/(c_\star^2 n d^2)$.

Randomized methods, larger dimension, simplex. Randomized methods are handled as described at the beginning of Section D.5; the errors are uniformly bounded by $2L\lambda R$ on K (as $\|\nabla h_a\|_2 \leq 1$), so the expectations can be exchanged. In dimension $d' > d$ let the function depend on the first d coordinates only. The projection of a feasible output has ℓ_1-norm at most R, the minimum is unchanged, and the extra coordinates of a global query reveal nothing. For the simplex use $\widetilde{f}_a(z) = f_a(-z)$: $z \in \Delta_d(R)$ gives $-z \in B_1^d(R)$, the minimizer $x^\star$ corresponds to the feasible point $z_i = (t_* - a_i)_+$, the reflection preserves smoothness and the oracle information, and (38) and the risk argument apply without change with the minimum of φ_a over $-\Delta_d(R)$, which equals $-t_*$.

Strong convexity. From $(t-a_i)_+ \geq t-a_i$ and the root equation, $t_* \leq R/d + \frac{1}{d}\sum_{i\geq 1} a_i$, so $\mathbb{E}\|x^\star\|_2^2 \leq R\,\mathbb{E}t_* \leq 2R^2/d$. Let $d' \geq d$ be the ambient dimension (the n of the statement), while the number of shifted coordinates is $n = 32N$; let $\Lambda = L/\mu_1 - d' \geq 0$ be the slack, and write $x = (u, v) \in \mathbb{R}^d \times \mathbb{R}^{d'-d}$. Put $\theta = Ld/(d+\Lambda) \in (0, L]$ and

$$\bar{f}_a(u,v) = (L-\theta)\lambda h_a(u) + \frac{\theta}{2}\|u\|_2^2 + \frac{L}{2}\|v\|_2^2$$

(no last term if $d' = d$). It is convex and L-smooth on $\mathbb{R}^{d'}$, and its curvature is at least θ in the first d coordinates and L in the others, so by Lemma F.7 in the ℓ_1 case it is μ_1-strongly convex in ℓ_1, since $d/\theta + (d'-d)/L = (d+\Lambda+d'-d)/L = 1/\mu_1$. The known quadratic part is subtracted from the oracle answers and the extra coordinates of a global query reveal nothing, so the information argument is unchanged, and the first block $\widehat{u}$ of a feasible output satisfies $\|\widehat{u}\|_1 \leq R$. Comparing with the feasible point $(x^\star, 0)$, where $x^\star$ is the minimizer of φ_a on $B_1^d(R)$ (not necessarily a minimizer of $\bar{f}_a$), and dropping the nonnegative quadratic at the output,

$$\begin{aligned}\mathbb{E}[\bar{f}_a(\widehat{x}) - \min_K \bar{f}_a] &\geq (L-\theta)\lambda\,\mathbb{E}[h_a(\widehat{u}) - h_a(x^\star)] - \frac{\theta}{2}\mathbb{E}\|x^\star\|_2^2 \\ &\geq \frac{(L-\theta)R^2}{c_\star^2 nd^2} - \frac{\theta R^2}{d} = \frac{LR^2}{d+\Lambda}\Big[\frac{\Lambda}{c_\star^2 nd^2} - 1\Big],\end{aligned}$$

using $h_a(\widehat{u}) - h_a(x^\star) \geq \varphi_a(\widehat{u}) - \varphi_a(x^\star) - \lambda/2$ and $L - \theta = L\Lambda/(d+\Lambda)$. If $\Lambda \geq 2c_\star^2 nd^2$, then $\Lambda \geq d$, so $d + \Lambda \leq 2\Lambda$ and the bracket is at least $\Lambda/(2c_\star^2 nd^2)$, whence the expected error is at least $LR^2/(4c_\star^2 nd^2)$. With $n = 32N$, $d = 32N+1 \leq 33N$ and $32 \cdot 33^2 < 2^{16}$, the condition holds when $\Lambda \geq 2^{51}N^3$ and the bound is at least $LR^2/(2^{52}N^3)$. The reflection to the simplex preserves the quadratic part. (Padding with curvature θ instead of L would not give μ_1-strong convexity in ℓ_1.) □

D.5.3 Compact product kernels: proof of Proposition D.8

Proof of Proposition D.8. As in Section D.5.2, n denotes the number of shifted coordinates and $d = n+1$ the ambient dimension, and $c_\star = 2^{17}$.

Smoothing. Let $\rho > 0$, let $\mathcal{S}_{\rho,k}$ be the compact product smoothing of Lemma D.3 on $\mathbb{R}^d$, and put $H_{k,\nu} = 2^{1-\nu}(k\sqrt{10})^{s-1}$. For a convex $\varphi : \mathbb{R}^d \to \mathbb{R}$ that is 1-Lipschitz in ℓ_2 and in ℓ_∞, the function $\mathcal{S}_{\rho,k}\varphi$ is convex and C^{k+1}, and by (30) and (31),

$$\|D^k\mathcal{S}_{\rho,k}\varphi(x) - D^k\mathcal{S}_{\rho,k}\varphi(y)\| \leq H_{k,\nu}\rho^{-(s-1)}\|x-y\|_2^\nu, \qquad 0 \leq \mathcal{S}_{\rho,k}\varphi - \varphi \leq \rho. \tag{39}$$

Moreover $\mathcal{S}_{\rho,k}\varphi = \varphi * \chi_h^{*k}$, where the density χ_h^{*k} is supported in $[-\rho,\rho]^d$ and its derivatives up to order k are integrable (each derivative can be put on a different factor). Hence the value and the derivatives up to order k of $\mathcal{S}_{\rho,k}\varphi$ at x are integrals of $\varphi(x-w)$, $w \in [-\rho,\rho]^d$, against fixed kernels.

Locality. Let $\varphi_a(x) = \max_{0\leq i\leq n}(x_i - a_i)$, which is 1-Lipschitz in ℓ_2 and in ℓ_∞. For a query x let $m_x = \max_i(x_i - a_i)$ and $I_x = \{i : x_i - a_i \geq m_x - 2\rho\}$. For $w \in [-\rho,\rho]^d$, $\varphi_a(x-w) = \max_{i\in I_x}(x_i - w_i - a_i)$: an excluded coordinate is below $m_x - \rho$ after the perturbation, while a coordinate attaining m_x stays at least $m_x - \rho$. Hence the order-k answer for $\mathcal{S}_{\rho,k}\varphi_a$ at x is determined by x, I_x and the values a_i for $i \in I_x$. We pass to the stronger oracle that returns I_x, m_x and the exact a_i for $i \in I_x$; the other parameters then satisfy $a_i > x_i - m_x + 2\rho$.

Exponential family. Let $a_0 = 0$ and $a_1, \ldots, a_n$ be independent exponential with mean $\delta = R/(n+1)$, and set $\rho = \delta/(c_\star n)$, $n \geq 32N$. By the argument of Section D.5.2, conditionally on any extended history the unrevealed parameters are independent, with $a_i = b_i + E_i$, where b_i is the maximum of 0 and the past thresholds $x_i - m_x + 2\rho$ and E_i is exponential with mean δ. The expected number of newly revealed parameters at a query is at most $1 + n(1 - e^{-2\rho/\delta}) \leq 1 + 2n\rho/\delta \leq 2$: conditionally on the maximizer and on m_x, and after the truncation from below, any other unrevealed score $x_i - b_i - E_i$ falls in the window of length 2ρ below the maximum with probability at most $1 - e^{-2\rho/\delta}$. Hence after N queries $\mathbb{E}\#\{i \text{ unrevealed} : b_i \leq \delta/4\} \geq n(1 - e^{-1/4}) - 2N \geq n/8$.

Residual uncertainty. The minimizer of φ_a over $B_1^d(R)$, the ℓ_1-distance bound (38), and the sensitivity of $x_i^\star$ to a_i are as in Section D.5.2 (the root $t_*(a_i)$ is continuous, piecewise linear with slope $1/\sigma \leq 1/2$, where $\sigma \geq 2$ is the number of active coordinates); with the events A_1, A_2 of that proof,

Algorithm 3 SP-Cut: deep cuts at approximate Steiner points

Input: known polytope $K \subseteq c + B_1^n(R)$; subgradient oracle with $\|g\|_2 \le G$; cut depth $\varepsilon > 0$; selector accuracy $\eta = \varepsilon/(2G)$

$C_0 \leftarrow K$
for $t = 1, 2, \dots$ **do**
  $x_t \leftarrow$ a point of C_{t-1} with $\|x_t - \mathrm{st}(C_{t-1})\|_2 \le \eta$ ▷ Section C.3; $x_1 = 0$ if $K = B_1^n(R)$
  query $g_t \in \partial f(x_t)$
  $C_t \leftarrow C_{t-1} \cap \{y : \langle g_t, y - x_t\rangle \le -\varepsilon\}$
  if $C_t = \varnothing$ **then**
    solve the linear program $\widehat{\varepsilon} = \min_{\lambda \in \Delta_t} \{\sum_{j \le t} \lambda_j \langle g_j, x_j\rangle + h_K(-\sum_{j \le t} \lambda_j g_j)\}$ with minimizer λ
    return $\widehat{x} = \sum_{j \le t} \lambda_j x_j$ and the certificate $\widehat{\varepsilon}$

$\mathbb{E}|\widehat{x}_i - X_i^\star| \ge \delta/2048$ for every unrevealed i with $b_i \le \delta/4$, so $\mathbb{E}\|\widehat{x} - x^\star\|_1 \ge n\delta/16384$ and $\mathbb{E}[\varphi_a(\widehat{x}) - \min_K \varphi_a] \ge \delta/65536$.

The bound. Let $f_a = \frac{L_{k,\nu}\rho^{s-1}}{H_{k,\nu}} \mathcal{S}_{\rho,k}\varphi_a$; by (39) it is convex, C^{k+1} and satisfies (12) on $\mathbb{R}^d$. By the second inequality in (39), $\mathbb{E}[\mathcal{S}_{\rho,k}\varphi_a(\widehat{x}) - \min_K \mathcal{S}_{\rho,k}\varphi_a] \ge \mathbb{E}[\varphi_a(\widehat{x}) - \min_K \varphi_a] - \rho \ge \delta/65536 - \delta/(c_\star n) \ge \delta/c_\star$, and multiplying by $L_{k,\nu}\rho^{s-1}/H_{k,\nu}$ with $\delta = R/(n+1)$ and $\rho = \delta/(c_\star n)$ gives

$$\mathbb{E}[f_a(\widehat{x}) - \min_K f_a] \;\ge\; \frac{L_{k,\nu}R^s}{c_\star^s H_{k,\nu}\, n^{s-1}(n+1)^s} \;\ge\; \frac{L_{k,\nu}R^s}{c_\star^s H_{k,\nu} 32^{s-1}33^s}\, N^{-(2s-1)} \qquad (n = 32N),$$

that is, $c'_{k,\nu} = (c_\star^s H_{k,\nu} 32^{s-1}33^s)^{-1}$. The passage from the average over a (for each fixed seed) to a fixed function and the extension to larger dimensions (the function depends on the first $32N+1$ coordinates) are as in Section D.5.2. So is the simplex, through $\widetilde{f}_a(u) = f_a(-u)$: the jet at u is the jet at $-u$ with the factor $(-1)^j$ on the order-j derivative, $-x^\star \in \Delta_d(R)$, and $\mathcal{S}_{\rho,k}\varphi_a(-u) - \min_{v \in \Delta_d(R)} \mathcal{S}_{\rho,k}\varphi_a(-v) \ge \varphi_a(-u) - \varphi_a(x^\star) - \rho$. □

E Nonsmooth, Hölder and higher-order objectives

This section treats the minimization classes of Table 1 other than the smooth one with a first-order oracle: nonsmooth objectives with deep cuts (Section E.1), Hölder continuous gradients (Section E.2), oracles of higher order (Section E.3), and gradient oracles without function values (Sections E.4 and E.5). As before, $K \subseteq c + B_1^n(R)$ is a known polytope, all queries are feasible, and b_N, c_n, P_N, $\eta_N = P_N/(4N)$ and $\delta_N = 4P_N/N$ are as in Section 3.

E.1 Nonsmooth objectives: deep cuts and a certificate

Let f be convex and G-Lipschitz on K and let the oracle return, at a feasible x, one subgradient $g \in \partial f(x)$ with $\|g\|_2 \le G$; function values are not used. Algorithm 3 queries an approximate Steiner point of the current set and cuts *deep*: it keeps only the points at which the supporting hyperplane predicts an improvement of at least ε. Such cuts may remove every minimizer, but the first *empty* intersection yields, by linear-programming duality, a convex combination of the queried points whose error is below ε, certified from the stored oracle answers alone.

Proposition E.1 (SP-Cut). *Let f be convex on $K \subseteq c + B_1^n(R)$ with all returned subgradients of Euclidean norm at most G, let $N \ge 1$, and run Algorithm 3 with $\varepsilon = \varepsilon_N = 4GP_N/N = 16GR\sqrt{\mathsf{b}_N\mathsf{c}_n}/N$. The method stops after at most N queries and returns $\widehat{x} \in K$ with*

$$f(\widehat{x}) - \min_K f \;\le\; \widehat{\varepsilon} \;<\; \varepsilon_N, \qquad \text{and moreover } \widehat{\varepsilon} \le 2GR \ (\widehat{\varepsilon} \le GR \text{ if } K = B_1^n(R)).$$

Consequently, for every $0 < \varepsilon \le GR$ an ε-solution of a G-Lipschitz convex function on $B_1^n(R)$ is found with $O\big(\frac{GR}{\varepsilon}\sqrt{\log(2n)\log(2 + \frac{GR}{\varepsilon}\sqrt{\log(2n)})}\big)$ subgradient queries.

Proof. Certificate. Suppose $C_t = \varnothing$. Then $\max_{y \in K} \min_{j \le t} \langle g_j, x_j - y\rangle < \varepsilon$: the maximum is attained on the compact set K, and a point y with all $\langle g_j, x_j - y\rangle \ge \varepsilon$ would belong to C_t. The

function $(y,\lambda)\mapsto\sum_j\lambda_j\langle g_j,x_j-y\rangle$ is bilinear on the compact convex sets $K\times\Delta_t$, so the minimax theorem gives

$$\max_{y\in K}\min_{j\le t}\langle g_j,x_j-y\rangle=\min_{\lambda\in\Delta_t}\max_{y\in K}\sum_j\lambda_j\langle g_j,x_j-y\rangle$$
$$=\min_{\lambda\in\Delta_t}\Big\{\sum_j\lambda_j\langle g_j,x_j\rangle+h_K\Big(-\sum_j\lambda_j g_j\Big)\Big\}=\widehat{\varepsilon},$$

a linear program over the known polytope K (for $K=B_1^n(R)$ its last term is $R\|\sum_j\lambda_j g_j\|_\infty$). For every $y\in K$, convexity and the subgradient inequalities give $f(\widehat{x})-f(y)\le\sum_j\lambda_j(f(x_j)-f(y))\le\sum_j\lambda_j\langle g_j,x_j-y\rangle\le\widehat{\varepsilon}<\varepsilon$. Concentrating λ on $j=1$ gives $\widehat{\varepsilon}\le\max_{y\in K}\langle g_1,x_1-y\rangle\le G\cdot\mathrm{diam}_2(K)\le 2GR$, and $\widehat{\varepsilon}\le R\|g_1\|_\infty\le GR$ when $x_1=0$ and $K=B_1^n(R)$.

Budget. Suppose that $C_0,\dots,C_N$ are all nonempty. Since $\mathrm{st}(C_t)\in C_t$, the cut gives $\langle g_t,\mathrm{st}(C_t)-x_t\rangle\le-\varepsilon$, so $\|\mathrm{st}(C_t)-x_t\|_2\ge\varepsilon/G$ and, by the choice of η, $\|\mathrm{st}(C_t)-\mathrm{st}(C_{t-1})\|_2\ge\varepsilon/G-\eta=\varepsilon/(2G)$ for $t=1,\dots,N$. Summing and applying Lemma 3.1 yields $N\varepsilon/(2G)\le P_N$, which contradicts $\varepsilon=4GP_N/N$. Hence some C_t with $t\le N$ is empty, and the method stops after at most N queries; the point $\mathrm{st}(C_N)$ is used only in this argument and is never computed.

Query count. Given ε, the smallest N with $4GP_N/N\le\varepsilon$ has the stated order by Lemma E.2 below. □

The bound $\Omega(GR/N)$ of Proposition D.1(a) (and its randomized version, Proposition D.6) shows that the exponent is optimal; we do not know whether the logarithmic factor is necessary. The proof uses only $\|g_t\|_2\le G$ and $f(y)\ge f(x_t)+\langle g_t,y-x_t\rangle$ on K, so the oracle may return any subgradient satisfying this norm bound, and f need be Lipschitz only on K; the same certificate therefore also covers bounded monotone operators and convex–concave saddle problems (Section G).

Lemma E.2. *Let $0<\varepsilon\le GR$, $Q=GR/\varepsilon\ge 1$ and $A=1+\log(2n)$. If $N=\lceil 32Q\sqrt{A[1+\log(16Q\sqrt{A})]}\rceil$, then $16GR\sqrt{(1+\log N)(1+\log(2n))}/N\le\varepsilon$; in particular Algorithm 3 with depth ε stops after at most $N=O\big(Q\sqrt{\log(2n)\log(2+Q\sqrt{\log(2n)})}\big)$ queries; the same holds, with N doubled, for $\mathsf{b}_N\mathsf{c}_n$ in place of $(1+\log N)(1+\log(2n))$.*

Proof. Put $c=16Q\sqrt{A}\ge 16$ and $u=2c\sqrt{1+\log c}=32Q\sqrt{A[1+\log(16Q\sqrt{A})]}$. Since $\log(1+\log c)\le\log c$,

$$1+\log u=1+\log 2+\log c+\tfrac12\log(1+\log c)<4(1+\log c),$$

so $u/\sqrt{1+\log u}>c$. The function $t\mapsto t/\sqrt{1+\log t}$ is increasing for $t\ge 1$, so $N=\lceil u\rceil$ satisfies $N/\sqrt{1+\log N}\ge c=16GR\sqrt{A}/\varepsilon$, which is equivalent to $16GR\sqrt{(1+\log N)A}/N\le\varepsilon$. For the condition with $\mathsf{b}_N\mathsf{c}_n$, note that $\mathsf{b}_N\le 2(1+\log N)$ and $\mathsf{c}_n\le 2(1+\log(2n))$, since $1+\lceil\log_2 t\rceil\le 2(1+\log t)$ for $t\ge 2$, so it suffices that $N/\sqrt{1+\log N}\ge 2c$. This holds for $N=\lceil 2u\rceil$, because $1+\log(2u)=1+2\log 2+\log c+\frac12\log(1+\log c)\le 2.4+1.5\log c<4(1+\log c)$ and hence $2u/\sqrt{1+\log(2u)}>2c$. □

E.2 Hölder continuous gradients

We use the following consequence of Lemma 3.1. If in a run of a level method the sets $C_0\supseteq\cdots\supseteq C_N$ are nonempty and every z_t is an η_N-approximate Steiner point of C_t, then, with $d_t=\|z_t-z_{t-1}\|_2$,

$$\sum_{t=1}^N d_t\le\tfrac32 P_N,\qquad \#\{t:d_t>\delta_N\}<\tfrac{3N}{8},\qquad \#\{t:d_t\le\delta_N\}>\tfrac{N}{2}. \tag{40}$$

The first bound is (8); the other two follow as in the proof of Proposition 3.2.

Let $0\le\nu\le 1$ and let f be convex on K such that the oracle returns, at each feasible x, a subgradient $g_x\in\partial f(x)$ with

$$0\le f(y)-f(x)-\langle g_x,y-x\rangle\le\frac{L_{1,\nu}}{1+\nu}\|y-x\|_2^{1+\nu}\qquad(x,y\in K). \tag{41}$$

For $0 < \nu \le 1$ this holds with $g_x = \nabla f(x)$ whenever ∇f is ν-Hölder with constant $L_{1,\nu}$; for $\nu = 0$ it holds with $L_{1,0} = 2G$ for every G-Lipschitz convex f and any subgradient selection with $\|g_x\|_2 \le G$. We run Algorithm 1 with L replaced by 1 in the weight recursion, $a_t = \frac{1}{2}(1 + \sqrt{1 + 4A_{t-1}})$, and with the certificate (42) below in place of (6); the cuts, the selector accuracy and the acceptance rule are unchanged.

Proposition E.3 (Hölder gradients)**.** *Under (41), the level method with horizon N either certifies $\min_K f > \ell$ or returns $y_N \in K$ with*

$$f(y_N) \le \ell + \frac{24 \cdot 4^\nu}{1+\nu} \cdot \frac{L_{1,\nu} P_N^{1+\nu}}{N^{1+2\nu}} \le \ell + \frac{48\, L_{1,\nu} P_N^{1+\nu}}{N^{1+2\nu}}. \tag{42}$$

With the bracket search of Algorithm 2, its horizon rule (22) replaced by $N(w) = \min\{2^j : e_{2^j} \le w/4\}$, where e_N is the computable bound (44) below, which dominates the right-hand side of (42) and avoids real powers, and with a computable initial width of at most $4L_{1,\nu}(2R)^{1+\nu}$, an ε-solution is obtained with

$$O\Big(1 + \Big[\frac{L_{1,\nu} R^{1+\nu}}{\varepsilon} \Big\{ \log(2n) \log\Big(2 + \frac{L_{1,\nu} R^{1+\nu}}{\varepsilon} \log(2n)\Big) \Big\}^{\frac{1+\nu}{2}} \Big]^{\frac{1}{1+2\nu}} \Big)$$

queries, uniformly in $\nu \in [0,1]$, and with a budget of $T \ge 385$ queries the error is $O\big(L_{1,\nu} R^{1+\nu} [\log(2T)\log(2n)]^{(1+\nu)/2} / T^{1+2\nu}\big)$.

The exponent $1+2\nu$ interpolates between the nonsmooth rate N^{-1} of Proposition E.1 and the smooth rate N^{-3} of Proposition C.2; it matches the lower bound $\nu + (1+\nu)/p = 1+2\nu$ of Proposition D.1(b) at $p = 1$, and the randomized bound of Proposition D.8 with $k = 1$. In the matched Euclidean setting the exponent is $(1+3\nu)/2$ (Nesterov, 2015). The level solver does not use ν or $L_{1,\nu}$; only the horizon rule does.

Proof of Proposition E.3. We write $M = L_{1,\nu}$, $s = 1 + \nu$ and $\sigma = 1 + 2\nu$.

Level method. The weights satisfy $a_t = \frac{1}{2}(1 + \sqrt{1 + 4A_{t-1}})$, so $a_t^2 = A_{t-1} + a_t = B_t$. Consider an accepted trial with $A = A_{t-1}$, $a = a_t$, $B = B_t$, $x = x_t$, $z = z_{t-1}$, $z^+ = z_t$, $y = y_{t-1}$, $y^+ = y_t$, and $d = \|z^+ - z\|_2$. Since $y^+ - x = (a/B)(z^+ - z)$, the remainder bound (41) at x gives $Bf(y^+) \le Bf(x) + a\,\langle g_x, z^+ - z\rangle + B\frac{M}{s}(a/B)^s d^s$. As in Lemma C.1, convexity and $z^+ \in C_t$ (the level cut) bound the first two terms by $Af(y) + a\ell$. Hence, using $B = a^2$,

$$B[f(y^+) - \ell] \le A[f(y) - \ell] + \frac{M}{s}\frac{a^s}{B^\nu} d^s = A[f(y) - \ell] + \frac{M}{s} a^{1-\nu} d^s. \tag{43}$$

Number the accepted updates $j = 1, \dots, k$. Then $a_1 = 1$, $A_j = a_j^2$, and $a_{j+1} = \frac{1}{2}(1 + \sqrt{1 + 4a_j^2}) \le a_j + 1$ because $1 + 4a_j^2 \le (2a_j + 1)^2$; hence $a_j \le j \le N$. Also $\sqrt{A_j} = a_j \ge \sqrt{A_{j-1}} + \frac{1}{2}$, so $A_k \ge k^2/4 \ge N^2/16$ by (40). Rejected trials do not change (A, y). Telescoping (43) from $A_0 = 0$ and using $d_t \le \delta_N$ on accepted trials and (40),

$$A_k[f(y_N) - \ell] \le \frac{M}{s} N^{1-\nu} \sum_{t\ \mathrm{acc.}} d_t^{1+\nu} \le \frac{M}{s} N^{1-\nu} \delta_N^\nu \sum_{t=1}^{N} d_t \le \frac{3 \cdot 4^\nu M}{2s} P_N^s N^{1-2\nu}$$

(for $\nu = 0$ read $\delta_N^0 = 1$). Dividing by $A_k \ge N^2/16$ gives the first bound in (42); the second follows because $\nu \mapsto 4^\nu/(1+\nu)$ is increasing on $[0,1]$ with value 2 at $\nu = 1$. The certificate branch is as in Proposition 3.2.

Finite parameters. For $u > 0$ let $D_\nu(u) = 2^{\lceil \nu \lceil \log_2 u \rceil \rceil}$, computed by binary scaling and integer comparisons. Since $\nu\lceil \log_2 u \rceil \in [\nu \log_2 u, \nu \log_2 u + \nu]$, we have $u^\nu \le D_\nu(u) < 2^{\nu+1} u^\nu \le 4u^\nu$, and D_ν is nondecreasing. Put

$$e_N = \frac{48 M P_N}{N} D_\nu\Big(\frac{P_N}{N^2}\Big), \qquad \text{so that} \qquad \frac{48 M P_N^s}{N^\sigma} \le e_N \le \frac{192 M P_N^s}{N^\sigma}, \tag{44}$$

because $(P_N/N)(P_N/N^2)^\nu = P_N^s/N^\sigma$. Thus a nonempty run returns $f(y_N) \le \ell + e_N$, and the test $e_N \le w/4$ needs only finite arithmetic.

Bracket. Query $(f(x_0), g_0)$ at $x_0 \in K$ and let $v \in \arg\min_{u\in K} \langle g_0, u\rangle$. Then $\underline{f} = f(x_0) + \langle g_0, v - x_0\rangle \le \min_K f$ by the lower bound in (41), and by its upper bound $f(v) \le \underline{f} + \frac{M}{s}\|v - x_0\|_2^s \le \underline{f} + \frac{M}{s}(2R)^s \le \underline{f} + W_0$ with the computable width $W_0 = M(2R)D_\nu(2R) \le 4M(2R)^s$ (for $K = B_1^n(R)$, $x_0 = 0$, one may take $W_0 = MRD_\nu(R) \le 4MR^s$). If $\varepsilon \ge W_0$, the method returns v.

Bracket search. With width $w > \varepsilon$ take $\ell = (\underline{f} + \overline{f})/2$ and $N(w) = \min\{2^j : e_{2^j} \le w/4\}$, which exists since $e_N \to 0$ along powers of two. An empty run raises $\underline{f}$ to ℓ (width $w/2$); a nonempty run stores y_N and lowers $\overline{f}$ to $\ell + w/4$ (width $3w/4$). Let $w_0 > \cdots > w_{m-1} > \varepsilon$ be the widths, $N_j = N(w_j) \le N_{m-1} \le N(\varepsilon)$. All $N_j > 1$, since $P_1 \ge 4R$ gives $e_1 \ge 48M(4R)^s \ge 12 \cdot 4M(2R)^s \ge 12W_0 > w_j/4$. Since $N_j/2$ violates the rule and $P_{N_j/2} \le P_{N_j}$,

$$\frac{w_j}{4} < e_{N_j/2} \le \frac{192 \cdot 2^\sigma M P_{N_j}^s}{N_j^\sigma} \le \frac{1536\, M P_{N_j}^s}{N_j^\sigma}, \qquad N_{m-1}^\sigma \ge \frac{192\, M P_{N_{m-1}}^s}{w_{m-1}},$$

where P_N is the path budget at horizon N and we used $2^\sigma \le 8$. As $P_{N_j} \le P_{N_{m-1}}$ and $w_j \ge (4/3)^{m-1-j} w_{m-1}$, dividing gives $N_j < 32^{1/\sigma}(3/4)^{(m-1-j)/\sigma} N_{m-1}$, hence, including the initial query,

$$N_{\text{total}} \le 1 + \frac{32^{1/\sigma}}{1 - (3/4)^{1/\sigma}} N(\varepsilon) < 1 + 384 N(\varepsilon),$$

uniformly for $1 \le \sigma \le 3$, since $32^{1/\sigma} \le 32$ and $(3/4)^{1/\sigma} \le (3/4)^{1/3} < 11/12$.

Order of $N(\varepsilon)$. Let $Q = MR^s/\varepsilon$; in the active branch $\varepsilon < W_0 \le 4M(2R)^s$, so Q is bounded below by a constant. Put $H = 768 \cdot 4^s Q \mathsf{c}_n^{s/2} \ge 2$ and let N be the smallest power of two with $N \ge N_0 := [16H(1 + \log_2 H)^{s/2}]^{1/\sigma}$. Then $N < 2N_0$, so $\log_2 N < 5 + \log_2 H + \log_2(1 + \log_2 H) \le 5 + 2\log_2 H$ and $\mathsf{b}_N \le 7(1 + \log_2 H)$; hence $N^\sigma \ge 16H(1 + \log_2 H)^{s/2} \ge H\mathsf{b}_N^{s/2}$ because $7^{s/2} \le 7 < 16$; this is equivalent to $192M(4R)^s(\mathsf{b}_N \mathsf{c}_n)^{s/2}/N^\sigma \le \varepsilon/4$ and gives $e_N \le \varepsilon/4$. Therefore $N(\varepsilon) \le N < 2N_0$, and since $\mathsf{c}_n = O(\log(2n))$ and $\log H = O(\log(2 + Q\log(2n)))$, this is the stated order, with a constant that can be chosen uniformly in $\nu \in [0, 1]$.

Fixed horizon. For $T \ge 385$ let Q_T be the largest power of two with $384Q_T \le T - 1$ and $\varepsilon_T = 4e_{Q_T}$. If $\varepsilon_T \ge W_0$ the method returns the initial v. Otherwise $N(\varepsilon_T) \le Q_T$, the count is at most $1 + 384Q_T \le T$, and the error is at most $\min\{W_0, 4e_{Q_T}\} \le \min\{4M(2R)^s, 768MP_{Q_T}^s/Q_T^\sigma\}$. Since $Q_T > (T-1)/768$ and $\mathsf{b}_{Q_T} \le \mathsf{b}_T$, this is $O(MR^s[\log(2T)\log(2n)]^{s/2}/T^\sigma)$. □

E.3 Oracles of higher order

Let $k \ge 2$, $0 \le \nu \le 1$, $s = k + \nu$, and let f be convex and C^k on a neighborhood of K with ν-Hölder kth derivative in the sense of (12) with constant $L_{k,\nu}$. The oracle returns the Taylor polynomial $T_k(x; u) = \sum_{j\le k} \frac{1}{j!} D^j f(x)[(u - x)^j]$ at a feasible x. Put

$$c_{k,\nu} = \prod_{j=1}^{k-1} \frac{1}{\nu + j} = \frac{\Gamma(\nu + 1)}{\Gamma(k + \nu)}, \qquad m_x(u) = T_k(x; u) + \frac{2c_{k,\nu}L_{k,\nu}}{s}\|u - x\|_2^s, \tag{45}$$

so that $|f(u) - T_k(x; u)| \le \frac{c_{k,\nu}L_{k,\nu}}{s}\|u - x\|_2^s$ and $\|\nabla f(u) - \nabla_u T_k(x; u)\|_2 \le c_{k,\nu}L_{k,\nu}\|u - x\|_2^{s-1}$ on K. The model m_x is known and C^1 but need not be convex; it is never used to form a cut. A trial of SP-Taylor (Algorithm 4) makes two queries. The first is the jet at the mixed point $x = (Ay + az)/B$; from it a finite search over the known model computes a point $v \in K$ with the *checkable* approximate stationarity $\max_{u\in K} \langle \nabla m_x(v), v - u\rangle \le \eta_{\text{mod}}$. The second is the jet at v, of which only $f(v)$ and $\nabla f(v)$ are used; they give the level cut $\{u : f(v) + \langle \nabla f(v), u - v\rangle \le \ell\}$, a supporting half-space of f. If the approximate Steiner point z^+ of the cut set is within δ_N of z, the state becomes $(y, A) \leftarrow (v, B)$ with the integer weights $a_j = j^k$; otherwise the cut is kept and the state is not updated.

Proposition E.4 (Higher-order oracles). *Let $\Xi_k = 3^{k+2}8^k(k + 1)^{k+1}$. With the parameters above, the level method of SP-Taylor makes at most $2N$ order-k queries and either certifies $\min_K f > \ell$ or returns $y \in K$ with*

$$f(y) \le \ell + e_N, \qquad e_N \text{ as in (50)}, \qquad 2\Xi_k c_{k,\nu}L_{k,\nu}\frac{P_N^s}{N^{2s-1}} \le e_N \le 8\Xi_k c_{k,\nu}L_{k,\nu}\frac{P_N^s}{N^{2s-1}}. \tag{46}$$

Algorithm 4 SP-Taylor: level method with a regularized Taylor model (level ℓ, horizon N)

Input: known polytope $K \subseteq c + B_1^n(R)$; order-k oracle; $M = L_{k,\nu}$; level ℓ; horizon N; geometric accuracy η_N, threshold δ_N; model accuracy $\eta_{\mathrm{mod}} = e_N/(2N)$, where the certificate e_N is defined in (50) below

$C \leftarrow K$; $z \leftarrow$ an η_N-approximate Steiner point of K; $y \leftarrow$ any point of K; $A \leftarrow 0$; $j \leftarrow 1$
for $t = 1, \dots, N$ **do**
  $a \leftarrow j^k$; $B \leftarrow A + a$; $x \leftarrow (Ay + az)/B$
  query the jet at x; form m_x from (45)
  find $v \in K$ with $\max_{u \in K} \langle \nabla m_x(v), v - u \rangle \le \eta_{\mathrm{mod}}$ (finite search, Lemma E.6)
  query the jet at v; use $f(v)$ and $g = \nabla f(v)$
  $C \leftarrow C \cap \{u : f(v) + \langle g, u - v \rangle \le \ell\}$; **if** $C = \varnothing$ **return** "$\min_K f > \ell$"
  $z^+ \leftarrow$ an η_N-approximate Steiner point of C
  if $\|z^+ - z\|_2 \le \delta_N$ **then** $y \leftarrow v$; $A \leftarrow B$; $j \leftarrow j + 1$ ▷ accept
  $z \leftarrow z^+$
return y with the certificate $f(y) \le \ell + e_N$

The bracket search, with the initial width obtained from one jet at x_0 as in the proof below, returns an ε-solution after at most $1 + 32(2k+1)N(\varepsilon)$ queries, where $N(\varepsilon) = \min\{2^j : e_{2^j} \le \varepsilon/4\} = O_k\big(1 + [\frac{L_{k,\nu}R^s}{\varepsilon}\{\log(2n)\log(2 + \frac{L_{k,\nu}R^s}{\varepsilon}\log(2n))\}^{s/2}]^{1/(2s-1)}\big)$ uniformly in $\nu \in [0,1]$, and with a budget of $T \ge 32(2k+1)+1$ queries the error is

$$O_k\Big(\frac{L_{k,\nu}R^{k+\nu}\,[\log(2T)\log(2n)]^{(k+\nu)/2}}{T^{2(k+\nu)-1}}\Big). \tag{47}$$

For $0 < \nu \le 1$ the exponent $2(k+\nu)-1$ equals the lower bound $s(1+1/p)-1$ of Proposition D.4 at $p = 1$ and the randomized bound of Proposition D.8. For a Hessian with Lipschitz constant $L_{2,1}$ the rate is $\widetilde{O}(L_{2,1}R^3/N^5)$, compared with the Euclidean $N^{-7/2}$ (Arjevani et al., 2019; Gasnikov et al., 2019; Kovalev & Gasnikov, 2022). Regularized Taylor models and the mixing of two states are also used in non-Euclidean higher-order acceleration (Contreras et al., 2025); here no inner problem is solved to optimality against an unknown function, the model may be nonconvex, and the geometry enters only through Lemma 3.1.

In this subsection $M = L_{k,\nu}$, $s = k + \nu$ and $\sigma = 2s - 1$.

Lemma E.5 (Taylor remainders)**.** *Let $k \ge 2$, $0 \le \nu \le 1$, and let f be C^k on a convex set containing x, u with (12) on the segment $[x,u]$. Then $|f(u) - T_k(x;u)| \le \frac{c_{k,\nu}M}{s}\|u - x\|_2^s$ and $\|\nabla f(u) - \nabla_u T_k(x;u)\|_2 \le c_{k,\nu}M\|u - x\|_2^{s-1}$.*

Proof. With $h = u - x$, the integral form of the remainder gives $f(u) - T_k(x;u) = \int_0^1 \frac{(1-t)^{k-1}}{(k-1)!}\big(D^k f(x + th) - D^k f(x)\big)[h^k]\,dt$, whose absolute value is at most $M\|h\|_2^s \int_0^1 \frac{(1-t)^{k-1}t^\nu}{(k-1)!}dt = M\|h\|_2^s \frac{\Gamma(\nu+1)}{\Gamma(k+\nu+1)} = \frac{c_{k,\nu}M}{s}\|h\|_2^s$. Similarly, $\nabla f(u) - \nabla_u T_k(x;u) = \int_0^1 \frac{(1-t)^{k-2}}{(k-2)!}\big(D^k f(x + th) - D^k f(x)\big)[h^{k-1}, \cdot\,]\,dt$, of norm at most $M\|h\|_2^{s-1}\frac{\Gamma(\nu+1)}{\Gamma(k+\nu)} = c_{k,\nu}M\|h\|_2^{s-1}$. Here we used the Beta integral $\int_0^1 (1-t)^{a-1}t^\nu dt = \Gamma(\nu+1)\Gamma(a)/\Gamma(a+\nu+1)$ and the recursion $\Gamma(\nu+1)/\Gamma(k+\nu) = \prod_{j=1}^{k-1}(\nu+j)^{-1}$; the argument also covers $\nu = 0$. □

Lemma E.6 (Finite model search)**.** *For every $\eta > 0$ a point $v \in K$ with $\mathcal{G}(v) := \max_{u \in K}\langle \nabla m_x(v), v - u\rangle \le \eta$ can be found by a finite enumeration of rational convex combinations of the vertices of K, using only certified approximations of known scalar powers.*

Proof. The model m_x is C^1 because $s \ge 2$, so it attains its minimum over the compact set K at some $v^\star$. Optimality along every segment $[v^\star, u] \subseteq K$ gives $\mathcal{G}(v^\star) = 0$, and $\mathcal{G}$ is continuous. Rational convex combinations of the vertices are dense in K, and we enumerate them. At each candidate v we compute a vector h with $\|h - \nabla m_x(v)\|_2 \le \eta/(8R')$, where R' bounds the Euclidean diameter of K (here $R' = 2R$), compute $\widehat{\mathcal{G}} = \max_{u \in K}\langle h, v - u\rangle$ by a linear program, and accept v if $\widehat{\mathcal{G}} \le \eta/2$.

Since $\max_{u\in K}\|v-u\|_2 \le R'$, the two objectives differ by at most $\eta/8$, so $\mathcal{G}(v) \le \widehat{\mathcal{G}} + \eta/8 \le \eta$. Conversely, for all candidates close enough to $v^\star$, $\mathcal{G}(v) < \eta/8$, hence $\widehat{\mathcal{G}} < \eta/4 < \eta/2$, so the enumeration stops. For rational ν the powers in m_x are algebraic; for a general known $\nu \in [0,1]$, certified rational enclosures of $\|v-x\|_2^{s-2}$ suffice because the test has positive slack. □

Lemma E.7 (One trial)**.** *Let $A \ge 0$, $a > 0$, $B = A + a$, $y, z \in K$, $x = (Ay + az)/B$, let $v \in K$ satisfy $\mathcal{G}(v) \le \eta$, let $g = \nabla f(v)$, and let $z^+ \in K$ satisfy $f(v) + \langle g, z^+ - v\rangle \le \ell$. Then, with $r = v - x$ and $d = z^+ - z$,*

$$B[f(v) - \ell] \le A[f(y) - \ell] + B\eta + 3^{k+1} c_{k,\nu} M \frac{a^s}{B^{s-1}} \|d\|_2^s. \tag{48}$$

Proof. Put $W = Bv - Ay - az^+ = Br - ad$. Convexity at v, $Af(y) \ge Af(v) + A\langle g, y - v\rangle$, and the cut, $a\ell \ge af(v) + a\langle g, z^+ - v\rangle$, give $B[f(v) - \ell] - A[f(y) - \ell] = Bf(v) - Af(y) - a\ell \le \langle g, W\rangle$. Since $(Ay + az^+)/B \in K$ and $W = B(v - (Ay + az^+)/B)$, the model condition gives $\langle \nabla m_x(v), W\rangle \le B\eta$. By Lemma E.5 and (45), $g - \nabla m_x(v) = e - 2c_{k,\nu}M\|r\|_2^{s-2} r$ with $\|e\|_2 \le c_{k,\nu} M \|r\|_2^{s-1}$ (for $r = 0$ the second term is 0). Hence

$$\begin{aligned}\langle g, W\rangle &\le B\eta - 2c_{k,\nu}M\|r\|_2^{s-2}\langle r, W\rangle + \langle e, W\rangle \\ &\le B\eta - 2c_{k,\nu}M\|r\|_2^{s-2}(B\|r\|_2^2 - a\|r\|_2\|d\|_2) + c_{k,\nu}M\|r\|_2^{s-1}(B\|r\|_2 + a\|d\|_2),\end{aligned}$$

that is, $\langle g, W\rangle \le B\eta - c_{k,\nu}MB\|r\|_2^s + 3c_{k,\nu}Ma\|r\|_2^{s-1}\|d\|_2$. For $\tau \ge 0$ the maximum of $-B\tau^s + 3a\|d\|_2\tau^{s-1}$ is attained at $\tau = 3(s-1)a\|d\|_2/(sB)$ (it is 0 if $d = 0$) and equals $3^s(s-1)^{s-1}s^{-s}a^s\|d\|_2^s/B^{s-1} \le 3^{k+1}a^s\|d\|_2^s/B^{s-1}$. □

Proof of Proposition E.4. Level method. Feasibility and the certificate branch are as in Proposition 3.2, since every cut is a supporting half-space of f and therefore contains every u with $f(u) \le \ell$. Suppose all N sets are nonempty. By (40) at least $m > N/2$ trials are accepted. The accepted weights are $a_j = j^k$ and $A_j = \sum_{i\le j} i^k \ge j^{k+1}/(k+1)$, so

$$\begin{aligned}\frac{a_j^s}{A_j^{s-1}} &\le (k+1)^{s-1} j^{ks-(k+1)(s-1)} = (k+1)^{s-1} j^{k+1-s} \\ &\le (k+1)^k N^{1-\nu}, \qquad A_m \ge \frac{N^{k+1}}{2^{k+1}(k+1)}.\end{aligned}$$

Telescoping (48) over accepted trials (rejected ones do not change (A, y)), using $B_t \le A_m$ for the model terms, $\|d_t\|_2 \le \delta_N$ on accepted trials and (40),

$$A_m[f(y) - \ell] \le NA_m\eta_{\mathrm{mod}} + 3^{k+1}c_{k,\nu}M(k+1)^k N^{1-\nu}\Big(\frac{4P_N}{N}\Big)^{s-1}\frac{3P_N}{2},$$

and dividing by A_m gives, with $4^{s-1} \le 4^k$ and $2^{k+1}\cdot\frac{3}{2}\cdot 3^{k+1} \le 3^{k+2}2^k$,

$$f(y) - \ell \le \Xi_k c_{k,\nu} M \frac{P_N^s}{N^{2s-1}} + N\eta_{\mathrm{mod}}. \tag{49}$$

Define, with D_ν from Section E.2,

$$e_N = 2\Xi_k c_{k,\nu} M \frac{P_N^k}{N^{2k-1}} D_\nu\Big(\frac{P_N}{N^2}\Big), \qquad \eta_{\mathrm{mod}} = \frac{e_N}{2N}, \tag{50}$$

so that $2\Xi_k c_{k,\nu} M P_N^s/N^\sigma \le e_N \le 8\Xi_k c_{k,\nu} M P_N^s/N^\sigma$; both terms of (49) are at most $e_N/2$, which proves (46). Each trial makes two queries, so the level method makes at most $2N$.

Bracket from one jet. Query the jet at $x_0 \in K$ and put $q_-(u) = T_k(x_0; u) - \frac{c_{k,\nu}M}{s}\|u - x_0\|_2^s \le f(u)$ on K (Lemma E.5). Let $E = \frac{c_{k,\nu}M}{s}(2R)^k D_\nu(2R)$, so $\frac{c_{k,\nu}M}{s}(2R)^s \le E \le \frac{4c_{k,\nu}M}{s}(2R)^s$ (for $K = B_1^n(R)$, $x_0 = 0$, use R in place of $2R$). By a finite grid search over rational convex combinations of the vertices (with the known Lipschitz constant of q_- on K, obtained from the coefficients of $T_k(x_0;\cdot)$) one finds $\underline{q} \le \min_K q_-$ and $v \in K$ with $q_-(v) \le \underline{q} + E$. Then $\underline{q} \le \min_K f \le f(v) \le q_-(v) + \frac{2c_{k,\nu}M}{s}\|v - x_0\|_2^s \le \underline{q} + 3E$, so the initial width is $W_0 = 3E \le \frac{12c_{k,\nu}M}{s}(2R)^s$, without any query at v.

Bracket search. We proceed as in Section E.2 with $N(w) = \min\{2^j : e_{2^j} \le w/4\}$; every phase uses at most $2N(w)$ queries. All $N_j > 1$ because $e_1 \ge 2\Xi_k c_{k,\nu} M(4R)^s > W_0/4$. Since $N_j/2$ violates the rule, $w_j/4 < e_{N_j/2} \le 8\Xi_k c_{k,\nu} M\, 2^\sigma P^s_{N_j}/N_j^\sigma$, while $N_{m-1}^\sigma \ge 8\Xi_k c_{k,\nu} M P^s_{N_{m-1}}/w_{m-1}$; hence $N_j < 2 \cdot 4^{1/\sigma}(3/4)^{(m-1-j)/\sigma} N_{m-1}$ and

$$N_{\text{total}} \le 1 + 2\sum_j N_j \le 1 + \frac{4 \cdot 4^{1/\sigma}}{1-(3/4)^{1/\sigma}} N(\varepsilon) \le 1 + 32(2k+1)N(\varepsilon),$$

using $\sigma \ge 3$, $4^{1/\sigma} < 2$ and $(3/4)^{1/\sigma} \le 1 - 1/(4\sigma)$ (Bernoulli's inequality), so that the fraction is at most $8 \cdot 4\sigma = 32(2s-1) \le 32(2k+1)$. The order of $N(\varepsilon)$ follows from $e_N \le 8\Xi_k c_{k,\nu} M P^s_N/N^\sigma$ as in Section E.2, with $Q = MR^s/\varepsilon$ bounded below by a constant depending on k in the active branch; the numerical constants 7 and 16 used there for $s \le 2$ are replaced by constants depending on k.

Fixed horizon. For $T \ge 32(2k+1)+1$ let Q_T be the largest power of two with $32(2k+1)Q_T \le T-1$ and $\varepsilon_T = 4e_{Q_T}$. If $\varepsilon_T \ge W_0$ the method returns the initial v, and otherwise it runs the bracket search with accuracy ε_T. The count is at most T and the error at most $\min\{W_0, 4e_{Q_T}\} = O_k(MR^s[\log(2T)\log(2n)]^{s/2}/T^{2s-1})$ since $Q_T > (T-1)/(64(2k+1))$ and $\mathsf{b}_{Q_T} \le \mathsf{b}_T$. □

E.4 Gradient oracles: proximal deep cuts

Suppose now that a query at $x \in K$ returns $\nabla f(x)$ only, for a convex f with L-Lipschitz gradient on $\mathbb{R}^n$. Level cuts are unavailable, and deep cuts $\{u : \langle \nabla f(x), x-u\rangle \ge \delta\}$ alone give the nonsmooth rate. The method SP-Prox (Algorithm 5) cuts deep at approximate *proximal points* w of f, taken with respect to points x of the segment $[y, z_t] \subseteq K$ between the incumbent y and the current approximate Steiner point z_t. The analysis uses the Moreau envelope

$$\begin{aligned} E(x) &= \min_{u\in K} f_x(u), & p(x) &= \arg\min_{u\in K} f_x(u), \\ f_x(u) &= f(u) + \frac{L}{2}\|u-x\|_2^2, & \nabla E(x) &= L\big(x - p(x)\big). \end{aligned} \tag{51}$$

The envelope is never evaluated. The method calls two gradient-only subroutines (Lemmas E.9 and E.10 below). $\textsc{Prox}(x,k)$ runs k projected-gradient steps with step $1/(2L)$ on the strongly convex function $f(u) + \frac{L}{2}\|u-x\|_2^2$ starting from $u_0 = x$, queries the gradient at the last point $w = u_k$, and returns w together with $g = \nabla f(w)$, a vector v of the normal cone $N_K(w) = \{u' : \langle u', u-w\rangle \le 0\ \forall u \in K\}$ read off the last projection, and the residual $r = g + L(w-x) + v$. The projected-gradient map is a contraction with factor $\frac{1}{2}$, so $\|r\|_2 \le 6LR\,2^{-k}$ after $k+1$ queries. $\textsc{Search}(y,z,\eta)$ bisects the segment $[y,z] \subseteq K$ on the sign of the approximate directional derivative $\langle L(x-w), z-y\rangle$ of E and returns $x \in [y,z]$ with the output of $\textsc{Prox}$ at x, such that

$$E(x) \le E(y) + \eta, \qquad \langle \nabla E(x), x-z\rangle \le \eta, \qquad \|L(x-w) - \nabla E(x)\|_2 \le \rho := \frac{\eta}{4R}, \tag{52}$$

using at most $(\bar h + 1)(k+1)$ gradients, $\bar h = \min\{h \ge 0 : 4LR^2 \cdot 2^{-h} \le \eta\}$, $k = \min\{j \ge 1 : 6LR \cdot 2^{-j} \le \rho\}$. Function values appear only in the analysis; neither routine queries them.

Proposition E.8 (Gradient-only oracle)**.** *Let f be convex with L-Lipschitz gradient on $\mathbb{R}^n$, $L > 0$, and let $K \subseteq c + B_1^n(R)$ be a known polytope. For every $\varepsilon > 0$, Algorithm 5 makes feasible gradient queries only, never uses a function value, and returns $y \in K$ with $f(y) - \min_K f \le \varepsilon$. Every stage ends with an empty cut within its $T = 10m$ trials, and the total number of gradient queries is at most*

$$1 + 60\, m_J B_J (k_J + 1) = O\Big(1 + \Big[\frac{LR^2 \mathsf{c}_n}{\varepsilon}\Big]^{1/3} \Big[\log\Big(2 + \frac{LR^2\mathsf{c}_n}{\varepsilon}\Big)\Big]^{7/3}\Big), \tag{53}$$

where $\delta_J = U_0 2^{-J}$ and m_J are the parameters of the last stage, $B_J = 1 + \lceil \log_2(64LR^2 m_J/\delta_J)\rceil$ bounds the number of $\textsc{Prox}$ *calls of one trial and $k_J + 1$, $k_J = \max\{1, \lceil \log_2(384LR^2 m_J/\delta_J)\rceil\}$, the number of gradients of one* $\textsc{Prox}$ *call, so that $B_J(k_J+1)$ bounds the gradients of one trial. For $\varepsilon \ge U_0$ the point y of the first line suffices and the count is 1.*

Unlike in SP-Level, a large approximate proximal step guarantees a decrease of f that the method does not observe, and a small one is compatible with a nonempty deep cut only if the Steiner point

Algorithm 5 SP-Prox: proximal deep cuts at approximate Steiner points (gradient oracle)

Input: known polytope $K \subseteq c + B_1^n(R)$; gradient oracle of an L-smooth convex f; accuracy $\varepsilon > 0$

query $g_a = \nabla f(a)$ at a point $a \in K$; $y \leftarrow \mathrm{proj}_K(a - g_a/L)$; $U_0 \leftarrow 2LR^2$ ▷ $a = 0$, $U_0 = LR^2/2$ if $K = B_1^n(R)$
for $j = 1, \dots, J = \lceil \log_2(U_0/\varepsilon) \rceil$ **do** ▷ stage j: $f(y) - \min_K f \le 2\delta$ is known
  $\delta \leftarrow U_0 2^{-j}$; $m \leftarrow \min\{2^i : 2^{3i} \ge 64LR^2 \mathsf{b}_{10\cdot 2^i} \mathsf{c}_n/\delta\}$; $T \leftarrow 10m$; $\eta \leftarrow \delta/(16m)$
  $C_0 \leftarrow K$; $z_0 \leftarrow$ a point of K within $P_T/(4T)$ of $\mathrm{st}(K)$
  for $t = 0, 1, \dots, T-1$ **do**
    $(x_t, w_t, g_t) \leftarrow$ SEARCH(y, z_t, η) ▷ $x_t \in [y, z_t]$, $w_t \in K$, $g_t = \nabla f(w_t)$
    $C_{t+1} \leftarrow C_t \cap \{u : \langle g_t, w_t - u \rangle \ge \delta\}$
    if $C_{t+1} = \varnothing$ **then**
      $\lambda \leftarrow$ minimizer of the certificate program of Algorithm 3 for $(w_i, g_i)_{i \le t}$; $y \leftarrow \sum_{i \le t} \lambda_i w_i$; **break** ▷ $f(y) - \min_K f < \delta$
    $z_{t+1} \leftarrow$ a point of C_{t+1} within $P_T/(4T)$ of $\mathrm{st}(C_{t+1})$
    if $\|x_t - w_t\|_2^2 \ge \delta/(Lm)$ **then** $y \leftarrow w_t$ ▷ accept: f decreased by $\delta/(4m)$
return y ▷ $f(y) - \min_K f \le \varepsilon$

moves far. Compared with (23), the count (53) has an extra factor $\log^2$ from the inner searches. Since Propositions D.1 and D.7 apply a fortiori to gradient oracles, N^{-3} is also the minimax exponent for the gradient oracle on $B_1^n(R)$ and on $\Delta_n(R)$, and for f that is L-smooth on $\mathbb{R}^n$ the strongly convex reductions of Section F.2 apply to this oracle as well, with the counts of Proposition E.8.

In the proofs below, $\mathrm{diam}_2(K) \le 2R$ because $K \subseteq c + B_1^n(R)$, and we use the co-coercivity of the gradient,

$$\|\nabla f(a) - \nabla f(b)\|_2^2 \le L \langle \nabla f(a) - \nabla f(b), a - b \rangle, \quad \text{hence} \quad \Big\| a - b - \frac{\nabla f(a) - \nabla f(b)}{L} \Big\|_2 \le \|a - b\|_2, \tag{54}$$

which follows by adding the descent lemma for $f - \langle \nabla f(b), \cdot \rangle$ (minimized at b) to the same inequality with a and b exchanged; expanding the square gives the second inequality.

The envelope. Let E and p be as in (51). The function $f_x(u) = f(u) + \frac{L}{2}\|u - x\|_2^2$ is L-strongly convex, so $p(x)$ is unique; E is the Moreau envelope of $f + \iota_K$ (ι_K the indicator of K) with parameter $1/L$, hence convex and differentiable with $\nabla E(x) = L(x - p(x))$ and L-Lipschitz gradient (for two points x, x' the monotonicity of $\partial(f + \iota_K)$ at $p(x), p(x')$ gives $\langle q_x - q_{x'}, x - x' \rangle \ge \|q_x - q_{x'}\|_2^2/L$ for $q_x = L(x - p(x))$). For $x \in K$, $\|\nabla E(x)\|_2 = L\|x - p(x)\|_2 \le 2LR$.

Lemma E.9 (Approximate proximal point)**.** *Fix $x \in K$ and $k \ge 1$, let $u_0 = x$ and $u_{j+1} = \mathrm{proj}_K\big(u_j - [\nabla f(u_j) + L(u_j - x)]/(2L)\big)$ for $j < k$, query $\nabla f(u_k)$, and put $w = u_k$, $g = \nabla f(w)$,*

$$v = 2L(u_{k-1} - u_k) - \nabla f(u_{k-1}) - L(u_{k-1} - x),$$
$$r = g + L(w - x) + v = \nabla f(u_k) - \nabla f(u_{k-1}) - L(u_k - u_{k-1}).$$

This uses $k + 1$ gradient queries, all in K; $v \in N_K(w)$, $\|r\|_2 \le 6LR \cdot 2^{-k}$, and if $\|r\|_2 \le \rho$ then

$$\|w - p(x)\|_2 \le \rho/L, \qquad \|L(x - w) - \nabla E(x)\|_2 \le \rho, \qquad 0 \le f(w) + \tfrac{L}{2}\|w - x\|_2^2 - E(x) \le \frac{\rho^2}{2L}. \tag{55}$$

Proof. The map $u \mapsto \frac{1}{2}(u - \nabla f(u)/L) + \frac{x}{2}$ is a contraction with factor $\frac{1}{2}$ by (54), and the projection is nonexpansive, so the iteration is a contraction with factor $\frac{1}{2}$; its fixed point is the unique minimizer $p(x)$ of f_x over K (the fixed-point equation is the optimality condition of a projected-gradient step with step $1/(2L)$ on f_x). Hence $\|u_j - p(x)\|_2 \le 2^{-j}\|x - p(x)\|_2 \le 2R\,2^{-j}$ and $\|u_k - u_{k-1}\|_2 \le 6R\,2^{-k}$; (54) applied to the pair u_k, u_{k-1} gives $\|r\|_2 \le L\|u_k - u_{k-1}\|_2 \le 6LR \cdot 2^{-k}$. The optimality condition of the last projection says that $u_{k-1} - [\nabla f(u_{k-1}) + L(u_{k-1} - x)]/(2L) - u_k \in N_K(u_k)$, which is $v/(2L) \in N_K(w)$; the two expressions for r agree by substitution.

For (55), note that $\nabla f_x(w) = g + L(w - x) = r - v$, so L-strong convexity of f_x and $\langle v, u - w \rangle \le 0$ give, for all $u \in K$, $f_x(u) \ge f_x(w) + \langle r, u - w \rangle + \frac{L}{2}\|u - w\|_2^2$. Taking $u = p(x)$ and using

$f_x(w) \geq f_x(p(x)) + \frac{L}{2}\|w-p(x)\|_2^2$ (minimality of $p(x)$ and strong convexity) yields $L\|w-p(x)\|_2^2 \leq \rho\|w-p(x)\|_2$, i.e. the first bound, and then $f_x(w) - E(x) \leq \rho\|w-p(x)\|_2 - \frac{L}{2}\|w-p(x)\|_2^2 \leq \rho^2/(2L)$; the second bound is $L\|w-p(x)\|_2 \leq \rho$. □

Lemma E.10 (Bisection without values)**.** *Let $y, z \in K$, $\eta > 0$, $\rho = \eta/(4R)$, $k = \min\{j \geq 1 : 6LR \cdot 2^{-j} \leq \rho\}$ and $\bar{h} = \min\{h \geq 0 : 4LR^2 \cdot 2^{-h} \leq \eta\}$. Put $d = z - y$ and $\varphi(t) = E(y + td)$. Starting from $[a, b] = [0, 1]$, while $b - a > 2^{-\bar{h}}$: let $t = (a + b)/2$, run Lemma E.9 at $x_t = y + td$ with k steps, and compute $\widetilde{s} = \langle L(x_t - w), d\rangle$; if $\widetilde{s} > \eta/2$ set $b = t$, if $\widetilde{s} < -\eta/2$ set $a = t$, otherwise stop and return t with the output (w, g, v, r) of Lemma E.9 at x_t. If the loop ends by length, return $t = b$ with the output stored at b (running Lemma E.9 once at z if $b = 1$). The returned $x = x_t \in [y, z]$ and (w, g, v, r) satisfy $\|r\|_2 \leq \rho$, (55), and*

$$E(x) \leq E(y) + \eta, \qquad \langle \nabla E(x), x - z\rangle \leq \eta, \qquad \langle L(x - w), x - z\rangle \leq \eta/2, \tag{56}$$

at a cost of at most $(\bar{h} + 1)(k + 1)$ gradient queries.

Proof. The function φ is convex and C^1 with $\varphi'(t) = \langle \nabla E(x_t), d\rangle$ and $|\varphi'| \leq 2LR \cdot 2R = 4LR^2$. By (55), $|\widetilde{s} - \varphi'(t)| \leq \rho\|d\|_2 \leq 2R\rho = \eta/2$. Hence a positive test certifies $\varphi'(t) > 0$ and a negative one $\varphi'(t) < 0$, so $[a, b]$ always contains a minimizer t_* of φ on $[0, 1]$, and $\varphi(a) \leq \varphi(0)$ because φ is nonincreasing on $[0, t_*]$. If the loop stops at t, then $|\widetilde{s}| \leq \eta/2$, so $|\varphi'(t)| \leq \eta$, and convexity gives $\varphi(t) - \varphi(0) \leq t\varphi'(t) \leq \eta$ and $\langle \nabla E(x_t), x_t - z\rangle = (t - 1)\varphi'(t) \leq \eta$; likewise $\langle L(x_t - w), x_t - z\rangle = (t - 1)\widetilde{s} \leq \eta/2$. If it ends by length, $\varphi(b) - \varphi(0) \leq \varphi(b) - \varphi(t_*) \leq 4LR^2(b - a) \leq 4LR^2 \cdot 2^{-\bar{h}} \leq \eta$; if $b < 1$ the last test at b was positive, so $\varphi'(b) > 0$, $\widetilde{s}(b) > 0$ and both inner products with $x_b - z = (b - 1)d$ are negative, while for $b = 1$ they vanish. The loop performs at most $\bar{h}$ iterations and at most one further call is made. □

Lemma E.11 (One stage)**.** *Let $0 < \delta \leq LR^2$ and $y \in K$ with $f(y) - \min_K f \leq 2\delta$. Let m be the smallest power of two with $m^3 \geq 64LR^2\mathsf{b}_{10m}\mathsf{c}_n/\delta$, $T = 10m$, $\eta = \delta/(16m)$, $\rho = \eta/(4R)$, and run the inner loop of Algorithm 5 with these parameters. Then some cut C_{t+1} with $t \leq 9m-1$ is empty, the returned mixture y satisfies $f(y) - \min_K f < \delta$, and the stage uses at most $T(\bar{h}+1)(k+1)$ gradient queries, where $\bar{h} + 1 \leq B = 1 + \lceil \log_2(64LR^2m/\delta)\rceil$ and $k = \max\{1, \lceil \log_2(384LR^2m/\delta)\rceil\}$.*

Proof. The values of $\bar{h}$ and k follow from Lemma E.10 with $\eta = \delta/(16m)$ and $\rho = \delta/(64Rm)$ ($4LR^2/\eta = 64LR^2m/\delta$ and $6LR/\rho = 384LR^2m/\delta$). Consider a trial t with a nonempty cut, write $x = x_t$, $w = w_t$, $g = g_t$, $q = x - w$, $z = z_t$, $z^+ = z_{t+1}$, $d = z^+ - z$.

The dichotomy. Since $z^+ \in C_{t+1}$, $\delta \leq \langle g, w - z^+\rangle$. Substituting $g = Lq + r - v$ and using $\langle v, z^+ - w\rangle \leq 0$ ($v \in N_K(w)$, $z^+ \in K$) and $\langle r, w - z^+\rangle \leq 2R\rho$,

$$\delta \leq L\langle q, w - z^+\rangle + 2R\rho = L\langle q, x - z\rangle - L\|q\|_2^2 - L\langle q, d\rangle + 2R\rho.$$

By (56) and (55), $L\langle q, x - z\rangle \leq \langle \nabla E(x), x - z\rangle + \rho\|x - z\|_2 \leq \eta + 2R\rho$. Hence $\delta \leq L\|q\|_2\|d\|_2 - L\|q\|_2^2 + \eta + 4R\rho$, and $\eta + 4R\rho = 2\eta = \delta/(8m) \leq \delta/8$ gives

$$\tfrac{3}{4}\delta + L\|q\|_2^2 \;\leq\; L\|q\|_2\|d\|_2. \tag{57}$$

In particular $q \neq 0$.

Large steps. By (55), (56) and $E(y) \leq f(y)$ (take $u = y$ in (51)), $f(w) \leq E(x) + \rho^2/(2L) - \frac{L}{2}\|q\|_2^2 \leq f(y) - \frac{L}{2}\|q\|_2^2 + \eta + \rho^2/(2L)$. If $\|q\|_2^2 \geq \delta/(Lm)$, then, since $\rho^2/(2L) = \delta^2/(8192LR^2m^2) \leq \delta/(8192m)$ by $\delta \leq LR^2$,

$$f(w) \leq f(y) - \frac{\delta}{2m} + \frac{\delta}{16m} + \frac{\delta}{8192m} < f(y) - \frac{\delta}{4m}.$$

The algorithm replaces y by w if and only if $\|q\|_2^2 \geq \delta/(Lm)$, so each accepted trial decreases $f(y)$ by more than $\delta/(4m)$. Since $f(y) \geq \min_K f$ and $f(y) - \min_K f \leq 2\delta$ initially, there are at most $8m$ accepted trials.

Small steps. If $\|q\|_2^2 < \delta/(Lm)$, (57) gives $\|d\|_2 \geq \frac{3\delta}{4L\|q\|_2} + \|q\|_2 > \frac{3}{4}\sqrt{m\delta/L}$. If s trials have a nonempty cut, the sets $C_0 \supseteq \cdots \supseteq C_s$ are nonempty and nested in $c + B_1^n(R)$ with $s \leq T$, so Lemma 3.1 and the selector accuracy $P_T/(4T)$ give $\sum_{t<s}\|z_{t+1} - z_t\|_2 \leq P_T + 2T \cdot P_T/(4T) = \frac{3}{2}P_T$.

The number of small nonempty trials is therefore less than $\frac{3}{2}P_T/\big(\frac{3}{4}\sqrt{m\delta/L}\big) = 2P_T\sqrt{L/(m\delta)} \le m$, where the last inequality is $4P_T^2 L \le m^3\delta$, i.e. $64LR^2\mathsf{b}_T\mathsf{c}_n \le m^3\delta$ with $\mathsf{b}_T = \mathsf{b}_{10m}$, the defining property of m.

Termination and certificate. At most $8m + (m-1)$ trials have a nonempty cut, so the cut of some trial $t \le 9m - 1 < T$ is empty. Then $\max_{u\in K}\min_{i\le t}\langle g_i, w_i - u\rangle < \delta$, and the minimax argument of Proposition E.1 gives $\lambda \in \Delta_{t+1}$ with $\max_{u\in K}\sum_i \lambda_i \langle g_i, w_i - u\rangle < \delta$, computed by the linear program of Algorithm 5; for $\bar w = \sum_i \lambda_i w_i \in K$ and any $u \in K$, $f(\bar w) - f(u) \le \sum_i \lambda_i[f(w_i) - f(u)] \le \sum_i \lambda_i \langle g_i, w_i - u\rangle < \delta$. Each trial makes at most $(\bar h + 1)(k+1)$ queries by Lemma E.10, and the mixture needs no query. □

Proof of Proposition E.8. Initialization. The point $y_0 = \mathrm{proj}_K(a - g_a/L)$ minimizes the model $\psi(u) = f(a) + \langle g_a, u - a\rangle + \frac{L}{2}\|u-a\|_2^2$ over K, and for a minimizer $x^\star$, $f(y_0) \le \psi(y_0) \le \psi(x^\star) \le f(x^\star) + \frac{L}{2}\|x^\star - a\|_2^2$ by smoothness at a and convexity; hence $f(y_0) - \min_K f \le \frac{L}{2}\operatorname{diam}_2(K)^2 \le 2LR^2 = U_0$, and for $K = B_1^n(R)$, $a = 0$, $f(y_0) - \min_K f \le \frac{L}{2}\|x^\star\|_2^2 \le LR^2/2$. No value is computed.

Stages. Let $\delta_j = U_0 2^{-j}$ and $J = \lceil \log_2(U_0/\varepsilon)\rceil$, so that $\delta_J \le \varepsilon < 2\delta_J$ and $\delta_j \le U_0/2 \le LR^2$ for $j \ge 1$. Stage 1 receives y_0 with $f(y_0) - \min_K f \le U_0 = 2\delta_1$; by Lemma E.11 stage j returns a point with error below $\delta_j = 2\delta_{j+1}$, which is the hypothesis of stage $j+1$. The output of stage J has error below $\delta_J \le \varepsilon$. All queries are at points of K.

Count. Stage j uses at most $10m_jB_j(k_j+1)$ queries, and B_j, k_j are nondecreasing in j because m_j/δ_j is. Let $\phi(m) = m^3/(64LR^2\mathsf{b}_{10m}\mathsf{c}_n)$ for powers of two m; m_j is the smallest m with $\phi(m) \ge 1/\delta_j$, and since $\phi(2m)/\phi(m) = 8\mathsf{b}_{10m}/\mathsf{b}_{20m} = 8\mathsf{b}_{10m}/(\mathsf{b}_{10m}+1) \in (4,8)$, ϕ is increasing and $m_j \ge 2$ (as $\phi(1) < 1/(LR^2) \le 1/\delta_j$). A given m is used if and only if $1/\phi(m) \le \delta_j < 1/\phi(m/2)$; the endpoints of this interval have ratio less than 8, so it contains at most three of the values δ_j, which halve from stage to stage. Consequently $\sum_j m_j \le 3(m_J + m_J/2 + \dots) \le 6m_J$ and the total count is at most $1 + 10B_J(k_J+1)\sum_j m_j \le 1 + 60m_JB_J(k_J+1)$.

Order. Put $Q = LR^2\mathsf{c}_n/\varepsilon$. From the minimality of m_J and $\delta_J > \varepsilon/2$, $(m_J/2)^3 < 64LR^2\mathsf{b}_{5m_J}\mathsf{c}_n/\delta_J < 128Q\,\mathsf{b}_{5m_J}$, and $\mathsf{b}_{5m_J} \le 5 + \log_2 m_J$; as $m_J^3 < 1024Q(5+\log_2 m_J)$ forces $m_J \le C(1+Q)$, we get $m_J = O(1+[Q\log(2+Q)]^{1/3})$. Also $LR^2m_J/\delta_J \le 2LR^2m_J/\varepsilon \le 2Qm_J$ gives $B_J, k_J = O(\log(2+Q))$. Multiplying proves (53). □

E.5 A CENTER-OF-GRAVITY METHOD FOR SMALL DIMENSION

When n is small compared with the count (53), the classical center-of-gravity method with deep cuts and the same mixture certificate needs fewer queries, and it also uses gradients only. We state it for an operator F, so that it also applies to variational inequalities (Section G); for $F = \nabla f$ the certificate bounds $f(\bar w) - \min_K f$.

Proposition E.12 (Center-of-gravity branch)**.** *Let $F : B_1^n(R) \to \mathbb{R}^n$ be L-Lipschitz in ℓ_2 and $\varepsilon > 0$. CG-Cut (Algorithm 6) queries $F(0)$, then $F(y)$ at a vertex y minimizing $\langle F(0), u\rangle$ over the ball, and then F at the centroid of a known polytope that it cuts by the half-space $\{u : \langle F(w), w - u\rangle \ge \varepsilon/2\}$, until the deeper cuts $\{u : \langle F(w_i), w_i - u\rangle \ge \varepsilon\}$ have empty intersection with the ball. It stops after at most*

$$3 + \left\lfloor \frac{n\log(2 + 8LR^2/\varepsilon)}{-\log(1 - 1/e)} \right\rfloor = O\big(1 + n\log(2 + LR^2/\varepsilon)\big) \tag{58}$$

calls and returns weights $\lambda \in \Delta_{t+1}$ on the queried points $w_0 = y, w_1, \dots, w_t$ (the centroids) and $\bar w = \sum_i \lambda_i w_i \in B_1^n(R)$ with $\max_{u\in B_1^n(R)}\sum_i \lambda_i \langle F(w_i), w_i - u\rangle \le \varepsilon$. If $F = \nabla f$ for a convex f, then $f(\bar w) - \min_{B_1^n(R)} f \le \varepsilon$; if F is monotone, the weak gap of $\bar w$ is at most ε. No bound on $\|F(0)\|$ is needed, and $L > 0$ is assumed (for $L = 0$ the operator is constant and the point y above is exact).

Proof of Proposition E.12. Two initial queries. If $b = 0$ the certificate of the single point 0 is $\max_u \langle 0, 0 - u\rangle = 0$. Otherwise y minimizes $\langle b, u\rangle$ over K, so $\langle b, u - y\rangle \ge 0$ on K, and $\|h - b\|_2 \le L\|y\|_2 \le LR$. For $u \in K$,

$$a_y(u) = \langle h, y - u\rangle = -\langle b, u - y\rangle + \langle h - b, y - u\rangle \le -\langle b, u - y\rangle + 2LR^2, \tag{59}$$

Algorithm 6 CG-Cut: centroid cuts with a global certificate (operator or gradient oracle)

Input: $K = B_1^n(R)$; oracle F that is L-Lipschitz on K; accuracy $\varepsilon > 0$

query $b = F(0)$; **if** $b = 0$ **return** $\bar{w} = 0$, $\lambda = (1)$
$y \leftarrow -R\,\mathrm{sign}(b_i)e_i$ for an index i with $|b_i| = \|b\|_\infty$; query $h = F(y)$; $w_0 \leftarrow y$
$\widehat{C} \leftarrow \{u \in K : \langle b, u - y\rangle \le 2LR^2\}$; $a_w(u) := \langle F(w), w - u\rangle$ for stored pairs
for $t = 0, 1, 2, \ldots$ **do**
  $C'_t \leftarrow K \cap \bigcap_{i\le t}\{u : a_{w_i}(u) \ge \varepsilon\}$; $C_t \leftarrow \widehat{C} \cap \bigcap_{i\le t}\{u : a_{w_i}(u) \ge \varepsilon/2\}$
  if $C'_t = \varnothing$ **then**
    solve $\min_{\lambda\in\Delta_{t+1}}\{\sum_{i\le t}\lambda_i \langle F(w_i), w_i\rangle + R\|\sum_{i\le t}\lambda_i F(w_i)\|_\infty\}$; **return** $\bar{w} = \sum_{i\le t}\lambda_i w_i$, λ
  $w_{t+1} \leftarrow$ centroid of C_t; query $F(w_{t+1})$

so $a_y(u) \ge \varepsilon$ implies $u \in \widehat{C}$, and hence $C'_t \subseteq \widehat{C}$ for every t. Since $2LR^2 > 0$, the known polytope $\widehat{C}$ is full-dimensional: it contains a neighborhood of the interior point $(1-s)y$ of K for small $s > 0$, where $\langle b, (1-s)y - y\rangle = sR\|b\|_\infty < 2LR^2$. For $w, v \in \widehat{C}$,

$$a_w(v) = \langle b, w - v\rangle + \langle F(w) - b, w - v\rangle \ge -2LR^2 - LR \cdot 2R = -4LR^2 =: -U, \tag{60}$$

because $\langle b, w - y\rangle \ge 0$, $\langle b, v - y\rangle \le 2LR^2$ and $\|F(w) - b\|_2 \le L\|w\|_2 \le LR$.

Volume. Let $\theta = \varepsilon/(2(U + \varepsilon))$. If $C'_t \ne \varnothing$, pick $u \in C'_t \subseteq \widehat{C}$; for $v \in \widehat{C}$ and $i \le t$ the affine function a_{w_i} satisfies $a_{w_i}(u) \ge \varepsilon$ and $a_{w_i}(v) \ge -U$, so $a_{w_i}((1-\theta)u + \theta v) \ge (1-\theta)\varepsilon - \theta U = \varepsilon/2$. Hence $(1-\theta)u + \theta\widehat{C} \subseteq C_t$ and $\mathrm{vol}(C_t) \ge \theta^n \mathrm{vol}(\widehat{C}) > 0$; in particular the centroid w_{t+1} is defined. If $F(w_{t+1}) = 0$ the new constraint $a_{w_{t+1}} \ge \varepsilon/2$ cannot be satisfied, so $C_{t+1} = C'_{t+1} = \varnothing$ and the method stops. Otherwise the half-space $\{u : a_{w_{t+1}}(u) \ge \varepsilon/2\}$ is contained in the half-space $\{u : \langle F(w_{t+1}), w_{t+1} - u\rangle \ge 0\}$ through the centroid of C_t, and Grünbaum's inequality (Grünbaum, 1960) gives $\mathrm{vol}(C_{t+1}) \le (1 - 1/e)\,\mathrm{vol}(C_t)$; also $C_0 \subseteq \widehat{C}$. Let $T = 1 + \lfloor n\log(2 + 8LR^2/\varepsilon)/(-\log(1 - 1/e))\rfloor$. If $C'_T \ne \varnothing$, then $\theta^n\,\mathrm{vol}(\widehat{C}) \le \mathrm{vol}(C_T) \le (1-1/e)^T\,\mathrm{vol}(\widehat{C})$, i.e. $T \le n\log(1/\theta)/(-\log(1-1/e))$ with $1/\theta = 2 + 8LR^2/\varepsilon$, which contradicts the choice of T. So the method stops with $C'_t = \varnothing$ for some $t \le T$, after the two initial queries and at most T centroid queries; this is (58).

Certificate. If $C'_t = \varnothing$ then $\max_{u\in K}\min_{i\le t} a_{w_i}(u) < \varepsilon$ by compactness, and the minimax theorem, as in Proposition E.1, gives $\lambda \in \Delta_{t+1}$ with $\max_{u\in K}\sum_i \lambda_i a_{w_i}(u) < \varepsilon$; the maximum equals $\sum_i \lambda_i \langle F(w_i), w_i\rangle + R\|\sum_i \lambda_i F(w_i)\|_\infty$ on the ball, a linear program. Every w_i is a point of K at which F was queried ($w_0 = y$ with $h = F(y)$; $F(0)$ is not used in the mixture), so $\bar{w} \in K$. For $F = \nabla f$, convexity gives $f(\bar{w}) - f(u) \le \sum_i \lambda_i \langle \nabla f(w_i), w_i - u\rangle \le \varepsilon$ for all $u \in K$; for monotone F, $\langle F(u), \bar{w} - u\rangle = \sum_i \lambda_i \langle F(u), w_i - u\rangle \le \sum_i \lambda_i \langle F(w_i), w_i - u\rangle \le \varepsilon$ (Lemma G.1). □

The proof uses only the diameter, the linear minimization over K, the emptiness tests and the centroids of known polytopes. On a general known polytope K of dimension d the same argument therefore gives $O(1 + d\log(2 + LR^2/\varepsilon))$ calls, with the centroid taken in the affine hull of K, an arbitrary initial point $a \in K$ in place of 0, and the absolute constants in the diameter bounds doubled.

Choosing between Proposition E.8 and Proposition E.12 by their guaranteed query counts gives, for the gradient oracle on $B_1^n(R)$, $O\big(1 + \min\{[\frac{LR^2\mathsf{c}_n}{\varepsilon}]^{1/3}\log^{7/3}(2 + \frac{LR^2\mathsf{c}_n}{\varepsilon}),\, n\log(2 + \frac{LR^2}{\varepsilon})\}\big)$ queries.

F Strong convexity

This section gives the strongly convex versions of the upper bounds, first by restarts with the condition number $\kappa_1 = L/\mu_1$ (Section F.2), and then, for smooth functions on the ball, a sharper analysis in terms of the slack $\Lambda = \kappa_1 - n$ that determines the deterministic complexity (Sections F.3 and F.4). We begin with the constraint that relates the two constants.

F.1 Norm compatibility

Euclidean smoothness and ℓ_p strong convexity cannot be combined arbitrarily: comparing the two quadratic bounds at x and $x + tv$ gives the following inequality.

Proposition F.1 (Compatibility inequality)**.** *If f is L-smooth in ℓ_2 and μ_p-strongly convex in ℓ_p on a nonempty open convex subset of $\mathbb{R}^n$, then $L \geq \mu_p n^{\beta_p}$, with equality for $f(x) = \frac{L}{2}\|x\|_2^2$.*

Proof. For x in the set, $v \in \mathbb{R}^n$ and small $t > 0$, the upper and the lower quadratic model at x, evaluated at $x + tv$, give $\mu_p\|v\|_p^2 \leq L\|v\|_2^2$ after the affine terms cancel. The largest value of $\|v\|_p^2/\|v\|_2^2$ is $n^{2/p-1} = n^{\beta_p}$, attained at vectors with equal absolute coordinates. For $f = \frac{L}{2}\|x\|_2^2$ the inequality $\|v\|_2^2 \geq n^{-\beta_p}\|v\|_p^2$ shows that $\mu_p = Ln^{-\beta_p}$ is admissible. □

In particular $\kappa_1 = L/\mu_1 \geq n$, and we call $\Lambda = \kappa_1 - n \geq 0$ the *slack*.

F.2 Restarts

Suppose that f is in addition μ_1-strongly convex with respect to $\|\cdot\|_1$ on K (a boundary minimizer is allowed). Then $f(x) - f(x^\star) \geq \frac{\mu_1}{2}\|x - x^\star\|_1^2$ for all $x \in K$, so an ε-solution x_j localizes $x^\star$ in the known polytope $K_{j+1} = K \cap (x_j + B_1^n(r_{j+1}))$ with $r_{j+1} = \sqrt{2\varepsilon/\mu_1}$. SP-Restart solves the problem on $K_0 = K$ to accuracy $\mu_1 R^2/8$ and then, with $r_j = R2^{-j}$, on K_j to accuracy $\mu_1 r_j^2/8$, each time by Algorithm 2 restricted to K_j. The radius cancels in the required horizon, which is therefore the same in every phase.

Proposition F.2 (Restarts)**.** *Let f be convex, L-smooth and μ_1-strongly convex in ℓ_1 on $K \subseteq c + B_1^n(R)$, $\kappa_1 = L/\mu_1$, and let $N_\star$ be the smallest power of two with $N_\star^3 \geq 24576\,\kappa_1 \mathsf{c}_n(1+\log_2 N_\star)$. For every $\varepsilon > 0$, SP-Restart returns an ε-solution after at most $(J+1)(1+24N_\star)$ first-order queries, where $J = \big\lceil \log_4(\mu_1 R^2/(8\varepsilon))\big\rceil_+$; that is,*

$$O\Big(\big[1 + \{\kappa_1 \log(2n)\log(2+\kappa_1\log(2n))\}^{1/3}\big]\big[1 + \log_+ \tfrac{\mu_1 R^2}{\varepsilon}\big]\Big). \tag{61}$$

If f is G-Lipschitz and μ_1-strongly convex in ℓ_1 on K (subgradient oracle without values), the same restarts with Algorithm 3 return an ε-solution after $O\big([1 + \log_+ \frac{\mu_1 R^2}{\varepsilon}] + \frac{G\sqrt{\log(2n)}}{\sqrt{\mu_1\varepsilon}}\sqrt{\log(2 + \frac{G\sqrt{\log(2n)}}{\sqrt{\mu_1\varepsilon}})}\big)$ queries. If ∇f is ν-Hölder with constant $M = L_{1,\nu}$, $0 < \nu < 1$, the restarts with the level method of Proposition E.3 need $\widetilde{O}_\nu\big(1 + (M/(\mu_1 R^{1-\nu}))^{1/(1+2\nu)} + \log_+ \frac{\mu_1 R^2}{\varepsilon} + M^{1/(1+2\nu)}\mu_1^{-(1+\nu)/(2+4\nu)}\varepsilon^{-(1-\nu)/(2+4\nu)}\big)$ queries.

The Euclidean bound is $O(\sqrt{\kappa_2}\log(1/\varepsilon))$ with $\kappa_2 = L/\mu_2$, where $\mu_2 \geq \mu_1$ is the Euclidean modulus, so the new bound is smaller when μ_1 is comparable to μ_2 and κ_1 is large; it requires no regularity of the Hessian. For deterministic methods and $n \geq \Lambda^{1/3}$, Proposition F.9 below shows that the power $\kappa_1^{1/3}$ cannot be improved when Λ is comparable to κ_1. The same restarts apply to the method of higher order of Proposition E.4.

Proof of Proposition F.2. Localization. Let $x^\star \in \arg\min_K f$. Strong convexity along the segment from $x^\star$ to $x \in K$ and minimality of $x^\star$ give, for $0 < t < 1$, $f(x^\star + t(x - x^\star)) \leq (1-t)f(x^\star) + tf(x) - \frac{\mu_1}{2}t(1-t)\|x - x^\star\|_1^2$ and $f(x^\star + t(x - x^\star)) \geq f(x^\star)$, hence $f(x) - f(x^\star) \geq \frac{\mu_1}{2}(1-t)\|x - x^\star\|_1^2$; letting $t \downarrow 0$ gives $f(x) - f(x^\star) \geq \frac{\mu_1}{2}\|x - x^\star\|_1^2$, so the minimizer is unique (possibly on the boundary of K). Consequently, if $f(x_j) - f(x^\star) \leq \mu_1 r_j^2/8$ then $\|x_j - x^\star\|_1 \leq r_j/2 = r_{j+1}$, and $K_{j+1} = K \cap (x_j + B_1^n(r_{j+1}))$ is a known polytope with $x^\star \in K_{j+1} \subseteq x_j + B_1^n(r_{j+1})$.

Smooth case. Apply Algorithm 2 on $K_j \subseteq x_j' + B_1^n(r_j)$ ($x_j' = x_{j-1}$ for $j \geq 1$, $x_0' = c$, $r_0 = R$) to accuracy $\delta_j = \mu_1 r_j^2/8$. By Proposition C.2 for general polytopes (Section C.2), this uses at most $1 + 24N_j$ queries, where N_j is the smallest power of two with $768Lr_j^2\mathsf{b}_N\mathsf{c}_n/N^3 \leq \delta_j/4 = \mu_1 r_j^2/32$, i.e. $N^3 \geq 24576\kappa_1\mathsf{b}_N\mathsf{c}_n$; the radius cancels, so $N_j = N_\star$ for all j (for a power of two, $\mathsf{b}_N = 1 + \log_2 N$). After $J+1$ phases, where J is the smallest index with $\mu_1 R^2 4^{-J}/8 \leq \varepsilon$,

the output x_J satisfies $f(x_J) - \min_K f \le \delta_J \le \varepsilon$; the total is at most $(J+1)(1+24N_\star)$. Finally $N_\star = O(1+[\kappa_1 \mathsf{c}_n \log(2+\kappa_1 \mathsf{c}_n)]^{1/3})$, because for a large absolute constant C the value $N = C\max\{1, [\kappa_1 \mathsf{c}_n \log(2+\kappa_1 \mathsf{c}_n)]^{1/3}\}$ satisfies the defining inequality (as $1+\log_2 N = O(\log(2+\kappa_1 \mathsf{c}_n))$ for such N), and rounding up to a power of two at most doubles it. This gives (61).

Nonsmooth case. Apply Algorithm 3 on K_j with depth δ_j; by Proposition E.1 it stops after at most $Q(r_j, \delta_j)$ queries, where $Q(r,\delta)$ is the smallest $m \ge 1$ with $16Gr\sqrt{\mathsf{b}_m \mathsf{c}_n}/m \le \delta$, and returns a point with error at most δ_j on K_j. Since $\mathsf{b}_m \mathsf{c}_n \le 4(1+\log m)(1+\log(2n))$, it suffices that $m/\sqrt{1+\log m} \ge A_j/2$, where $A_j = 64Gr_j\sqrt{1+\log(2n)}/\delta_j = 2^j A_0$ and $A_0 = 512G\sqrt{1+\log(2n)}/(\mu_1 R)$; the argument in the proof of Lemma E.2 with $c = \max\{16, A_j/2\}$ gives $Q(r_j,\delta_j) = O(\max\{1, A_j\}\sqrt{1+\log(2+A_j)})$. These bounds grow geometrically in j, so $\sum_{j\le J} Q(r_j,\delta_j) = O(J+1+A_J\sqrt{1+\log(2+A_J)})$, and since $\varepsilon/4 < \delta_J \le \varepsilon$ we have $r_J \asymp \sqrt{\varepsilon/\mu_1}$ and $A_J \asymp G\sqrt{1+\log(2n)}/\sqrt{\mu_1\varepsilon}$, which gives the stated count. If $\varepsilon \ge \mu_1 R^2/8$, a single run of Algorithm 3 with depth ε suffices.

Hölder case. With the level method of Proposition E.3, $M = L_{1,\nu}$, $s = 1+\nu$ and $\sigma = 1+2\nu$, phase j uses $\widetilde{O}(1+(Mr_j^s/\delta_j)^{1/\sigma}) = \widetilde{O}(1+(M/(\mu_1 r_j^{2-s}))^{1/\sigma})$ queries. For $\nu < 1$ this grows geometrically with the factor $2^{(2-s)/\sigma} > 1$, so the sum is dominated by the last phase, where $r_J \asymp \sqrt{\varepsilon/\mu_1}$, and the total is $\widetilde{O}_\nu\big(1+(M/(\mu_1 R^{2-s}))^{1/\sigma} + \log_+(\mu_1 R^2/\varepsilon) + M^{1/\sigma}\mu_1^{-s/(2\sigma)}\varepsilon^{-(2-s)/(2\sigma)}\big)$. For $\nu = 1$ all phases have the same bound and one recovers the smooth case. □

Corollary F.3 (Restarts with higher-order oracles)**.** *Let $k \ge 2$, $0 \le \nu \le 1$, $s = k+\nu$ and $\sigma = 2s-1$. Let f satisfy the assumptions of Proposition E.4 and be μ_1-strongly convex in ℓ_1 on $K \subseteq c + B_1^n(R)$, and put $\bar\kappa = L_{k,\nu}R^{s-2}/\mu_1$ and $\Pi_0 = [\mathsf{c}_n \log(2+\bar\kappa \mathsf{c}_n^{s/2})]^{s/(2\sigma)}$. Restarting the bracket search of Proposition E.4 on $K_0 = K$ and $K_{j+1} = K \cap (x_j + B_1^n(R2^{-j-1}))$, with accuracy $\mu_1 R^2 4^{-j}/8$ in phase $j = 0, \dots, J$, $J = \big\lceil \log_4(\mu_1 R^2/(8\varepsilon))\big\rceil_+$, returns $x_J \in K$ with $f(x_J) - \min_K f \le \varepsilon$ after*

$$O_k\Big(\Big[1+\log_+\frac{\mu_1 R^2}{\varepsilon}\Big]\big[1+\bar\kappa^{1/\sigma}\Pi_0\big]\Big) \quad \text{if } s = 2,$$
$$O_k\Big(1+\log_+\frac{\mu_1 R^2}{\varepsilon} + \frac{\bar\kappa^{1/\sigma}\Pi_0}{1-2^{-(s-2)/\sigma}}\Big) \quad \text{if } s > 2$$

order-k queries, where O_k hides a factor that depends only on k.

For a Lipschitz Hessian ($s = 3$) the count is $\widetilde{O}(1+\log_+(\mu_1 R^2/\varepsilon) + (L_{2,1}R/\mu_1)^{1/5})$. In Euclidean geometry the condition-number term of second-order methods is of order $(L_{2,1}R/\mu_2)^{2/7}$ (Arjevani et al., 2019), and near the solution Newton steps reduce the dependence on ε to $\log\log(1/\varepsilon)$; the corollary improves the first term when μ_1 is comparable to μ_2, but not the second.

Proof of Corollary F.3. Let $x^\star$ be the minimizer of f on K; strong convexity gives $f(x) - f(x^\star) \ge \frac{\mu_1}{2}\|x - x^\star\|_1^2$ on K. Let $J = \big\lceil \log_4(\mu_1 R^2/(8\varepsilon))\big\rceil_+$, $\rho_j = R2^{-j}$ and $\delta_j = \mu_1\rho_j^2/8$. If K_j contains $x^\star$, the bracket search of Proposition E.4 on K_j, started from one jet on K_j as in the proof of Proposition E.4, returns $x_j \in K_j$ with $f(x_j) - f(x^\star) \le \delta_j$, hence $\|x_j - x^\star\|_1^2 \le 2\delta_j/\mu_1 = \rho_{j+1}^2$, and K_{j+1} contains $x^\star$. After the phases $j = 0, \dots, J$ the gap is at most $\delta_J \le \varepsilon$.

On a domain of radius ρ_j the certificate (50) satisfies $e_N \le 8\Xi_k c_{k,\nu} L_{k,\nu}(4\rho_j)^s(\mathsf{b}_N \mathsf{c}_n)^{s/2}/N^\sigma$. Let N_j be the smallest power of two with $N_j^\sigma \ge 256\Xi_k c_{k,\nu} 4^s (L_{k,\nu}/\mu_1)\rho_j^{s-2}(\mathsf{b}_{N_j}\mathsf{c}_n)^{s/2}$; then $e_{N_j} \le \mu_1\rho_j^2/32 = \delta_j/4$, so the horizon used at accuracy δ_j is at most N_j and phase j uses at most $1 + 32(2k+1)N_j$ queries, for a total of $(J+1) + 32(2k+1)\sum_{j\le J} N_j$. With $\bar\kappa_j = L_{k,\nu}\rho_j^{s-2}/\mu_1 = \bar\kappa\, 2^{-j(s-2)}$, the dyadic horizon estimate in the proof of Proposition E.4 gives $N_j = O_k(1+\bar\kappa_j^{1/\sigma}\{\mathsf{c}_n \log(2+\bar\kappa_j \mathsf{c}_n^{s/2})\}^{s/(2\sigma)}) = O_k(1+\bar\kappa^{1/\sigma}\Pi_0\, 2^{-j(s-2)/\sigma})$, since $s \ge 2$. For $s = 2$ the bound does not depend on j; for $s > 2$ the geometric series is at most $(1-2^{-(s-2)/\sigma})^{-1}$. Since $J+1 \le 2 + \log_+(\mu_1 R^2/\varepsilon)$, this proves the corollary. □

F.3 THE SLACK AND LOCALIZATION BY ONE GRADIENT

By Proposition F.1, $\kappa_1 = n + \Lambda \geq n$, so the factor $\kappa_1^{1/3}$ of Proposition F.2 is at least $n^{1/3}$ even when the slack $\Lambda = \kappa_1 - n$ is small. On the ball the appropriate parameter is Λ, because one gradient localizes the minimizer in an ℓ_1 ball whose radius is controlled by Λ instead of κ_1.

Lemma F.4 (Localization by one gradient)**.** *Let $K = B_1^n(R)$, let f be convex and differentiable on an open set containing K, L-smooth in ℓ_2 and μ_1-strongly convex in ℓ_1 on K, with $\Lambda = L/\mu_1 - n > 0$, and let $x^\star$ be its minimizer on K. For $x \in K$ put $x^+ = \mathrm{proj}_K(x - \nabla f(x)/L)$. Then*

$$\|x^+ - x^\star\|_1^2 \leq \frac{16\Lambda\,[1 + \log(1+\Lambda)]}{L}\,\big[f(x) - f(x^\star)\big], \tag{62}$$

and if smoothness and strong convexity hold on all of $\mathbb{R}^n$, then $\|x^+ - x^\star\|_1^2 \leq \frac{8\Lambda}{L}[f(x) - f(x^\star)]$.

Proposition F.5 (Restarts governed by the slack)**.** *Under the assumptions of Lemma F.4, let $\Lambda_* = 8\Lambda$ in the global case and $\Lambda_* = 16\Lambda(1 + \log(1+\Lambda))$ in the domain case, and let N_Λ be the smallest power of two with $N_\Lambda^3 \geq 6144\Lambda_*\,\mathsf{b}_{N_\Lambda}\mathsf{c}_n$. For every $\varepsilon > 0$, the restarted method returns an ε-solution after at most $1 + J(2 + 24N_\Lambda)$ first-order queries, $J = \big\lceil \log_2(LR^2/(2\varepsilon))\big\rceil_+$; that is,*

$$O\Big(\big[1 + \{\Lambda_* \log(2n)\log(2 + \Lambda_*\log(2n))\}^{1/3}\big]\big[1 + \log_+ \tfrac{LR^2}{\varepsilon}\big]\Big). \tag{63}$$

In the global case, with the gradient-only method SP-Prox in each phase, the count is $O([1 + (\Lambda\log(2n))^{1/3}\log^{7/3}(2 + \Lambda\log(2n))][1 + \log_+ \frac{LR^2}{\varepsilon}])$ gradient queries, and no function value is used. If instead $n \leq \Lambda^{1/3}$, CG-Cut with $F = \nabla f$ gives $O(1 + n\log(2 + LR^2/\varepsilon))$ gradient queries.

Lemma F.6 (The ℓ_1 ball projection in ℓ_1)**.** $\|\mathrm{proj}_K(u) - \mathrm{proj}_K(v)\|_1 \leq 2\|u - v\|_1$ *for all $u, v \in \mathbb{R}^n$.*

Proof. For $\|u\|_1 \leq R$, $\mathrm{proj}_K(u) = u$. For $\|u\|_1 > R$, $\mathrm{proj}_K(u)_i = \mathrm{sign}(u_i)(|u_i| - \lambda)_+$ with the unique $\lambda > 0$ such that $\sum_i (|u_i| - \lambda)_+ = R$. The map is continuous and piecewise affine on a finite polyhedral subdivision of $\mathbb{R}^n$. On a piece with active set I ($|I| = m$, the coordinates with $|u_i| \geq \lambda$) and signs $s_i = \mathrm{sign}\,u_i$ we have $\lambda = (\sum_{i\in I}|u_i| - R)/m$, so the Jacobian has the block $I_m - ss^\top/m$ on the active coordinates and zero columns for the inactive ones. The ℓ_1-norm of every column of $I_m - ss^\top/m$ is $(1 - 1/m) + (m-1)/m \leq 2$, so the Jacobian has $\ell_1 \to \ell_1$ operator norm at most 2 on every piece. Splitting the segment $[u, v]$ into finitely many pieces and adding gives the claim; on the ball itself the constant is 1. □

Proof of Lemma F.4, global case. Let f be L-smooth and μ_1-strongly convex in ℓ_1 on $\mathbb{R}^n$. Then f is coercive and strictly convex, $f^*(g) = \sup_u\{\langle g, u\rangle - f(u)\}$ is finite, and $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is a homeomorphism with $\nabla f^*(\nabla f(x)) = x$; moreover ∇f^* is Lipschitz, because ∇f is strongly monotone in ℓ_2 (with modulus at least μ_1, since $\|v\|_1 \geq \|v\|_2$). For $x = \nabla f^*(g)$ and every h, the smoothness upper model and the strong-convexity lower model of f at x give

$$f^*(g) + \langle x, h\rangle + \frac{\|h\|_2^2}{2L} \;\leq\; f^*(g+h) \;\leq\; f^*(g) + \langle x, h\rangle + \frac{\|h\|_\infty^2}{2\mu_1} \tag{64}$$

(for the lower bound take $u = x + h/L$ in the supremum; for the upper one use $f(u) \geq f(x) + \langle g, u - x\rangle + \frac{\mu_1}{2}\|u - x\|_1^2$ and $\sup_v\{\langle h, v\rangle - \frac{\mu_1}{2}\|v\|_1^2\} = \|h\|_\infty^2/(2\mu_1)$). Put $\psi(g) = f^*(g) - \|g\|_2^2/(2L)$. By (64), ψ is convex with $\nabla\psi(g) = x - g/L$, and ∇f^* is Lipschitz, so by Rademacher's theorem ∇f^* is differentiable almost everywhere. At such a point g the second-order expansion of (64) gives $B := \nabla^2 f^*(g) - I/L \succeq 0$ and $s^\top B s \leq 1/\mu_1 - n/L = \Lambda/L$ for every sign vector $s \in \{-1, 1\}^n$. Since $h \mapsto h^\top B h$ is convex, its maximum over the cube $[-1, 1]^n$ is attained at a vertex, so $h^\top B h \leq (\Lambda/L)\|h\|_\infty^2$ for all h. Mollifying ψ with a smooth compactly supported kernel gives C^∞ functions ψ_ϵ whose Hessians are averages of the matrices B and satisfy the same two-sided bound. Hence $\psi_\epsilon(g + h) \leq \psi_\epsilon(g) + \langle\nabla\psi_\epsilon(g), h\rangle + \frac{\Lambda}{2L}\|h\|_\infty^2$, and letting $\epsilon \to 0$ (local uniform convergence of ψ_ϵ and of $\nabla\psi_\epsilon$, as $\nabla\psi$ is continuous) shows that ψ is (Λ/L)-smooth with respect to $\|\cdot\|_\infty$ on $\mathbb{R}^n$.

For a convex ψ that is γ-smooth with respect to a norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$, $D_\psi(u, v) \geq \|\nabla\psi(u) - \nabla\psi(v)\|_*^2/(2\gamma)$ (apply the smoothness upper model of $\psi - \langle\nabla\psi(v), \cdot\rangle$, which is minimized at v, at u in a direction attaining the dual norm). With $g = \nabla f(x)$, $g^\star = \nabla f(x^\star)$, the conjugate

identity $D_{f^*}(g^\star, g) = D_f(x, x^\star)$ and $D_{\|\cdot\|_2^2/(2L)}(g^\star, g) = \|g - g^\star\|_2^2/(2L)$ give $D_\psi(g^\star, g) = D_f(x, x^\star) - \|g - g^\star\|_2^2/(2L) \le D_f(x, x^\star) \le f(x) - f(x^\star)$, the last step because $\langle g^\star, x - x^\star\rangle \ge 0$ for $x \in K$ by optimality. Since $\nabla\psi(g) - \nabla\psi(g^\star) = x - x^\star - (g - g^\star)/L$, we obtain

$$\Big\|x - x^\star - \frac{\nabla f(x) - \nabla f(x^\star)}{L}\Big\|_1^2 \le \frac{2\Lambda}{L}\big[f(x) - f(x^\star)\big]. \tag{65}$$

Finally $x^\star = \mathrm{proj}_K(x^\star - \nabla f(x^\star)/L)$ by the optimality condition, so Lemma F.6 gives $\|x^+ - x^\star\|_1 \le 2\|x - x^\star - (\nabla f(x) - \nabla f(x^\star))/L\|_1$, and squaring proves the global bound $8\Lambda/L$. □

Proof of Lemma F.4, domain case. First let f be C^2 on a neighborhood of the segment $[z, x]$ with $\mu_1\|v\|_1^2 \le v^\top H_t v \le L\|v\|_2^2$ for $H_t = \nabla^2 f(z + td)$, $d = x - z$, $t \in [0, 1]$ (only these bounds are used, so the argument applies to the mollified functions below). Put $J_t = I - H_t/L$ and $A_t = H_t^{-1} - I/L \succeq 0$. Maximizing $\langle s, v\rangle - \frac{1}{2}v^\top H_t v$ over v and using $\langle s, v\rangle \le \|s\|_\infty\|v\|_1$ gives $\frac{1}{2}s^\top H_t^{-1} s \le \|s\|_\infty^2/(2\mu_1)$, so $s^\top A_t s \le \Lambda/L$ for every sign vector s; averaging over uniformly random signs gives $\mathrm{tr}\, A_t \le \Lambda/L$, hence $\lambda_{\max}(H_t^{-1}) \le (1 + \Lambda)/L$ and $H_t \succeq \frac{L}{1+\Lambda} I$. The matrices H_t, J_t, A_t commute, and on an eigenvalue $0 < \lambda \le L$ of H_t one checks $(1 - \lambda/L)^2/\lambda \le 1/\lambda - 1/L$ and $(1 - \lambda/L)^2 \le L(1/\lambda - 1/L)$; hence $J_t H_t^{-1} J_t \preceq A_t$ and $J_t^2 \preceq L A_t$. For a sign vector s, the Cauchy–Schwarz inequality in the H_t-metric and in the Euclidean metric gives $\langle s, J_t d\rangle \le \sqrt{s^\top J_t H_t^{-1} J_t s}\sqrt{d^\top H_t d}$ and $\langle s, J_t d\rangle \le \sqrt{s^\top J_t^2 s}\,\|d\|_2$, so

$$\|J_t d\|_1^2 \le \frac{\Lambda}{L}\, d^\top H_t d, \qquad \|J_t d\|_1^2 \le \Lambda\|d\|_2^2. \tag{66}$$

Now $\beta := D_f(x, z) = \int_0^1 (1 - t)\, d^\top H_t d\, dt$ and $x - z - (\nabla f(x) - \nabla f(z))/L = \int_0^1 J_t d\, dt$. Let $\eta = 1/(1 + \Lambda)$. On $[0, 1 - \eta]$ the first bound of (66) and the Cauchy–Schwarz inequality with the weight $(1 - t)$ give

$$\Big\|\int_0^{1-\eta} J_t d\, dt\Big\|_1 \le \sqrt{\frac{\Lambda}{L}}\int_0^{1-\eta} \frac{\sqrt{(1-t)d^\top H_t d}}{\sqrt{1-t}}dt \le \sqrt{\frac{\Lambda\beta}{L}\log\frac{1}{\eta}},$$

and on $[1 - \eta, 1]$ the second bound and $\beta \ge \frac{L}{2(1+\Lambda)}\|d\|_2^2$ give $\|\int_{1-\eta}^1 J_t d\, dt\|_1 \le \eta\sqrt{\Lambda}\|d\|_2 \le \sqrt{2\Lambda\beta/(L(1 + \Lambda))}$. By $(a + b)^2 \le 2a^2 + 2b^2$,

$$\Big\|x - z - \frac{\nabla f(x) - \nabla f(z)}{L}\Big\|_1^2 \le \frac{\Lambda}{L}\Big[2\log(1+\Lambda) + \frac{4}{1+\Lambda}\Big] D_f(x, z) \le \frac{4\Lambda[1 + \log(1 + \Lambda)]}{L} D_f(x, z). \tag{67}$$

For a C^1 function f satisfying the two inequalities on K only, we apply (67) to mollifications of f. For interior x, z the segment has positive distance to the boundary, and the convolution with a smooth kernel of smaller radius is C^∞, L-smooth and μ_1-strongly convex near the segment (both inequalities are averaged over translates that stay in K). Its values and gradients converge locally uniformly to those of f, so (67) holds for interior points. For boundary points we replace x, z by $(1-\sigma)x, (1-\sigma)z$ and let $\sigma \downarrow 0$, using the continuity of f and ∇f on K. With $z = x^\star$, $D_f(x, x^\star) \le f(x) - f(x^\star)$ by optimality, and Lemma F.6 with $x^\star = \mathrm{proj}_K(x^\star - \nabla f(x^\star)/L)$ gives (62). □

Proof of Proposition F.5. Initialization. One query at 0 gives $g_0 = \nabla f(0)$ and a vertex $v \in \arg\min_{u \in K}\langle g_0, u\rangle$. Smoothness at 0 gives $f(v) \le f(0) + \langle g_0, v\rangle + \frac{L}{2}\|v\|_2^2$, and convexity with the choice of v gives $f(x^\star) \ge f(0) + \langle g_0, x^\star\rangle \ge f(0) + \langle g_0, v\rangle$; hence $f(v) - f(x^\star) \le \frac{L}{2}\|v\|_2^2 = LR^2/2 =: \Delta_0$ without evaluating f. If $\Delta_0 \le \varepsilon$, the method returns v.

One phase. Let $x \in K$ satisfy the certified bound $f(x) - f(x^\star) \le \Delta$. Query $\nabla f(x)$ and put $x^+ = \mathrm{proj}_K(x - \nabla f(x)/L)$, $r = \min\{2R, \sqrt{\Lambda_*\Delta/L}\}$ and $K' = K \cap (x^+ + B_1^n(r))$, a known polytope contained in $x^+ + B_1^n(r)$. By Lemma F.4 (the global or the domain bound, according to the case), $x^\star \in K'$, so $\min_{K'} f = f(x^\star)$. We run Algorithm 2 on K' with accuracy $\Delta/2$. By Proposition C.2 for general polytopes, this uses at most $1 + 24N'$ queries, where N' is the smallest power of two with $N'^3 \ge 3072Lr^2\mathsf{b}_{N'}\mathsf{c}_n/(\Delta/2)$ (if the initial bracket width of the subproblem is at most $\Delta/2$, one query suffices). Since $Lr^2/(\Delta/2) \le 2\Lambda_*$, every N satisfying $N^3 \ge 6144\Lambda_*\mathsf{b}_N\mathsf{c}_n$

satisfies the defining inequality of N', so $N' \le N_\Lambda$. The phase returns a point $x' \in K' \subseteq K$ with $f(x') - f(x^\star) \le \Delta/2$ and uses at most $2 + 24N_\Lambda$ queries.

Phases. Starting from $x = v$, $\Delta = \Delta_0$, we run $J = \lceil \log_2(LR^2/(2\varepsilon)) \rceil_+$ phases, halving Δ each time; the last certified bound is $\Delta_0 2^{-J} \le \varepsilon$. The total is at most $1 + J(2 + 24N_\Lambda)$, and $N_\Lambda = O(1 + [\Lambda_* \mathsf{c}_n \log(2 + \Lambda_* \mathsf{c}_n)]^{1/3})$ by the argument in the proof of Proposition F.2 (substitute $N = C \max\{1, [\Lambda_* \mathsf{c}_n \log(2 + \Lambda_* \mathsf{c}_n)]^{1/3}\}$), which gives (63).

Gradient oracle (global case). Here f is L-smooth on $\mathbb{R}^n$, as Proposition E.8 requires. We replace Algorithm 2 by Algorithm 5 on K' with accuracy $\Delta/2$, started at $a = x^+$. Its initial certified bound is $U_0 = 2Lr^2 \le 2\Lambda_* \Delta$, and Proposition E.8 on the polytope $K' \subseteq x^+ + B_1^n(r)$ with the ratio $Lr^2 \mathsf{c}_n/(\Delta/2) \le 2\Lambda_* \mathsf{c}_n$ bounds its cost by $O(1 + (\Lambda_* \mathsf{c}_n)^{1/3} \log^{7/3}(2 + \Lambda_* \mathsf{c}_n))$, with $\Lambda_* = 8\Lambda$; the localization step uses one gradient and no value.

Dimension branch. Proposition E.12 with $F = \nabla f$ on K gives the last claim, since $f(\bar{w}) - \min_K f \le \varepsilon$ for its output. □

F.4 The deterministic complexity for $\Lambda \ge 8$

The lower bounds need two ingredients: the randomized bound of Proposition D.7 with strong convexity, which gives the cubic-rate phase, and a transfer of a classical finite-dimensional bound, which gives the high-accuracy phase.

Lemma F.7 (Weighted norm comparison)**.** *Let $1 \le p < 2$, $\beta_p = 2/p - 1$, and $a_i > 0$. Then*

$$\|h\|_p^2 \le \left(\sum_{i=1}^n a_i^{-1/\beta_p} \right)^{\beta_p} \sum_{i=1}^n a_i h_i^2. \tag{68}$$

The constant is sharp for the diagonal quadratic form.

Proof. Apply Hölder with exponents $2/p$ and $2/(2-p)$ to $|h_i|^p = (a_i h_i^2)^{p/2} a_i^{-p/2}$, then raise to $2/p$. Equality holds when $|h_i|$ is proportional to $a_i^{-1/(2-p)}$. □

Proposition F.8 (High accuracy at fixed dimension)**.** *Let $n \ge 2$, $L, \mu_1, R > 0$, $\Lambda = L/\mu_1 - n$, and let d be an integer with $2 \le d \le n$ and $d^3 - d \le \Lambda$. Consider the class of convex $C^{1,1}$ functions on $\mathbb{R}^n$ that are L-smooth in ℓ_2 and μ_1-strongly convex in ℓ_1, with global first-order queries and output in $B_1^n(R)$. If a deterministic method with at most N queries guarantees error at most ε on this class, where $0 < \varepsilon < LR^2/(2d^5)$, then*

$$N + 2 \;\ge\; c_{\mathrm{NY}} \, \frac{d}{\log d} \, \log \frac{LR^2}{2d^5 \varepsilon}, \tag{69}$$

where $c_{\mathrm{NY}} > 0$ is the absolute constant of Nemirovsky & Yudin (1983, Theorem 7.2.7): every deterministic method that solves all Euclidean L-smooth and (L/d^2)-strongly convex functions on $\mathbb{R}^d$ from the start 0 to relative accuracy $\varepsilon_{\mathrm{rel}} \in (0,1)$ has laboriousness (number of queries plus one for the output) at least $c_{\mathrm{NY}}(d/\log d)\log(1/\varepsilon_{\mathrm{rel}})$.

The proof normalizes an unknown d-dimensional instance by one query, so that its minimizer lies in the ball, and pads the remaining $n - d$ coordinates with the known curvature L. The padded function is μ_1-strongly convex in ℓ_1 because $d/(L/d^2) + (n-d)/L = (d^3 + n - d)/L \le 1/\mu_1$. With $d \asymp \min\{n, \Lambda^{1/3}\}$ this gives the high-accuracy term of Proposition F.9 below.

Proof of Proposition F.8. Let $\mu_2 = L/d^2$ and let $h : \mathbb{R}^d \to \mathbb{R}$ be any convex $C^{1,1}$ function that is L-smooth and μ_2-strongly convex in ℓ_2 on $\mathbb{R}^d$, with unique minimizer $u^\star$. We describe a deterministic method $\mathcal{B}$ for h that simulates the given method $\mathcal{A}$ for the mixed class and has laboriousness at most $N + 2$ in the convention of Nemirovsky & Yudin (1983), in which the queries and the final output are counted.

Normalization by one query. $\mathcal{B}$ queries $h(0)$ and $g_0 = \nabla h(0)$. If $g_0 = 0$ it outputs 0, which is exact. Otherwise it puts $a = R\mu_2/(\sqrt{d}\|g_0\|_2) > 0$. Strong monotonicity of ∇h between 0 and $u^\star$ gives $\mu_2 \|u^\star\|_2^2 \le \langle g_0, -u^\star \rangle \le \|g_0\|_2 \|u^\star\|_2$, so $\|u^\star\|_2 \le \|g_0\|_2/\mu_2$ and $\|a u^\star\|_1 \le \sqrt{d}\, a \|u^\star\|_2 \le R$.

Padding. Write $x = (u, v) \in \mathbb{R}^d \times \mathbb{R}^{n-d}$ and define

$$f(u, v) = a^2 h(u/a) + \tfrac{L}{2}\|v\|_2^2$$

(no second term if $d = n$). Then f is convex and L-smooth on $\mathbb{R}^n$, and for increments (ξ_u, ξ_v) its strong-convexity remainder is at least $\frac{1}{2}(\mu_2\|\xi_u\|_2^2 + L\|\xi_v\|_2^2)$. By Lemma F.7 in the ℓ_1 case ($\beta_1 = 1$), $\|(\xi_u, \xi_v)\|_1^2 \le (d/\mu_2 + (n-d)/L)(\mu_2\|\xi_u\|_2^2 + L\|\xi_v\|_2^2)$, and $d/\mu_2 + (n-d)/L = (d^3 + n - d)/L \le (n + \Lambda)/L = 1/\mu_1$ by the hypothesis $d^3 - d \le \Lambda$. Hence f is μ_1-strongly convex in ℓ_1 and belongs to the mixed class; its global minimizer $(au^\star, 0)$ lies in $B_1^n(R)$, so $\min_{B_1^n(R)} f = f(au^\star, 0) = a^2 h(u^\star)$.

Simulation. A query of $\mathcal{A}$ at any $(u, v) \in \mathbb{R}^n$ is answered from one query of h at u/a: $f(u, v) = a^2 h(u/a) + \frac{L}{2}\|v\|_2^2$ and $\nabla f(u, v) = (a\nabla h(u/a), Lv)$. Thus $\mathcal{B}$ makes at most $N + 1$ queries of h in total, and it is deterministic. When $\mathcal{A}$ outputs $(\widehat{u}, \widehat{v}) \in B_1^n(R)$, $\mathcal{B}$ outputs $\widehat{u}/a$.

Accuracy. Smoothness at 0 gives $h(0) - h(u^\star) \ge \|g_0\|_2^2/(2L)$ (compare the value of the upper model at $-g_0/L$ with the minimum), so $a^2[h(0) - h(u^\star)] \ge R^2\mu_2^2/(2Ld) = LR^2/(2d^5)$. The guarantee of $\mathcal{A}$ and the nonnegativity of the padding give $a^2[h(\widehat{u}/a) - h(u^\star)] \le f(\widehat{u}, \widehat{v}) - \min_{B_1^n(R)} f \le \varepsilon$, hence the relative error of $\mathcal{B}$ is at most $\varepsilon_{\mathrm{rel}} := 2d^5\varepsilon/(LR^2) \in (0, 1)$. The cited theorem, applied to $\mathcal{B}$ with the condition number $L/\mu_2 = d^2$, gives $N + 2 \ge c_{\mathrm{NY}}(d/\log d)\log(1/\varepsilon_{\mathrm{rel}})$, which is (69). □

Proposition F.5 and these lower bounds determine the deterministic complexity of the smooth strongly convex class on the ball for every $n \ge 2$ and $\Lambda \ge 8$, up to factors polylogarithmic in n and Λ, and for both first-order oracles. To state the result, let

$$\bar{n} = \min\{n, \Lambda^{1/3}\}, \quad W(\varepsilon) = \min\Big\{\bar{n}, \Big(\frac{LR^2}{\varepsilon}\Big)^{1/3}\Big\}, \quad T(\varepsilon) = \log_+ \frac{LR^2}{\bar{n}^3\varepsilon}, \quad \vartheta_{\bar{n}} = \log(2 + \mathsf{c}_n\bar{n}^3), \tag{70}$$

and let $N_{\mathrm{J}}(\varepsilon)$ and $N_{\mathrm{G}}(\varepsilon)$ be the smallest numbers of queries with which a deterministic method, using global queries of $(f, \nabla f)$ or of ∇f respectively, guarantees $f(\widehat{x}) - \min_{B_1^n(R)} f \le \varepsilon$ on every convex $C^{1,1}$ function on $\mathbb{R}^n$ that is L-smooth in ℓ_2 and μ_1-strongly convex in ℓ_1, with output $\widehat{x} \in B_1^n(R)$.

Proposition F.9 (Deterministic complexity for $\Lambda \ge 8$)**.** *There are absolute constants $c, C > 0$ such that for all $n \ge 2$, $L, \mu_1, R > 0$ with $\Lambda = L/\mu_1 - n \ge 8$, $K = B_1^n(R)$ and all $\varepsilon > 0$, with c_{NY} from Proposition F.8,*

$$\begin{aligned} c\,\min\{1, c_{\mathrm{NY}}\}\,\frac{1 + W(\varepsilon) + \bar{n}T(\varepsilon)}{\log(2 + \bar{n})} \;\le\; N_{\mathrm{J}}(\varepsilon) \;\le\; N_{\mathrm{G}}(\varepsilon) \;\le\; C\,\Pi_{\mathrm{G}}\big[1 + W(\varepsilon) + \bar{n}T(\varepsilon)\big], \\ N_{\mathrm{J}}(\varepsilon) \;\le\; C\,\Pi_{\mathrm{J}}\big[1 + W(\varepsilon) + \bar{n}T(\varepsilon)\big], \end{aligned} \tag{71}$$

where $\Pi_{\mathrm{J}} = 1 + \log(2 + n) + (\mathsf{c}_n\vartheta_{\bar{n}})^{1/3}$ and $\Pi_{\mathrm{G}} = 1 + \log(2 + n) + \mathsf{c}_n^{1/3}\vartheta_{\bar{n}}^{7/3}$ are polylogarithmic in n and Λ. The lower bound $c(1 + W(\varepsilon))$ holds for randomized methods as well.

The two terms of (71) correspond to two regimes: a cubic-rate phase of $\min\{\bar{n}, (LR^2/\varepsilon)^{1/3}\}$ queries, which lasts until the accuracy $LR^2/\bar{n}^3$, and a high-accuracy phase of $\bar{n}\log(LR^2/(\bar{n}^3\varepsilon))$ queries, linear in the logarithm of the accuracy with the factor $\min\{n, \Lambda^{1/3}\}$. In particular, when $n \ge \Lambda^{1/3}$ and $\Lambda \ge n$, the count $\widetilde{O}(\kappa_1^{1/3}\log(1/\varepsilon))$ of Proposition F.2 is optimal for deterministic methods up to polylogarithmic factors. For randomized methods at fixed dimension, and for $\Lambda < 8$, the complexity is not determined here.

Proof of Proposition F.9. Write $A = LR^2$ and $S = 1 + W + \bar{n}T$ with $W = W(\varepsilon)$, $T = T(\varepsilon)$. The inequality $N_{\mathrm{J}} \le N_{\mathrm{G}}$ holds because the joint oracle is more informative. All upper bounds below use feasible queries only, and $\Lambda \ge 8$ with $n \ge 2$ gives $\bar{n} \ge 2$.

Upper bounds, $\varepsilon \ge A/\bar{n}^3$. Then $T = 0$ and $W = (A/\varepsilon)^{1/3} \le \bar{n}$. Proposition C.2 (joint oracle) gives $N_{\mathrm{J}} \le 1 + 24N(\varepsilon) = O(1 + [(A/\varepsilon)\mathsf{c}_n\log(2 + (A/\varepsilon)\mathsf{c}_n)]^{1/3}) \le C(1 + W(\mathsf{c}_n\vartheta_{\bar{n}})^{1/3}) \le C\Pi_{\mathrm{J}}S$, because $A/\varepsilon \le \bar{n}^3$; and Proposition E.8 gives $N_{\mathrm{G}} = O(1 + [(A/\varepsilon)\mathsf{c}_n]^{1/3}\log^{7/3}(2 + (A/\varepsilon)\mathsf{c}_n)) \le C(1 + W\mathsf{c}_n^{1/3}\vartheta_{\bar{n}}^{7/3}) \le C\Pi_{\mathrm{G}}S$.

Upper bounds, $\varepsilon < A/\bar{n}^3$. Then $W = \bar{n}$ and $T > 0$. If $\bar{n} = \Lambda^{1/3} \le n$, we first reach the accuracy $e_0 = A/\Lambda = A/\bar{n}^3$ with the convex methods, with the count just computed for $W = \bar{n}$, i.e. $O(\Pi_{\mathrm{J}}\bar{n})$

or $O(\Pi_{\mathrm{G}}\bar{n})$. We then run the phases of Proposition F.5 (global case, $\Lambda_* = 8\Lambda$) from the certified bound e_0 until it is at most ε. This takes $\lceil \log_2(e_0/\varepsilon) \rceil \le 1 + T/\log 2$ phases, each using $O(1 + (\Lambda \mathsf{c}_n \log(2+\Lambda \mathsf{c}_n))^{1/3}) = O(1+\bar{n}(\mathsf{c}_n \vartheta_{\bar{n}})^{1/3}) \le C\Pi_{\mathrm{J}}\bar{n}$ joint queries or $O(1+(\Lambda \mathsf{c}_n)^{1/3}\log^{7/3}(2+\Lambda \mathsf{c}_n)) \le C\Pi_{\mathrm{G}}\bar{n}$ gradient queries. The total is $O(\Pi_{\mathrm{J}}(1+\bar{n}+\bar{n}T))$, resp. $O(\Pi_{\mathrm{G}}(1+\bar{n}+\bar{n}T))$. If $\bar{n} = n \le \Lambda^{1/3}$, Proposition E.12 with $F = \nabla f$ gives $N_{\mathrm{G}} \le C[1 + n\log(2 + A/\varepsilon)]$, and $\log(2 + A/\varepsilon) \le \log(2+\bar{n}^3) + T \le 3\log(2+\bar{n}) + T$, so $N_{\mathrm{G}} \le C[1 + \bar{n}(3\log(2+\bar{n}) + T)] \le C'(1 + \log(2+n))S$, which is at most $C'\Pi_{\mathrm{J}}S$ and $C'\Pi_{\mathrm{G}}S$.

Lower bound $N_{\mathrm{J}} \ge 1$. The class contains $f_b(x) = \frac{L}{2}\|x\|_2^2 + \langle b, x\rangle$ for every b (it is L/n-strongly convex in ℓ_1, and $L/n \ge \mu_1$). A method with no query returns a point $\widehat{x}$ (random or not) that does not depend on b. For $b = \mp t e_1$, $t > 0$, $\min_K f_b \le f_b(\pm R e_1) = LR^2/2 - tR$ and $f_b(\widehat{x}) \ge \mp t\widehat{x}_1$, so $f_b(\widehat{x}) - \min_K f_b \ge t(R \mp \widehat{x}_1) - LR^2/2$; the sum over the two signs is $2tR - LR^2$, so for one of them the (expected) error is at least $tR - LR^2/2$, which exceeds ε for large t. Hence $N_{\mathrm{J}} \ge 1$, also for randomized methods.

Lower bound of order W. Let $c_0 = 2^{-19}$ and $N = \lfloor c_0 W \rfloor$. If $N \ge 1$, then $W \ge 2^{19}$, so $n \ge W \ge 2^{19}$ and $32N + 1 \le 32c_0 n + 1 \le n$, and $2^{51}N^3 \le 2^{51}c_0^3 W^3 \le 2^{-6}\Lambda \le \Lambda$ since $W^3 \le \bar{n}^3 \le \Lambda$. Hence Proposition D.7 applies with this N and gives, for every method with at most N queries (randomized methods included), a function of the class on which the expected error is at least $A/(2^{52}N^3) \ge A/(2^{52}c_0^3 W^3) = 32A/W^3 \ge 32\varepsilon > \varepsilon$, using $W^3 \le A/\varepsilon$. Therefore every method that guarantees ε makes more than $\lfloor c_0 W \rfloor$ queries, and with the previous paragraph $N_{\mathrm{J}} \ge \max\{1, c_0 W\} \ge \frac{1}{2}(1 + c_0 W)$.

Lower bound of order $\bar{n}T/\log(2+\bar{n})$. Let $d = \lfloor \bar{n} \rfloor$, so $2 \le d \le \bar{n} \le n$, $\bar{n}/2 \le d$, and $d^3 - d \le \bar{n}^3 \le \Lambda$. Suppose first $T \ge 6\log(2+\bar{n})$. Then $A/(2d^5\varepsilon) = (A/(\bar{n}^3\varepsilon))\,\bar{n}^3/(2d^5) \ge e^T\bar{n}^3/(2\bar{n}^5) = e^T/(2\bar{n}^2) > 1$, so Proposition F.8 applies, and $\log \frac{A}{2d^5\varepsilon} \ge T - 2\log\bar{n} - \log 2 \ge T - 3\log(2+\bar{n}) \ge T/2$. Hence $N_{\mathrm{J}} + 2 \ge c_{\mathrm{NY}} \frac{d}{\log d} \cdot \frac{T}{2} \ge c_{\mathrm{NY}} \frac{\bar{n}T}{4\log(2+\bar{n})}$, and since $N_{\mathrm{J}} \ge 1$, $N_{\mathrm{J}} \ge \frac{1}{3}(N_{\mathrm{J}} + 2) \ge c_{\mathrm{NY}} \frac{\bar{n}T}{12\log(2+\bar{n})}$. If instead $0 < T < 6\log(2+\bar{n})$, then $W = \bar{n}$ (as $\varepsilon < A/\bar{n}^3$) and $\bar{n}T/\log(2+\bar{n}) < 6\bar{n} = 6W$, so the previous paragraph gives $N_{\mathrm{J}} \ge \frac{c_0}{6} \cdot \frac{\bar{n}T}{\log(2+\bar{n})}$. If $T = 0$ there is nothing to prove. Thus $N_{\mathrm{J}} \ge \max\{1, c_0 W, \min\{c_{\mathrm{NY}}/12, c_0/6\}\,\bar{n}T/\log(2+\bar{n})\}$, so $N_{\mathrm{J}} \ge \frac{1}{3}\min\{1, c_{\mathrm{NY}}\}\frac{c_0}{12}\big(1 + W + \bar{n}T/\log(2+\bar{n})\big)$, and $S/\log(2+\bar{n}) \le 1 + W + \bar{n}T/\log(2+\bar{n})$ because $\log(2+\bar{n}) \ge 1$. This proves the left-hand side of (71) with $c = 2^{-25}$. □

G Variational inequalities and saddle-point problems

This section extends the deep-cut methods to monotone operators: Lipschitz operators (Section G.1), operators with a model whose error is a power of the distance, including Hölder operators and Taylor models of higher order (Section G.2), and the matching lower bound for Lipschitz operators (Section G.3). As before, $K \subseteq c + B_1^n(R)$ is a known polytope.

G.1 Lipschitz operators

Let $F : K \to \mathbb{R}^n$ and consider the variational inequality of finding $x^\star \in K$ with $\langle F(x^\star), x - x^\star\rangle \ge 0$ for all $x \in K$. For a monotone F we measure accuracy by the *weak (Minty) gap* $\sup_{u\in K}\langle F(u), \bar{w} - u\rangle$; for a convex–concave Φ on $K = X \times Y$ with $F = (\nabla_x\Phi, -\nabla_y\Phi)$ we measure the *saddle gap* $\max_{y\in Y}\Phi(\bar{x}, y) - \min_{x\in X}\Phi(x, \bar{y})$. Both are certified by the mixture certificate of Algorithm 3. If $\alpha \in \Delta_T$ satisfies $\max_{u\in K}\sum_t \alpha_t \langle F(w_t), w_t - u\rangle \le \varepsilon$, then $\bar{w} = \sum_t \alpha_t w_t$ has weak gap at most ε when F is monotone, and saddle gap at most ε in the convex–concave case (Lemma G.1). Two methods produce such certificates.

Lemma G.1 (Mixture certificate)**.** *Let $w_1, \dots, w_T \in K$, $\alpha \in \Delta_T$, $\bar{w} = \sum_t \alpha_t w_t$, and suppose $\sum_t \alpha_t \langle F(w_t), w_t - u\rangle \le \varepsilon$ for all $u \in K$.*

(a) If F is monotone on K, then $\langle F(u), \bar{w} - u\rangle \le \varepsilon$ for all $u \in K$.

(b) If $K = X \times Y$, $F(x, y) = (g, -h)$ with $g \in \partial_x\Phi(x, y)$ and $-h \in \partial_y(-\Phi)(x, y)$ for a function Φ that is convex in x and concave in y, then $\sup_{y\in Y}\Phi(\bar{x}, y) - \inf_{x\in X}\Phi(x, \bar{y}) \le \varepsilon$.

Algorithm 7 SP-VI: extragradient step with deep cuts (horizon N)

Input: known polytope $K \subseteq c + B_1^n(R)$; operator oracle, F L-Lipschitz on K; horizon N; $\varepsilon_N = 6LP_N^2/N^2$; $\eta_N = P_N/(4N)$

$C_0 \leftarrow K$; $z_0 \leftarrow$ an η_N-approximate Steiner point of K
for $t = 0, 1, \dots$ **do**
  query $F(z_t)$; $w_t \leftarrow \mathrm{proj}_K\big(z_t - F(z_t)/(2L)\big)$
  query $F(w_t)$; $C_{t+1} \leftarrow C_t \cap \{u : \langle F(w_t), w_t - u\rangle \ge \varepsilon_N\}$
  if $C_{t+1} = \varnothing$ **then**
    solve $\min_{\alpha \in \Delta_{t+1}} \max_{u \in K} \sum_{j \le t} \alpha_j \langle F(w_j), w_j - u\rangle$ (a linear program)
    return $\bar{w} = \sum_{j \le t} \alpha_j w_j$
  $z_{t+1} \leftarrow$ an η_N-approximate Steiner point of C_{t+1}

Proof. (a) Monotonicity gives $\langle F(u), w_t - u\rangle \le \langle F(w_t), w_t - u\rangle$; multiplying by α_t and summing, with $\sum_t \alpha_t(w_t - u) = \bar{w} - u$, gives the claim. (b) For $(x, y) \in K$ and $w_t = (x_t, y_t)$, convexity in x and concavity in y give $\Phi(x_t, y) - \Phi(x, y_t) \le \langle g_t, x_t - x\rangle - \langle h_t, y_t - y\rangle = \langle F(w_t), w_t - (x, y)\rangle$. Multiplying by α_t, summing, and using convexity in x and concavity in y on the left ($\sum_t \alpha_t \Phi(x_t, y) \ge \Phi(\bar{x}, y)$ and $\sum_t \alpha_t \Phi(x, y_t) \le \Phi(x, \bar{y})$ respectively) gives $\Phi(\bar{x}, y) - \Phi(x, \bar{y}) \le \varepsilon$ for all $(x, y) \in K$. □

Proposition G.2 (Bounded and Lipschitz operators)**.** *Let $K \subseteq c + B_1^n(R)$ be a known polytope.*

(a) If $\|F(x)\|_2 \le G$ on K, then Algorithm 3 with the cuts $\{y : \langle F(x_t), x_t - y\rangle \ge \varepsilon_N\}$, $\varepsilon_N = 4GP_N/N$, stops after at most N operator calls and returns $\bar{w} \in K$ with the certificate above. No monotonicity is needed for the certificate.

(b) If F is L-Lipschitz in ℓ_2 on K, then SP-VI (Algorithm 7), which at the current approximate Steiner point z_t computes $w_t = \mathrm{proj}_K(z_t - F(z_t)/(2L))$ and cuts with $\{u : \langle F(w_t), w_t - u\rangle \ge \varepsilon_N\}$, $\varepsilon_N = 6LP_N^2/N^2 = 96LR^2\mathsf{b}_N\mathsf{c}_n/N^2$, stops after at most $2N$ operator calls with the certificate above. With a budget of $Q \ge 2$ calls the weak gap of a monotone operator, and the saddle gap of a convex–concave problem, is $O(LR^2(1 + \log Q)(1 + \log(2n))/Q^2)$.

(c) If moreover $\langle F(x) - F(y), x - y\rangle \ge \mu_1 \|x - y\|_1^2$ on K, then restarting (b) on the localized polytopes $K \cap (\bar{w}_j + B_1^n(R2^{-j-1}))$ returns a point at ℓ_1-distance at most $d_\star$ from the unique solution after $O\big([1 + \{\kappa_1 \log(2n) \log(2 + \kappa_1 \log(2n))\}^{1/2}][1 + \log_+(R/d_\star)]\big)$ calls, $\kappa_1 = L/\mu_1$.

The method of part (b) is an extragradient step (Korpelevich, 1976) whose second point is used as a deep cut. In the mismatched setting its rate $\widetilde{O}(LR^2/N^2)$ improves on the Euclidean $O(LR^2/N)$ of Mirror-Prox (Nemirovski, 2004). For $F = \nabla f$ the weak gap of $\bar{w}$ is at most $f(\bar{w}) - \min_K f$, so lower bounds for minimization do not transfer to the weak gap. The skew-symmetric construction of Proposition G.7 gives the matching deterministic bound $LR^2/(8(2N + 3)^2)$ for $n \ge 4N + 6$; hence the exponent N^{-2} of (b) is optimal up to the factor $\log(2N) \log(2n)$. The saddle gap of a bilinear game on $\Delta_{n_x} \times \Delta_{n_y}$ with a Euclidean operator bound is a special case of (b), and Proposition G.7(b) shows that its rate is also optimal in this sense. In small dimension the center-of-gravity branch CG-Cut of Proposition E.12 applies to operators without change. It returns a mixture with the same certificate after $O(1 + n \log(2 + LR^2/\varepsilon))$ operator calls, without any bound on $\|F(0)\|$, so the weak gap of a monotone L-Lipschitz operator on $B_1^n(R)$ can be brought below ε with

$$O\Big(1 + \min\Big\{\Big[\frac{LR^2\mathsf{c}_n}{\varepsilon} \log\Big(2 + \frac{LR^2\mathsf{c}_n}{\varepsilon}\Big)\Big]^{1/2},\ n \log\Big(2 + \frac{LR^2}{\varepsilon}\Big)\Big\}\Big)$$

calls.

Proof of Proposition G.2. (a) We run Algorithm 3 with the cut $C_t = C_{t-1} \cap \{y : \langle F(x_t), x_t - y\rangle \ge \varepsilon_N\}$ and selector accuracy $\eta = \varepsilon_N/(2G)$. If $C_t \ne \varnothing$ then $\mathrm{st}(C_t) \in C_t$ gives $\langle F(x_t), x_t - \mathrm{st}(C_t)\rangle \ge \varepsilon_N$, so $\|\mathrm{st}(C_t) - x_t\|_2 \ge \varepsilon_N/G$ and $\|\mathrm{st}(C_t) - \mathrm{st}(C_{t-1})\|_2 \ge \varepsilon_N/(2G)$; as in Proposition E.1, Lemma 3.1 forces an empty set within N queries. Emptiness of C_T means $\max_{u \in K} \min_{t \le T} \langle F(x_t), x_t - u\rangle < \varepsilon_N$, and the minimax theorem gives $\alpha \in \Delta_T$ with

$\max_{u\in K}\sum_t \alpha_t \langle F(x_t), x_t - u\rangle < \varepsilon_N$, computed by the linear program of Algorithm 3. Lemma G.1 gives the gaps.

(b) *One nonempty cut.* We drop the index t and write $r = w - z$, $d = z' - z$ for the next selected point $z' \in C_{t+1}$, and $v = -F(z) - 2Lr$. By the optimality condition of the projection, v belongs to the normal cone $N_K(w) = \{g : \langle g, u - w\rangle \le 0\ \forall u \in K\}$. With $e = F(w) - F(z)$, $\|e\|_2 \le L\|r\|_2$, we have $F(w) = -2Lr + e - v$. Since $z' \in K$, $\langle -v, w - z'\rangle \le 0$, and the cut gives

$$\begin{aligned}\varepsilon_N \le \langle F(w), w - z'\rangle \le \langle -2Lr + e, r - d\rangle &\le -2L\|r\|_2^2 + 2L\|r\|_2\|d\|_2 + L\|r\|_2^2 + L\|r\|_2\|d\|_2 \\ &= -L\|r\|_2^2 + 3L\|r\|_2\|d\|_2 \le \tfrac{9}{4}L\|d\|_2^2,\end{aligned}$$

using $w - z' = r - d$ and $-\tau^2 + 3\tau\sigma \le \frac{9}{4}\sigma^2$. Hence every nonempty transition has $\|z_{t+1} - z_t\|_2 \ge \frac{2}{3}\sqrt{\varepsilon_N/L}$. If $C_1, \dots, C_N$ were all nonempty, (40) would give $\frac{2N}{3}\sqrt{\varepsilon_N/L} \le \frac{3}{2}P_N$, i.e. $\varepsilon_N \le \frac{81}{16}LP_N^2/N^2$, contradicting $\varepsilon_N = 6LP_N^2/N^2$. Hence the method stops within N trials and $2N$ calls, and the certificate and the gaps follow as in (a). For a budget $Q \ge 2$ we take $N = \lfloor Q/2\rfloor$; then $\varepsilon_N = 96LR^2\mathsf{b}_N\mathsf{c}_n/N^2 = O(LR^2(1+\log Q)(1+\log(2n))/Q^2)$.

(c) Existence of a solution $x^\star$ follows from continuity and compactness, and uniqueness from strong monotonicity. Let $K' \subseteq K$ be a known polytope containing $x^\star$ and let the certificate of (b) on K' hold with right-hand side δ. Testing it at $u = x^\star$ and using $\langle F(w_t), w_t - x^\star\rangle \ge \langle F(x^\star), w_t - x^\star\rangle + \mu_1\|w_t - x^\star\|_1^2 \ge \mu_1\|w_t - x^\star\|_1^2$ (by strong monotonicity and the variational inequality at $x^\star$, as $w_t \in K' \subseteq K$) and the convexity of $\|\cdot\|_1^2$,

$$\delta \ge \sum_t \alpha_t \langle F(w_t), w_t - x^\star\rangle \ge \mu_1 \sum_t \alpha_t\|w_t - x^\star\|_1^2 \ge \mu_1\|\bar w - x^\star\|_1^2.$$

With $\rho_j = R2^{-j}$, $K_0 = K$ and $\delta_j = \mu_1\rho_j^2/4$, the output $\bar w_j$ of phase j satisfies $\|\bar w_j - x^\star\|_1 \le \rho_j/2 = \rho_{j+1}$, so $K_{j+1} = K \cap (\bar w_j + B_1^n(\rho_{j+1}))$ contains $x^\star$; the projection in phase j is onto K_j. The horizon $N_\star$, the smallest power of two with $96L\rho_j^2\mathsf{b}_N\mathsf{c}_n/N^2 \le \mu_1\rho_j^2/4$, i.e. $N^2 \ge 384\kappa_1\mathsf{b}_N\mathsf{c}_n$, does not depend on j, and $N_\star = O(1+[\kappa_1\mathsf{c}_n\log(2+\kappa_1\mathsf{c}_n)]^{1/2})$ as in the proof of Proposition F.2. For the target $d_\star > 0$ let J be the smallest index with $R2^{-J-1} \le d_\star$; after $J+1$ phases $\|\bar w_J - x^\star\|_1 \le d_\star$ with at most $2(J+1)N_\star$ calls. □

G.2 Operator models with power-law error

The deep-cut argument of Proposition G.2(b) applies to any operator model whose error is a power of the distance. Let $F : K \to \mathbb{R}^n$ be continuous and monotone, and suppose that one oracle call at $z \in K$ returns $F(z)$ and a continuous model $T_z : K \to \mathbb{R}^n$, which can be evaluated to any prescribed accuracy, such that

$$\|F(w) - T_z(w)\|_2 \le M\|w - z\|_2^r \qquad (z, w \in K) \tag{72}$$

for some $M, r > 0$; computations with T_z are not counted as calls. Two classes satisfy (72). If F is ν-Hölder with constant $L_{0,\nu}$, $0 < \nu \le 1$, take $T_z \equiv F(z)$, $r = \nu$ and $M = L_{0,\nu}$. If $k \ge 1$, F is C^k on a neighborhood of K and D^kF is ν-Hölder with constant $L_{k,\nu}$ in the Euclidean operator norm, $0 \le \nu \le 1$, an order-k call returns $(F(z), DF(z), \dots, D^kF(z))$, and the Taylor model $T_z(w) = \sum_{j\le k}\frac{1}{j!}D^jF(z)[(w-z)^j]$ satisfies (72) with $r = k + \nu$ and $M = L_{k,\nu}\prod_{j=1}^k(\nu + j)^{-1}$ (see the paragraph on Taylor models below).

Proposition G.3 (Operator models)**.** *Under (72), the variant of SP-VI described below makes at most $2N$ feasible calls and returns the points $w_1, \dots, w_T \in K$ of its $T \le N$ trials and weights $\alpha \in \Delta_T$ with*

$$\max_{u\in K}\sum_{t=1}^T \alpha_t \langle F(w_t), w_t - u\rangle \le \varepsilon_N = 2M\Big(\frac{6P_N}{N}\Big)^{r+1} = O_r\Big(\frac{MR^{r+1}[\log(2N)\log(2n)]^{(r+1)/2}}{N^{r+1}}\Big).$$

Hence $\bar w = \sum_t \alpha_t w_t$ has weak gap at most ε_N, and saddle gap at most ε_N for a convex–concave saddle operator. Each trial uses a finite model search whose arithmetic cost is not bounded.

For $r = 1$ this recovers the rate N^{-2} of Proposition G.2(b), with a worse constant. For ν-Hölder operators ($k = 0$) the Euclidean weak gap of Mirror-Prox-type methods is $O(N^{-(1+\nu)/2})$ (Stonyakin

et al., 2022), so on the ℓ_1 ball the exponent doubles. For Lipschitz D^kF, $k \geq 1$, extragradient-type methods reach weak gap $O(N^{-(k+2)/2})$ (Monteiro & Svaiter, 2012; Bullins & Lai, 2022; Lin & Jordan, 2025), and Zhang et al. (2026a) obtain the tangent residual $\widetilde{O}(N^{-(3k+2)/2})$, with a matching lower bound in Euclidean geometry; after multiplication by the diameter this bounds the weak gap, and against it the exponent $k+2$ of Proposition G.3 is larger only for $k=1$. A matching lower bound for Proposition G.3 is known only for $T_z \equiv F(z)$ and $r=1$ (Proposition G.7); for $r=\nu<1$ the affine instance of Proposition G.7 gives only $\Omega(N^{-2})$, and for Taylor models it is identified by one call that returns F and DF.

Proposition G.4 (Restarts under ℓ_1-strong monotonicity). *Assume (72) and $\langle F(x)-F(y), x-y\rangle \geq \mu_1\|x-y\|_1^2$ on K with $\mu_1>0$, and let $x^\star$ be the solution of the variational inequality. Let $0<d_\star<R$, $J=\lceil \log_2(R/d_\star)\rceil$ and $\rho_j = R2^{-j}$, and let N_j be the smallest power of two with*

$$N_j^{r+1} \geq 8\cdot 24^{r+1}\frac{M}{\mu_1}\rho_j^{r-1}(\mathsf{b}_{N_j}\mathsf{c}_n)^{(r+1)/2}. \tag{73}$$

Restarting Proposition G.3 on $K_0=K$ and $K_{j+1}=K\cap(\bar{w}_j+B_1^n(\rho_{j+1}))$ returns $x\in K$ with $\|x-x^\star\|_1\leq d_\star$ after at most $2\sum_{j<J}N_j$ feasible calls. With the condition number $\bar\kappa(\rho)=M\rho^{r-1}/\mu_1$ at scale ρ, $\bar\kappa_\star=\bar\kappa(d_\star)$ for $r<1$ and $\bar\kappa_\star=\bar\kappa(R)$ for $r\geq 1$, and $\mathcal{H}_\star=[\mathsf{c}_n\log(2+\bar\kappa_\star\mathsf{c}_n^{(r+1)/2})]^{1/2}$, this number is $O_r(J+\mathcal{H}_\star\bar\kappa(d_\star)^{1/(r+1)})$ for $r<1$, $O(J[1+\mathcal{H}_\star(M/\mu_1)^{1/2}])$ for $r=1$, and $O_r(J+\mathcal{H}_\star\bar\kappa(R)^{1/(r+1)})$ for $r>1$.

Taylor models. Let $k\geq 1$ and let F be C^k on a neighborhood of K with $\|D^kF(w)-D^kF(z)\|\leq L_{k,\nu}\|w-z\|_2^\nu$, and let $h=w-z$. The integral form of the remainder,

$$F(w)-T_z(w)=\int_0^1\frac{(1-t)^{k-1}}{(k-1)!}\big(D^kF(z+th)-D^kF(z)\big)[h^k]\,dt,$$

and the beta integral of Lemma E.5 give $\|F(w)-T_z(w)\|_2 \leq L_{k,\nu}\|h\|_2^{k+\nu}\int_0^1\frac{(1-t)^{k-1}t^\nu}{(k-1)!}\,dt = L_{k,\nu}\prod_{j=1}^k(\nu+j)^{-1}\|h\|_2^{k+\nu}$, which is (72) with $r=k+\nu$.

The model step. Let $\psi_r(a)=\|a\|_2^{r-1}a$ for $a\neq 0$ and $\psi_r(0)=0$; this map is continuous for every $r>0$. At a selector point z put $\widetilde{T}_z(w)=T_z(w)+2M\psi_r(w-z)$, the regularized model of higher-order methods for variational inequalities (Monteiro & Svaiter, 2012; Bullins & Lai, 2022; Lin & Jordan, 2025), and, for $\eta>0$, find $w\in K$ with

$$\max_{u\in K}\left\langle \widetilde{T}_z(w), w-u\right\rangle \leq \eta. \tag{74}$$

This condition concerns the model only and is checked by one linear program over K. The method then queries $F(w)$ and uses $F(w)$, not $\widetilde{T}_z(w)$, in the cut.

Lemma G.5 (Model search). *For every $z\in K$ and $\eta>0$, a point satisfying (74) exists and is found by a finite enumeration of rational convex combinations of the vertices of K. Monotonicity of $\widetilde{T}_z$ is not needed.*

Proof. The map $w\mapsto \mathrm{proj}_K(w-\widetilde{T}_z(w))$ sends the compact convex set K continuously into itself, so by Brouwer's theorem it has a fixed point w^*. The optimality condition of the projection gives $\left\langle \widetilde{T}_z(w^*), u-w^*\right\rangle \geq 0$ for all $u\in K$, so the left-hand side of (74) vanishes at w^*. This left-hand side is the maximum over the vertices u of K of the continuous functions $w\mapsto\left\langle \widetilde{T}_z(w), w-u\right\rangle$, hence continuous, and rational convex combinations of the vertices are dense in K. Enumerate them, evaluate the residual with absolute error at most $\eta/8$, and accept the first point whose computed residual is at most $\eta/2$. An accepted point has residual at most $5\eta/8$, and every candidate close enough to w^* has residual below $\eta/8$ and is accepted, so the enumeration stops. □

Lemma G.6 (Displacement forced by a cut). *Assume (72) and (74). If $z'\in K$ and $\langle F(w), w-z'\rangle \geq \varepsilon$, then*

$$\varepsilon \leq \eta + M\big(3\|z'-z\|_2\big)^{r+1}. \tag{75}$$

Proof. Put $a = w - z$, $b = z' - z$ and $e = F(w) - T_z(w)$, so that $F(w) = \widetilde{T}_z(w) - 2M\psi_r(a) + e$, $\|e\|_2 \le M\|a\|_2^r$ and $w - z' = a - b$. By (74) with $u = z'$ and the Cauchy–Schwarz inequality,

$$\begin{aligned}\varepsilon \le \langle F(w), a - b\rangle &\le \eta - 2M\|a\|_2^{r+1} + 2M\|a\|_2^r\|b\|_2 + M\|a\|_2^r(\|a\|_2 + \|b\|_2)\\ &= \eta + M\|a\|_2^r\big(3\|b\|_2 - \|a\|_2\big).\end{aligned}$$

If $\|a\|_2 \ge 3\|b\|_2$, the last term is nonpositive. Otherwise $\|a\|_2^r \le (3\|b\|_2)^r$ and $3\|b\|_2 - \|a\|_2 \le 3\|b\|_2$. □

Proof of Proposition G.3. We run Algorithm 7 with the same nested sets and selector accuracy $P_N/(4N)$, with the projected step replaced by a point w_t satisfying (74) at $z = z_t$ with $\eta = \varepsilon_N/2$, and with the deep cut $C_{t+1} = C_t \cap \{u : \langle F(w_t), w_t - u\rangle \ge \varepsilon_N\}$. Each trial makes one call at z_t, which returns the model, and one at w_t; the model search and the selector make no calls. If $C_1, \dots, C_N$ were all nonempty, Lemma G.6 with $z' = z_{t+1}$ would give $M(6P_N/N)^{r+1} = \varepsilon_N/2 \le M(3\|z_{t+1} - z_t\|_2)^{r+1}$, so every displacement would be at least $2P_N/N$ and their sum at least $2P_N$, contrary to the bound $\frac{3}{2}P_N$ of (40). Hence some C_T, $T \le N$, is empty, and the minimax theorem gives $\alpha \in \Delta_T$ with the stated certificate, computed by a linear program as in Proposition G.2(a). Substituting $P_N = 4R\sqrt{\mathsf{b}_N \mathsf{c}_n}$ gives the rate, and Lemma G.1 gives the weak and the saddle gap. □

Proof of Proposition G.4. By Brouwer's theorem, applied as in Lemma G.5 with F in place of $\widetilde{T}_z$, the variational inequality has a solution, which is unique by strong monotonicity. Suppose that $K_j \subseteq K$ is a known polytope containing $x^\star$ and contained in an ℓ_1 ball of radius ρ_j (for $j = 0$, the ball $c + B_1^n(R)$). Proposition G.3 on K_j with horizon N_j gives a certificate with right-hand side $2M(24\rho_j\sqrt{\mathsf{b}_{N_j}\mathsf{c}_n}/N_j)^{r+1}$, which is at most $\mu_1\rho_j^2/4$ by (73). As in the proof of Proposition G.2(c), testing the certificate at $u = x^\star$ gives $\mu_1\|\bar{w}_j - x^\star\|_1^2 \le \mu_1\rho_j^2/4$, so $\|\bar{w}_j - x^\star\|_1 \le \rho_{j+1}$ and K_{j+1} contains $x^\star$. After J phases the distance is at most $\rho_J \le d_\star$, and phase j uses at most $2N_j$ calls.

For the explicit count, put $\bar{\kappa}_j = \bar{\kappa}(\rho_j)$. Bounding $\mathsf{b}_N = O(1 + \log N)$ as in the proof of Proposition E.3 shows that (73) holds for every power of two above a suitable multiple, depending only on r, of $1 + \bar{\kappa}_j^{1/(r+1)}(\mathsf{c}_n \log(2 + \bar{\kappa}_j \mathsf{c}_n^{(r+1)/2}))^{1/2}$; since $\bar{\kappa}_j \le \bar{\kappa}_\star$ (for $r < 1$ because $\rho_{J-1} > d_\star$), this gives $N_j = O_r(1 + \bar{\kappa}_j^{1/(r+1)}\mathcal{H}_\star)$. For $r = 1$ all $\bar{\kappa}_j$ equal M/μ_1. For $r < 1$ the sequence $\bar{\kappa}_j^{1/(r+1)}$ increases geometrically with ratio $2^{(1-r)/(r+1)}$ and its sum is at most a constant times $\bar{\kappa}(d_\star)^{1/(r+1)}$; for $r > 1$ it decreases geometrically and its sum is at most a constant times $\bar{\kappa}(R)^{1/(r+1)}$. The constants contain $(1 - 2^{-|r-1|/(r+1)})^{-1}$ and are not uniform as $r \to 1$. □

G.3 A LOWER BOUND FOR MONOTONE OPERATORS

As noted after Proposition G.2, lower bounds for minimization do not transfer to the weak gap, so a separate construction is needed. It is affine and skew-symmetric, so the weak and the strong gap coincide, and it uses a difference matrix whose rotations are revealed two columns per query.

Proposition G.7 (Lower bound for monotone operators)**.** *Let $L, R, R_x, R_y > 0$, $N \ge 1$ and $m = 2N + 3$.*

(a) Let $n \ge 2m = 4N + 6$. For every deterministic method that makes at most N global queries of the operator and returns $\widehat{z} \in B_1^n(R)$ there is an affine operator $F(z) = Hz + b$ with $H^\top = -H$ and $\|H\|_{2\to 2} \le L$, hence monotone and L-Lipschitz on $\mathbb{R}^n$, with a zero $z^\star$ of F in $B_1^n(R/2)$, such that

$$\max_{u \in B_1^n(R)} \langle F(u), \widehat{z} - u\rangle = \max_{u \in B_1^n(R)} \langle F(\widehat{z}), \widehat{z} - u\rangle \ge \frac{LR^2}{8(2N+3)^2} \ge \frac{LR^2}{200N^2}.$$

(b) Let $n_x, n_y \ge m$. For every deterministic method that makes at most N global queries of $(\Phi, \nabla_x\Phi, \nabla_y\Phi)$ and returns $(\widehat{x}, \widehat{y}) \in B_1^{n_x}(R_x) \times B_1^{n_y}(R_y)$ there is a bilinear-affine $\Phi(x, y) = \frac{L}{2}\langle x, Ay - \gamma e_1\rangle$ with $\|A\|_{2\to 2} \le 2$, whose operator $(\nabla_x\Phi, -\nabla_y\Phi)$ is L-Lipschitz, such that the saddle gap $\max_{y \in B_1^{n_y}(R_y)} \Phi(\widehat{x}, y) - \min_{x \in B_1^{n_x}(R_x)} \Phi(x, \widehat{y})$ is at least $LR_xR_y/(4(2N+3)^2)$. On $\Delta_{2m}(R_x) \times \Delta_{2m}(R_y)$ the function $\widetilde{\Phi}(p, q) = \frac{1}{2}\Phi(Pp, Pq)$ with $P = (I_m, -I_m)$ has an L-Lipschitz operator and saddle gap at least $LR_xR_y/(8(2N+3)^2)$ for every deterministic method with at most N global queries.

Part (a) shows that the exponent of the rate $\widetilde{O}(LR^2/N^2)$ of SP-VI (Proposition G.2) is optimal for deterministic methods when $n \geq 4N + 6$. The logarithmic gap remains open, and the construction gives no randomized bound, since its last columns depend on the whole transcript of the method. Part (b) gives the same for the saddle gap of bilinear games on products of ℓ_1 balls and of simplices, with the operator measured in the spectral norm. The constant 8 on simplices comes from the factor $\frac{1}{2}$ in $\widetilde{\Phi}$, which keeps the lifted operator L-Lipschitz because $\|(I, -I)\|_{2\to 2} = \sqrt{2}$. At fixed dimension and large N no such bound can hold, since an affine operator is identified by $n + 1$ queries.

Proof of Proposition G.7. Let $m = 2N + 3$ and let $B \in \mathbb{R}^{m\times m}$ be the lower bidiagonal difference matrix, $(Bv)_i = v_i - v_{i-1}$ with $v_0 = 0$. Then $B = I - S$ for the shift S with $\|S\|_{2\to 2} \leq 1$, so $\|B\|_{2\to 2} \leq 2$; $B\mathbf{1} = e_1$; and $\sum_i (Bv)_i = v_m$ by telescoping.

A resisting pair of rotations. Fix a deterministic method with N queries $(x_t, y_t) \in \mathbb{R}^m \times \mathbb{R}^m$, which receives answers of the form $\big(c(Ay_t - \gamma e_1),\ -cA^\top x_t,\ c\,\langle x_t, Ay_t - \gamma e_1\rangle\big)$ with known $c, \gamma > 0$ and $A = UBV^\top$, and returns $(\widehat{x}, \widehat{y})$. We construct orthogonal U, V with $Ue_1 = Ve_1 = e_1$ column by column. Put $u_1 = v_1 = e_1$. Before query t the columns $u_1, \dots, u_{2t-1}$ and $v_1, \dots, v_{2t-1}$ are fixed, and the transcript so far determines x_t, y_t. Let u_{2t} be the normalized component of x_t orthogonal to $\operatorname{span}\{u_1, \dots, u_{2t-1}\}$ (any unit vector orthogonal to that span if the component vanishes), and v_{2t} likewise for y_t; then $x_t \in \operatorname{span}\{u_1, \dots, u_{2t}\}$ and $y_t \in \operatorname{span}\{v_1, \dots, v_{2t}\}$. Choose u_{2t+1}, v_{2t+1} as arbitrary further orthonormal vectors. Now $V^\top y_t$ is supported on the first $2t$ coordinates, so $BV^\top y_t$ is supported on the first $2t + 1$ coordinates and $Ay_t = U(BV^\top y_t)$ uses only $u_1, \dots, u_{2t+1}$; similarly $U^\top x_t$ is supported on the first $2t$ coordinates, $B^\top U^\top x_t$ as well, and $A^\top x_t$ uses only $v_1, \dots, v_{2t}$. Hence the answer to query t, including the value, is determined by the columns already fixed and is unchanged by any later completion. After N queries, $2N + 1 = m - 2$ columns of each matrix are fixed and the output $(\widehat{x}, \widehat{y})$ is determined; choose u_{m-1}, v_{m-1} from the orthogonal components of $\widehat{x}, \widehat{y}$ as above, and complete with u_m, v_m. The resulting single matrix A reproduces the entire transcript (the method, being deterministic, makes the same queries and the same output against A), and $\widehat{y} \in \operatorname{span}\{v_1, \dots, v_{m-1}\}$, i.e. $(V^\top \widehat{y})_m = 0$. A method that stops early is padded with ignored queries.

The key inequality. Put $\widehat{v} = V^\top \widehat{y}$. Since $\widehat{v}_m = 0$, $\sum_i (B\widehat{v} - \gamma e_1)_i = \widehat{v}_m - \gamma = -\gamma$, so $\|B\widehat{v} - \gamma e_1\|_1 \geq \gamma$ and $\|B\widehat{v} - \gamma e_1\|_2 \geq \gamma/\sqrt{m}$. As $Ue_1 = e_1$, $A\widehat{y} - \gamma e_1 = U(B\widehat{v} - \gamma e_1)$, and therefore

$$\|A\widehat{y} - \gamma e_1\|_\infty \ \geq\ \frac{\|A\widehat{y} - \gamma e_1\|_2}{\sqrt{m}} \ \geq\ \frac{\gamma}{m}. \tag{76}$$

(b) Saddle gap on two balls. Let $c = L/2$, $\gamma = R_y/(2m)$ and $\Phi(x, y) = c\,\langle x, Ay - \gamma e_1\rangle$ on $\mathbb{R}^m \times \mathbb{R}^m$, which is bilinear-affine, hence convex–concave. Its operator is $F(x, y) = H(x, y) + b$ with $H = c\left(\begin{smallmatrix} 0 & A \\ -A^\top & 0 \end{smallmatrix}\right)$, $H^\top = -H$ and $\|H\|_{2\to 2} = c\|A\|_{2\to 2} \leq L$, so F is monotone and L-Lipschitz. The point $(x^\star, y^\star) = (0, \gamma V\mathbf{1})$ satisfies $Ay^\star = \gamma UBV^\top V\mathbf{1} = \gamma UB\mathbf{1} = \gamma e_1$, so $\Phi(x, y^\star) = 0 = \Phi(0, y)$ for all x, y: it is a saddle point of value 0, and $\|y^\star\|_1 \leq \sqrt{m}\|y^\star\|_2 = m\gamma = R_y/2$, so it is feasible. For the output of the method (whose queries receive the answers used in the construction),

$$\begin{aligned}\max_{y\in B_1^m(R_y)} \Phi(\widehat{x}, y) - \min_{x\in B_1^m(R_x)} \Phi(x, \widehat{y}) &\geq \Phi(\widehat{x}, y^\star) - \min_{x\in B_1^m(R_x)} c\,\langle x, A\widehat{y} - \gamma e_1\rangle \\ &= cR_x\|A\widehat{y} - \gamma e_1\|_\infty \ \geq\ \frac{LR_xR_y}{4m^2}\end{aligned}$$

by (76). If $n_x > m$ or $n_y > m$, let Φ depend on the first m coordinates of each block only: a global query is answered from its first m coordinates, the padded operator is still L-Lipschitz, and the output's first m coordinates are feasible. Since $2N + 3 \leq 5N$, the bound is at least $LR_xR_y/(100N^2)$.

Simplices. Let $P = (I_m, -I_m)$, so that $P\Delta_{2m}(r) = B_1^m(r)$ for every $r > 0$ and $\|P\|_{2\to 2} = \sqrt{2}$, and let $\widetilde{\Phi}(p, q) = \frac{1}{2}\Phi(Pp, Pq) = \frac{L}{4}\,\langle Pp, APq - \gamma e_1\rangle$. Its operator is $\frac{L}{4}\left(\begin{smallmatrix} 0 & P^\top AP \\ -P^\top A^\top P & 0 \end{smallmatrix}\right)(p, q) + \text{const}$, of spectral norm at most $\frac{L}{4} \cdot 2 \cdot 2 = L$. A method for $\widetilde{\Phi}$ with global queries (p_t, q_t) defines a method for Φ with the queries (Pp_t, Pq_t), since the answers for $\widetilde{\Phi}$ are those for Φ multiplied by known matrices; an output $(\widehat{p}, \widehat{q})$ on the simplices gives the feasible output $(P\widehat{p}, P\widehat{q})$ on the balls. Because P maps each simplex onto the corresponding ball, the saddle gap of $(\widehat{p}, \widehat{q})$ for $\widetilde{\Phi}$ equals half the saddle

gap of $(P\widehat{p}, P\widehat{q})$ for Φ, which is at least $LR_xR_y/(8m^2)$ by the previous paragraph applied to the simulating method.

(a) Weak gap on one ball. Let $K = B_1^{2m}(R)$, $c = L/2$, $\gamma = R/(2m)$, and let $F(z) = Hz + b$ be the operator of Φ above, now on the single ball. The point $z^\star = (0, \gamma V\mathbf{1})$ has $F(z^\star) = 0$ and $\|z^\star\|_1 \le R/2$. For any affine skew-symmetric operator, $\langle Hu, u\rangle = 0$ and $\langle Hu, z\rangle = -\langle Hz, u\rangle$ give

$$\max_{u\in K} \langle F(u), z-u\rangle = \langle b, z\rangle + \max_{u\in K} \langle -(Hz+b), u\rangle = \langle b, z\rangle + R\|F(z)\|_\infty = \max_{u\in K} \langle F(z), z-u\rangle ,$$

so the weak and the strong gap coincide. As $b = -Hz^\star$ and $\langle z^\star, b\rangle = -\langle z^\star, Hz^\star\rangle = 0$, $\langle b, z\rangle = \langle z^\star, Hz\rangle = \langle z^\star, F(z)\rangle$, hence $\max_{u\in K}\langle F(u), z-u\rangle \ge (R - \|z^\star\|_1)\|F(z)\|_\infty \ge \frac{R}{2}\|F(z)\|_\infty$. The method queries the full operator, which is the answer form used in the construction with x_t, y_t the two blocks of the query; its output $\widehat{z} = (\widehat{x}, \widehat{y})$ satisfies (76), and the first block of $F(\widehat{z})$ is $c(A\widehat{y} - \gamma e_1)$, so $\|F(\widehat{z})\|_\infty \ge c\gamma/m$ and the weak gap is at least $\frac{R}{2}\cdot\frac{L}{2}\cdot\frac{R}{2m^2} = \frac{LR^2}{8m^2} \ge \frac{LR^2}{200N^2}$. For $n > 2m$ let F act on the first $2m$ coordinates and vanish on the others; the operator remains affine, skew-symmetric, monotone and L-Lipschitz, the support function of $B_1^n(R)$ is still $R\|\cdot\|_\infty$, so the gap identity and the bound hold with the first $2m$ coordinates of the output. □

H Convex quadratics

The Steiner-point methods settle the general smooth class on the ℓ_1 ball up to logarithmic factors. For convex quadratics a different mechanism, *curvature learning*, gives the rate $O_p(LR^2/N^{1+2/p})$ without logarithms, on every ℓ_p ball with $1 \le p < 2$ and in every dimension. The matching lower bound, and with it optimality of the exponent, needs $n \ge 3N + 2$ (Proposition H.9); some such condition is necessary, since in dimension n a quadratic is identified by $n+1$ gradient queries (Arjevani et al., 2016). Throughout this section $1 \le p < 2$, $K \subseteq B_p^n(1)$ (the general radius is recovered by scaling), and

$$f(x) = \tfrac{1}{2}\langle x, Qx\rangle + \langle b, x\rangle , \qquad 0 \preceq Q \preceq LI, \tag{77}$$

with L known and Q unknown. The method needs b and products Qd, both obtained from gradients: $b = \nabla f(0)$, and $Qd = \nabla f(d) - \nabla f(0)$ for the global quadratic. Thus T products require $T+1$ first-order queries. The query at 0 and the probe points d need not lie in K; these are the only infeasible queries used by an upper bound in this paper; on $B_p^n(R)$ itself the query at 0 and the scaled probe hd with $h\|d\|_p \le R$ are feasible.

H.1 Mechanism

An accelerated estimate sequence needs a regularizer with quadratic curvature, while the comparison with the minimizer needs a regularizer that is bounded on $B_p^n(1)$. For $p < 2$ the function $\frac{1}{2}\|x\|_2^2$ is bounded on the ball but gives only the Euclidean rate. We use, for a threshold $0 < \tau \le 1$ and $c_p = \frac{2-p}{2p}$, the *power–Huber function*

$$\psi_{\tau,p}(s) = \begin{cases} \frac{1}{2}\tau^{p-2}s^2, & |s| \le \tau, \\ \frac{1}{p}|s|^p - c_p\tau^p, & |s| > \tau, \end{cases} \qquad \Psi_I(x) = \sum_{j\notin I}\psi_{\tau,p}(x_j) + \frac{1}{p}\sum_{j\in I}|x_j|^p, \tag{78}$$

where $I \subseteq \{1, \dots, n\}$ is the set of *promoted* coordinates.

Lemma H.1 (Power–Huber regularizer)**.** *The function $\psi_{\tau,p}$ is convex and continuously differentiable, Ψ_I is convex, and on $B_p^n(1)$:* (i) $0 \le \Psi_I(x) \le 1/p$; (ii) $\psi_{\tau,p}(s) \ge s^2/2$ *for* $|s| \le 1$; (iii) $\psi_{\tau,p}$ *has curvature* τ^{p-2} *on* $[-\tau, \tau]$; (iv) *if* $|x_j| \ge \tau$, *moving* j *into* I *increases* $\Psi_I(x)$ *by exactly* $c_p\tau^p$ *and does not change* $\partial\Psi_I(x)$.

Acceleration uses property (iii) and the comparison with the minimizer uses property (i); the two are compatible only if coordinates that leave $[-\tau, \tau]$ are promoted. By (iv) a promotion keeps the current model minimizer optimal and increases the potential by $c_p\tau^p$, so at most $O(\tau^{-p})$ promotions can occur. The method *learns* curvature: at every stage it queries the product Qx_{t-1} at the previous iterate and, after a promotion, the product Qe_j for every newly promoted j, and it updates a lower

Algorithm 8 CLA: curvature-learning acceleration for quadratics on $B_p^n(1)$

Input: compact convex $K \subseteq B_p^n(1)$, $1 \le p < 2$; L and b; product oracle $d \mapsto Qd$; budget T with $T+1 \ge 6/(2-p)$

$\tau \leftarrow [6/((2-p)(T+1))]^{1/p}$, $L_\tau \leftarrow L\tau^{2-p}$, $\Gamma_0 \leftarrow \tau^{p-2}$, $H_0 \leftarrow 0$, calls $\leftarrow 0$
$x_0 \leftarrow$ a minimizer of $L_\tau \Gamma_0 \Psi_\varnothing(x) + \langle b, x\rangle$ over K; $\bar{x}_0 \leftarrow x_0$; $I_0 \leftarrow U_0 \leftarrow \{j : |x_{0,j}| \ge \tau\}$
for $t = 1, 2, \dots$ **do**
  $H \leftarrow H_{t-1}$
  for $d = e_j$, $j \in U_{t-1}$, and then $d = x_{t-1}$ **do**
    query Qd; calls $\leftarrow$ calls $+ 1$ ▷ redundant directions are also counted
    if calls $= T$ **then return** $\bar{x}_{t-1}$
    update H by (79)
  $H_t \leftarrow H$; $\bar{\gamma}_t \leftarrow$ the root in $(0,1)$ of $\bar{\gamma}^2 = \Gamma_{t-1}(1-\bar{\gamma})$
  if some $x \in X_t(0)$ has $r_t(x) \ge \tau$ **then**
    $\gamma_t \leftarrow 0$, $x_t \leftarrow x$ ▷ promotion step
  else if $Y_t(\bar{\gamma}_t) \ne \varnothing$ **then**
    $\gamma_t \leftarrow \bar{\gamma}_t$, $x_t \in Y_t(\bar{\gamma}_t)$ ▷ full accelerated step
  else
    $\gamma_t \leftarrow \max\{\gamma \in [0, \bar{\gamma}_t] : Y_t(\gamma) \ne \varnothing\}$; $x_t \in Y_t(\gamma_t)$ with $r_t(x_t) = \tau$ ▷ partial step
  $\Gamma_t \leftarrow (1-\gamma_t)\Gamma_{t-1}$; $\bar{x}_t \leftarrow (1-\gamma_t)\bar{x}_{t-1} + \gamma_t x_t$
  $U_t \leftarrow \{j \notin I_{t-1} : |x_{t,j}| \ge \tau\}$ if $\gamma_t < \bar{\gamma}_t$, else $U_t \leftarrow \varnothing$; $I_t \leftarrow I_{t-1} \cup U_t$

model $0 \preceq H \preceq Q$ by the positive semidefinite SR1 step

$$v = (Q-H)d, \qquad H^+ = \begin{cases} H + \dfrac{vv^\top}{\langle d, v\rangle}, & \langle d, v\rangle > 0, \\ H, & \langle d, v\rangle = 0. \end{cases} \tag{79}$$

The update preserves $0 \preceq H \preceq H^+ \preceq Q$, makes $(Q - H^+)d = 0$, and keeps all previously learned null directions of $Q - H$. Consequently the residual $f - h_t$, $h_t(x) = \frac{1}{2}\langle x, H_t x\rangle + \langle b, x\rangle$, is a positive semidefinite quadratic that vanishes on the span of the history, including the accelerated averages at which the estimate sequence evaluates it. Every stage without a full accelerated step promotes a coordinate, and every promoted coordinate costs one product, so T products and m promoted coordinates leave at least $T - 2m - 1$ accelerated steps; the threshold is tuned so that $\tau^p \asymp 1/N$ and $m \le N/3$:

$$\underbrace{N^{-2}}_{\text{acceleration}} \cdot \underbrace{\tau^{2-p}}_{\text{inverse curvature of } \psi_{\tau,p}} \asymp N^{-2} \cdot N^{-(2-p)/p} = N^{-(1+2/p)}. \tag{80}$$

At $p = 1$ this is the inverse-cubic mechanism of Ouyang (2026), whose regularizer is the Huber function; for $p > 1$ the power branch $|s|^p/p$ makes (i) and (iv) hold simultaneously.

H.2 The method and its rate

For a stage $t \ge 1$, a step parameter $\gamma \in [0, 1)$ and the current promoted set, the *model* is

$$m_t(\gamma; x) = L_\tau\, \Gamma_{t-1}(1-\gamma)\, \Psi_{I_{t-1}}(x) + h_t(x), \qquad r_t(x) = \max_{j \notin I_{t-1}} |x_j|, \tag{81}$$

with $\max \varnothing = 0$, $L_\tau = L\tau^{2-p}$, and we write $X_t(\gamma) = \arg\min_K m_t(\gamma; \cdot)$ and $Y_t(\gamma) = \{x \in X_t(\gamma) : r_t(x) \le \tau\}$. Algorithm 8 states the method.

The step selection is well defined (Section H.4.3), and every stage with $\gamma_t < \bar{\gamma}_t$ promotes at least one new coordinate. The algorithm assumes exact minimization of the known models and an exact choice of the largest admissible γ; both assumptions are removed below.

Proposition H.2 (Rate of CLA). *Let $1 \le p < 2$ and let T be an integer with $T + 1 \ge 6/(2-p)$. After T products, Algorithm 8 returns $\bar{x} \in K$ with*

$$f(\bar{x}) - \min_K f \;\le\; \frac{C_p L}{(T+1)^{2/p}(T-2)}, \qquad C_p = \frac{9}{2p}\Big(\frac{6}{2-p}\Big)^{(2-p)/p}. \tag{82}$$

Table 3: Quadratics on $B_p^n(R)$: error after N queries, and queries per halving of the error for strongly convex quadratics.

p	1	4/3	3/2	2
convex rate	N^{-3}	$N^{-5/2}$	$N^{-7/3}$	N^{-2}
queries per halving	$\kappa_1^{1/3}$	$\kappa_{4/3}^{2/5}$	$\kappa_{3/2}^{3/7}$	$\kappa_2^{1/2}$

For $K \subseteq B_p(c,R)$ and a quadratic with Euclidean smoothness L, the method applied to $u \mapsto f(c+Ru)$ uses at most $T+1$ first-order queries and the bound becomes $C_pLR^2/[(T+1)^{2/p}(T-2)]$. The quadratic lower bound of Proposition H.9 (dimension $n \geq 3N+2$) shows that the exponent $1+2/p$ is optimal for every fixed $p<2$.

At $p=1$, $C_1=27$ and (82) is the bound of Ouyang (2026). The constant tends to $9/4$ as $p \to 2$, while the smallest admissible T, $\lceil 6/(2-p) \rceil - 1$, diverges; at $p=2$ Euclidean acceleration gives the matching N^{-2}. Table 3 lists some exponents: the upper bounds are Proposition H.2 and Corollary H.5 (Euclidean acceleration for $p=2$), and the matching lower bounds are Propositions H.9 and H.11.

A certified implementation. Exact model minimization and the exact largest admissible step can be replaced by finite procedures without changing the exponent. The certified version of CLA (Algorithm 9 in Section H.4.5) makes three changes: (i) each model, stabilized by $\frac{\zeta}{2}\|x\|_2^2$, is solved only to an η-certificate (Definition H.3); (ii) the step is selected by a bisection that stops as soon as the largest unpromoted coordinate lies in the band $[(1-\sigma)\tau, \tau]$, $\sigma = c_p\tau/16$, and coordinates are promoted from $(1-\sigma)\tau$ on; (iii) the SR1 matrix is replaced by the ridge model

$$H_\delta(S) = QV_S\big(V_S^\top QV_S + \delta I\big)^{-1}V_S^\top Q, \tag{83}$$

where V_S is a matrix whose columns form an orthonormal basis of the learned span S. Unlike (79), the ridge model needs no exact test $\langle d, v\rangle = 0$; it satisfies $0 \preceq H_\delta(S) \preceq Q$, increases with S, and $\langle v, (Q - H_\delta(S))v\rangle \leq \delta\|v\|_2^2$ for $v \in S$ (Lemma H.8). The clipped-coordinate inequality (91) shows that an objective certificate controls the clipped coordinates, which are all the threshold test needs, although $\psi_{\tau,p}$ has little curvature far from the origin.

For inexact inner minimization we use the following certificate, which can be checked by one linear minimization over K.

Definition H.3 (Certified inner solution)**.** Let $\mathcal{J}$ be a known convex function on K and $\eta \geq 0$. A triple $(x, g, \underline{s})$ with $x \in K$, $g \in \partial\mathcal{J}(x)$ and $\underline{s} \leq \min_{y\in K}\langle g, y\rangle$ is an η-certificate for $\mathcal{J}$ if $\langle g, x\rangle - \underline{s} \leq \eta$. It implies $\mathcal{J}(x) - \min_K \mathcal{J} \leq \eta$.

Proposition H.4 (Certified implementation)**.** *Let $1 \leq p < 2$, let $T+1 \geq 24/(2-p)$, and assume the certificate interface of Definition H.3 is available on K. With the parameters of (90), the certified version of CLA performs finitely many certified inner solves and exact-real comparisons, uses T products, and returns $\bar{x} \in K$ with*

$$f(\bar{x}) - \min_K f \leq \frac{\widetilde{C}_pL}{(T+1)^{2/p}(T-2)}, \qquad \widetilde{C}_p = \frac{18}{p}\Big(\frac{24}{2-p}\Big)^{(2-p)/p}, \qquad \widetilde{C}_1 = 432. \tag{84}$$

H.3 Restarts and the quadratic assumption

Restarts under quadratic growth. Suppose that f satisfies the quadratic growth condition

$$f(x) - f^\star \geq \frac{\mu_p}{2}\operatorname{dist}_p(x, X^\star)^2 \qquad (x \in K), \tag{85}$$

where $X^\star = \arg\min_K f$ and $\operatorname{dist}_p(x, X^\star) = \min_{y\in X^\star}\|x-y\|_p$; strong convexity in ℓ_p implies it. If $f(z) - f^\star \leq \Delta$, every point of $X^\star$ closest to z lies in $B_p(z,r)$ with $r^2 = 2\Delta/\mu_p$, so CLA can be restarted on $K \cap B_p(z,r)$.

Corollary H.5 (Restarted CLA)**.** *Assume (85), a point $z_0 \in K$ with $f(z_0) - f^\star \leq \Delta_0$, and $0 < \varepsilon < \Delta_0$. Let $\alpha_p = 1 + 2/p$ and*

$$M_p = \Big\lceil \max\Big\{\frac{6}{2-p} - 1,\ \big(8C_pL/\mu_p\big)^{1/\alpha_p}\Big\}\Big\rceil.$$

Running Algorithm 8 with budget M_p on $K \cap B_p(z_j, r_j)$, $r_j^2 = 2^{1-j}\Delta_0/\mu_p$, and setting z_{j+1} to its output, gives an ε-solution after at most $(M_p+1)\lceil \log_2(\Delta_0/\varepsilon)\rceil = O_p\big((1+\kappa_p^{p/(p+2)})\log(2\Delta_0/\varepsilon)\big)$ first-order queries. No uniqueness of the minimizer is needed.

By Proposition H.11, the power $\kappa_p^{p/(p+2)}$ cannot be improved for a constant-factor reduction. Restarting under quadratic growth is standard (Fercoq & Qu, 2019).

Why only quadratics. The proof of Proposition H.2 uses that the residual $f - h_t$ is a fixed positive semidefinite quadratic vanishing on every learned direction. Off the quadratic class this invariant can fail in one dimension: for $f = \log\cosh$ the secant slope learned between 0 and $a > 0$ is $H = \tanh(a)/a$, while $f''(a) = \operatorname{sech}^2 a < H$, so $f - \frac{1}{2}Hx^2$ is not convex near a. The Steiner-point methods do not model curvature and store only first-order cuts, so the difficulty does not arise for them.

H.4 Proofs for curvature learning

Throughout this subsection $1 \le p < 2$, $K \subseteq B_p^n(1)$ is nonempty, compact and convex, and $f(x) = \frac{1}{2}\langle x, Qx\rangle + \langle b, x\rangle$ with $0 \preceq Q \preceq LI$. The vector b and the bound L are known, and one call returns a product Qd. For a global quadratic one call takes one gradient query, and obtaining b takes one more (see the beginning of this section). Exact minimization of known models and the exact step selection of Algorithm 8 are internal operations; Section H.4.5 replaces them by finite procedures.

H.4.1 The power–Huber regularizer

Proof of Lemma H.1. The two branches of $\psi_{\tau,p}$ have the same value and the same derivative $\pm\tau^{p-1}$ at $s = \pm\tau$, and both are convex; hence $\psi_{\tau,p}$ is convex and C^1, and Ψ_I is convex as a sum of convex functions. Property (iii) is immediate.

For (i), write $|s| = \tau u$ with $0 \le u \le 1$ inside the threshold; then $\frac{1}{2}\tau^{p-2}s^2 = \frac{1}{2}\tau^p u^2 \le \frac{1}{p}\tau^p u^p = |s|^p/p$ because $u^2 \le u^p$ and $\frac{1}{2} \le \frac{1}{p}$. Outside the threshold $\psi_{\tau,p}(s) = |s|^p/p - c_p\tau^p \le |s|^p/p$. Hence $0 \le \Psi_I(x) \le \sum_j |x_j|^p/p \le 1/p$ on $B_p^n(1)$; nonnegativity is clear on both branches, since $\tau^p/p - c_p\tau^p = \tau^p/2 > 0$ and $|s|^p/p$ increases.

For (ii), note that $\tau^{p-2} \ge 1$ inside the threshold. On $[\tau, 1]$ the function $\psi_{\tau,p}(s) - s^2/2$ has derivative $s^{p-1} - s \ge 0$ and value $(\tau^p - \tau^2)/2 \ge 0$ at $s = \tau$. The case $s < 0$ follows by symmetry.

For (iv), if $|x_j| \ge \tau$ then $|x_j|^p/p - \psi_{\tau,p}(x_j) = c_p\tau^p$ (at $|x_j| = \tau$ both formulas give $\tau^p/2$), and the derivatives of the two functions coincide at x_j; the other terms of Ψ_I do not change. □

H.4.2 The positive semidefinite SR1 update

Lemma H.6. *Let $0 \preceq H \preceq Q$, $d \in \mathbb{R}^n$, $v = (Q-H)d$, and define H^+ by (79). Then $0 \preceq H \preceq H^+ \preceq Q$, $(Q-H^+)d = 0$, and $(Q-H)u = 0$ implies $(Q-H^+)u = 0$.*

Proof. Write $E = Q - H \succeq 0$. If $\langle d, v\rangle = \langle d, Ed\rangle = 0$, the Cauchy–Schwarz inequality for the semidefinite form E gives $Ed = 0$, so $v = 0$ and $H^+ = H$. Otherwise $E^+ = E - Edd^\top E/\langle d, Ed\rangle$. Cauchy–Schwarz gives $(u^\top Ed)^2 \le (u^\top Eu)(d^\top Ed)$, i.e. $E^+ \succeq 0$, and $E^+ \preceq E$ is clear; thus $H \preceq H^+ \preceq Q$. Finally $E^+d = 0$, and $Eu = 0$ implies $E^+u = 0$. □

H.4.3 Well-posedness of the step selection

Fix a stage t and write $X(\gamma) = X_t(\gamma)$, $Y(\gamma) = Y_t(\gamma)$ and $r = r_t$. The graph of X over $[0, \bar\gamma_t]$ is compact, because K is compact and the model is jointly continuous in (γ, x). If one minimizer $x \in X(\gamma)$ satisfies $r(x) < \tau$, then every minimizer has the same unpromoted coordinates. Along the segment to another minimizer the model is constant, so each of its convex terms is affine there; but the unpromoted power–Huber terms are strictly convex near x in every direction that changes an unpromoted coordinate (their coefficient $L_\tau\Gamma_{t-1}(1-\gamma)$ is positive). By compactness, the strict inequality $r < \tau$ on $X(\gamma)$ persists for all γ' close to γ. Consequently, if $X(0)$ contains no point with

$r \geq \tau$, the set $\{\gamma : Y(\gamma) \neq \varnothing\}$ is compact and contains a neighborhood of 0 in $[0, \bar{\gamma}_t]$. If it does not contain $\bar{\gamma}_t$, its largest element γ_t lies in $(0, \bar{\gamma}_t)$, and some $x \in Y(\gamma_t)$ has $r(x) = \tau$, since otherwise the strict inequality would persist beyond γ_t. Hence every stage with $\gamma_t < \bar{\gamma}_t$ promotes at least one new coordinate.

H.4.4 The estimate sequence and the proof of the rate

Let $\lambda_t = L_\tau \Gamma_t$ with $L_\tau = L\tau^{2-p}$.

Lemma H.7 (Estimate sequence)**.** *At every completed stage $t \geq 0$ of Algorithm 8,*

$$f(\bar{x}_t) - f^\star \leq L\tau^{2-p}\Gamma_t\Big(\frac{1}{p} - c_p\tau^p|I_t|\Big), \qquad |I_t| \leq \frac{2}{(2-p)\tau^p}. \tag{86}$$

Proof. By Lemma H.1(iv), promotion does not change the subdifferential of the model at the current point, so x_t minimizes $\lambda_t \Psi_{I_t} + h_t$ over K. Put $W_t = \lambda_t \Psi_{I_t}(x_t) + h_t(x_t)$. If x and x_t have all unpromoted coordinates in $[-\tau, \tau]$, convex optimality and the quadratic branch of $\psi_{\tau,p}$ give

$$\lambda_t \Psi_{I_t}(x) + h_t(x) - W_t \;\geq\; \frac{L\Gamma_t}{2}\|(x - x_t)_{I_t^c}\|_2^2. \tag{87}$$

Let $\gamma_t > 0$; we apply (87) at stage $t-1$ with $x = x_t$. The increase $H_t - H_{t-1} \succeq 0$ does not decrease the quadratic part, and promotion increases the regularizer at x_t by $c_p\tau^p|U_t|$. Multiplying by $1 - \gamma_t$, adding $\gamma_t h_t(x_t)$ and using $\Gamma_t \geq \gamma_t^2$, we obtain

$$W_t \;\geq\; (1-\gamma_t)W_{t-1} + \gamma_t h_t(x_t) + \frac{L\gamma_t^2}{2}\|(x_t - x_{t-1})_{I_{t-1}^c}\|_2^2 + Lc_p\tau^2\Gamma_t|U_t|. \tag{88}$$

For $\gamma_t = 0$ the same inequality follows by minimizing the old model, using $H_t \succeq H_{t-1}$ and adding the promotion increment.

The residual $Q - H_t$ annihilates $x_0, \ldots, x_{t-1}$ and e_j for $j \in I_{t-1}$ (Lemma H.6), hence also $\bar{x}_{t-1}$. Expanding the quadratic form of $Q - H_t$ at $\bar{x}_t = \bar{x}_{t-1} + \gamma_t(x_t - x_{t-1}) + \gamma_t(x_{t-1} - \bar{x}_{t-1})$ and using convexity of h_t gives

$$f(\bar{x}_t) \;\leq\; (1-\gamma_t)f(\bar{x}_{t-1}) + \gamma_t h_t(x_t) + \frac{L\gamma_t^2}{2}\|(x_t - x_{t-1})_{I_{t-1}^c}\|_2^2. \tag{89}$$

(The promoted part of $x_t - x_{t-1}$ lies in the span of the learned coordinates and does not contribute.) We subtract (89) from (88) and divide by Γ_t. At initialization, Lemma H.1(ii) and $Q \preceq LI$ give

$$\frac{W_0 - f(x_0)}{\Gamma_0} = \frac{L\Psi_\varnothing(x_0) - \frac{1}{2}\langle x_0, Qx_0\rangle}{\Gamma_0} + Lc_p\tau^2|I_0| \;\geq\; Lc_p\tau^2|I_0|.$$

Telescoping yields $W_t - f(\bar{x}_t) \geq Lc_p\tau^2\Gamma_t|I_t|$. On the other hand, optimality of x_t, $H_t \preceq Q$ and $\Psi_{I_t}(x^\star) \leq 1/p$ give $W_t \leq f^\star + L\tau^{2-p}\Gamma_t/p$. The two inequalities give the gap bound, and nonnegativity of its left side gives the cardinality bound. □

Proof of Proposition H.2. Let J be the last completed stage, $m = |I_J|$, and let r be the number of completed stages with $\gamma_t = \bar{\gamma}_t$. Every other completed stage promotes a new coordinate, so $J - r \leq m$. Before stopping, at most $J + 1$ iterate directions and m coordinate directions were queried; hence

$$T \leq J + m + 1, \qquad r \geq T - 2m - 1, \qquad m \leq \frac{T+1}{3},$$

the last inequality by (86) and the choice $\tau^p = 6/((2-p)(T+1))$. If $m = (T+1)/3$, then (86) gives zero error. Otherwise $m \leq T/3$ and $T \geq 5$ imply $r \geq 1$. At a full stage, $\Gamma_t = \bar{\gamma}_t^2$ and

$$\frac{1}{\sqrt{\Gamma_t}} = \frac{1 + \sqrt{1 + 4/\Gamma_{t-1}}}{2} \;\geq\; \frac{1}{\sqrt{\Gamma_{t-1}}} + \frac{1}{2},$$

and the first full stage gives $\Gamma_t < 1$; at the other stages Γ does not increase. Thus $\Gamma_J \leq 4/(r+1)^2 \leq 4/(T-2m)^2$. Put $t_0 = T + 1 - 3m > 0$ and $u = T - 2 > 0$. By (86),

$$f(\bar{x}_J) - f^\star \;\leq\; \frac{4L\tau^{2-p}}{p(T+1)} \cdot \frac{t_0}{(T-2m)^2} \;\leq\; \frac{9L\tau^{2-p}}{2p(T+1)(T-2)},$$

where the last step uses $T - 2m = (u + 2t_0)/3$ and $(u + 2t_0)^2 \geq 8ut_0$. Substituting τ proves (82), since $1 + (2 - p)/p = 2/p$.

For $K \subseteq B_p(c, R)$ we substitute $x = c + Rz$; the Hessian becomes $R^2Q \preceq LR^2I$ and the linear coefficient $R\nabla f(c)$. One query at c gives this coefficient, and each product requires one query at $c + Rd$, since $R(\nabla f(c + Rd) - \nabla f(c)) = R^2Qd$. Optimality of the exponent follows from Proposition H.9, whose hard function is a convex quadratic. □

H.4.5 A CERTIFIED IMPLEMENTATION

In the certified version of Algorithm 8 the inner problems are solved only up to the certificate of Definition H.3, the step is chosen by a finite bisection, and exact null directions are replaced by a ridge-regularized model. Products Qd are exact. "Finite" refers to the number of inner calls and exact-real comparisons, not to the arithmetic cost of the certificates.

Inner certificates. If $(x, g, \underline{s})$ is an η-certificate for a convex $\mathcal{J}$, then for every $y \in K$, $\mathcal{J}(y) \geq \mathcal{J}(x) + \langle g, y - x\rangle \geq \mathcal{J}(x) - \eta$. For the stabilized objective $\widetilde{\mathcal{J}} = \mathcal{J} + \frac{\zeta}{2}\|x\|_2^2$, $\zeta > 0$, the exact minimizer X is unique and a certified point satisfies $\|x - X\|_2 \leq \sqrt{2\eta/\zeta}$. Removing the stabilizer increases the certificate by at most $\zeta/4$, because $\langle x, y - x\rangle \leq \|x\|_2 - \|x\|_2^2 \leq \frac{1}{4}$ for $x, y \in B_2(1)$. On the full balls and on the simplex such certificates can be found by finite enumeration of rational points, because the continuous VI gap tends to zero along a sequence converging to the minimizer. On a general compact convex K the certificate interface is assumed.

The ridge model. Let S be the span of the learned directions, V_S a matrix whose columns form an orthonormal basis of S, and $\delta > 0$ fixed for the whole run. Define $H_\delta(S)$ by (83) ($H_\delta(\{0\}) = 0$) and $h_S(x) = \frac{1}{2}\langle x, H_\delta(S)x\rangle + \langle b, x\rangle$. The model is determined by the products QV_S, and each new orthogonalized direction needs one new product.

Lemma H.8 (Ridge model)**.** *Let $0 \preceq Q$ and $\delta > 0$. Then $H_\delta(S)$ does not depend on the choice of the basis V_S, $0 \preceq H_\delta(S) \preceq Q$, $H_\delta(S) \preceq H_\delta(S')$ whenever $S \subseteq S'$, and $0 \leq \langle v, (Q - H_\delta(S))v\rangle \leq \delta\|v\|_2^2$ for every $v \in S$.*

Proof. Let $A = Q^{1/2}$ and $B = AV_S$. The push-through identity gives

$$H_\delta(S) = A^\top B\big(B^\top B + \delta I\big)^{-1}B^\top A = A^\top\big[I - \delta\big(BB^\top + \delta I\big)^{-1}\big]A, \qquad BB^\top = A\Pi_S A,$$

where $\Pi_S = V_S V_S^\top$ is the orthogonal projector onto S. This depends on S only. The matrix in brackets lies between 0 and I, which gives $0 \preceq H_\delta(S) \preceq A^\top A = Q$. If $S \subseteq S'$, then $\Pi_S \preceq \Pi_{S'}$, hence $(A\Pi_S A + \delta I)^{-1} \succeq (A\Pi_{S'} A + \delta I)^{-1}$ and $H_\delta(S) \preceq H_\delta(S')$. Finally, for $v = V_S y \in S$ put $G = B^\top B \succeq 0$; then $\|y\|_2 = \|v\|_2$ and

$$\langle v, (Q - H_\delta(S))v\rangle = y^\top\big[G - G(G + \delta I)^{-1}G\big]y = \delta\, y^\top G(G + \delta I)^{-1}y \;\leq\; \delta\|y\|_2^2,$$

since $0 \preceq G(G + \delta I)^{-1} \preceq I$. □

Parameters. For an integer T with $T + 1 \geq 24/(2 - p)$ we use

$$\begin{aligned} &\tau^p = \frac{24}{(2-p)(T+1)}, \quad L_\tau = 2L\tau^{2-p}, \quad \sigma = \frac{c_p\tau}{16}, \quad a = (1-\sigma)\tau, \qquad \xi = \frac{\sigma\tau}{8}, \\ &e = \frac{L_\tau}{p(T+1)^3}, \qquad \zeta = e, \qquad \delta = \frac{e}{18}, \quad \eta = \min\Big\{\frac{e}{4}, \frac{L\xi^2}{(T+1)^2}\Big\}. \end{aligned} \tag{90}$$

Here a is the lower promotion threshold and ξ a positional tolerance. For the current learned model $h_t = h_S$ let

$$J_t(\gamma; x) = L_\tau\Gamma_{t-1}(1 - \gamma)\Psi_{I_{t-1}}(x) + h_t(x) + \frac{\zeta}{2}\|x\|_2^2,$$

and let $\textsc{Solve}_t(\gamma)$ return an η-certified point for $J_t(\gamma; \cdot)$. Algorithm 9 states the certified version of CLA.

Algorithm 9 Certified implementation of CLA

Input: as in Algorithm 8, with $T+1 \geq 24/(2-p)$; exact products; the certificate interface for the known models

set the parameters (90); $S \leftarrow \{0\}$, calls $\leftarrow 0$, $\Gamma_0 \leftarrow \tau^{p-2}$
$x_0 \leftarrow$ an η-certified minimizer of $L_\tau \Gamma_0 \Psi_\varnothing + \langle b, \cdot \rangle + \frac{\zeta}{2}\|\cdot\|_2^2$; $\bar{x}_0 \leftarrow x_0$; $I_0 \leftarrow U_0 \leftarrow \{j : |x_{0,j}| \geq a\}$
for $t = 1, 2, \ldots$ **do**
  for $d = e_j$, $j \in U_{t-1}$, and then $d = x_{t-1}$ **do**
    query Qd; calls $\leftarrow$ calls $+ 1$;
    if calls $= T$ **then return** $\bar{x}_{t-1}$
    $S \leftarrow \operatorname{span}(S \cup \{d\})$
  build $h_t = h_S$ from the stored products; $\bar{\gamma}_t \leftarrow$ root of $\bar{\gamma}^2 = \Gamma_{t-1}(1-\bar{\gamma})$
  $\gamma_t \leftarrow 0$; $x_t \leftarrow \textsc{Solve}_t(0)$
  if $r_t(x_t) < a$ **then**
    $\gamma_t \leftarrow \bar{\gamma}_t$; $x_t \leftarrow \textsc{Solve}_t(\bar{\gamma}_t)$
    if $r_t(x_t) > \tau$ **then**
      $\gamma_- \leftarrow 0$, $\gamma_+ \leftarrow \bar{\gamma}_t$
      repeat
        $\gamma_t \leftarrow (\gamma_- + \gamma_+)/2$; $x_t \leftarrow \textsc{Solve}_t(\gamma_t)$
        if $r_t(x_t) < a$ then $\gamma_- \leftarrow \gamma_t$; if $r_t(x_t) > \tau$ then $\gamma_+ \leftarrow \gamma_t$
      until $a \leq r_t(x_t) \leq \tau$
  $\Gamma_t \leftarrow (1-\gamma_t)\Gamma_{t-1}$; $\bar{x}_t \leftarrow (1-\gamma_t)\bar{x}_{t-1} + \gamma_t x_t$
  $U_t \leftarrow \{j \notin I_{t-1} : |x_{t,j}| \geq a\}$ if $\gamma_t < \bar{\gamma}_t$, else $U_t \leftarrow \varnothing$; $I_t \leftarrow I_{t-1} \cup U_t$

Clipped coordinates. Let $\operatorname{clip}_\tau(s) = \max\{-\tau, \min\{s, \tau\}\}$. The Huber function $\tau\psi_{\tau,1}(s) = s\operatorname{clip}_\tau(s) - \operatorname{clip}_\tau(s)^2/2$ has derivative clip_τ, and checking the cases y inside, above and below the threshold interval gives

$$\tau D_{\psi_{\tau,1}}(y, x) = \tfrac{1}{2}\big(\operatorname{clip}_\tau(y) - \operatorname{clip}_\tau(x)\big)^2 + \big(y - \operatorname{clip}_\tau(y)\big)\big(\operatorname{clip}_\tau(y) - \operatorname{clip}_\tau(x)\big),$$

where the second term is nonnegative. The difference $\psi_{\tau,p} - \tau^{p-1}\psi_{\tau,1}$ is convex, since it vanishes on $[-\tau, \tau]$ and its derivative $\operatorname{sign}(s)(|s|^{p-1} - \tau^{p-1})$ is nondecreasing outside. Hence

$$D_{\psi_{\tau,p}}(y, x) \;\geq\; \tfrac{1}{2}\tau^{p-2}\big|\operatorname{clip}_\tau(y) - \operatorname{clip}_\tau(x)\big|^2. \tag{91}$$

Every tested step γ has $\Gamma_{\text{trial}} := \Gamma_{t-1}(1-\gamma) \geq (T+1)^{-2}$, because $\Gamma_{\text{trial}}^{-1/2}$ increases by at most one per stage and is at most one initially. Thus the coefficient of the unpromoted quadratic curvature is at least $L_\tau \Gamma_{\text{trial}} \tau^{p-2} \geq 2L/(T+1)^2$, and (91) together with the choice of η shows that the clipped unpromoted coordinates of a certified point are within Euclidean distance ξ of those of the exact stabilized minimizer $X(\gamma)$.

The bisection is finite. Subgradients of Ψ_I on $B_p^n(1)$ have Euclidean norm at most $n^{1/p-1/2}$. Strong monotonicity of the stabilized optimality condition gives $\|X(\gamma) - X(\gamma')\|_2 \leq L_\tau \Gamma_{t-1} n^{1/p-1/2}|\gamma - \gamma'|/\zeta$, and the clipped maximum $\widehat{r}(X) = \min\{r_t(X), \tau\}$ is Lipschitz with the same constant. A sampled left endpoint has $\widehat{r}(X) < a + \xi$ and a sampled right endpoint has $\widehat{r}(X) \geq \tau - \xi$; these differ by more than $\sigma\tau - 2\xi > 0$, so an interval shorter than $\zeta(\sigma\tau - 2\xi)/(L_\tau \Gamma_{t-1} n^{1/p-1/2})$ cannot have endpoints of both types, and the bisection stops. With rational thresholds $a_{\text{rat}} \in (a, a + \sigma\tau/8)$ and $t_{\text{rat}} \in (\tau - \sigma\tau/8, \tau)$ all comparisons become rational and the same argument applies.

Proof of Proposition H.4. For $|s| \geq a$, promotion adds at least $\pi_\sigma \tau^p$ to the regularizer, where $\pi_\sigma = \frac{(1-\sigma)^p}{p} - \frac{(1-\sigma)^2}{2} \geq c_p - \frac{\sigma^2}{2}$, because on $[1-\sigma, 1]$ the derivative of $r^p/p - r^2/2$ lies in $[0, 1-r]$. Promotion does not change the subgradient above τ and changes it by at most $\sigma\tau^{p-1}$ inside the band. Hence the returned point has VI residual, for the unstabilized promoted model, at most

$$\bar{\eta}_t = \alpha + 2\lambda_t \sigma \tau^{p-1}|U_t|, \qquad \alpha = \eta + \zeta/4 \leq e/2, \tag{92}$$

where $\lambda_t = L_\tau \Gamma_t$ and $\|y - x\|_\infty \le 2$ was used.

Let $W_t = \lambda_t \Psi_{I_t}(x_t) + h_t(x_t)$ and $u_t = (x_t - x_{t-1})_{I_{t-1}^c}$. For a positive step, the local curvature inequality with error $\bar{\eta}_{t-1}$ and the monotonicity of the ridge models give

$$W_t \;\ge\; (1-\gamma_t) W_{t-1} + \gamma_t h_t(x_t) + L\gamma_t^2 \|u_t\|_2^2 + L_\tau \Gamma_t \pi_\sigma \tau^p |U_t| - (1-\gamma_t)\bar{\eta}_{t-1},$$

where we used $\Gamma_t \ge \gamma_t^2$; the curvature coefficient is now $2L$. For a zero step the same inequality follows from the global VI residual.

The learned span contains all previous iterates and all e_j, $j \in I_{t-1}$. Hence $v = \bar{x}_t - \gamma_t u_t \in S$ and $\|v\|_2 \le 1 + 2\gamma_t \le 3$. Young's inequality in the seminorm of $Q - H_\delta(S)$ and the defect bound of Lemma H.8 give $\frac{1}{2}\langle \bar{x}_t, (Q - H_\delta(S))\bar{x}_t \rangle \le L\gamma_t^2 \|u_t\|_2^2 + 9\delta$, and therefore

$$f(\bar{x}_t) \;\le\; (1-\gamma_t) f(\bar{x}_{t-1}) + \gamma_t h_t(x_t) + L\gamma_t^2 \|u_t\|_2^2 + 9\delta, \qquad 9\delta = \frac{e}{2}.$$

At initialization $\Psi_\varnothing(x_0) \ge \|x_0\|_2^2/2$ and $Q \preceq LI$ give $W_0 - f(x_0) \ge L_\tau \Gamma_0 \pi_\sigma \tau^p |I_0|$. Subtracting the two inequalities, dividing by Γ_t, telescoping and using $W_J \le f^\star + L_\tau \Gamma_J / p + \bar{\eta}_J$, we get

$$f(\bar{x}_J) - f^\star \;\le\; L_\tau \Gamma_J \Big(\frac{1}{p} - \pi_\sigma \tau^p m \Big) + \Gamma_J \sum_{j=0}^{J} \frac{\bar{\eta}_j}{\Gamma_j} + \Gamma_J \sum_{t=1}^{J} \frac{9\delta}{\Gamma_t}, \qquad m = |I_J|. \tag{93}$$

The promotion part of $\bar{\eta}_j$ contributes $2L_\tau \Gamma_J \sigma \tau^{p-1} m$, and $\pi_\sigma - 2\sigma/\tau \ge c_p/2$. For every allowed step $1/\sqrt{\Gamma_t} \le 1/\sqrt{\Gamma_{t-1}} + 1$, so $\Gamma_j \ge (T+1)^{-2}$ for $j \le J \le T$, and the remaining normalized errors sum to at most $e(T+1)^3 = L_\tau / p$. Hence

$$f(\bar{x}_J) - f^\star \;\le\; \frac{2L_\tau \Gamma_J}{p} \Big(1 - \frac{3m}{T+1} \Big),$$

so $m \le (T+1)/3$, and the counting argument of Proposition H.2 gives $\Gamma_J \le 4/(r+1)^2$ and the bound $9L_\tau / [p(T+1)(T-2)]$. Substituting L_τ and τ proves (84). □

H.4.6 Restarts under quadratic growth

Proof of Corollary H.5. Let $\Delta_j = 2^{-j}\Delta_0$ and assume $f(z_j) - f^\star \le \Delta_j$. For the analysis only, let $x_j^\star \in X^\star$ be a point closest to z_j in ℓ_p. By (85), $\|z_j - x_j^\star\|_p \le r_j$ with $r_j^2 = 2\Delta_j/\mu_p$, so $K_j = K \cap B_p(z_j, r_j)$ contains a global minimizer and $\min_{K_j} f = f^\star$. Proposition H.2 with $T = M_p$ applies on K_j, and $M_p \ge 6/(2-p) - 1 \ge 5$. For $T \ge 5$, $(T+1)^{2/p}(T-2) \ge T^{\alpha_p}/2$, so the error is at most

$$\frac{2C_p L r_j^2}{M_p^{\alpha_p}} = \frac{4C_p L \Delta_j}{\mu_p M_p^{\alpha_p}} \;\le\; \frac{\Delta_j}{2}.$$

Each stage uses M_p products and one query for the linear coefficient. After $\lceil \log_2(\Delta_0/\varepsilon) \rceil$ stages the gap is at most ε. Only the common value $f^\star$ enters the induction, so the minimizer need not be unique. □

H.5 Lower bounds for quadratics

Proposition H.9 (Quadratic lower bound)**.** *Let $N \ge 1$, $n \ge 3N+2$, $1 \le p \le 2$ and $R, L > 0$, and put $n_0 = 3N+2$, $m = 2N+1$ and $S_m = \sum_{j=1}^m j^2$. For every deterministic method with at most N global first-order queries and output $\widehat{x} \in B_p^n(R)$ there is a convex quadratic f with $0 \preceq \nabla^2 f \preceq LI$ and a global minimizer in $B_p^n(R)$ such that*

$$f(\widehat{x}) - \min_{B_p^n(R)} f \;\ge\; \frac{L(N+1)R^2 n_0^{1-2/p}}{8 S_m} \;\ge\; \frac{LR^2}{8 \cdot 3^{2+2/p} (N+1)^{1+2/p}}. \tag{94}$$

Proof. Let $B_m z = (z_1 - z_2, \dots, z_m - z_{m+1})$ with $z_{m+1} = 0$; it is invertible and $\|B_m\|_{2\to 2} \le 2$. For $a > 0$ put $\Phi(z) = \frac{1}{2}\|B_m z\|_2^2 - a z_1$. In the variables $w = B_m z$ we have $z_1 = \sum_i w_i$, so every term $\frac{1}{2}w_i^2 - a w_i$ is minimized at $w_i = a$, and

$$z_i^\star = (m+1-i)a, \qquad \min \Phi = -\frac{m}{2}a^2, \qquad \min_{z_j = 0 \ (j>t)} \Phi = -\frac{t}{2}a^2. \tag{95}$$

Resisting oracle. All directions are chosen in the span of the first n_0 coordinate vectors. At the query x_t, choose a unit vector v_t in this span perpendicular to $x_1, \dots, x_t$ and $v_1, \dots, v_{t-1}$; these are at most $2t-1 < n_0$ linear constraints, and later directions are also perpendicular to all earlier queries. Hence $\langle v_j, x_t \rangle = 0$ for $j \geq t$, and since $B_m^\top B_m$ is tridiagonal, the gradient of Φ at $V^\top x_t$ involves only the directions $v_1, \dots, v_t$ chosen so far. The value and the gradient at x_t can therefore be returned consistently with every completion. After the output $\widehat{x}$, complete $v_1, \dots, v_N$ by $N+1$ orthonormal directions perpendicular to all queries, to the output and to the first N directions; the available dimension is at least $n_0 - (2N+1) = N+1$.

The hard function. Let $V = [v_1 \ \cdots \ v_m]$, $r = Rn_0^{1/2-1/p}$, $a = r/\sqrt{S_m}$ and $f(x) = \frac{L}{4}\Phi(V^\top x)$. Since $V^\top V = I$ and $\|B_m\|_{2\to 2} \leq 2$, f is a convex quadratic with $0 \preceq \nabla^2 f \preceq LI$, and it reproduces every answer. Its minimizer $x^\star = Vz^\star$ has Euclidean norm $a\sqrt{S_m} = r$ and at most n_0 nonzero coordinates, so $\|x^\star\|_p \leq n_0^{1/p-1/2} r = R$ by (11). The last $m - N = N+1$ chain coordinates $\langle v_j, \widehat{x} \rangle$ of the output vanish, so (95) gives

$$f(\widehat{x}) - \min f \ \geq \ \frac{L}{4} \cdot \frac{(m-N)a^2}{2} = \frac{L(N+1)r^2}{8S_m},$$

which is the first bound in (94). Finally $n_0 \leq 3(N+1)$, $1 - 2/p \leq 0$ and $S_m \leq m^3 \leq [3(N+1)]^3$ give the second. □

For strongly convex quadratics, a tridiagonal chain with an explicit inverse gives an exponential tail.

Proposition H.10 (Explicit exponential first-order tail)**.** *Let $L > \mu_2 > 0$, $R > 0$, $n \geq 3N+2$, $1 \leq p \leq 2$ and $N \geq 1$. Put $n_0 = 3N+2$ and $\rho = (\sqrt{L/\mu_2} - 1)/(\sqrt{L/\mu_2} + 1) \in (0,1)$. For every deterministic N-query first-order algorithm on $B_p^n(R)$ there is a globally L-smooth, globally Euclidean μ_2-strongly convex quadratic with a feasible global minimizer, on which the error of the algorithm is at least*

$$\frac{\mu_2 R^2 n_0^{1-2/p}}{4}(1+\rho)^2 \rho^{2N}(1-\rho^{2N+2}). \tag{96}$$

Proof. Set $\theta = (L-\mu_2)/4$, $m = 2N+1$, $H_j = \mu_2 I_j + \theta B_j^\top B_j$, and $s_j = e_1^\top H_j^{-1} e_1$. The diagonal of H_j is $(\mu_2+\theta, \mu_2+2\theta, \dots, \mu_2+2\theta)$ and its neighboring off-diagonal entries are $-\theta$, so H_N is the leading principal block of H_m, also when $N = 1$. The identity $\mu_2 = \theta(1-\rho)^2/\rho$ and substitution into this tridiagonal system give, for $1 \leq i \leq j$,

$$(H_j^{-1} e_1)_i = \frac{c}{1+\rho^{2j+1}}(\rho^{i-1} - \rho^{2j-i+1}), \qquad c = \frac{\rho}{\theta(1-\rho)}.$$

To check this, note that the two geometric sequences solve each interior recurrence, their difference has zero coordinate $j+1$, and the first row equals one with the displayed coefficient. Thus

$$s_j = c\frac{1-\rho^{2j}}{1+\rho^{2j+1}}, \qquad s_m - s_N = \frac{c(1+\rho)\rho^{2N}(1-\rho^{2(m-N)})}{(1+\rho^{2m+1})(1+\rho^{2N+1})}.$$

Let $w = H_m^{-1} e_1$, $r = Rn_0^{1/2-1/p}$, and choose the linear coefficient $b = r/\|w\|_2$ in $f(x) = \theta\|B_m V^\top x\|_2^2/2 - b\langle v_1, x\rangle + \mu_2\|x\|_2^2/2$. Its global minimizer is bVw, of norm r, hence feasible. The resisting oracle is that of Proposition H.9, and the hidden chain coordinates of the output vanish. The full and truncated unconstrained minima are $-b^2 s_m/2$ and $-b^2 s_N/2$, respectively, so the error is at least $r^2(s_m - s_N)/(2\|w\|_2^2)$. Finally,

$$\|w\|_2^2 \leq \frac{c^2}{(1+\rho^{2m+1})^2(1-\rho^2)}, \qquad \frac{1+\rho^{2m+1}}{1+\rho^{2N+1}} \geq \frac{1}{2}.$$

Combining these inequalities and using $c^{-1}(1-\rho^2)(1+\rho) = \mu_2(1+\rho)^2$ proves (96), since $m - N = N+1$. □

H.5.1 A MATCHING CONDITION-NUMBER POWER FOR CONSTANT REDUCTION

The explicit chain also bounds the initial gap, which gives a condition-number power matching restarted CLA for a constant-factor reduction.

Proposition H.11 (Strong quadratic constant-reduction lower bound)**.** *Fix $1 \le p \le 2$, $N \ge 1$, $R > 0$, and $\mu_p > 0$, and put*

$$n_0 = n = 3N + 2, \qquad L_N = \mu_p n_0^{\beta_p} (2N+1)^2.$$

For every deterministic algorithm using at most N first-order queries on $B_p^{n_0}(R)$, there is a globally L_N-smooth quadratic that is globally μ_p-strongly convex in $\|\cdot\|_p$, has a feasible unconstrained minimizer, and satisfies

$$f(0) - f^\star \le \frac{100}{27}\mu_p R^2 < 4\mu_p R^2, \qquad f(\widehat{x}) - f^\star \ge \frac{3}{64}\mu_p R^2. \tag{97}$$

The algorithm may be given the feasible start 0 and the certificate $\Delta_0 = 4\mu_p R^2$ in advance. All gradient queries, including any query at zero, are counted; queries anywhere in $\mathbb{R}^{n_0}$ are allowed. Consequently N queries cannot guarantee error at most $\Delta_0/128$.

Proof. Set $\mu_2 = \mu_p n_0^{\beta_p}$, $m = 2N+1$, $\theta = (L_N - \mu_2)/4$, and $r = R n_0^{-\beta_p/2}$. Use the construction of Proposition H.10, with $H_m = \mu_2 I + \theta B_m^\top B_m$, $w = H_m^{-1} e_1$, $b = r/\|w\|_2$, and

$$f(x) = \tfrac{\theta}{2}\|B_m V^\top x\|_2^2 - b\,\langle v_1, x\rangle + \tfrac{\mu_2}{2}\|x\|_2^2.$$

Its minimizer bVw has Euclidean norm r and p-norm at most R. Furthermore, $\mu_2\|h\|_2^2 \ge \mu_2 n_0^{-\beta_p}\|h\|_p^2 = \mu_p\|h\|_p^2$, proving the claimed global strong convexity. The Hessian upper bound is $\mu_2 + 4\theta = L_N$.

The parameter of Proposition H.10 is now $\rho = N/(N+1)$. For every integer $N \ge 1$,

$$\rho \ge \tfrac{1}{2}, \qquad \rho^{2N} \ge \tfrac{1}{9}, \qquad \rho^{2N+2} \le \tfrac{1}{4}. \tag{98}$$

The binomial theorem and $k! \ge 2^{k-1}$ for $k \ge 1$ give $(1+1/N)^N \le \sum_{k=0}^{N} 1/k! < 3$, and Bernoulli's inequality gives $(1+1/N)^{N+1} \ge 1 + (N+1)/N > 2$. Substituting (98) into (96), and using $\mu_2 r^2 = \mu_p R^2$, yields

$$f(\widehat{x}) - f^\star \ge \frac{\mu_p R^2}{4}(1+\rho)^2\rho^{2N}(1-\rho^{2N+2}) \ge \frac{3}{64}\mu_p R^2.$$

It remains to bound the initial gap without a condition-number factor. The inverse formula in the preceding proof, with $c = (1-\rho)/\mu_2$, gives

$$w_i = \frac{c(\rho^{i-1} - \rho^{2m-i+1})}{1+\rho^{2m+1}}, \qquad s_m = w_1 \le c, \qquad f(0) - f^\star = \frac{r^2 s_m}{2\|w\|_2^2}.$$

For $1 \le i \le N+1$, the ratio of the subtracted geometric term to the first is $\rho^{2m-2i+2} \le \rho^{2N+2} \le 1/4$. Also $\rho^{2m+1} = \rho^{4N+3} \le 1/4$. Consequently $w_i \ge (3/5)c\rho^{i-1}$, and

$$\|w\|_2^2 \ge \frac{9}{25}c^2 \sum_{j=0}^{N} \rho^{2j} \ge \frac{27c^2}{100(1-\rho^2)}.$$

It follows that

$$f(0) - f^\star \le \frac{50 r^2(1-\rho^2)}{27c} = \frac{50}{27}\mu_2 r^2(1+\rho) \le \frac{100}{27}\mu_p R^2.$$

This certificate is independent of the hidden resisting rotation. Finally $\Delta_0/128 = \mu_p R^2/32 < 3\mu_p R^2/64$, proving the constant-reduction assertion. □

The condition $n = n_0$ is used here, because the conversion to ℓ_p strong convexity holds only in n_0 ambient coordinates.

I EXPERIMENTAL DETAILS

Except for the study of Section I.8, all runs used Python 3.11 with NumPy, SciPy (HiGHS) and Matplotlib on a two-core 2.8 GHz Intel Xeon with 7 GB of memory. The randomized runs use fixed seeds; only in the largest shared-direction runs can a repeated run differ, by up to a few percent, because HiGHS may return different optimal support points.

Each budget N defines its own instance, including the dimension, so slopes against N describe budget-indexed families, not the trajectory of one method on one objective. The sparse quadratic family is deterministic, the spread quadratic family uses three rotations per budget, and the Steiner experiments use one seed per budget; these runs do not estimate a worst-case expectation over the randomized selector. The study of Section I.8 adds fixed instances, repeated instance seeds and more baselines.

I.1 SETTING AND FURTHER RESULTS

Instances, constants and baselines are those of Section 6; the quadratic comparison also uses $p = 4/3$ and $3/2$. AGD-E is FISTA (Beck & Teboulle, 2009), AGD-Ent is the accelerated method with the entropy prox-function on the simplex lift (Dvurechensky et al., 2017), PSG and MD-Ent return their best iterate, and EG is the extragradient method of Korpelevich (1976). "PSG + mixture LP" and "EG + mixture LP" apply the certificate program of Algorithm 3 to the points of PSG and EG.

Steiner paths. Figure 1(a) uses the sign chain $C_t = K \cap \{x : x_i \le 0,\ i \le t\}$, $t = 0, \dots, n$, in the unit balls K of ℓ_1 and ℓ_2, and the faces of Proposition B.4 in $B_1^n(1)$. For the sign chain, $\mathrm{st}(C_t) = \alpha_t(e_1 + \cdots + e_t)$ by symmetry, with scalars α_t that we estimate from 40 000 Gaussian samples; the path length is then $\sum_t ((t-1)(\alpha_t - \alpha_{t-1})^2 + \alpha_t^2)^{1/2}$. For the faces the path is computed exactly. Over $n = 8, 16, \dots, 4096$ the sign chain travels between 0.687 and 0.710 in the ℓ_1 ball and between $0.467\sqrt{n}$ and $0.491\sqrt{n}$ in the ℓ_2 ball. The faces travel 8.337 at $n = 4096$, between the bounds 7.90 and 8.89 of Proposition B.4, and the path budget at $N = n = 4096$ is $P_N = 53.96$.

Smooth chains. The instances (Section I.3) are Nesterov's chains, quadratic or nonquadratic, in dimension $n = 2N + 1$, rotated so that the minimizer is a fixed 5-sparse point. Figure 1(b) shows the quadratic chain. SP-Level runs at the exact level $\ell = f^\star$ with horizon N and $m = 256$ directions, which separates the rate of the level method from the search for the level. On both chains AGD-E, AGD-Ent and PG remove a small fraction of the initial gap, and so does CLA on the quadratic chain at these budgets. SP-Level falls below the best error attainable in the span of the first N chain coordinates at $N = 64$. Its fitted slope is -2.32 on the quadratic chain ($8 \le N \le 256$) and -2.31 on the nonquadratic one ($8 \le N \le 128$), and at $N = 256$ its error is $6.90 \cdot 10^{-8}$, against $2.52 \cdot 10^{-7}$ for AGD-E. The nonquadratic chain is nearly quadratic near its minimizer: its initial gaps and baseline errors differ from those of the quadratic chain by a relative amount of at most 10^{-4}, and the errors of SP-Level on the two chains agree within 5%. At $N = 256$ the guarantee of Proposition 3.2 is $1.7 \cdot 10^4$ times the initial gap.

Nonsmooth objectives. On ℓ_∞ regression (Fig. 2a) the error of SP-Cut with $m = 32$ directions decays with the exponent -1.8, against -0.7 for PSG and -0.3 for MD-Ent; at $N = 256$ ($n = 1025$) it is $2.1 \cdot 10^{-4}$, against $6.6 \cdot 10^{-3}$ for PSG. The mixture certificate improves PSG by less than 10% for $N \ge 32$, so the gain comes from the query points and not from the certificate. The certified value $\widehat{\varepsilon}$ of SP-Cut bounds the true error on every run and is within a factor 1.25 of it for $N \ge 64$. The best-iterate baselines use additional function values to select their output.

Saddle problems. On the bilinear games (Fig. 2b) the saddle gap of SP-VI decays faster than Q^{-2} for $16 \le Q \le 256$ (fitted exponent -3.7, local slope -2.4 over the last doubling). At $Q = 256$ it is $8 \cdot 10^{-7}$, against $5.6 \cdot 10^{-3}$ for EG, whose gap decays like Q^{-1}; the mixture certificate improves EG by a factor between 1.4 and 2.2. Localization probably benefits from the unique interior saddle point of these games. Because the objective is bilinear, the certificate equals the saddle gap of the returned mixture.

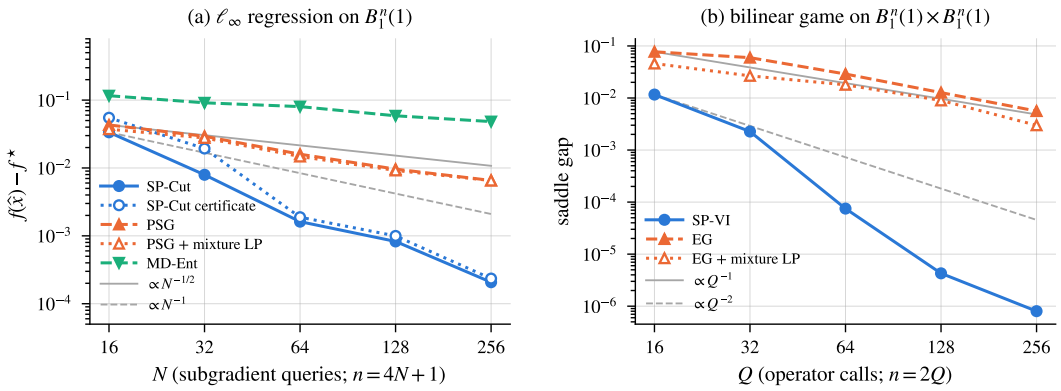

Figure 2: (a) ℓ_∞ regression, $n = 4N + 1$: error after N subgradient queries. (b) Bilinear games, $n = 2Q$: saddle gap after Q operator calls.

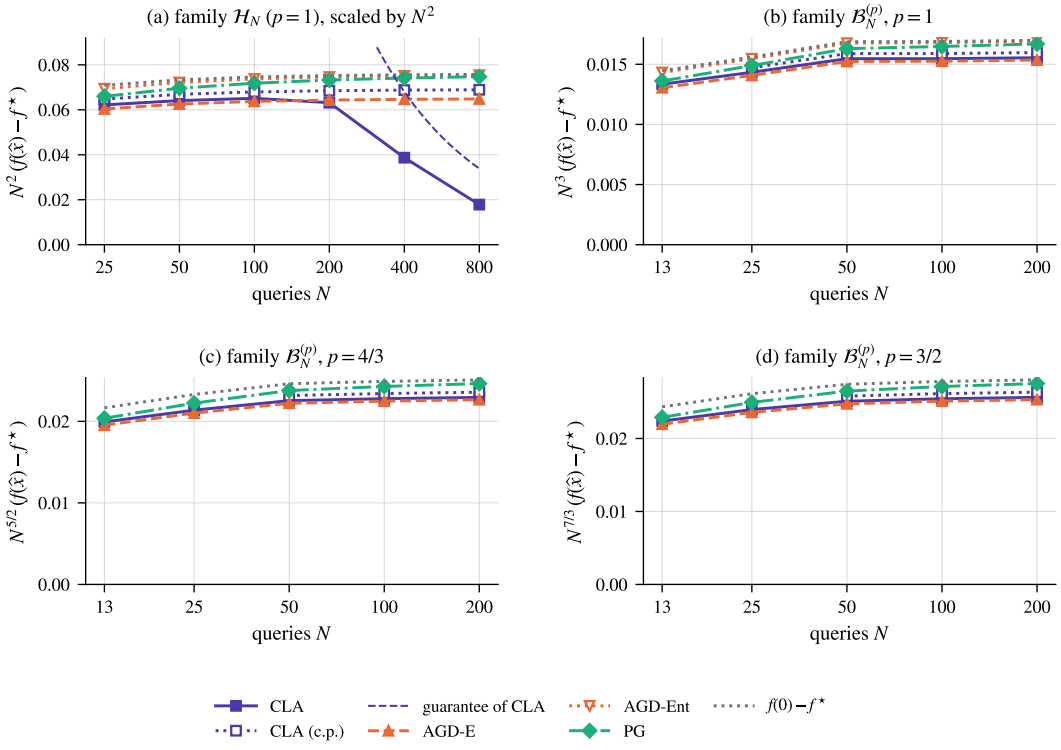

Figure 3: Quadratics: error after N gradient queries, multiplied by N^2 on $\mathcal{H}_N$ (a) and by $N^{1+2/p}$ on $\mathcal{B}_N^{(p)}$ (b–d).

Quadratics on ℓ_p balls. Figure 3 uses two quadratic families with $L = 1$ (Section I.6): $\mathcal{H}_N$, a rotated chain with the 1-sparse minimizer $0.9e_1$ and $n = 2N + 1$, and $\mathcal{B}_N^{(p)}$, the chain of Proposition H.9 in dimension $n = 3N + 2$ under a random rotation, with a minimizer of ℓ_p-norm 0.9 spread over all coordinates. CLA (c.p.), for "certified parameters", is CLA with the parameters of Proposition H.4. On $\mathcal{H}_N$ the Euclidean methods stay close to the initial gap, and the entropic method does not help. The bound of Proposition H.2, which becomes $27/(N - 3)$ after multiplication by N^2 (dashed in Fig. 3(a)), falls below the plateau $N^2(f(0) - f^\star) \approx 0.076$ near $N = 360$; the error of CLA is already slightly below that of AGD-E at $N = 200$, and at $N = 800$ it is 0.275 times that of AGD-E. On $\mathcal{B}_N^{(p)}$ no method removes more than 10% of the initial gap, which is dotted in Fig. 3(b–d), and the fitted exponents of all methods agree with $1 + 2/p$ (Table 4), as the lower bound

Table 4: Family $\mathcal{B}_N^{(p)}$, $25 \le N \le 200$: fitted slopes of $\log(f(\widehat{x}) - f^\star)$ against $\log N$, and the median of $(f(\widehat{x}) - f^\star)/(f(0) - f^\star)$.

	predicted	CLA	CLA (c.p.)	AGD-E	AGD-Ent	PG	$f(0) - f^\star$
$p = 1$	-3	-2.97	-2.97	-2.97	-2.96	-2.95	-2.97
$p = 4/3$	-2.5	-2.47	-2.49	-2.47	–	-2.45	-2.47
$p = 3/2$	-2.33	-2.30	-2.32	-2.30	–	-2.29	-2.30
relative error		0.92	0.94	0.90	0.99	0.97	1

predicts. CLA (c.p.) runs only for $N \ge 24/(2-p)$, so its slopes for $p = 4/3$ and $3/2$ are fitted on $N \in \{50, 100, 200\}$; the relative errors in the last row agree for the three values of p up to 0.01.

I.2 Implementation of the Steiner-point methods

Support points. A localization set is $C = K \cap \{x : \langle a_j, x\rangle \le \beta_j,\ j \le k\}$, where K is the unit ℓ_1 ball or a product of two such balls. The support problem $\max_{x \in C} \langle z, x\rangle$ is solved through its dual linear program, which has one variable per cut and per ball and $2n$ constraints; the support point is read off the dual multipliers. HiGHS works in floating point, so support points and certificates are numerical approximations, which we check against the primal problem. An unbounded dual means that C is empty; a primal feasibility problem confirms it. One support problem takes about 40 ms at $n = 257$ with $k \le 128$ cuts and about 0.4 s at $n = 1025$ with $k \le 256$.

Selector. The approximate Steiner point is the average $\bar{x} = \frac{1}{m}\sum_i V_i$ of Proposition C.4 over m Gaussian directions. Within one level call, or one run of SP-Cut or SP-VI, all centers use the same m directions. The average stays feasible up to LP errors, and sharing the directions reduces fluctuations between successive centers (Section I.7). Because later localization sets then depend on the directions, the guarantee of Proposition C.4, which uses fresh samples, does not apply. Its schedule m_N would require at least $32N^2/(\mathsf{b}_N \mathsf{c}_n) \approx 1.9 \cdot 10^4$ directions per center at $N = 256$; the level runs use $m = 256$.

SP-Level and SP-Accel. SP-Level is Algorithm 1 with the threshold $\delta_N = 4P_N/N$; at $n = 2N+1$ it equals 10.6, 6.3, 3.7, 2.1, 1.2 and 0.65 for $N = 8, \dots, 256$, so it exceeds the diameter 2 of K for $N \le 64$. Every trial was accepted at every budget. At $N = 256$ the largest displacement of the center is 0.063 and its total path is 2.09, against the path budget $P_N = 41.6$. SP-Accel is Algorithm 2 with three changes that keep every certificate valid. (i) After each level call the value of the returned point is queried and counted; $\overline{f}$ is the best measured value, and $\underline{f}$ is the largest of the certificates and of the bundle lower bound $\min_{u \in K} \max_j [f(x_j) + \langle g_j, u - x_j\rangle]$ over the stored queries. (ii) The horizon is doubled whenever a level call at width w returns a point with value above $\ell + w/4$. The level method succeeds once the horizon reaches (22), so this costs at most twice that horizon per phase plus one query per retry. (iii) Each level call starts from K intersected with the level cuts of all stored queries, and a level whose initial polytope is empty is certified without a query. The output is the best measured point.

SP-Cut and SP-VI. Both methods choose the cut depth from the certificate. After every query the mixture certificate $\widehat{\varepsilon}$ of the stored cuts is recomputed (a linear program with one variable per stored point), the depth ε is replaced by $\min\{\varepsilon, \widehat{\varepsilon}/2\}$, and the cuts are rebuilt when the depth changes. By the minimax identity of Proposition E.1, the localization set is never empty at a depth below $\widehat{\varepsilon}$, and the returned mixture is the certified one, so the certificate bounds the error or gap whatever the accuracy of the selector. The depth $4GP_N/N$ of the analysis exceeds GR for $N \le 128$ and cannot be used at these budgets. SP-VI uses the step $w = \mathrm{proj}_K(z - F(z)/(2L))$ of Algorithm 7, the cut $\langle F(w), w - u\rangle \ge \varepsilon$ and $m = 32$ directions.

I.3 Instances

Smooth chains. Let B be the chain difference operator $(By)_i = y_i - y_{i+1}$, $i < n$, $(By)_n = y_n$, used in the proof of Proposition H.9, and $A = \frac{1}{4}B^\top B$, so that $0 \prec A \preceq I$. Let w be the unit

Table 5: SP-Accel with $m = 64$ after N joint queries on the chains, $n = 2N + 1$.

chain	N	n	$f(0) - f^\star$	AGD-E	SP-Accel (best)	last level call
quadratic	32	65	$1.80 \cdot 10^{-5}$	$1.54 \cdot 10^{-5}$	$1.80 \cdot 10^{-5}$	$5.68 \cdot 10^{-5}$
quadratic	64	129	$4.63 \cdot 10^{-6}$	$3.95 \cdot 10^{-6}$	$4.63 \cdot 10^{-6}$	$9.64 \cdot 10^{-6}$
quadratic	128	257	$1.17 \cdot 10^{-6}$	$1.00 \cdot 10^{-6}$	$1.17 \cdot 10^{-6}$	$1.52 \cdot 10^{-6}$
nonquadratic	32	65	$1.80 \cdot 10^{-5}$	$1.54 \cdot 10^{-5}$	$1.80 \cdot 10^{-5}$	$5.77 \cdot 10^{-5}$
nonquadratic	64	129	$4.63 \cdot 10^{-6}$	$3.95 \cdot 10^{-6}$	$4.63 \cdot 10^{-6}$	$9.93 \cdot 10^{-6}$
nonquadratic	128	257	$1.17 \cdot 10^{-6}$	$1.00 \cdot 10^{-6}$	$1.17 \cdot 10^{-6}$	$1.52 \cdot 10^{-6}$

vector proportional to $A^{-1}e_1$, the chain minimizer, for which Bw is a constant vector; let u be the unit vector proportional to $(1, -0.7, 0.5, -0.3, 0.2, 0, \dots, 0)$ and U the Householder reflection with $Uw = u$. With $x^\star = 0.9u/\|u\|_1$, the objective is

$$f(x) = \tfrac{1}{4}\sum_i \psi((BUx)_i) - \langle g_{\text{lin}}, x\rangle\,, \qquad g_{\text{lin}} = \tfrac{1}{4}U^\top B^\top \psi'(BUx^\star),$$

so that $x^\star$ is the unconstrained minimizer and $\|x^\star\|_1 = 0.9$. For $\psi(t) = t^2/2$ this is Nesterov's chain $\frac{1}{2}\langle x, Qx\rangle - \langle Qx^\star, x\rangle$ with $Q = UAU$; for $\psi(t) = \sqrt{1+t^2} - 1$ the Hessian is at most I. In both cases $\psi'(BUx^\star)$ is a constant vector, so g_{lin} is proportional to $U^\top e_1$, and for every method whose iterates lie in the span of the past gradients, the kth gradient is supported on the first k coordinates of Ux. The minimum over the span of the first N of these coordinates has the gap $\frac{n-N}{n}(f(0) - f^\star)$, the dotted line of Fig. 1(b); it bounds only methods whose outputs stay in that span. The nonquadratic objective is nearly quadratic on the region explored: $BUx^\star = t_\star \mathbf{1}$ with $t_\star \approx 2 \cdot 10^{-4}$ at $N = 128$, where $\psi''(t_\star) \approx 1 - 5 \cdot 10^{-8}$.

ℓ_∞ regression. $f(x) = \|M(x - x^\star)\|_\infty$, where $M \in \mathbb{R}^{4n\times n}$ is Gaussian with unit rows ($G = 1$), $x^\star = (0.5, -0.3, 0.1, 0, \dots, 0)$ and $n = 4N + 1$; the oracle returns $\pm M_i^\top$ for a row i of largest residual. PSG and MD-Ent use the steps $R/(G\sqrt{N})$ and $R\sqrt{2\log(2n)}/(G\sqrt{N})$, the latter on the lift $w \mapsto R(w_+ - w_-)$ of Δ_{2n}.

Bilinear games. $\Phi(x, y) = \langle x, My\rangle + \langle b, x\rangle - \langle c, y\rangle$ on $B_1^n(1) \times B_1^n(1)$, where M is Gaussian scaled to $\|M\|_2 = 1$ (so $L = 1$), $b = -My^\star$ and $c = M^\top x^\star$ for the 3-sparse pair $x^\star = (0.5, -0.3, 0.1, 0, \dots, 0)$, $y^\star = (-0.4, 0.35, 0.15, 0, \dots, 0)$, which is then the unique saddle point; $n = 2Q$ for a budget of Q calls. The saddle gap of $(\bar{x}, \bar{y})$ is $\langle b, \bar{x}\rangle + \langle c, \bar{y}\rangle + \|M^\top \bar{x} - c\|_\infty + \|M\bar{y} + b\|_\infty$. EG uses the step $1/(2L)$ and Euclidean projections and returns the average of its extrapolated points.

I.4 The complete bracket search

Table 5 reports the errors of SP-Accel, which also searches for the level, of AGD-E after the same number of queries, and of the output of the last level call. The initial point has a small gap, while the bracket starts with the width $R\|g_0\|_\infty$, several orders of magnitude larger. The phases with horizons $2, 4, \dots$ reduce the width, but the outputs of their level calls stay above the initial gap, so the best measured point is the initial point, whose gap exceeds the error of AGD-E by about 17%. The last level call, with horizon $N/2$, was stopped by the budget after 12, 27 and 58 trials for $N = 32, 64, 128$; its output approaches the initial gap as N grows and is within 30% of it at $N = 128$. With $m = 64$ directions even the exact-level method stays near 0.75 of the initial gap (Section I.7), so the overhead of the search and the effect of the sample size cannot be separated here; with $m = 256$ at fixed $n = 65$ the complete method does separate from the baselines (Section I.8.1). Larger budgets were not run because of the cost of the linear programs.

I.5 The gradient-only method

Table 6 reports SP-Prox (Algorithm 5) with 32 directions shared by the centers of a stage and the Euclidean projection in the inner projected-gradient loop, on the nonquadratic chain with $n = 65$ and $L = R = 1$, started at 0 ($f(0) - f^\star = 1.80 \cdot 10^{-5}$). For each stage j it lists the cut depth $\delta_j = 2^{-j-1}$, the numbers of trials and of accepted proximal steps, the gradient queries of the stage and in total, and the certificate and the true error of the stage output. The method uses no function

Table 6: SP-Prox on the nonquadratic chain with $n = 65$, stages $j \geq 12$.

j	δ_j	trials	accepted	gradients	total	certificate	$f(y) - f^\star$
12	$1.22 \cdot 10^{-4}$	2	0	1 600	5 709	$1.03 \cdot 10^{-4}$	$1.74 \cdot 10^{-5}$
13	$6.10 \cdot 10^{-5}$	4	1	2 574	8 283	$5.81 \cdot 10^{-5}$	$1.67 \cdot 10^{-5}$
14	$3.05 \cdot 10^{-5}$	15	14	10 360	18 643	$3.01 \cdot 10^{-5}$	$1.29 \cdot 10^{-5}$
15	$1.53 \cdot 10^{-5}$	24	22	16 164	34 807	$1.52 \cdot 10^{-5}$	$6.32 \cdot 10^{-6}$
16	$7.63 \cdot 10^{-6}$	27	26	21 014	55 821	$7.30 \cdot 10^{-6}$	$2.77 \cdot 10^{-6}$
17	$3.81 \cdot 10^{-6}$	24	23	19 539	75 360	$3.74 \cdot 10^{-6}$	$1.29 \cdot 10^{-6}$
18	$1.91 \cdot 10^{-6}$	27	25	23 280	98 640	$1.87 \cdot 10^{-6}$	$5.80 \cdot 10^{-7}$
19	$9.54 \cdot 10^{-7}$	28	27	27 132	125 772	$9.42 \cdot 10^{-7}$	$2.47 \cdot 10^{-7}$
20	$4.77 \cdot 10^{-7}$	31	30	30 616	156 388	$4.63 \cdot 10^{-7}$	$9.66 \cdot 10^{-8}$

values; the true errors are computed afterward. The stages $j \leq 11$ end at their first trial with an empty cut, which certifies that the current point is δ_j-optimal without comparing values. Every stage ended with an empty cut after at most 31 trials, far below the worst-case allowance, and almost every trial with a nonempty cut was an accepted proximal step, so the small-step case of (57), which the path bound controls, was almost never used. The certificate bounds the true error at every stage, by a factor between 2.3 and 4.8 for $j \geq 14$. The cost lies in the inner loops: each trial runs a bisection whose steps are loops of projected-gradient steps, and at $n = 65$ the stages with more than one trial use between $6 \cdot 10^2$ and 10^3 gradients per trial. The final error 10^{-7} takes $1.6 \cdot 10^5$ gradients at $n = 65$; at $n = 17$ and 33, 16 and 18 stages take $5.0 \cdot 10^4$ and $8.9 \cdot 10^4$ gradients and reach $1.5 \cdot 10^{-6}$ and $3.8 \cdot 10^{-7}$. In gradient evaluations the method is therefore not competitive with the joint-oracle methods or with AGD-E; the experiment supports the mechanism and the certificates, not a practical advantage.

I.6 Implementation of CLA and the quadratic families

Instance families. Both families use the matrix A of Section I.3 in dimension n. For $\mathcal{H}_N$, $n = 2N + 1$, $Q = UAU$, where U is the Householder reflection that maps $A^{-1}e_1/\|A^{-1}e_1\|_2$ to e_1, and $f(x) = \frac{1}{2}\langle x, Qx\rangle - \langle Qx^\star, x\rangle$ with $x^\star = 0.9e_1$; in the coordinates Ux the linear term is a multiple of e_1, so Krylov-type methods have error of order $\|x^\star\|_2^2/N^2$ for $N \leq (n-1)/2$. For $\mathcal{B}_N^{(p)}$, $n = 3N+2$, $Q = UAU^\top$ with a random orthogonal U, and $x^\star$ is proportional to $UA^{-1}e_1$ with $\|x^\star\|_p = 0.9$; then $f(0) - f^\star$ decays like $N^{-(1+2/p)}$. Three rotations are used for each pair (p, N), and medians are reported. All methods receive exactly N gradient queries; for CLA these are the query at 0 and $N - 1$ products.

Learned model. The directions requested by Algorithm 8 are orthonormalized against the current basis V_S of the learned span S by two Gram–Schmidt passes; a direction whose residual has norm below 10^{-10} is redundant but still counted as a query. For a new unit vector d, the product $Qd = (\nabla f(hd) - \nabla f(0))/h$ with $h = 1/\max\{1, \|d\|_p\}$ uses the feasible query $hd \in B_p^n(1)$. The learned matrix is the ridge model (83), stored as $EE^\top$ with $E \in \mathbb{R}^{n \times \dim S}$.

Model minimization. Every model has the form $J(x) = \lambda\Psi_I(x) + \frac{1}{2}\|E^\top x\|_2^2 + \langle b, x\rangle + \frac{\zeta}{2}\|x\|_2^2$ on $B_p^n(1)$ with a weight $\lambda > 0$. We maximize the concave dual function

$$\varphi(u) = -\tfrac{1}{2}\|u\|_2^2 + \min_{x \in B_p^n(1)} \{\lambda\Psi_I(x) + \langle b + Eu, x\rangle + \tfrac{\zeta}{2}\|x\|_2^2\}, \qquad u \in \mathbb{R}^{\dim S},$$

by Newton's method with backtracking. The inner problem is separable up to the ball constraint and is solved by a scalar search for the multiplier of the constraint, with closed-form coordinate roots for $p \in \{1, \frac{4}{3}, \frac{3}{2}\}$; the generalized Jacobian of its solution map gives the Newton matrix. At the pair $(x(u), u)$ the duality gap equals $\frac{1}{2}\|E^\top x(u) - u\|_2^2$, and it bounds $J(x(u)) - \min J$ when the inner minimization is exact. CLA stops when this gap is at most $10^{-13}\max\{1, |\varphi(u)|\}$, and CLA (c.p.) when it is at most η of (90); consecutive solves are warm started. This rule does not check the residual required by Definition H.3, so CLA (c.p.) is not a certified realization of Proposition H.4.

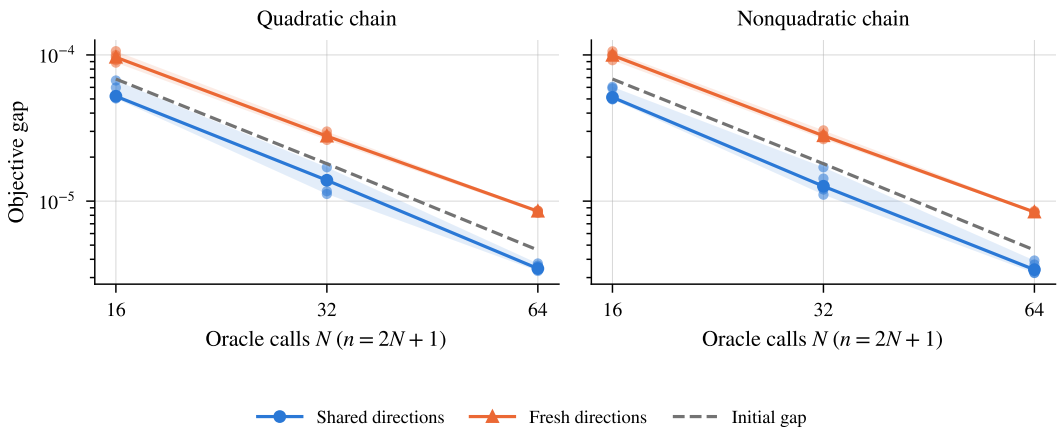

Figure 4: Exact-level SP-Level with 64 shared or fresh directions per center: medians, ranges and all runs over five seeds.

Step selection and parameters. The step parameter is selected as in Algorithms 8 and 9: the model at $\gamma = 0$ is solved first; if its largest unpromoted coordinate is at least $(1 - \sigma)\tau$, the stage promotes; otherwise the model at $\bar{\gamma}_t$ is solved, and if its largest unpromoted coordinate exceeds τ, a bisection on γ stops once this coordinate lies in $[(1-\sigma)\tau, \tau]$. Both variants use $\zeta = L_\tau/(p(T+1)^3)$ and the ridge parameter $\zeta/18$ of (90). CLA (c.p.) uses all parameters of (90); CLA uses $\tau^p = 6/((2-p)(T+1))$ and $L_\tau = L\tau^{2-p}$ of Algorithm 8 and $\sigma = 10^{-6}$. The stabilizer then changes the objective by at most $\zeta/2 = O(L\tau^{2-p}T^{-3})$ on $B_p^n(1)$, below the guarantee of Proposition H.2. On $\mathcal{H}_N$, CLA made only full steps for $N \le 100$; for $N \ge 200$ one stage was a partial step that promoted one coordinate. On one core, a run on $\mathcal{H}_{800}$ took about 21 minutes for CLA and 2.5 minutes for CLA (c.p.).

Baselines. AGD-E uses the indicator of $B_p^n(1)$, the step $1/L$ with $L = \lambda_{\max}(Q)$, and its last iterate as output. The projection onto $B_1^n(1)$ uses sorting (Duchi et al., 2008), and the projection onto $B_p^n(1)$, $p > 1$, the separable solver above. PG is $x_{k+1} = \mathrm{proj}_K(x_k - \nabla f(x_k)/L)$ from $x_0 = 0$. AGD-Ent is the similar-triangles method with $\theta_k = 2/(k+2)$ on the lift $w \mapsto w_+ - w_-$ of Δ_{2n} onto $B_1^n(1)$, with entropic mirror steps of size $1/(L_1\theta_k)$ from the uniform distribution, where $L_1 = \max_{ij} |Q_{ij}|$ is the ℓ_1 smoothness constant of the lifted function; on the nonquadratic chain it uses the constant of the quadratic chain, which is larger.

I.7 Sensitivity to selector sampling

We repeat the exact-level runs on both chains at $N \in \{16, 32, 64\}$, $n = 2N + 1$, with 64 directions per center and five seeds. Each seed is used twice: with directions shared by all centers of the level call, and with a fresh batch at every center; every run returns its last accelerated iterate. All 60 runs use exactly N queries, stay feasible to within 10^{-10} and accept every trial.

Figure 4 shows a strong sampling effect. At $N = 64$ on the quadratic chain the median gap is $3.46 \cdot 10^{-6}$ with shared directions (range 3.33–$3.75 \cdot 10^{-6}$) and $8.53 \cdot 10^{-6}$ with fresh ones (8.31–$8.62 \cdot 10^{-6}$), for an initial gap of $4.63 \cdot 10^{-6}$; on the nonquadratic chain the medians are $3.40 \cdot 10^{-6}$ and $8.42 \cdot 10^{-6}$. With fresh directions the last iterate is worse than the initial point at every budget and for every seed; with shared directions the median stays between 0.70 and 0.77 of the initial gap. Fresh batches satisfy the independence assumption of Proposition C.4, but 64 directions are far below its schedule. Shared directions shorten the path of the centers (median 2.06 against 4.98 at $N = 64$) and couple later localization sets to the sample. At this sample size neither variant shows the rate of the level method.

Larger fresh samples. Table 7 repeats the quadratic-chain runs with m fresh directions per center and three seeds. The relative error decreases with m, and for $m = 1024$ also with N: the fitted slope over $N = 16, 32, 64$ is -2.29, against -2.08 for $m = 256$ and -2.42 for the shared runs with $m = 256$ of Fig. 1(b). The spread over seeds is at most 0.09, and every trial was accepted.

Table 7: Exact-level SP-Level on the quadratic chain: median of $(f(\widehat{x}) - f^\star)/(f(0) - f^\star)$ with m fresh directions per center.

directions per center	$N = 16$	$N = 32$	$N = 64$
64 (five seeds)	1.41	1.54	1.84
256	0.85	0.81	0.71
1024	0.67	0.57	0.42
4096	0.66	0.50	–
256, shared	0.71	0.55	0.37

With $N + 1$ selector calls and failure probability 0.1, the schedule of Proposition C.4 requires about $1.3 \cdot 10^3$, $4.1 \cdot 10^3$ and $1.4 \cdot 10^4$ directions for $N = 16, 32, 64$, so only $m = 4096$ at $N = 16$ meets it.

I.8 Controlled study at fixed dimension

This study complements the budget-indexed runs: it uses fixed instances, three instance seeds at selected endpoints, more baselines and measured times. It ran on a Linux server with Python 3.12, NumPy 2.4 and SciPy 1.17 (HiGHS), with twelve tasks in parallel and one BLAS thread each; times are elapsed times of a task under this load. We report medians with observed ranges, or single runs, not confidence intervals. The relative error of an output $\widehat{x}$ is $r = (f(\widehat{x}) - f^\star)/(f(0) - f^\star)$; for games, r is the ratio of saddle gaps.

Baselines and oracles. The study adds Frank–Wolfe (FW) and its away-step and pairwise variants (AFW, PFW) (Lacoste-Julien & Jaggi, 2015), all with the short step clipped to $[0, 1]$; AGD-E-feasible, the similar-triangles method of Dvurechensky et al. (2017) with the Euclidean prox-function, whose queries are convex combinations of feasible points; and E-Bundle, a Euclidean bundle-level method (Lemaréchal et al., 1995) that projects onto a level set of the stored cuts and returns its best evaluated point. PSG-mix and EG-mix are the methods "PSG + mixture LP" and "EG + mixture LP" of Section I.1, where CLA (c.p.) is also defined. PG, the accelerated methods, the Frank–Wolfe variants and CLA use gradients only; E-Bundle and the Steiner-point methods also use values, and SP-Level is given $f^\star$. A smaller error at equal N therefore does not mean a lower total cost.

I.8.1 Fixed smooth instances

The instances are the chains of Section I.3 with $n = 65$, a 5-sparse minimizer of ℓ_1-norm 0.9 and budgets $N = 16, 32, 64$. The nonquadratic chain of Section I.3 is almost quadratic near its minimizer, so here the nonquadratic chain uses

$$\psi_s(t) = s^2\big(\sqrt{1 + (t/s)^2} - 1\big), \qquad s = \tfrac{1}{n}\textstyle\sum_i |(BUx^\star)_i|,$$

whose curvature $\psi_s''(t) = (1 + (t/s)^2)^{-3/2}$ varies on the scale of the solution, while $L = 1$ still holds. The Steiner-point methods use $m = 256$ shared directions. Figure 5 shows seed 0 at $N = 16, 32$ (open markers) and the median and range over three seeds at $N = 64$ (filled markers).

On the nonquadratic chain SP-Accel leaves $r = 0.111$ (range 0.100–0.111), against 0.742 for AGD-E-feasible and 0.583 (0.573–0.757) for E-Bundle; on the quadratic chain the medians are 0.229, 0.709 and 0.388 (Table 8). SP-Level, although given $f^\star$, leaves 0.193 on the nonquadratic chain, so knowing the level does not guarantee a smaller error at a fixed budget; CLA was run only on the quadratic chain. SP-Accel takes about 94 s on the nonquadratic chain, against 1.18 s for E-Bundle and milliseconds for the accelerated methods. The advantage depends on the dimension: at $n = 129$ and $N = 64$, on one nonquadratic instance, SP-Accel leaves $r = 0.853$, SP-Level 0.564, AGD-E-feasible 0.870 and E-Bundle 1.00.

I.8.2 Sparsity and dimension in curvature learning

Figure 6 uses instance seed 0 and a minimizer of ℓ_1-norm 0.9; each marker is one run. At $n = 65$, $N = 128$ and $p = 1$, CLA leaves $r = 0.0706$ for a 1-sparse minimizer, against 0.424 for AGD-E; for 5-sparse minimizers the medians over three instance seeds are 0.456 and 0.424, and on the

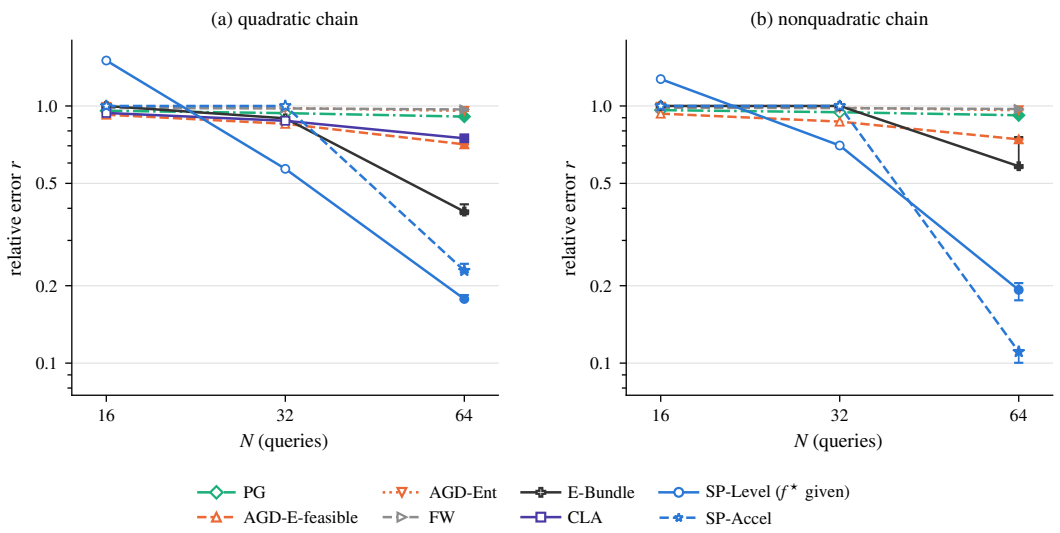

Figure 5: Relative error r on fixed chains with $n = 65$.

Table 8: Median relative error r and time at $n = 65$, $N = 64$, over three seeds ($\star$: given $f^\star$).

Method	r, quadratic	Time (s)	r, nonquadratic	Time (s)
PG	0.909	0.00413	0.92	0.00547
AGD-E	0.709	0.00441	0.742	0.00555
AGD-E-feasible	0.709	0.00584	0.742	0.00752
AGD-Ent	0.957	0.00565	0.963	0.00708
FW	0.969	0.00597	0.973	0.008
AFW	0.969	0.00983	0.973	0.0107
PFW	0.97	0.0092	0.973	0.0108
E-Bundle	0.388	3.73	0.583	1.18
CLA	0.749	0.234	—	—
SP-Level*	0.178	92.4	0.193	90.6
SP-Accel	0.229	74.1	0.111	93.6

dense rotated chain CLA gives 0.511, AGD-E 0.424 and E-Bundle 0.138. The advantage of CLA is therefore tied to sparse minimizers. For a 1-sparse minimizer, increasing n from 65 to 129 and 513 changes r from 0.0706 to 0.400 and 0.936 for CLA, against 0.424, 0.710 and 0.927 for AGD-E; for $p = 4/3$ and $3/2$ at $n = 65$, CLA gives 0.385 and 0.481, against 0.424 for AGD-E. At $n = 65$, $N = 128$ and a 5-sparse minimizer, tightening the inner tolerance of CLA and CLA (c.p.) from 10^{-11} to 10^{-15} changes r by less than $6 \cdot 10^{-8}$ for each p.

Table 9 lists the gaps of CLA and AGD-E, their ratio, the times and the number of inner solves of CLA that miss their tolerance. It confirms the separation seen in Fig. 3 up to $N = 800$. At $N = 1600$ the gap of CLA is 0.109 times that of AGD-E, but 47 inner solves miss their tolerance, so this point is not evidence for the exact method; the run took $25\,614$ s, against 10.2 s for AGD-E, and no fitted exponent uses it. The queries of CLA stay in the ball, while those of AGD-E reach the ℓ_1-norm 1.000754 at $N = 1600$.

I.8.3 Selector sample size, nonsmooth problems and games

Selector sample size. On one nonquadratic instance of Section I.8.1 with $n = 65$, a 5-sparse minimizer and $N = 32$, SP-Level with $m = 32$, 128 and 512 directions and three selector seeds reaches the median relative errors 0.9994, 0.7502 and 0.6726 with shared directions and 1.9143, 1.0236 and 0.8331 with fresh ones; the shared runs take 3.9, 15.1 and 60.5 s. Larger samples help, at a cost linear in m, and shared directions do better than fresh ones on this instance. All 18 corresponding

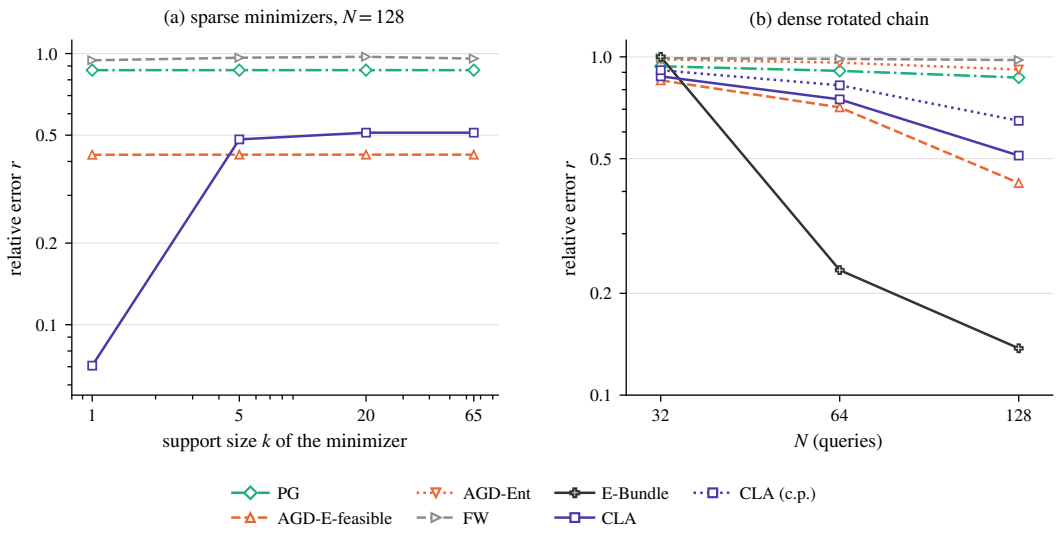

Figure 6: Relative error at $n = 65$, $p = 1$: (a) against the support size k of the minimizer, $N = 128$; (b) on the dense rotated chain.

Table 9: CLA and AGD-E on $\mathcal{H}_N$, $p = 1$, $n = 2N + 1$.

N	AGD-E gap	CLA gap	Ratio	AGD-E (s)	CLA (s)	Misses
100	$6.37 \cdot 10^{-6}$	$6.51 \cdot 10^{-6}$	1.02	0.00353	0.854	0
200	$1.61 \cdot 10^{-6}$	$1.58 \cdot 10^{-6}$	0.982	0.0129	9.08	0
400	$4.04 \cdot 10^{-7}$	$2.42 \cdot 10^{-7}$	0.599	0.0809	96.9	0
800	$1.01 \cdot 10^{-7}$	$2.79 \cdot 10^{-8}$	0.275	0.7	$1.32 \cdot 10^{3}$	0
1600	$2.54 \cdot 10^{-8}$	$2.76 \cdot 10^{-9}$	0.109	10.2	$2.56 \cdot 10^{4}$	47

runs of SP-Accel end with $r = 1$: at this budget the overhead of the search for the level outweighs the gain.

ℓ_∞ regression and bilinear games. Table 10 reports medians over three instances with 64 oracle calls, and 128 operator calls for games, where n is the total dimension of both blocks. On ℓ_∞ regression with $n = 32$, SP-Cut reaches $r = 0.00391$, against 0.101 for PSG and 0.0268 for MD-Ent, but E-Bundle, which also uses values, reaches $3.79 \cdot 10^{-5}$ in 0.368 s, against 15.4 s for SP-Cut; at $n = 128$ E-Bundle is again more accurate (0.006 against 0.0128). On bilinear games with $n = 128$, SP-VI reaches $1.96 \cdot 10^{-5}$, against 0.07 for EG and 0.0266 for EG-mix, in 46 s against 0.0129 s for EG; at $n = 32$ it reaches $4.88 \cdot 10^{-7}$, against 0.0377 for EG. The Steiner-point methods thus save oracle calls at a large cost in linear programs.

Table 10: ℓ_∞ regression (first five rows) and bilinear games (last three rows): median relative error r and time over three instances.

Method	r, $n = 32$	Time (s)	r, $n = 128$	Time (s)
PSG	0.101	0.00407	0.115	0.00735
PSG-mix	0.0936	0.0114	0.115	0.0269
MD-Ent	0.0268	0.00352	0.441	0.00628
E-Bundle	$3.79 \cdot 10^{-5}$	0.368	0.006	0.837
SP-Cut	0.00391	15.4	0.0128	37
EG	0.0377	0.0104	0.07	0.0129
EG-mix	0.0138	0.0175	0.0266	0.0271
SP-VI	$4.88 \cdot 10^{-7}$	15.9	$1.96 \cdot 10^{-5}$	46

Other diagnostics. With a memory of 32 cuts, E-Bundle was not monotone in its inner tolerance on one instance with $n = 33$ and $N = 64$: the tolerances 10^{-7}, 10^{-9} and 10^{-11} gave $r = 0.0468$, 0.0669 and 0.136. The four runs of SP-Prox, on both chains with $n = 33$ and budgets of 64 and 256 gradients, stopped at the budget before completing their four stages: with 64 gradients no stage was completed, and with 256 two, leaving $r = 0.978$ on the quadratic chain and 0.990 on the nonquadratic one. These budgets are too small for the method; in Section I.5, 18 stages at $n = 33$ take $8.9 \cdot 10^4$ gradients.

J CONCURRENT WORK

Four preprints posted on arXiv between 17 and 20 September 2026 study closely related questions. We discuss them for completeness; our proofs do not use their results or constructions. We have not verified the proofs of these preprints and do not vouch for their correctness. Three of them consider the domain $B_p^n(R)$ with Lipschitz or Hölder constants measured in the ℓ_q norm; in this appendix q denotes this norm, not the conjugate exponent of p. Our setting is $p = 1$, $q = 2$. Their rates are stated through the exponent $\rho_{p,q} = 1/p - \max\{1/q - 1/2, 0\}$, which equals 1 at $(p, q) = (1, 2)$ and $1/2$ when $p = q \le 2$.

Lipschitz convex optimization. For convex objectives that are G-Lipschitz in ℓ_q, with values and subgradients, Martínez-Rubio et al. (2026a, Corollary 11) prove the deterministic oracle bound $\widetilde{O}_{p,q}(GR/N^{\rho_{p,q}})$ for $1 \le p \le q \le \infty$; their implementation may take time exponential in the dimension and the number of queries. Martínez-Rubio & Guzmán (2026, Theorems 3 and 12) attain the same rate for $p < \min\{q, 2\}$ with feasible queries and, with high probability, polynomial computation, and they bound the movement of their centers along nested subsets of $B_p^n(R)$ (their Theorem 2). At $(p, q) = (1, 2)$ both give $\widetilde{O}(GR/N)$, the rate of Proposition E.1. Section 4 of Martínez-Rubio & Guzmán (2026) uses, at $(p, q) = (1, 2)$, the Steiner point of the sublevel sets of a bundle model, deep cuts, a movement bound deduced from Bubeck et al. (2020), and a Monte Carlo implementation with one linear program per sample; this is the mechanism of SP-Cut. Our version uses subgradients without function values, returns a certificate computed from them, and applies to every known polytope in $c + B_1^n(R)$, including the simplex. Appendix G of Martínez-Rubio & Guzmán (2026) proves randomized lower bounds for the Lipschitz class in dimension $n \gtrsim \varepsilon^{-2+\delta}$ for any fixed $\delta > 0$; Proposition D.6 needs $n \ge 8N$.

Smooth and Hölder convex optimization. For gradients that are ν-Hölder in ℓ_q with constant L, Martínez-Rubio et al. (2026b, Theorem 1) obtain, after $N \le n$ queries of values and gradients, a feasible point with error $\widetilde{O}_{\nu,p,q}(LR^{1+\nu}/N^{(1+\nu)(1+\rho_{p,q})-1})$ for $p < \min\{q, 2\}$ and in the matched cases $p = q \le 2$, with high probability and in polynomial time; they also state a deterministic version without an efficiency guarantee. At $(p, q) = (1, 2)$ the exponent is $1 + 2\nu$, so for $N \le n$ their result has the exponents of Theorem 1 and Proposition E.3, for a wider range of norm pairs. Their method combines a stable center, a line search on the segment between the incumbent and the center, and a Moreau envelope for feasibility. SP-Level uses the Steiner point in place of the prox step of an accelerated level method, and SP-Prox uses a related envelope search without function values. Our upper bounds hold for all N and n; the bounds for oracles of higher order, the randomized lower bounds for smooth functions and the results on strong convexity have no counterpart there.

Variational inequalities of higher order. Zhang et al. (2026b) study monotone variational inequalities in Euclidean geometry with an oracle that returns F and its derivatives up to order k, where $D^k F$ is Lipschitz with constant L. For $k \ge 1$ their Theorem 3.1 finds a point x with $\operatorname{dist}(0, F(x) + N_K(x)) \le \varepsilon$ after $O_k((LD_0^{k+1}/\varepsilon)^{2/(3k+2)})$ calls, where D_0 is the initial Euclidean distance to a solution and $N_K(x)$ is the normal cone of K at x. A point with this residual has weak gap at most ε times the Euclidean diameter of K, so on the ℓ_1 ball their method gives the weak gap $O(LR^{k+2}/N^{(3k+2)/2})$ after N calls. For the same oracle, Proposition G.3 with $r = k + 1$ gives $\widetilde{O}(LR^{k+2}/N^{k+2})$. Our exponent is larger for $k = 1$ (3 against $5/2$), the two coincide for $k = 2$, and theirs is larger for $k \ge 3$; the gain at $k = 1$ comes from the ℓ_1 geometry, while their method applies to general convex domains. Zhang et al. (2026a) prove the same Euclidean rate with a matching lower bound (Section G.2).