Mathematical Programming (2020) 184:71–120
https://doi.org/10.1007/s10107-019-01406-y

FULL LENGTH PAPER

Series A

Lower bounds for finding stationary points I

Yair Carmon[1] · John C. Duchi[2] · Oliver Hinder[3] · Aaron Sidford[3]

Received: 13 December 2017 / Accepted: 27 May 2019 / Published online: 7 June 2019

Abstract

We prove lower bounds on the complexity of finding ϵ-stationary points (points x such that $\|\nabla f(x)\| \leq \epsilon$) of smooth, high-dimensional, and potentially non-convex functions f. We consider oracle-based complexity measures, where an algorithm is given access to the value and all derivatives of f at a query point x. We show that for any (potentially randomized) algorithm A, there exists a function f with Lipschitz pth order derivatives such that A requires at least $\epsilon^{-(p+1)/p}$ queries to find an ϵ-stationary point. Our lower bounds are sharp to within constants, and they show that gradient descent, cubic-regularized Newton's method, and generalized pth order regularization are worst-case optimal within their natural function classes.

Keywords Non-convex optimization · Information-based complexity · Dimension-free rates · Gradient descent · Cubic regularization of Newton's method

Mathematics Subject Classification 90C06 · 90C26 · 90C30 · 90C60 · 68Q25

OH was supported by the PACCAR INC fellowship. YC and JCD were partially supported by the SAIL-Toyota Center for AI Research, NSF-CAREER award 1553086, and a Sloan Foundation Fellowship in Mathematics. YC was partially supported by the Stanford Graduate Fellowship and the Numerical Technologies Fellowship.

✉ Yair Carmon
yairc@stanford.edu

John C. Duchi
jduchi@stanford.edu

Oliver Hinder
ohinder@stanford.edu

Aaron Sidford
sidford@stanford.edu

[1] Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA

[2] Departments of Statistics and Electrical Engineering, Stanford University, Stanford, CA 94305, USA

[3] Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, USA

1 Introduction

Consider the optimization problem

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \ f(x)$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is smooth, but possibly non-convex. In general, it is intractable to even approximately minimize such f [33,35], so—following an established line of research—we consider the problem of finding an ϵ-stationary point of f, meaning some $x \in \mathbb{R}^d$ such that

$$\|\nabla f(x)\| \le \epsilon. \tag{1}$$

We prove lower bounds on the number of function and derivative evaluations required for algorithms to find a point x satisfying inequality (1). While for arbitrary smooth f, a near-stationary point (1) is certainly insufficient for any type of optimality, there are a number of reasons to study algorithms and complexity for finding stationary points. In several statistical and engineering problems, including regression models with non-convex penalties and objectives [30,31], phase retrieval [12,42], and non-convex (low-rank) reformulations of semidefinite programs and matrix completion [8,11,27], it is possible to show that all first- or second-order stationary points are (near) global minima. The strong empirical success of local search strategies for such problems, as well as for neural networks [28], motivates a growing body of work on algorithms with strong complexity guarantees for finding stationary points [2,7,13,15, 40]. In contrast to this algorithmic progress, algorithm-independent lower bounds for finding stationary points are largely unexplored.

Even for non-convex functions f, it is possible to find ϵ-stationary points for which the number of function and derivative evaluations is polynomial in $1/\epsilon$ and the dimension d of $\operatorname{dom} f$. Of particular interest are methods for which the number of function and derivative evaluations does not depend on d, but instead depends on measures of f's regularity. The best-known method with such a *dimension-free* convergence guarantee is classical gradient descent: for every (non-convex) function f with L_1-Lipschitz gradient satisfying $f(x^{(0)}) - \inf_x f(x) \le \Delta$ at the initial point $x^{(0)}$, gradient descent finds an ϵ-stationary point in at most $2L_1\Delta\epsilon^{-2}$ iterations [37]. Under the additional assumption that f has Lipschitz continuous Hessian, our work [15] and Agarwal et al. [2] exhibit randomized first-order methods that find an ϵ-stationary point in time scaling as $\epsilon^{-7/4} \log \frac{d}{\epsilon}$ (igoring other problem-dependent constants). In subsequent work [13], we show a different deterministic accelerated gradient method that achieves dimension-free complexity $\epsilon^{-7/4} \log \frac{1}{\epsilon}$, and if f additionally has Lipschitz third derivatives, then $\epsilon^{-5/3} \log \frac{1}{\epsilon}$ iterations suffice to find an ϵ-stationary point.

By evaluation of higher order derivatives, such as the Hessian, it is possible to achieve better ϵ dependence. Nesterov and Polyak's cubic regularization of Newton's method [16,40] guarantees ϵ-stationarity (1) in $\epsilon^{-3/2}$ iterations, but each iteration may be expensive when the dimension d is large. More generally, pth-order regularization methods iterate by sequentially minimizing models of f based on order p Taylor approximations, and Birgin et al. [7] show that these methods converge in $\epsilon^{-(p+1)/p}$ iterations. Each iteration requires finding an approximate stationary point of a high-dimensional, potentially non-convex, degree $p+1$ polynomial, which suggests that the methods will be practically challenging for $p > 2$. The methods nonetheless provide fundamental upper complexity bounds.

In this paper and its companion [14], we focus on the converse problem: providing dimension-free complexity lower bounds for finding ϵ-stationary points. We show fundamental limits on the best achievable ϵ dependence, as well as dependence on other problem parameters. Together with known upper bounds, our results shed light on the optimal rates of convergence for finding stationary points.

1.1 Related lower bounds

In the case of *convex* optimization, we have a deep understanding of the complexity of finding ϵ-suboptimal points, that is, x satisfying $f(x) \le f(x^\star) + \epsilon$ for some $\epsilon > 0$, where $x^\star \in \arg\min_x f(x)$. Here we review only the dimension-free optimal rates, as those are most relevant for our results. Given a point $x^{(0)}$ satisfying $\|x^{(0)} - x^\star\| \le D < \infty$, if f is convex with L_1-Lipschitz gradient, Nesterov's accelerated gradient method finds an ϵ-suboptimal point in $\sqrt{L_1}D\epsilon^{-1/2}$ gradient evaluations, which is optimal even among randomized, higher-order algorithms [35–37,45].[1] For non-smooth problems, that is, when f is L_0-Lipschitz, subgradient methods achieve the optimal rate of $L_0^2 D^2/\epsilon^2$ subgradient evaluations (cf. [10,35,37]). In Part II of this paper [14], we consider the impact of convexity on the difficulty of finding stationary points using first-order methods.

Globally optimizing smooth non-convex functions is of course intractable: Nemirovski and Yudin [35, §1.6] show that for functions $f : \mathbb{R}^d \to \mathbb{R}$ with Lipschitz 1st through pth derivatives, and algorithms receiving all derivatives of f at the query point x, the worst case complexity of finding ϵ-suboptimal points scales at least as $(1/\epsilon)^{d/p}$. This exponential scaling in d shows that dimension-free guarantees for achieving near-optimality in smooth non-convex functions are impossible to obtain.

Less is known about lower bounds for finding stationary points for $f : \mathbb{R}^d \to \mathbb{R}$. Nesterov [39] proposes lower bounds for finding stationary points under a box constraint, but his construction does not extend to the unconstrained case when $f(x^{(0)}) - \inf_x f(x)$ is bounded. Vavasis [44] considers the complexity of finding ϵ-stationary points of functions with Lipschitz derivatives in a first-order (gradient and function-value) oracle model. For such problems, he proves a lower bound of $\epsilon^{-1/2}$ oracle queries that applies to any deterministic algorithm operating on certain two-dimensional functions. This appears to be the first algorithm-independent lower bound for approximating stationary points of non-convex functions, but it is unclear if the bound is tight, even for functions on $\mathbb{R}^2$.

A related line of work considers algorithm-dependent lower bounds, describing functions that are challenging for common classes of algorithms, such as Newton's

[1] Higher order methods can yield improvements under additional smoothness: if in addition f has L_2-Lipschitz Hessian and $\epsilon \le L_1^{7/3} L_2^{-4/3} D^{2/3}$, an accelerated Newton method achieves the (optimal) rate $(L_2 D^3/\epsilon)^{2/7}$ [4,32].

method and gradient descent. In this vein, Jarre [25] shows that the Chebyshev–Rosenbrock function is difficult to optimize, and that any algorithm that employs line search to determine the step size will require an exponential (in ϵ) number of iterations to find an ϵ-suboptimal point, even though the Chebyshev–Rosenbrock function has only a single stationary point. While this appears to contradict the polynomial complexity guarantees mentioned above, Cartis et al. [19] explain this by showing that the difficult Chebyshev–Rosenbrock instances have ϵ-stationary point with function value that is $\omega(\epsilon)$-suboptimal. Cartis et al. also develop algorithm-specific lower bounds on the iteration complexity of finding approximate stationary points. Their works [16,17] show that the performance guarantees for gradient descent and cubic regularization of Newton's method are tight for two-dimensional functions they construct, and they also extend these results to certain structured classes of methods [18,20].

1.2 The importance of high-dimensional constructions

To tightly characterize the algorithm- and dimension-independent complexity of finding ϵ-stationary points, one *must* construct hard instances whose domain has dimension that grows with $1/\epsilon$. The reason for this is simple: there exist algorithms with complexity that trades dependence on dimension d in favor of better $1/\epsilon$ dependence. Indeed, Vavasis [44] gives a grid-search method that, for functions with Lipschitz gradient, finds an ϵ-stationary point in $\max\{2^d, \epsilon^{-2d/(d+2)}\}$ gradient and function evaluations. Moreover, Hinder [24] exhibits a cutting-plane method that, for functions with Lipschitz first and third derivatives, finds an ϵ-stationary point in $d \cdot \epsilon^{-4/3} \log \frac{1}{\epsilon}$ gradient and function evaluations.

High-dimensional constructions are similarly unavoidable when developing lower bounds in convex optimization. There, the center-of-gravity cutting plane method (cf. [37]) finds an ϵ-suboptimal point in $d \log \frac{1}{\epsilon}$ (sub)gradient evaluations, for any continuous convex function with bounded distance to optimality. Consequently, proofs of the dimension-free lower bound for convex optimization (as we cite in the previous section) all rely on constructions whose dimensionality grows polynomially in $1/\epsilon$.

Our paper continues this well-established practice, and our lower bounds apply in the following order of quantifiers: for all $\epsilon > 0$, there exists a dimension $d \in \mathbb{N}$ such that for any $d' \geq d$ and algorithm A, there is some $f : \mathbb{R}^{d'} \to \mathbb{R}$ such that A requires at least $T(\epsilon)$ oracle queries to find an ϵ-stationary point of f. Our bounds on deterministic algorithms require dimension $d = 1 + 2T(\epsilon)$, while our bounds on all randomized algorithms require $d = c \cdot T(\epsilon)^2 \log T(\epsilon)$ for a numerical constant $c < \infty$. In contrast, the results of Vavasis [44] and Cartis et al. [16–18,20] hold with $d = 2$ independent of ϵ. Inevitably, they do so at a cost; the lower bound [44] is loose, while the lower bounds [16–18,20] apply only to certain algorithm classes (based on Taylor models) that exclude the aforementioned grid-search and cutting-plane algorithms.

1.3 Our contributions

In this paper, we consider the class of all randomized algorithms that access the function f through an *information oracle* that returns the function value, gradient, Hessian and all higher-order derivatives of f at a queried point x. Our main result (Theorem 2 in Sect. 5) is as follows. Let $p \in \mathbb{N}$ and Δ, L_p, and $\epsilon > 0$. Then, for any randomized algorithm A based on the oracle described above, there exists a function f that has L_p-Lipschitz pth derivative, satisfies $f(x^{(0)}) - f(x^\star) \le \Delta$, and is such that, with high probability, A requires at least

$$c_p \cdot \Delta L_p^{1/p} \epsilon^{-(p+1)/p}$$

oracle queries to find an ϵ-stationary point of f, where $c_p > 0$ is a constant decreasing at most polynomially in p. As explained in the previous section, the domain of the constructed function f has dimension polynomial in $1/\epsilon$.

For every p, our lower bound matches (up to a constant) known upper bounds, thereby characterizing the optimal complexity of finding stationary points. For $p = 1$, our results imply that gradient descent [37,39] is optimal among all methods (even randomized, high-order methods) operating on functions with Lipschitz continuous gradient and bounded initial sub-optimality. Therefore, to strengthen the guarantees of gradient descent one *must* introduce additional assumptions, such as convexity of f or Lipschitz continuity of $\nabla^2 f$. Similarly, in the case $p = 2$ we establish that cubic regularization of Newton's method [16,40] achieves the optimal rate $\epsilon^{-3/2}$, and for general p we show that pth order Taylor-approximation methods [7] are optimal.

These results say little about the potential of first-order methods on functions with higher-order Lipschitz derivatives, where first-order methods attain rates better than ϵ^{-2} [13]. In Part II of this series [14], we address this issue and show lower bounds for deterministic algorithms using only first-order information. The lower bounds exhibit a fundamental gap between first- and second-order methods, and nearly match the known upper bounds [13].

1.4 Our approach and paper organization

In Sect. 2 we introduce the classes of functions and algorithms we consider as well as our notion of complexity. Then, in Sect. 3, we present the generic technique we use to prove lower bound for deterministic algorithms in both this paper and Part II [14]. While essentially present in previous work, our technique abstracts away and generalizes the central arguments in many lower bounds [4,34,35,45]. The technique applies to higher-order methods and provides lower bounds for general optimization goals, including finding stationary points (our main focus), approximate minimizers, and second-order stationary points. It is also independent of whether the functions under consideration are convex, applying to any function class with appropriate rotational invariance [35]. The key building blocks of the technique are Nesterov's notion of a "chain-like" function [37], which is difficult for a certain subclass of algorithms, and a "resisting oracle" [35,37] reduction that turns a lower bound for this subclass into a lower bound for all deterministic algorithms.

In Sect. 4 we apply this generic method to produce lower bounds for *deterministic* methods (Theorem 1). The deterministic results underpin our analysis for randomized algorithms, which culminates in Theorem 2 in Sect. 5. Following Woodworth and

Srebro [45], we consider random rotations of our deterministic construction, and show that for any algorithm such a randomly rotated function is, with high probability, difficult. For completeness, in Sect. 6 we provide lower bounds on finding stationary points of functions where $\|x^{(0)} - x^\star\|$ is bounded, rather than the function value gap $f(x^{(0)}) - f(x^\star)$; these bounds have the same ϵ dependence as their bounded function value counterparts.

Notation Before continuing, we provide the conventions we adopt throughout the paper. For a sequence of vectors, subscripts denote coordinate index, while parenthesized superscripts denote element index, e.g. $x_j^{(i)}$ is the jth coordinate of the ith entry in the sequence $x^{(1)}, x^{(2)}, \ldots$. For any $p \ge 1$ and p times continuously differentiable $f : \mathbb{R}^d \to \mathbb{R}$, we let $\nabla^p f(x)$ denote the tensor of pth order partial derivatives of f at point x, so $\nabla^p f(x)$ is an order p symmetric tensor with entries

$$\left[\nabla^p f(x)\right]_{i_1,\ldots,i_p} = \nabla^p_{i_1,\ldots,i_p} f(x) = \frac{\partial^p f}{\partial x_{i_1} \cdots \partial x_{i_p}}(x) \ \text{ for } i_j \in \{1, \ldots, d\}.$$

Equivalently, we may write $\nabla^p f(x)$ as a multilinear operator $\nabla^p f(x) : (\mathbb{R}^d)^p \to \mathbb{R}$,

$$\begin{aligned}&\nabla^p f(x)\left[v^{(1)}, \ldots, v^{(p)}\right] \\ &= \sum_{i_1=1}^{d} \cdots \sum_{i_p=1}^{d} v^{(1)}_{i_1} \cdots v^{(p)}_{i_p} \frac{\partial^p f}{\partial x_{i_1} \cdots \partial x_{i_p}}(x) = \left\langle \nabla^p f(x), v^{(1)} \otimes \cdots \otimes v^{(p)} \right\rangle,\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product on tensors, defined for order k tensors T and M by $\langle T, M\rangle = \sum_{i_1,\ldots,i_k} T_{i_1,\ldots,i_k} M_{i_1,\ldots,i_k}$, and $\otimes$ denotes the Kronecker product. We let $\otimes^k d$ denote $d \times \cdots \times d$, k times, so that $T \in \mathbb{R}^{\otimes^k d}$ denotes an order k tensor.

For a vector $v \in \mathbb{R}^d$ we let $\|v\| := \sqrt{\langle v, v\rangle}$ denote the Euclidean norm of v. For a tensor $T \in \mathbb{R}^{\otimes^k d}$, the Euclidean operator norm of T is

$$\begin{aligned}\|T\|_{\mathrm{op}} := \sup_{v^{(1)},\ldots,v^{(k)}} &\Big\{ \langle T, v^{(1)} \otimes \cdots \otimes v^{(k)} \rangle \\ &= \sum_{i_1,\ldots,i_k} T_{i_1,\ldots,i_k} v^{(1)}_{i_1} \cdots v^{(k)}_{i_k} \mid \|v^{(i)}\| \le 1, i = 1, \ldots, k \Big\}.\end{aligned}$$

If T is a symmetric order k tensor, meaning that $T_{i_1,\ldots,i_k}$ is invariant to permutations of the indices (for example, $\nabla^k f(x)$ is always symmetric), then Zhang et al. [47, Thm. 2.1] show that

$$\|T\|_{\mathrm{op}} = \sup_{\|v\|=1} \left|\langle T, v^{\otimes k}\rangle\right|, \quad \text{where} \quad v^{\otimes k} = \underbrace{v \otimes v \otimes \cdots \otimes v}_{k \text{ times}}. \tag{2}$$

For vectors, the Euclidean and operator norms are identical.

For any $n \in \mathbb{N}$, we let $[n] := \{1, \ldots, n\}$ denote the set of positive integers less than or equal to n. We let $\mathcal{C}^\infty$ denote the set of infinitely differentiable functions.

We denote the ith standard basis vector by $e^{(i)}$, and let $I_d \in \mathbb{R}^{d\times d}$ denote the $d \times d$ identity matrix; we drop the subscript d when it is clear from context. For any set $\mathcal{S}$ and functions $g, h : \mathcal{S} \to [0, \infty)$ we write $g \lesssim h$ or $g = O(h)$ if there exists $c > 0$ such that $g(s) \le c \cdot h(s)$ for every $s \in \mathcal{S}$. We write $g = \widetilde{O}(h)$ if $g \lesssim h \log(h + 2)$.

2 Preliminaries

We begin our development with definitions of the classes of functions (Sect. 2.1), classes of algorithms (Sect. 2.2), and notions of complexity (Sect. 2.3) that we study.

2.1 Function classes

Measures of function regularity are crucial for the design and analysis of optimization algorithms [9,35,37]. We focus on two types of regularity conditions: Lipschitzian properties of derivatives and bounds on function value.

We first list a few equivalent definitions of Lipschitz continuity. A function $f : \mathbb{R}^d \to \mathbb{R}$ has L_p-Lipschitz pth order derivatives if it is p times continuously differentiable, and for every $x \in \mathbb{R}^d$ and direction $v \in \mathbb{R}^d$, $\|v\| \le 1$, the directional projection $f_{x,v}(t) := f(x + t \cdot v)$ of f, defined for $t \in \mathbb{R}$, satisfies

$$\left| f_{x,v}^{(p)}(t) - f_{x,v}^{(p)}(t') \right| \le L_p \left| t - t' \right|$$

for all $t, t' \in \mathbb{R}$, where $f_{x,v}^{(p)}(\cdot)$ is the pth derivative of $t \mapsto f_{x,v}(t)$. If f is $p+1$ times continuously differentiable, this is equivalent to requiring

$$\left| f_{x,v}^{(p+1)}(0) \right| \le L_p \quad \text{or} \quad \left\| \nabla^{p+1} f(x) \right\|_{\text{op}} \le L_p$$

for all $x, v \in \mathbb{R}^d$, $\|v\| \le 1$. We occasionally refer to a function with Lipschitz pth order derivatives as pth order smooth.

Complexity guarantees for finding stationary points of non-convex functions f typically depend on the function value bound $f(x^{(0)}) - \inf_x f(x)$, where $x^{(0)}$ is a pre-specified point. Without loss of generality, we take the pre-specified point to be 0 for the remainder of the paper. With that in mind, we define the following classes of functions.

Definition 1 Let $p \ge 1$, $\Delta > 0$ and $L_p > 0$. Then the set

$$\mathcal{F}_p(\Delta, L_p)$$

denotes the union, over $d \in \mathbb{N}$, of the collection of $\mathcal{C}^\infty$ functions $f : \mathbb{R}^d \to \mathbb{R}$ with L_p-Lipschitz pth derivative and $f(0) - \inf_x f(x) \le \Delta$.

The function classes $\mathcal{F}_p(\Delta, L_p)$ include functions on $\mathbb{R}^d$ for all $d \in \mathbb{N}$, following the established study of "dimension free" convergence guarantees [35,37]. As explained in Sect. 1.2, we construct explicit functions $f : \mathbb{R}^d \to \mathbb{R}$ that are difficult to optimize, where the dimension d is finite, but our choice of d grows inversely in the desired accuracy of the solution.

For our results, we also require the following important invariance notion, proposed (in the context of optimization) by Nemirovski and Yudin [35, Ch. 7.2].

Definition 2 [*Orthogonal invariance*] A class of functions $\mathcal{F}$ is *orthogonally invariant* if for every $f \in \mathcal{F}$, $f : \mathbb{R}^d \to \mathbb{R}$, and every matrix $U \in \mathbb{R}^{d' \times d}$ such that $U^\top U = I_d$, the function $f_U : \mathbb{R}^{d'} \to \mathbb{R}$ defined by $f_U(x) = f(U^\top x)$ belongs to $\mathcal{F}$.

Every function class we consider is orthogonally invariant, as $f(0) - \inf_x f(x) = f_U(0) - \inf_x f_U(x)$ and f_U has the same Lipschitz constants to all orders as f, as their collections of associated directional projections are identical.

2.2 Algorithm classes

We also require careful definition of the classes of optimization algorithms we consider. For any dimension $d \in \mathbb{N}$, an *algorithm* A (also referred to as *method*) maps functions $f : \mathbb{R}^d \to \mathbb{R}$ to a sequence of *iterates* in $\mathbb{R}^d$; that is, A is defined separately for every finite d. We let

$$\mathsf{A}[f] = \{x^{(t)}\}_{t=1}^{\infty}$$

denote the sequence $x^{(t)} \in \mathbb{R}^d$ of iterates that A generates when operating on f.

To model the computational cost of an algorithm, we adopt the *information-based complexity* framework, which Nemirovski and Yudin [35] develop (see also [1,10,43]), and view every every iterate $x^{(t)}$ as a query to an *information oracle*. Typically, one places restrictions on the information the oracle returns (e.g. only the function value and gradient at the query point) and makes certain assumptions on how the algorithm uses this information (e.g. deterministically). Our approach is syntactically different but semantically identical: we build the oracle restriction, along with any other assumption, directly into the structure of the algorithm. To formalize this, we define

$$\nabla^{(0,\ldots,p)} f(x) := \{f(x), \nabla f(x), \nabla^2 f(x), \ldots, \nabla^p f(x)\}$$

as shorthand for the response of a pth order oracle to a query at point x. When $p = \infty$ this corresponds to an oracle that reveals all derivatives at x. Our algorithm classes follow.

Deterministic algorithms For any $p \geq 0$, a *pth-order deterministic algorithm* A operating on $f : \mathbb{R}^d \to \mathbb{R}$ is one producing iterates of the form

$$x^{(i)} = \mathsf{A}^{(i)}\left(\nabla^{(0,\ldots,p)} f(x^{(1)}), \ldots, \nabla^{(0,\ldots,p)} f(x^{(i-1)})\right) \text{ for } i \in \mathbb{N},$$

where $\mathsf{A}^{(i)}$ is a measurable mapping to $\mathbb{R}^d$ (the dependence on dimension d is implicit). We denote the class of pth-order deterministic algorithms by $\mathcal{A}_{\mathsf{det}}^{(p)}$ and let $\mathcal{A}_{\mathsf{det}} := \mathcal{A}_{\mathsf{det}}^{(\infty)}$ denote the class of all deterministic algorithms based on derivative information.

As a concrete example, for any $p \geq 1$ and $L > 0$ consider the algorithm $\mathsf{REG}_{p,L} \in \mathcal{A}_{\mathsf{det}}^{(p)}$ that produces iterates by minimizing the sum of a pth order Taylor expansion and an order $p+1$ proximal term:

$$x^{(k+1)} := \arg\min_x \left\{ f(x^{(k)}) + \sum_{q=1}^{p} \langle \nabla^q f(x^{(k)}), x^{\otimes q} \rangle + \frac{L}{(p+1)!} \|x - x^{(k)}\|^{p+1} \right\}. \tag{3}$$

For $p = 1$, $\mathsf{REG}_{p,L}$ is gradient descent with step-size $1/L$, for $p = 2$ it is cubic-regularized Newton's method [40], and for general p it is a simplified form of the scheme that Birgin et al. [7] propose.

Randomized algorithms (and function-informed processes) A *pth-order randomized algorithm* A is a distribution on pth-order deterministic algorithms. We can write any such algorithm as a deterministic algorithm given access to a random uniform variable on [0, 1] (i.e. infinitely many random bits). Thus the algorithm operates on f by drawing $\xi \sim \mathsf{Uni}[0, 1]$ (independently of f), then producing iterates of the form

$$x^{(i)} = \mathsf{A}^{(i)} \left(\xi, \nabla^{(0,\ldots,p)} f(x^{(1)}), \ldots, \nabla^{(0,\ldots,p)} f(x^{(i-1)}) \right) \quad \text{for } i \in \mathbb{N}, \tag{4}$$

where $\mathsf{A}^{(i)}$ are measurable mappings into $\mathbb{R}^d$. In this case, $\mathsf{A}[f]$ is a random sequence, and we call a random process $\{x^{(t)}\}_{t\in\mathbb{N}}$ *informed by* f if it has the same law as $\mathsf{A}[f]$ for some randomized algorithm A. We let $\mathcal{A}_{\mathsf{rand}}^{(p)}$ denote the class of pth-order randomized algorithms and $\mathcal{A}_{\mathsf{rand}} := \mathcal{A}_{\mathsf{rand}}^{(\infty)}$ denote the class of randomized algorithms that use derivative-based information.

Zero-respecting sequences and algorithms While deterministic and randomized algorithms are the natural collections for which we prove lower bounds, it is useful to define an additional structurally restricted class. This class forms the backbone of our lower bound strategy (Sect. 3), as it is both 'small' enough to uniformly underperform on a single function, and 'large' enough to imply lower bounds on the natural algorithm classes.

For $v \in \mathbb{R}^d$ we let $\operatorname{supp}\{v\} := \{i \in [d] \mid v_i \neq 0\}$ denote the support (non-zero indices) of v. We extend this to tensors as follows. Let $T \in \mathbb{R}^{\otimes^k d}$ be an order k tensor, and for $i \in \{1, \ldots, d\}$ let $T_i \in \mathbb{R}^{\otimes^{k-1} d}$ be the order $(k-1)$ tensor defined by $[T_i]_{j_1,\ldots,j_{k-1}} = T_{i,j_1,\ldots,j_{k-1}}$. With this notation, we define

$$\operatorname{supp}\{T\} := \{i \in \{1, \ldots, d\} \mid T_i \neq 0\}.$$

Then for $p \in \mathbb{N}$ and any $f : \mathbb{R}^d \to \mathbb{R}$, we say that the sequence $x^{(1)}, x^{(2)}, \ldots$ is *pth order zero-respecting with respect to* f if

$$\operatorname{supp}\left\{x^{(t)}\right\} \subseteq \bigcup_{q\in[p]} \bigcup_{s<t} \operatorname{supp}\left\{\nabla^q f(x^{(s)})\right\} \quad \text{for each } t \in \mathbb{N}. \tag{5}$$

The definition (5) says that $x_i^{(t)} = 0$ if all partial derivatives involving the ith coordinate of f (up to the pth order) are zero. For $p = 1$, this definition is equivalent to the requirement that for every t and $j \in [d]$, if $\nabla_j f(x^{(s)}) = 0$ for $s < t$, then $x_j^{(t)} = 0$. The requirement (5) implies that $x^{(1)} = 0$.

An algorithm $\mathsf{A} \in \mathcal{A}_{\mathsf{rand}}$ is *pth order zero-respecting* if for any $f : \mathbb{R}^d \to \mathbb{R}$, the (potentially random) iterate sequence $\mathsf{A}[f]$ is pth order zero respecting w.r.t. f. Informally, an algorithm is zero-respecting if it never explores coordinates which appear not to affect the function. When initialized at the origin, most common first- and second-order optimization methods are zero-respecting, including gradient descent (with and without Nesterov acceleration), conjugate gradient [23], BFGS and L-BFGS [29,41],[2] Newton's method (with and without cubic regularization [40]) and trust-region methods [22]. We denote the class of pth order zero-respecting algorithms by $\mathcal{A}_{\mathsf{zr}}^{(p)}$, and let $\mathcal{A}_{\mathsf{zr}} := \mathcal{A}_{\mathsf{zr}}^{(\infty)}$.

In the literature on lower bounds for first-order convex optimization, it is common to assume that methods only query points in the span of the gradients they observe [3,37]. Our notion of zero-respecting algorithms generalizes this assumption to higher-order methods, but even first-order zero-respecting algorithms are slightly more general. For example, coordinate descent methods [38] are zero-respecting, but they generally do not remain in the span of the gradients.

2.3 Complexity measures

With the definitions of function and algorithm class in hand, we turn to formalizing our notion of complexity: what is the best performance an algorithm in class $\mathcal{A}$ can achieve *for all* functions in class $\mathcal{F}$? As we consider finding stationary points of f, the natural performance measure is the number of iterations (oracle queries) required to find a point x such that $\|\nabla f(x)\| \le \epsilon$. Thus for a deterministic sequence $\{x^{(t)}\}_{t\in\mathbb{N}}$ we define

$$\mathsf{T}_\epsilon\big(\{x^{(t)}\}_{t\in\mathbb{N}}, f\big) := \inf\left\{t \in \mathbb{N} \mid \left\|\nabla f(x^{(t)})\right\| \le \epsilon\right\},$$

and refer to it as the *complexity of* $\{x^{(t)}\}_{t\in\mathbb{N}}$ *on* f. As we consider randomized algorithms as well, for a random process $\{x^{(t)}\}_{t\in\mathbb{N}}$ with probability distribution P, meaning for a set $A \subset (\mathbb{R}^d)^{\mathbb{N}}$ the probability that $\{x^{(t)}\}_{t\in\mathbb{N}} \in A$ is $P(A)$, we define

$$\mathsf{T}_\epsilon\big(P, f\big) := \inf\left\{t \in \mathbb{N} \mid P\left(\left\|\nabla f(x^{(s)})\right\| > \epsilon \text{ for all } s \le t\right) \le \frac{1}{2}\right\}. \tag{6}$$

The complexity $\mathsf{T}_\epsilon\big(P, f\big)$ is also the median of the random variable $\mathsf{T}_\epsilon\big(\{x^{(t)}\}_{t\in\mathbb{N}}, f\big)$ for $\{x^{(t)}\}_{t\in\mathbb{N}} \sim P$. By Markov's inequality, definition (6) provides a lower bound on expectation-based alternatives, as

[2] If the initial Hessian approximation is a diagonal matrix, as is typical.

$$\inf\left\{t\in\mathbb{N}\mid\mathbb{E}_P\left[\|\nabla f(x^{(t)})\|\right]\le\epsilon\right\}\ge\mathsf{T}_{2\epsilon}\big(P,f\big)\ \text{ and }\ \mathbb{E}_P\left[\mathsf{T}_\epsilon\big(\{x^{(t)}\}_{t\in\mathbb{N}},f\big)\right]\ge\frac{1}{2}\mathsf{T}_\epsilon\big(P,f\big).$$

(Here $\mathbb{E}_P$ denotes expectation taken according to the distribution P.)

To measure the performance of algorithm A on function f, we evaluate the iterates it produces from f, and with mild abuse of notation, we define

$$\mathsf{T}_\epsilon\big(\mathsf{A},f\big):=\mathsf{T}_\epsilon\big(\mathsf{A}[f],f\big)$$

as the complexity of A on f. With this setup, we define the *complexity of algorithm class $\mathcal{A}$ on function class $\mathcal{F}$* as

$$\mathcal{T}_\epsilon\big(\mathcal{A},\mathcal{F}\big):=\inf_{\mathsf{A}\in\mathcal{A}}\sup_{f\in\mathcal{F}}\mathsf{T}_\epsilon\big(\mathsf{A},f\big).\tag{7}$$

Many algorithms guarantee "dimension independent" convergence [37] and thus provide upper bounds for the quantity (7). A careful tracing of constants in the analysis of Birgin et al. [7] implies that the generalized regularization scheme $\mathsf{REG}_{p,L}$ defined by the recursion (3) guarantees

$$\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{det}}^{(p)}\cap\mathcal{A}_{\mathsf{zr}}^{(p)},\mathcal{F}_p(\Delta,L_p)\big)\le\sup_{f\in\mathcal{F}_p(\Delta,L_p)}\mathsf{T}_\epsilon\big(\mathsf{REG}_{p,L_p},f\big)\lesssim\Delta L_p^{1/p}\epsilon^{-(1+p)/p}\tag{8}$$

for all $p\in\mathbb{N}$. In this paper we prove these rates are sharp to within (p-dependent) constant factors.

While definition (7) is our primary notion of complexity, our proofs provide bounds on smaller quantities than (7) that also carry meaning. For zero-respecting algorithms, we exhibit a single function f and bound $\inf_{\mathsf{A}\in\mathcal{A}_{\mathsf{zr}}}\mathsf{T}_\epsilon\big(\mathsf{A},f\big)$ from below, in effect interchanging the inf and sup in (7). This implies that all zero-respecting algorithms share a common vulnerability. For randomized algorithms, we exhibit a distribution P supported on functions of a fixed dimension d, and we lower bound the average $\inf_{\mathsf{A}\in\mathcal{A}_{\mathsf{rand}}}\int\mathsf{T}_\epsilon\big(\mathsf{A},f\big)dP(f)$, bounding the *distributional complexity* [10,35], which is never greater than worst-case complexity (and is equal for randomized and deterministic algorithms). Even randomized algorithms share a common vulnerability: functions drawn from P.

3 Anatomy of a lower bound

In this section we present a generic approach to proving lower bounds for optimization algorithms. The basic techniques we use are well-known and applied extensively in the literature on lower bounds for convex optimization [4,35,37,45]. However, here we generalize and abstract away these techniques, showing how they apply to high-order methods, non-convex functions, and various optimization goals (e.g. ϵ-stationarity, ϵ-optimality).

3.1 Zero-chains

Nesterov [37, Chapter 2.1.2] proves lower bounds for smooth convex optimization problems using the "chain-like" quadratic function

$$f(x) := \frac{1}{2}(x_1 - 1)^2 + \frac{1}{2}\sum_{i=1}^{d-1}(x_i - x_{i+1})^2, \tag{9}$$

which he calls the "worst function in the world." The important property of f is that for every $i \in [d]$, $\nabla_i f(x) = 0$ whenever $x_{i-1} = x_i = x_{i+1} = 0$ (with $x_0 := 1$ and $x_{d+1} := 0$). Thus, if we "know" only the first $t-1$ coordinates of f, i.e. are able to query only vectors x such $x_t = x_{t+1} = \cdots = x_d = 0$, then any x we query satisfies $\nabla_s f(x) = 0$ for $s > t$; we only "discover" a single new coordinate t. We generalize this chain structure to higher-order derivatives as follows.

Definition 3 For $p \in \mathbb{N}$, a function $f : \mathbb{R}^d \to \mathbb{R}$ is a *pth-order zero-chain* if for every $x \in \mathbb{R}^d$,

$$\operatorname{supp}\{x\} \subseteq \{1, \ldots, i-1\} \text{ implies } \bigcup_{q \in [p]} \operatorname{supp}\left\{\nabla^q f(x)\right\} \subseteq \{1, \ldots, i\}.$$

We say f is a *zero-chain* if it is a pth-order zero-chain for every $p \in \mathbb{N}$.

In our terminology, Nesterov's function (9) is a first-order zero-chain but not a second-order zero-chain, as $\operatorname{supp}\left\{\nabla^2 f(0)\right\} = [d]$. Informally, at a point for which $x_{i-1} = x_i = \cdots = x_d = 0$, a zero-chain appears constant in $x_i, x_{i+1}, \ldots, x_d$. Zero-chains structurally limit the rate with which zero-respecting algorithms acquire information from derivatives. We formalize this in the following observation, whose proof is a straightforward induction; see Table 1 for an illustration.

Observation 1 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a pth order zero-chain and let $x^{(1)} = 0, x^{(2)}, \ldots$ be a pth order zero-respecting sequence with respect to f. Then $x_j^{(t)} = 0$ for $j \geq t$ and all $t \leq d$.*

Proof We show by induction on k that $\operatorname{supp}\left\{x^{(t)}\right\} \subseteq [t-1]$ for every $t \leq k$; the case $k = d$ is the required result. The case $k = 1$ holds since $x^{(1)} = 0$. If the hypothesis holds for some $k < d$ then by Definition 3 we have $\cup_{q \in [p]} \operatorname{supp}\left\{\nabla^q f(x^{(t)})\right\} \subseteq \{1, \ldots, t\}$ for every $t \leq k$. Therefore, by the zero-respecting property (5), we have $\operatorname{supp}\left\{x^{(k+1)}\right\} \subseteq \cup_{q \in [p]} \cup_{t < k+1} \operatorname{supp}\left\{\nabla^q f(x^{(t)})\right\} \subseteq [k]$, completing the induction. □

3.2 A lower bound strategy

The preceding discussion shows that zero-respecting algorithms take many iterations to "discover" all the coordinates of a zero-chain. In the following observation, we formalize how finding a suitable zero-chain provides a lower bound on the performance of zero-respecting algorithms.

Table 1 Illustration of Observation 1: a zero-respecting algorithm operating on a zero-chain

Iteration	Information	Coordinate $j = 1$	2	3	4	$\cdots$	$d - 1$	d
$t = 0$	$x^{(0)}$	0	0	0	0	$\cdots$	0	0
	$\nabla f(x^{(0)})$	$*$	0	0	0	$\cdots$	0	0
$t = 1$	$x^{(1)}$	$*$	0	0	0	$\cdots$	0	0
	$\nabla f(x^{(1)})$	$*$	$*$	0	0	$\cdots$	0	0
$t = 2$	$x^{(2)}$	$*$	$*$	0	0	$\cdots$	0	0
	$\nabla f(x^{(2)})$	$*$	$*$	$*$	0	$\cdots$	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$t = d - 1$	$x^{(d-1)}$	$*$	$*$	$*$	$*$	$\cdots$	$*$	0
	$\nabla f(x^{(d-1)})$	$*$	$*$	$*$	$*$	$\cdots$	$*$	$*$

We indicate the nonzero entries of the iterates and the gradients by $*$

Observation 2 *Consider $\epsilon > 0$, a function class $\mathcal{F}$, and $p, T \in \mathbb{N}$. If $f : \mathbb{R}^T \to \mathbb{R}$ satisfies*

i. *f is a pth-order zero-chain,*
ii. *f belongs to the function class, i.e. $f \in \mathcal{F}$, and*
iii. *$\|\nabla f(x)\| > \epsilon$ for every x such that $x_T = 0$;*[3]

then $\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{zr}}^{(p)}, \mathcal{F}\big) \geq \mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{zr}}^{(p)}, \{f\}\big) > T$.

Proof For $\mathsf{A} \in \mathcal{A}_{\mathsf{zr}}^{(p)}$ and $\{x^{(t)}\}_{t\in\mathbb{N}} = \mathsf{A}[f]$ we have by Observation 1 that $x_T^{(t)} = 0$ for all $t \leq T$ and the large gradient property (iii) then implies $\big\|\nabla f(x^{(t)})\big\| > \epsilon$ for all $t \leq T$. Therefore $\mathsf{T}_\epsilon\big(\mathsf{A}, f\big) > T$, and since this holds for any $\mathsf{A} \in \mathcal{A}_{\mathsf{zr}}^{(p)}$ we have

$$\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{zr}}^{(p)}, \mathcal{F}\big) = \inf_{\mathsf{A}\in\mathcal{A}_{\mathsf{zr}}^{(p)}} \sup_{\tilde{f}\in\mathcal{F}} \mathsf{T}_\epsilon\big(\mathsf{A}, \tilde{f}\big) \geq \sup_{\tilde{f}\in\mathcal{F}} \inf_{\mathsf{A}\in\mathcal{A}_{\mathsf{zr}}^{(p)}} \mathsf{T}_\epsilon\big(\mathsf{A}, \tilde{f}\big) \geq \inf_{\mathsf{A}\in\mathcal{A}_{\mathsf{zr}}^{(p)}} \mathsf{T}_\epsilon\big(\mathsf{A}, f\big) > T.$$

□

If f is a zero-chain, then so is the function $x \mapsto \mu f(x/\sigma)$ for any multiplier μ and scale parameter σ. This is useful for our development, as we construct zero-chains $\{g_T\}_{T\in\mathbb{N}}$ such that $\|\nabla g_T(x)\| > c$ for every x with $x_T = 0$ and some constant $c > 0$. By setting $f(x) = \mu g_T(x/\sigma)$, then choosing T, μ, and σ to satisfy conditions (ii) and (iii), we obtain a lower bound. As our choice of T is also the final lower bound, it must grow to infinity as ϵ tends to zero. Thus, the hard functions we construct are fundamentally high-dimensional, making this strategy suitable only for dimension-free lower bounds.

[3] We can readily adapt this property for lower bounds on other termination criteria, e.g. require $f(x) - \inf_y f(y) > \epsilon$ for every x such that $x_T = 0$.

3.3 From deterministic to zero-respecting algorithms

Zero-chains allow us to generate strong lower bounds for zero-respecting algorithms. The following reduction shows that these lower bounds are valid for deterministic algorithms as well.

Proposition 1 *Let $p \in \mathbb{N} \cup \{\infty\}$, $\mathcal{F}$ be an orthogonally invariant function class and $\epsilon > 0$. Then*

$$\mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{det}}, \mathcal{F}\big) \geq \mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{zr}}, \mathcal{F}\big).$$

We also give a variant of Proposition 1 that is tailored to lower bounds constructed by means of Observation 2 and allows explicit accounting of dimensionality.

Proposition 2 *Let $p \in \mathbb{N} \cup \{\infty\}$, $\mathcal{F}$ be an orthogonally invariant function class, $f \in \mathcal{F}$ with domain of dimension d, and $\epsilon > 0$. If $\mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{zr}}, \{f\}\big) \geq T$, then*

$$\mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{det}}, \mathcal{F}\big) \geq \mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{det}}, \{f_U \mid U \in \mathsf{O}(d+T, d)\}\big) \geq T,$$

where $f_U := f(U^\top z)$ and $\mathsf{O}(d+T, d)$ is the set of $(d+T) \times d$ orthogonal matrices, so that $\{f_U \mid U \in \mathsf{O}(d+T, d)\}$ contains only function with domain of dimension $d+T$.

The proofs of Propositions 1 and 2 , given in Appendix A, build on the classical notion of a *resisting oracle* [35,37], which we briefly sketch here. Let $\mathsf{A} \in \mathcal{A}_{\mathsf{det}}$, and let $f \in \mathcal{F}$, $f : \mathbb{R}^d \to \mathbb{R}$. We adversarially select an orthogonal matrix $U \in \mathbb{R}^{d' \times d}$ (for some finite $d' > d$) such that on the function $f_U := f(U^\top z) \in \mathcal{F}$ the algorithm A behaves as if it was a zero-respecting algorithm. In particular, U is sequentially constructed such that for the function $f_U(z)$ the sequence $U^\top \mathsf{A}[f_U] \subset \mathbb{R}^d$ is zero-respecting with respect to f. Thus, there exists an algorithm $\mathsf{Z}_{\mathsf{A}} \in \mathcal{A}_{\mathsf{det}} \cap \mathcal{A}_{\mathsf{zr}}$ such that $\mathsf{Z}_{\mathsf{A}}[f] = U^\top \mathsf{A}[f_U]$, implying $\mathsf{T}_\epsilon\big(\mathsf{A}, f_U\big) = \mathsf{T}_\epsilon\big(\mathsf{Z}_{\mathsf{A}}, f\big)$. Therefore,

$$\begin{aligned}\inf_{\mathsf{A} \in \mathcal{A}_{\mathsf{det}}} \sup_{f \in \mathcal{F}} \mathsf{T}_\epsilon\big(\mathsf{A}, f\big) &= \inf_{\mathsf{A} \in \mathcal{A}_{\mathsf{det}}} \sup_{f \in \mathcal{F}, U} \mathsf{T}_\epsilon\big(\mathsf{A}, f_U\big) \\ &= \inf_{\mathsf{A} \in \mathcal{A}_{\mathsf{det}}} \sup_{f \in \mathcal{F}} \mathsf{T}_\epsilon\big(\mathsf{Z}_{\mathsf{A}}, f\big) \geq \inf_{\mathsf{A} \in \mathcal{A}_{\mathsf{zr}}} \sup_{f \in \mathcal{F}} \mathsf{T}_\epsilon\big(\mathsf{A}, f\big),\end{aligned}$$

giving Propositions 1 and 2 follows similarly, and for it we may take $d' = d + T$.

The adversarial rotation argument that yields Propositions 1 and 2 is more or less apparent in the proofs of previous lower bounds in convex optimization [4,35,45] for deterministic algorithms. We believe it is instructive to separate the proof of lower bounds on $\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{zr}}, \mathcal{F}\big)$ and the reduction from $\mathcal{A}_{\mathsf{det}}$ to $\mathcal{A}_{\mathsf{zr}}$, as the latter holds in great generality. Indeed, Propositions 1 and 2 hold for any complexity measure $\mathsf{T}_\epsilon\big(\cdot, \cdot\big)$ that satisfies

1. Orthogonal invariance: for every $f : \mathbb{R}^d \to \mathbb{R}$, every $U \in \mathbb{R}^{d' \times d}$ such that $U^\top U = I_d$ and every sequence $\{z^{(t)}\}_{t \in \mathbb{N}} \subset \mathbb{R}^{d'}$, we have

$$\mathsf{T}_\epsilon\big(\{z^{(t)}\}_{t \in \mathbb{N}}, f(U^\top \cdot)\big) = \mathsf{T}_\epsilon\big(\{U^\top z^{(t)}\}_{t \in \mathbb{N}}, f\big).$$

2. "Stopping time" invariance: for any $T_0 \in \mathbb{N}$, if $\mathsf{T}_\epsilon\big(\{x^{(t)}\}_{t\in\mathbb{N}}, f\big) \le T_0$ then $\mathsf{T}_\epsilon\big(\{x^{(t)}\}_{t\in\mathbb{N}}, f\big) = \mathsf{T}_\epsilon\big(\{\hat{x}^{(t)}\}_{t\in\mathbb{N}}, f\big)$ for any sequence $\{\hat{x}^{(t)}\}_{t\in\mathbb{N}}$ such that $\hat{x}^{(t)} = x^{(t)}$ for $t \le T_0$.

These properties hold for the typical performance measures used in optimization. Examples include time to ϵ-optimality, in which case $\mathsf{T}_\epsilon\big(\{x^{(t)}\}_{t\in\mathbb{N}}, f\big) = \inf\{t \in \mathbb{N} \mid f(x^{(t)}) - \inf_x f(x) \le \epsilon\}$, and the second-order stationarity desired in many non-convex optimization problems [15,26,40], where for $\epsilon_1, \epsilon_2 > 0$ we define $\mathsf{T}_\epsilon\big(\{x^{(t)}\}_{t\in\mathbb{N}}, f\big) = \inf\{t \in \mathbb{N} \mid \|\nabla f(x^{(t)})\| \le \epsilon_1 \text{ and } \nabla^2 f(x^{(t)}) \succeq -\epsilon_2 I\}$.

3.4 Randomized algorithms

Propositions 1 and 2 do not apply to randomized algorithms, as they require the adversary (maximizing choice of f) to simulate the action of A on f. To handle randomized algorithms, we strengthen the notion of a zero-chain as follows.

Definition 4 A function $f : \mathbb{R}^d \to \mathbb{R}$ is a *robust zero-chain* if for every $x \in \mathbb{R}^d$,

$$|x_j| < 1/2,\ \forall j \ge i \ \text{ implies } \ f(y) \\ = f(y_1, \ldots, y_i, 0, \ldots, 0) \ \text{ for all } y \text{ in a neighborhood of } x.$$

A robust zero-chain is also an "ordinary" zero-chain. In Sect. 5 we replace the adversarial rotation U of Sect. 3.3 with an orthogonal matrix drawn uniformly at random, and consider the random function $f_U(x) = f(U^\top x)$, where f is a robust zero-chain. We adapt a lemma by Woodworth and Srebro [45], and use it to show that for every $\mathsf{A} \in \mathcal{A}_{\mathsf{rand}}$, $A[f_U]$ satisfies an approximate form of Observation 1 (w.h.p.) whenever the iterates $\mathsf{A}[f_U]$ have bounded norm. With further modification of f_U to handle unbounded iterates, our zero-chain strategy yields a strong distributional complexity lower bound on $\mathcal{A}_{\mathsf{rand}}$.

4 Lower bounds for zero-respecting and deterministic algorithms

For our first main results, we provide lower bounds on the complexity of all deterministic algorithms for finding stationary points of smooth, potentially non-convex functions. By Observation 2 and Proposition 1, to prove a lower bound on deterministic algorithms it is sufficient to construct a function that is difficult for zero-respecting algorithms. For fixed $T > 0$, we define the (unscaled) hard instance $\bar{f}_T : \mathbb{R}^d \to \mathbb{R}$ as

$$\bar{f}_T(x) = -\Psi(1)\,\Phi(x_1) + \sum_{i=2}^{T} \big[\Psi(-x_{i-1})\,\Phi(-x_i) - \Psi(x_{i-1})\,\Phi(x_i)\big], \quad (10)$$

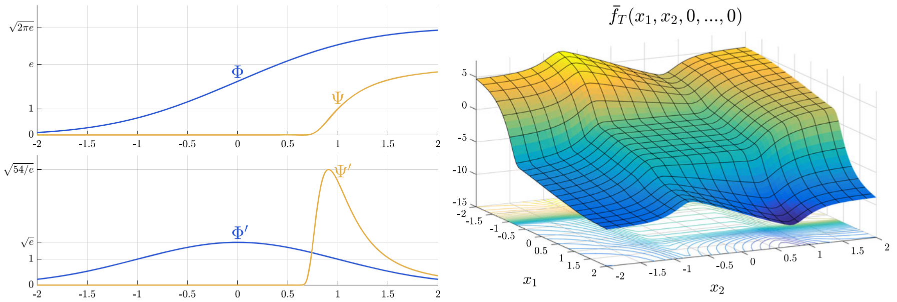

Fig. 1 Hard instance for full derivative information. Left: the functions Ψ and Φ (top) and their derivatives (bottom). Right: Surface and contour plot of a two-dimensional cross-section of the hard instance $\bar{f}_T$

where the component functions are

$$\Psi(x) := \begin{cases} 0 & x \le 1/2 \\ \exp\left(1 - \frac{1}{(2x-1)^2}\right) & x > 1/2 \end{cases} \quad \text{and} \quad \Phi(x) = \sqrt{e}\int_{-\infty}^{x} e^{-\frac{1}{2}t^2}dt.$$

Our construction, illustrated in Fig. 1, has two key properties. First is that f is a zero-chain (Observation 3 in the sequel). Second, as we show in Lemma 2, $\|\nabla \bar{f}_T(x)\|$ is large unless $|x_i| \ge 1$ for every $i \in [T]$. These properties make it hard for any zero-respecting method to find a stationary point of scaled versions of $\bar{f}_T$, and coupled with Proposition 1, this gives a lower bound for deterministic algorithms.

4.1 Properties of the hard instance

Before turning to the main theorem of this section, we catalogue the important properties of the functions Ψ, Φ and $\bar{f}_T$.

Lemma 1 *The functions Ψ and Φ satisfy the following.*

i. *For all $x \le \frac{1}{2}$ and all $k \in \mathbb{N}$, $\Psi^{(k)}(x) = 0$.*
ii. *For all $x \ge 1$ and $|y| < 1$, $\Psi(x)\Phi'(y) > 1$.*
iii. *Both Ψ and Φ are infinitely differentiable, and for all $k \in \mathbb{N}$ we have*

$$\sup_x |\Psi^{(k)}(x)| \le \exp\left(\frac{5k}{2}\log(4k)\right) \quad \textit{and} \quad \sup_x |\Phi^{(k)}(x)| \le \exp\left(\frac{3k}{2}\log\frac{3k}{2}\right).$$

iv. *The functions and derivatives Ψ, Ψ', Φ and Φ' are non-negative and bounded, with*

$$0 \le \Psi < e, \;\; 0 \le \Psi' \le \sqrt{54/e}, \;\; 0 < \Phi < \sqrt{2\pi e}, \;\; \textit{and} \;\; 0 < \Phi' \le \sqrt{e}.$$

We prove Lemma 1 in Appendix B.1. The remainder our development relies on Ψ and Φ only through Lemma 1. Therefore, the precise choice of Ψ, Φ is not particularly special; any two functions with properties similar to Lemma 1 will yield similar lower bounds.

The key consequence of Lemma 1.i is that the function f is a robust zero-chain (see Definition 4) and consequently also a zero-chain (Definition 3):

Observation 3 *For any $j > 1$, if $|x_{j-1}|, |x_j| < 1/2$ then $\bar{f}_T(y) = \bar{f}_T(y_1, \ldots, y_{j-1}, 0, y_{j+1}, \ldots, y_T)$ for all y in a neighborhood of x.*

Applying Observation 3 for $j = i+1, \ldots, T$ gives that $\bar{f}_T$ is a robust zero-chain by Definition 4. Taking derivatives of $\bar{f}_T(x_1, \ldots, x_i, 0, \ldots, 0)$ with respect to x_j, $j > i$, shows that $\bar{f}_T$ is also a zero-chain by Definition 3. Thus, Observation 1 shows that any zero-respecting algorithm operating on $\bar{f}_T$ requires $T+1$ iterations to find a point where $x_T \neq 0$.

Next, we establish the "large gradient property" that $\nabla \bar{f}_T(x)$ must be large if any coordinate of x is near zero.

Lemma 2 *If $|x_i| < 1$ for any $i \le T$, then there exists $j \le i$ such that $|x_j| < 1$ and*

$$\left\|\nabla \bar{f}_T(x)\right\| \ge \left|\frac{\partial}{\partial x_j} \bar{f}_T(x)\right| > 1.$$

Proof We take $j \le i$ to be the smallest j for which $|x_j| < 1$, so that $|x_{j-1}| \ge 1$ (where we use the shorthand $x_0 \equiv 1$). Therefore, we have

$$\begin{aligned}
&\frac{\partial \bar{f}_T}{\partial x_j}(x) \\
&\quad = -\Psi\left(-x_{j-1}\right)\Phi'\left(-x_j\right) - \Psi\left(x_{j-1}\right)\Phi'\left(x_j\right) - \Psi'\left(-x_j\right)\Phi\left(-x_{j+1}\right) \\
&\qquad - \Psi'\left(x_j\right)\Phi\left(x_{j+1}\right) \\
&\quad \overset{(i)}{\le} -\Psi\left(-x_{j-1}\right)\Phi'\left(-x_j\right) - \Psi\left(x_{j-1}\right)\Phi'\left(x_j\right) \\
&\quad \overset{(ii)}{=} -\Psi(|x_{j-1}|)\Phi'\left(x_j \operatorname{sign}(x_{j-1})\right) \overset{(iii)}{<} -1.
\end{aligned}$$

In the chain of inequalities, inequality (i) follows because $\Psi'(x)\Phi(y) \ge 0$ for every x, y; inequality (ii) follows because $\Psi(x) = 0$ for $x \le 1/2$, while equality (iii) follows from Lemma 1.ii and the pairing of $|x_j| < 1$ and $|x_{j-1}| \ge 1$. □

Finally, we verify that $\bar{f}_T$ meets the smoothness and boundedness requirements of the function classes we consider.

Lemma 3 *The function $\bar{f}_T$ satisfies the following.*

i. *We have $\bar{f}_T(0) - \inf_x \bar{f}_T(x) \le 12T$.*
ii. *For all $x \in \mathbb{R}^d$, $\left\|\nabla \bar{f}_T(x)\right\| \le 23\sqrt{T}$.*
iii. *For every $p \ge 1$, the p-th order derivatives of $\bar{f}_T$ are ℓ_p-Lipschitz continuous, where $\ell_p \le \exp(\frac{5}{2} p \log p + cp)$ for a numerical constant $c < \infty$.*

The proof of Lemma 3 is technical, so we defer it to Appendix B.2. In the lemma, Properties i and iii allow us to guarantee that appropriately scaled versions of $\bar{f}_T$ are in $\mathcal{F}_p(\Delta, L_p)$. Property is ii is necessary for analysis of the randomized construction in Sect. 5.

4.2 Lower bounds for zero-respecting and deterministic algorithms

We can now state and prove a lower bound for finding stationary points of pth order smooth functions using full derivative information and zero-respecting algorithms (the class $\mathcal{A}_{\mathsf{zr}}$). Proposition 1 transforms this bound into one on all deterministic algorithms (the class $\mathcal{A}_{\mathsf{det}}$).

Theorem 1 *There exist numerical constants $0 < c_0, c_1 < \infty$ such that the following lower bound holds. Let $p \geq 1$, $p \in \mathbb{N}$, and let Δ, L_p, and ϵ be positive. Then*

$$\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{det}}, \mathcal{F}_p(\Delta, L_p)\big) \geq \mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{zr}}, \mathcal{F}_p(\Delta, L_p)\big) \geq c_0 \Delta \left(\frac{L_p}{\ell_p}\right)^{1/p} \epsilon^{-\frac{1+p}{p}}$$

where $\ell_p \leq e^{\frac{5}{2}p\log p + c_1 p}$. The lower bound holds even if we restrict $\mathcal{F}_p(\Delta, L_p)$ to functions whose domain has dimension $1 + 2c_0\Delta(L_p/\ell_p)^{1/p}\epsilon^{-\frac{1+p}{p}}$.

Before we prove the theorem, a few remarks are in order. First, our lower bound matches the upper bound (8) that pth-order regularization schemes achieve [7], up to a constant depending polynomially on p. Thus, although our lower bound applies to algorithms given access to $\nabla^q f(x)$ for all $q \in \mathbb{N}$, only the first p derivatives are necessary to achieve minimax optimal scaling in Δ, L_p, and ϵ.

Second, inspection of the proof shows that we actually bound smaller quantities than the complexity defined in Eq. (7). Indeed, we show that taking $T \gtrsim \Delta(L_p/\ell_p)^{1/p}\epsilon^{-\frac{1+p}{p}}$ in the construction (10) and appropriately scaling $\bar{f}_T$ yields a function $f : \mathbb{R}^T \to \mathbb{R}$ that has L_p-Lipschitz continuous pth derivative, and for which *any* zero-respecting algorithm generates iterates such that $\|\nabla f(x^{(t)})\| > \epsilon$ for every $t \leq T$. That is,

$$\inf_{\mathsf{A} \in \mathcal{A}_{\mathsf{zr}}} \mathsf{T}_\epsilon\big(\mathsf{A}, f\big) > T \gtrsim \Delta L_p^{1/p} \epsilon^{-\frac{1+p}{p}},$$

which is stronger than a lower bound on $\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{zr}}, \mathcal{F}_p(\Delta, L_p)\big)$. Combined with the reduction in Proposition 2, this implies that for any deterministic algorithm $\mathsf{A} \in \mathcal{A}_{\mathsf{det}}$ there exists orthogonal $U \in \mathbb{R}^{(2T+1)\times T}$ for which $f_U(x) = f(U^\top x)$ is difficult, i.e. $\mathsf{T}_\epsilon\big(\mathsf{A}, f(U^\top \cdot)\big) > T$.

Finally, the scaling of ℓ_p with p may appear strange, or perhaps extraneous. We provide two viewpoints on this. First, one expects that the smoothness constants L_p should grow quickly as p grows; for $\mathcal{C}^\infty$ functions such as $\phi(t) = e^{-t^2}$ or $\phi(t) = \log(1 + e^t)$, $\sup_t |\phi^{(p)}(t)|$ grows super-exponentially in p. Indeed, ℓ_p is the Lipschitz constant of the pth derivative of $\bar{f}_T$. Second, the cases of main practical interest are

$p \in \{1, 2\}$, where $\ell_p^{1/p} \lesssim p^{\frac{5}{2}}$ can be considered a numerical constant. This is because, for $p \geq 3$, the only known methods with dimension-free rate of convergence $\epsilon^{-(p+1)/p}$ [7] require full access to third derivatives, which is generally impractical. Therefore, a realistic discussion of the complexity of finding stationary point with smoothness of order $p \geq 3$ must include additional restrictions on the algorithm class.

4.3 Proof of Theorem 1

To prove Theorem 1, we set up the hard instance $f : \mathbb{R}^T \to \mathbb{R}$ for some integer T by appropriately scaling f defined in Eq. (10),

$$f(x) := \frac{L_p \sigma^{p+1}}{\ell_p} \bar{f}_T(x/\sigma),$$

for some scale parameter $\sigma > 0$ to be determined, where $\ell_p \leq e^{2.5p \log p + c_1}$ is as in Lemma 3.iii. We wish to show f satisfies Observation 2. Observation 3 implies Observation 2.i (f is a zero-chain). Therefore it remains to show parts ii and iii of Observation 2. Consider any $x \in \mathbb{R}^T$ such that $x_T = 0$. Applying Lemma 2 guarantees that $\left\|\nabla \bar{f}_T(x/\sigma)\right\| > 1$, and therefore

$$\|\nabla f(x)\| = \frac{L_p \sigma^p}{\ell_p} \left\|\nabla \bar{f}_T(x/\sigma)\right\| > \frac{L_p \sigma^p}{\ell_p}. \tag{11}$$

It remains to choose T and σ based on ϵ such that $\|\nabla f(x)\| > \epsilon$ and $f \in \mathcal{F}_p(\Delta, L_p)$. By the lower bound (11), the choice $\sigma = (\ell_p \epsilon / L_p)^{1/p}$ guarantees $\|\nabla f(x)\| > \epsilon$. We note that $\nabla^{p+1} f(x) = (L_p/\ell_p) \nabla^{p+1} f(x/\sigma)$ and therefore by Lemma 3.iii we have that the p-th order derivatives of f are L_p-Lipschitz continuous. Thus, to ensure $f \in \mathcal{F}_p(\Delta, L_p)$ it suffices to show that $f(0) - \inf_x f(x) \leq \Delta$. By the first part of Lemma 3 we have

$$f(0) - \inf_x f(x) = \frac{L_p \sigma^{p+1}}{\ell_p} (\bar{f}_T(0) - \inf_x \bar{f}_T(x)) \leq \frac{12 L_p \sigma^{p+1}}{\ell_p} T = \frac{12 \ell_p^{1/p} \epsilon^{\frac{1+p}{p}}}{L_p^{1/p}} T,$$

where in the last transition we substituted $\sigma = (\ell_p \epsilon / L_p)^{1/p}$. We conclude that $f \in \mathcal{F}_p(\Delta, L_p)$ and $T = \left\lfloor \frac{\Delta L_p^{1/p}}{12 \ell_p^{1/p}} \epsilon^{-\frac{1+p}{p}} \right\rfloor$ so by Lemma 2, $\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{zr}}, \mathcal{F}_p(\Delta, L_p)\big) \geq \mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{zr}}, \{f\}\big) \geq 1 + T \geq \frac{\Delta L_p^{1/p}}{12 \ell_p^{1/p} \epsilon^{\frac{1+p}{p}}}$, with ℓ_p bounded from above as in Lemma 3.iii. By Proposition 2, this bound transfers to $\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{det}}, \mathcal{F}_p(\Delta, L_p)\big)$, where functions of dimension $2T + 1$ suffice to establish it.

5 Lower bounds for randomized algorithms

With our lower bounds on the complexity of deterministic algorithms established, we turn to the class of all randomized algorithms. We provide strong *distributional complexity* lower bounds by exhibiting a distribution on functions such that a function drawn from it is "difficult" for *any* randomized algorithm, with high probability. We do this via the composition of a random orthogonal transformation with the function $\bar{f}_T$ defined in (10).

The key steps in our deterministic bounds are (a) to show that any algorithm can "discover" at most one coordinate per iteration and (b) finding an approximate stationary point requires "discovering" T coordinates. In the context of randomized algorithms, we must elaborate this development in two ways. First, in Sect. 5.1 we provide a "robust" analogue of Observation 1 (step (a) above): we show that for a random orthogonal matrix U, any sequence of *bounded* iterates $\{x^{(t)}\}_{t\in\mathbb{N}}$ based on derivatives of $\bar{f}_T(U^\top \cdot)$ must (with high probability) satisfy that $|\langle x^{(t)}, u^{(j)}\rangle| \le \frac{1}{2}$ for all t and $j \ge t$, so that by Lemma 2, $\left\|\nabla \bar{f}_T(U^\top x^{(t)})\right\|$ must be large (step (b)). Second, in Sect. 5.2 we further augment our construction to force boundedness of the iterates by composing $\bar{f}_T(U^\top \cdot)$ with a soft projection, so that an algorithm cannot "cheat" with unbounded iterates. Finally, we present our general lower bounds in Sect. 5.3.

5.1 Random rotations and bounded iterates

To transform our hard instance (10) into a hard instance distribution, we introduce an orthogonal matrix $U \in \mathbb{R}^{d\times T}$ (with columns $u^{(1)}, \ldots, u^{(T)}$), and define

$$\tilde{f}_{T;U}(x) := \bar{f}_T(U^\top x) = \bar{f}_T(\langle u^{(1)}, x\rangle, \ldots, \langle u^{(T)}, x\rangle), \tag{12}$$

We assume throughout that U is chosen uniformly at random from the space of orthogonal matrices $\mathsf{O}(d, T) = \{V \in \mathbb{R}^{d\times T} \mid V^\top V = I_T\}$; unless otherwise stated, the probabilistic statements we give are respect to this uniform U in addition to any randomness in the algorithm that produces the iterates. With this definition, we have the following extension of Observation 1 to randomized iterates, which we prove for $\bar{f}_T$ but is valid for any robust zero-chain (Definition 4). Recall that a sequence is *informed by* f if it has the same distribution as $\mathsf{A}[f]$ for some randomized algorithm f (with iteration (4)).

Lemma 4 *Let $\delta > 0$ and $R \ge \sqrt{T}$, and let $x^{(1)}, \ldots, x^{(T)}$ be informed by $\tilde{f}_{T;U}$ and bounded, so that $\|x^{(t)}\| \le R$ for each T. If $d \ge 52TR^2 \log \frac{2T^2}{\delta}$ then with probability at least $1-\delta$, for all $t \le T$ and each $j \in \{t, \ldots, T\}$, we have*

$$|\langle u^{(j)}, x^{(t)}\rangle| < 1/2.$$

The result of Lemma 4 is identical (to constant factors) to an important result of Woodworth and Srebro [45, Lemma 7], but we must be careful with the sequential conditioning of randomness between the iterates $x^{(t)}$, the random orthogonal U, and how much information the sequentially computed derivatives may leak. Because of this additional care, we require a modification of their original proof,[4] which we provide in Sect. B.3, giving a rough outline here. For a fixed $t < T$, assume that $|\langle u^{(j)}, x^{(s)}\rangle| < 1/2$ holds for every pair $s \le t$ and $j \in \{s, \dots, T\}$; we argue that this (roughly) implies that $|\langle u^{(j)}, x^{(t+1)}\rangle| < 1/2$ for every $j \in \{t+1, \dots, T\}$ with high probability, completing the induction. When the assumption that $|\langle u^{(j)}, x^{(s)}\rangle| < 1/2$ holds, the robust zero-chain property of $\bar{f}_T$ (Definition 4 and Observation 3) implies that for every $s \le t$ we have

$$\tilde{f}_{T;U}(y) = \bar{f}_T(\langle u^{(1)}, y\rangle, \dots, \langle u^{(s)}, y\rangle, 0, \dots, 0)$$

for all y in a neighborhood of $x^{(s)}$. That is, we can compute all the derivatives of $\tilde{f}_{T;U}$ at $x^{(s)}$ from $x^{(s)}$ and $u^{(1)}, \dots, u^{(s)}$, as $\bar{f}_T$ is known. Therefore, given $u^{(1)}, x^{(1)}, \dots, u^{(t)}, x^{(t)}$ it is possible to reconstruct all the information the algorithm has collected up to iteration t. This means that beyond possibly revealing $u^{(1)}, \dots, u^{(t)}$, these derivatives contain no additional information on $u^{(t+1)}, \dots, u^{(T)}$. Consequently, any component of $x^{(t+1)}$ outside the span of $u^{(1)}, x^{(1)}, \dots, u^{(t)}, x^{(t)}$ is a complete "shot in the dark."

To give "shot in the dark" a more precise meaning, let $\hat{u}^{(j)}$ be the projection of $u^{(j)}$ to the orthogonal complement of $\mathrm{span}\{u^{(1)}, x^{(1)}, \dots, u^{(t)}, x^{(t)}\}$. We show that conditioned on $u^{(1)}, \dots, u^{(T)}$, and the induction hypothesis, $\hat{u}^{(j)}$ has a rotationally symmetric distribution in that subspace, and that it is independent of $x^{(t+1)}$. Therefore, by concentration of measure arguments on the sphere [5], we have $|\langle \hat{u}^{(j)}, x^{(t+1)}\rangle| \lesssim \|x^{(t+1)}\|/\sqrt{d} \le R/\sqrt{d}$ for any individual $j \ge t+1$, with high probability. Using an appropriate induction hypothesis, this is sufficient to guarantee that for every $t+1 \le j \le T$, $|\langle u^{(j)}, x^{(t+1)}\rangle| \lesssim R\sqrt{(T \log T)/d}$, which is bounded by $1/2$ for sufficiently large d.

5.2 Handling unbounded iterates

In the deterministic case, the adversary (choosing the hard function f) can choose the rotation matrix U to be exactly orthogonal to all past iterates; this is impossible for randomized algorithms. The construction (12) thus fails for unbounded random iterates, since as long as $x^{(t)}$ and $u^{(j)}$ are not exactly orthogonal, their inner product will exceed $1/2$ for sufficiently large $\|x^{(t)}\|$, thus breaching the "dead zone" of Ψ and providing the algorithm with information on $u^{(j)}$. To prevent this, we force the algorithm to only access $\tilde{f}_{T;U}$ at points with bounded norm, by first passing the iterates through a smooth mapping from $\mathbb{R}^d$ to a ball around the origin. We denote our final hard instance construction by $\hat{f}_{T;U} : \mathbb{R}^d \to \mathbb{R}$, and define it as

$$\hat{f}_{T;U}(x) = \tilde{f}_{T;U}(\rho(x)) + \frac{1}{10}\|x\|^2, \text{ where}$$
$$\rho(x) = \frac{x}{\sqrt{1 + \|x\|^2/R^2}} \text{ and } R = 230\sqrt{T}\,. \tag{13}$$

[4] In a recent note Woodworth and Srebro [46] independently provide a revision of their proof that is similar, but not identical, to the one we propose here.

The quadratic term in $\hat{f}_{T;U}$ guarantees that all points beyond a certain norm have a large gradient, which prevents the algorithm from trivially making the gradient small by increasing the norm of the iterates. The following lemma captures the hardness of $\hat{f}_{T;U}$ for randomized algorithms.

Lemma 5 *Let $\delta > 0$, and let $x^{(1)}, \ldots, x^{(T)}$ be informed by $\hat{f}_{T;U}$. If $d \geq 52 \cdot 230^2 \cdot T^2 \log \frac{2T^2}{\delta}$ then, with probability at least $1 - \delta$,*

$$\left\|\nabla \hat{f}_{T;U}(x^{(t)})\right\| > 1/2 \textit{ for all } t \leq T.$$

Proof For $t \leq T$, set $y^{(t)} := \rho(x^{(t)})$. For every $p \geq 0$ and $t \in \mathbb{N}$, the quantity $\nabla^p \hat{f}_{T;U}(x^{(t)})$ is measurable with respect $x^{(t)}$ and $\{\nabla^i \tilde{f}_{T;U}(y^{(t)})\}_{i=0}^{p}$ (the chain rule shows it can be computed from these variables without additional dependence on U, as ρ is fixed). Therefore, the process $y^{(1)}, \ldots, y^{(T)}$ is informed by $\tilde{f}_{T;U}$ (recall defining iteration (4)). Since $\|y^{(t)}\| = \|\rho(x^{(t)})\| \leq R$ for every t, we may apply Lemma 4 with $R = 230\sqrt{T}$ to obtain that with probability at least $1 - \delta$,

$$|\langle u^{(T)}, y^{(t)}\rangle| < 1/2 \text{ for every } t \leq T.$$

Therefore, by Lemma 2 with $i = T$, for each t there exists $j \leq T$ such that

$$\left|\left\langle u^{(j)}, y^{(t)}\right\rangle\right| < 1 \text{ and } \left|\left\langle u^{(j)}, \nabla \tilde{f}_{T;U}(y^{(t)})\right\rangle\right| > 1. \tag{14}$$

To show that $\|\nabla \hat{f}_{T;U}(x^{(t)})\|$ is also large, we consider separately the cases $\|x^{(t)}\| \leq R/2$ and $\|x^{(t)}\| \geq R/2$. For the first case, we use $\frac{\partial \rho}{\partial x}(x) = \frac{I - \rho(x)\rho(x)^\top/R^2}{\sqrt{1+\|x\|^2/R^2}}$ to write

$$\begin{aligned}\left\langle u^{(j)}, \nabla \hat{f}_{T;U}(x^{(t)})\right\rangle &= \left\langle u^{(j)}, \frac{\partial \rho}{\partial x}(x^{(t)}) \nabla \tilde{f}_{T;U}(y^{(t)})\right\rangle + \frac{1}{5}\left\langle u^{(j)}, x^{(t)}\right\rangle \\ &= \frac{\langle u^{(j)}, \nabla \tilde{f}_{T;U}(y^{(t)})\rangle - \langle u^{(j)}, y^{(t)}\rangle\langle y^{(t)}, \nabla \tilde{f}_{T;U}(y^{(t)})\rangle/R^2}{\sqrt{1+\|x^{(t)}\|^2/R^2}} \\ &\quad + \frac{1}{5}\langle u^{(j)}, y^{(t)}\rangle\sqrt{1+\|x^{(t)}\|^2/R^2}.\end{aligned}$$

Therefore, for $\|y^{(t)}\| \leq \|x^{(t)}\| \leq R/2$ we have

$$\begin{aligned}\left|\left\langle u^{(j)}, \nabla \hat{f}_{T;U}(x^{(t)})\right\rangle\right| \geq \frac{2}{\sqrt{5}}\left|\left\langle u^{(j)}, \nabla \tilde{f}_{T;U}(y^{(t)})\right\rangle\right| \\ - \left|\left\langle u^{(j)}, y^{(t)}\right\rangle\right|\left(\frac{\|\nabla \tilde{f}_{T;U}(y^{(t)})\|}{2R} + \frac{1}{2\sqrt{5}}\right).\end{aligned}$$

By Lemma 3.ii we have $\|\nabla \tilde{f}_{T;U}(y^{(t)})\| \leq 23\sqrt{T} = R/10$, which combined with (14) and the above display yields $\|\nabla \hat{f}_{T;U}(x^{(T)})\| \geq |\langle u^{(j)}, \nabla \hat{f}_{T;U}(x^{(T)})\rangle| \geq \frac{2}{\sqrt{5}} - \frac{1}{20} - \frac{1}{2\sqrt{5}} > \frac{1}{2}$.

In the second case, $\left\|x^{(t)}\right\| \geq R/2$, we have for any x satisfying $\|x\| \geq R/2$ and $y = \rho(x)$ that

$$\left\|\nabla \hat{f}_{T;U}(x)\right\| \geq \frac{1}{5}\|x\| - \left\|\frac{\partial \rho}{\partial x}(x)\right\|_{\mathrm{op}} \left\|\nabla \tilde{f}_{T;U}(y)\right\| \geq \frac{R}{10} - \frac{2}{\sqrt{5}}\frac{R}{10} > \sqrt{T} \geq 1, \quad (15)$$

where we used $\|\frac{\partial \rho}{\partial x}(x)\|_{\mathrm{op}} \leq \frac{1}{\sqrt{1+\|x\|^2/R^2}} \leq 2/\sqrt{5}$ and that $\|\nabla \tilde{f}_{T;U}(y)\| \leq 23\sqrt{T} = R/10$. □

As our lower bounds repose on appropriately scaling the function $\hat{f}_{T;U}$, it remains to verify that $\hat{f}_{T;U}$ satisfies the few boundedness properties we require. We do so in the following lemma.

Lemma 6 *The function $\hat{f}_{T;U}$ satisfies the following.*

i. *We have $\hat{f}_{T;U}(0) - \inf_x \hat{f}_{T;U}(x) \leq 12T$.*
ii. *For every $p \geq 1$, the pth order derivatives of $\hat{f}_{T;U}$ are $\hat{\ell}_p$-Lipschitz continuous, where $\hat{\ell}_p \leq \exp(cp\log p + c)$ for a numerical constant $c < \infty$.*

We defer the (computationally involved) proof of this Lemma to Sect. B.4.

5.3 Final lower bounds

With Lemmas 5 and 6 in hand, we can state our lower bound for all algorithms, randomized or otherwise, given access to all derivatives of a $\mathcal{C}^\infty$ function. Note that our construction also implies an identical lower bound for (slightly) more general algorithms that use any *local oracle* [10,35], meaning that the information the oracle returns about a function f when queried at a point x is identical to that it returns when a function g is queried at x whenever $f(z) = g(z)$ for all z in a neighborhood of x.

Theorem 2 *There exist numerical constants $0 < c_0, c_1 < \infty$ such that the following lower bound holds. Let $p \geq 1$, $p \in \mathbb{N}$, and let Δ, L_p, and ϵ be positive. Then*

$$\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{rand}}, \mathcal{F}_p(\Delta, L_p)\big) \geq c_0 \cdot \Delta \left(\frac{L_p}{\hat{\ell}_p}\right)^{1/p} \epsilon^{-\frac{1+p}{p}},$$

where $\hat{\ell}_p \leq e^{c_1 p \log p + c_1}$. The lower bound holds even if we restrict $\mathcal{F}_p(\Delta, L_p)$ to functions where the domain has dimension $1 + c_2 q\Big(\Delta\left(L_p/\ell_p\right)^{1/p}\epsilon^{-\frac{1+p}{p}}\Big)$ with c_2 a numerical constant and $q(x) = x^2\log(2x)$.

We return to the proof of Theorem 2 in Sect. 5.4, following the same outline as that of Theorem 1, and provide some commentary here. An inspection of the proof to come shows that we actually demonstrate a stronger result than that claimed in the theorem. For any $\delta \in (0,1)$ let $d \geq \left\lceil 52 \cdot (230)^2 \cdot T^2 \log(2T^2/\delta)\right\rceil$ where $T =$

$\lfloor c_0 \Delta (L_p/\hat{\ell}_p)^{1/p} \epsilon^{-\frac{1+p}{p}} \rfloor$ as in the claimed lower bound. In the proof we construct a probability measure μ on functions in $\mathcal{F}_p(\Delta, L_p)$, of fixed dimension d, such that

$$\inf_{\mathsf{A} \in \mathcal{A}_{\text{rand}}} \int \mathbb{P}_{\mathsf{A}} \left(\left\| \nabla f(x^{(t)}) \right\| > \epsilon \text{ for all } t \le T \mid f \right) d\mu(f) > 1 - \delta, \tag{16}$$

where the randomness in $\mathbb{P}_{\mathsf{A}}$ depends only on A. Therefore, by definition (6), for *any* $\mathsf{A} \in \mathcal{A}_{\text{rand}}$ a function f drawn from μ satisfies

$$\mathsf{T}_\epsilon \big(\mathsf{A}, f \big) > T \text{ with probability greater than } 1 - 2\delta, \tag{17}$$

implying Theorem 2 for any $\delta \ge 1/2$. Thus, we exhibit a randomized procedure for finding hard instances for any randomized algorithm that requires no knowledge of the algorithm itself.

Theorem 2 is stronger than Theorem 1 in that it applies to the broad class of all randomized algorithms. Our probabilistic analysis requires that the functions constructed to prove Theorem 2 have dimension scaling proportional to $T^2 \log(T)$ where T is the lower bound on the number of iterations. Contrast this to Theorem 1, which only requires dimension $2T + 1$. A similar gap exists in complexity results for convex optimization [45,46]. At present, it unclear if these gaps are fundamental or a consequence of our specific constructions.

5.4 Proof of Theorem 2

We set up our hard instance distribution $f_U : \mathbb{R}^d \to \mathbb{R}$, indexed by a uniformly distributed orthogonal matrix $U \in \mathsf{O}(d, T)$, by appropriately scaling $\hat{f}_{T;U}$ defined in (13),

$$f_U(x) := \frac{L_p \sigma^{p+1}}{\hat{\ell}_p} \hat{f}_{T;U}(x/\sigma),$$

where the integer T and scale parameter $\sigma > 0$ are to be determined, $d = \lceil 52 \cdot (230)^2 T^2 \log(4T^2) \rceil$, and the quantity $\hat{\ell}_p \le \exp(c_1 p \log p + c_1)$ for a numerical constant c_1 is defined in Lemma 6.ii.

Fix $\mathsf{A} \in \mathcal{A}_{\text{rand}}$ and let $x^{(1)}, x^{(2)}, \ldots, x^{(T)}$ be the iterates produced by A applied on f_U. Since f and $\hat{f}_{T;U}$ differ only by scaling, the iterates $x^{(1)}/\sigma, x^{(2)}/\sigma, \ldots, x^{(T)}/\sigma$ are informed by $\hat{f}_{T;U}$ (recall Sect. 2.2), and therefore we may apply Lemma 5 with $\delta = 1/2$ and our large enough choice of dimension d to conclude that

$$\mathbb{P}_{\mathsf{A},U} \left(\left\| \nabla \hat{f}_{T;U} \left(x^{(t)}/\sigma \right) \right\| > \frac{1}{2} \text{ for all } t \le T \right) > \frac{1}{2},$$

where the probability is taken over both the random orthogonal U and any randomness in A. As A is arbitrary, taking $\sigma = (2\hat{\ell}_p \epsilon / L_p)^{1/p}$, this inequality becomes the desired strong inequality (16) with $\delta = 1/2$ and μ induced by the distribution of U. Thus, by (17), for every $\mathsf{A} \in \mathcal{A}_{\mathsf{rand}}$ there exists $U_{\mathsf{A}} \in \mathsf{O}(d, T)$ such that $\mathsf{T}_\epsilon\big(\mathsf{A}, f_{U_{\mathsf{A}}}\big) \geq 1 + T$, so

$$\inf_{\mathsf{A} \in \mathcal{A}_{\mathsf{det}}} \sup_{U \in \mathsf{O}(d,T)} \mathsf{T}_\epsilon\big(\mathsf{A}, f_U\big) \geq 1 + T.$$

It remains to choose T to guarantee that f_U belongs to the relevant function class (bounded and smooth) for every orthogonal U. By Lemma 6.ii, f_U has L_p-Lipschitz continuous pth order derivatives. By Lemma 6.i, we have

$$\begin{aligned} f_U(0) - \inf_x f_U(x) &\leq \frac{L_p \sigma^{p+1}}{\hat{\ell}_p} \Big(\bar{f}_T(0) - \inf_x \bar{f}_T(x)\Big) \\ &\leq \frac{12 L_p \sigma^{p+1}}{\hat{\ell}_p} T = \frac{24(2\hat{\ell}_p)^{1/p} \epsilon^{\frac{p+1}{p}}}{L_p^{1/p}} T, \end{aligned}$$

where in the last transition we have substituted $\sigma = (2\ell_p \epsilon / L_p)^{1/p}$. Setting $T = \lfloor \frac{\Delta}{48} (L_p/\hat{\ell}_p)^{1/p} \epsilon^{-\frac{1+p}{p}} \rfloor$ gives $f_U(0) - \inf_x f_U(x) \leq \Delta$, and $f_U \in \mathcal{F}_p(\Delta, L_p)$, yielding the theorem.

6 Distance-based lower bounds

We have so far considered finding approximate stationary points of smooth functions with bounded sub-optimality at the origin, i.e. $f(0) - \inf_x f(x) \leq \Delta$. In convex optimization, it is common to consider instead functions with bounded distance between the origin and a global minimum. We may consider a similar restriction for non-convex functions; for $p \geq 1$ and positive L_p, D, let

$$\mathcal{F}_p^{\text{dist}}(D, L_p)$$

be the class of $\mathcal{C}^\infty$ functions with L_p-Lipschitz pth order derivatives satisfying

$$\sup_x \{\|x\| \mid x \in \arg\min f\} \leq D, \tag{18}$$

that is, all global minima have bounded distance to the origin.

In this section we give a lower bound on the complexity of this function class that has the same ϵ dependence as our bound for the class $\mathcal{F}_p(\Delta, L_p)$. This is in sharp contrast to convex optimization, where distance-bounded functions enjoy significantly better ϵ dependence than their value-bounded counterparts (see Sect. 3 in the companion [14]). Qualitatively, the reason for this difference is that the lack of convexity allows us to "hide" global minima close to the origin that are difficult to find for any algorithm with local function access [35].

We postpone the construction and proof to Appendix C, and move directly to the final bound.

Theorem 3 *There exist numerical constants $0 < c_0, c_1 < \infty$ such that the following lower bound holds. For any $p \geq 1$, let D, L_p, and ϵ be positive. Then*

$$\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{rand}}, \mathcal{F}_p^{\mathrm{dist}}(D, L_p)\big) \geq c_0 \cdot D^{1+p} \left(\frac{L_p}{\ell'_p}\right)^{\frac{1+p}{p}} \epsilon^{-\frac{1+p}{p}},$$

where $\ell'_p \leq e^{c_1 p \log p + c_1}$. The lower bound holds even if we restrict $\mathcal{F}_p^{\mathrm{dist}}(D, L_p)$ to functions with domain of dimension $1 + c_2 q\Big(D^{1+p}\left(L_p/\ell'_p\right)^{\frac{1+p}{p}} \epsilon^{-\frac{1+p}{p}}\Big)$, for a some numerical constant $c_2 < \infty$ and $q(x) = x^2 \log(2x)$.

We remark that a lower-dimensional construction suffices for proving the lower bound for deterministic algorithm, similarly to Theorem 1.

While we do not have a matching upper bound for Theorem 3, we can match its ϵ dependence in the smaller function class

$$\mathcal{F}_{1,p}^{\mathrm{dist}}(D, L_1, L_p) = \mathcal{F}_1^{\mathrm{dist}}(D, L_1) \cap \mathcal{F}_p^{\mathrm{dist}}(D, L_p),$$

due to the fact that for any $f : \mathbb{R}^d \to \mathbb{R}$ with L_1-Lipschitz continuous gradient and global minimizer $x^\star$, we have $f(x) - f(x^\star) \leq \frac{1}{2} L_1 \|x - x^\star\|^2$ for all $x \in \mathbb{R}^d$ [cf. 9, Eq. (9.13)]. Hence $\mathcal{F}_{1,p}^{\mathrm{dist}}(D, L_1, L_p) \subset \mathcal{F}_p(\Delta, L_p)$, with $\Delta := \frac{1}{2} L_1 D^2$, and consequently by the bound (8) we have

$$\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{det}}^{(p)} \cap \mathcal{A}_{\mathsf{zr}}^{(p)}, \mathcal{F}_{1,p}^{\mathrm{dist}}(D, L_1, L_p)\big) \lesssim D^2 L_1 L_p^{1/p} \epsilon^{-\frac{p+1}{p}}.$$

7 Conclusion

This work provides the first algorithm independent and tight lower bounds on the dimension-free complexity of finding stationary points. As a consequence, we have characterized the optimal rates of convergence to ϵ-stationarity, under the assumption of high dimension and an oracle that provides all derivatives. Yet, given the importance of high-dimensional problems, the picture is incomplete: high-order algorithms—even second-order method—are often impractical in large scale settings. We address this in the companion [14], which provides sharper lower bounds for the more restricted class of first-order methods. In [14] we also provide a full conclusion for this paper sequence, discussing in depth the implications and questions that arise from our results.

Acknowledgements OH was supported by the PACCAR INC fellowship. YC and JCD were partially supported by the SAIL-Toyota Center for AI Research, NSF-CAREER award 1553086, and a Sloan Foundation Fellowship in Mathematics. YC was partially supported by the Stanford Graduate Fellowship and the Numerical Technologies Fellowship.

A Proof of Propositions 1 and 2

The core of the proofs of Propositions 1 and 2 is the following construction.

Lemma 7 *Let $p \in \mathbb{N} \cup \{\infty\}$, $T_0 \in \mathbb{N}$ and $\mathsf{A} \in \mathcal{A}_{\mathsf{det}}^{(p)}$. There exists an algorithm $\mathsf{Z}_{\mathsf{A}} \in \mathcal{A}_{\mathsf{zr}}^{(p)}$ with the following property. For every $f : \mathbb{R}^d \to \mathbb{R}$ there exists an orthogonal matrix $U \in \mathbb{R}^{(d+T_0)\times d}$ such that, for every $\epsilon > 0$,*

$$T_\epsilon\big(\mathsf{A}, f_U\big) > T_0 \;\; or \;\; T_\epsilon\big(\mathsf{A}, f_U\big) = T_\epsilon\big(\mathsf{Z}_{\mathsf{A}}, f\big),$$

where $f_U(x) := f(U^\top x)$.

Proof We explicitly construct Z_{A} with the following slightly stronger property. For every every $f : \mathbb{R}^d \to \mathbb{R}$ in $\mathcal{F}$, there exists an orthogonal $U \in \mathbb{R}^{(d+T_0)\times d}$, $U^\top U = I_d$, such that $f_U(x) := f(U^\top x)$ satisfies that the first T_0 iterates in sequences $\mathsf{Z}_{\mathsf{A}}[f]$ and $U^\top \mathsf{A}[f_U]$ are identical. (Recall the notation $\mathsf{A}[f] = \{a^{(t)}\}_{t\in\mathbb{N}}$ where $a^{(t)}$ are the iterates of A on f, and we use the obvious shorthand $U^\top\{a^{(t)}\}_{t\in\mathbb{N}} = \{U^\top a^{(t)}\}_{t\in\mathbb{N}}$.)

Before explaining the construction of Z_{A}, let us see how its defining property implies the lemma. If $\mathsf{T}_\epsilon\big(\mathsf{A}, f_U\big) > T_0$, we are done. Otherwise, $\mathsf{T}_\epsilon\big(\mathsf{A}, f_U\big) \le T_0$ and we have

$$\mathsf{T}_\epsilon\big(\mathsf{A}, f_U\big) := \mathsf{T}_\epsilon\big(\mathsf{A}[f_U], f_U\big) \overset{(i)}{=} \mathsf{T}_\epsilon\big(U^\top \mathsf{A}[f_U], f\big) \overset{(ii)}{=} \mathsf{T}_\epsilon\big(\mathsf{Z}_{\mathsf{A}}, f\big), \tag{19}$$

as required. The equality (i) follows because $\|Ug\| = \|g\|$ for all orthogonal U, so for any sequence $\{a^{(t)}\}_{t\in\mathbb{N}}$

$$\begin{aligned}\mathsf{T}_\epsilon\big(\{a^{(t)}\}_{t\in\mathbb{N}}, f_U\big) &= \inf\Big\{t \in \mathbb{N} \mid \|\nabla f_U(a^{(t)})\| \le \epsilon\Big\} \\ &= \inf\Big\{t \in \mathbb{N} \mid \|\nabla f(U^\top a^{(t)})\| \le \epsilon\Big\} = \mathsf{T}_\epsilon\big(\{U^\top a^{(t)}\}_{t\in\mathbb{N}}, f\big)\end{aligned}$$

and in equality (i) we let $\{a^{(t)}\}_{t\in\mathbb{N}} = \mathsf{A}[f_U]$. The equality (ii) holds because $\mathsf{T}_\epsilon\big(\cdot, \cdot\big)$ is a "stopping time": if $\mathsf{T}_\epsilon\big(U^\top \mathsf{A}[f_U], f\big) \le T_0$ then the first T_0 iterates of $U^\top \mathsf{A}[f_U]$ determine $\mathsf{T}_\epsilon\big(U^\top \mathsf{A}[f_U], f\big)$, and these T_0 iterates are identical to the first T_0 iterates of $\mathsf{Z}_{\mathsf{A}}[f]$ by assumption.

It remains to construct the zero-respecting algorithm Z_{A} with iterates matching those of A under appropriate rotation. We do this by describing its operation inductively on any given $f : \mathbb{R}^d \to \mathbb{R}$, which we denote $\{z^{(t)}\}_{t\in\mathbb{N}} = \mathsf{Z}_{\mathsf{A}}[f]$. Letting $d' = d + T_0$, the state of the algorithm Z_{A} at iteration t is determined by a *support* $S_t \subseteq [d]$ and orthonormal vectors $\{u^{(i)}\}_{i\in S_t} \subset \mathbb{R}^{d'}$ identified with this support. The support condition (5) defines the set S_t,

$$S_t = \bigcup_{q\in[p]} \bigcup_{s<t} \mathrm{supp}\left\{\nabla^q f(z^{(s)})\right\},$$

so that $\emptyset = S_1 \subseteq S_2 \subseteq \cdots$ and the collection $\{u^{(i)}\}_{i\in S_t}$ grows with t. We let $U \in \mathbb{R}^{d'\times d}$ be the orthogonal matrix whose ith column is $u^{(i)}$—even though U may not be completely determined throughout the runtime of $\mathsf{Z_A}$, our partial knowledge of it will suffice to simulate the operation of A on $f_U(a) = f(U^\top a)$. Letting $\{a^{(t)}\}_{t\in\mathbb{N}} = \mathsf{A}[f_U]$, our requirements $\mathsf{Z_A}[f] = U^\top \mathsf{A}[f_U]$ and $\mathsf{Z_A} \in \mathcal{A}_{\mathsf{zr}}$ are equivalent to

$$z^{(t)} = U^\top a^{(t)} \text{ and } \mathrm{supp}\{z^{(t)}\} \subseteq S_t \tag{20}$$

for every $t \le T_0$ (we set $z^{(i)} = 0$ for every $i > T_0$ without loss of generality).

Let us proceed with the inductive argument. The iterate $a^{(1)} \in \mathbb{R}^{d'}$ is an arbitrary (but deterministic) vector in $\mathbb{R}^{d'}$. We thus satisfy (20) at $t = 1$ by requiring that $\langle u^{(j)}, a^{(1)}\rangle = 0$ for every $j \in [d]$, whence the first iterate of $\mathsf{Z_A}$ satisfies $z^{(1)} = 0 \in \mathbb{R}^d$. Assume now the equality and containment (20) holds for every $s < t$, where $t \le T_0$ (implying that $\mathsf{Z_A}$ has emulated the iterates $a^{(2)}, \ldots, a^{(t-1)}$ of A); we show how $\mathsf{Z_A}$ can emulate $a^{(t)}$, the t'th iterate of A, and from it can construct $z^{(t)}$ that satisfies (20). To obtain $a^{(t)}$, note that for every $q \le p$, and every $s < t$, the derivatives $\nabla^q f_U(a^{(s)})$ are a function of $\nabla^q f(z^{(s)})$ and orthonormal the vectors $\{u^{(i)}\}_{i\in S_{s+1}}$, because $\mathrm{supp}\{\nabla^q f(z^{(s)})\} \subseteq S_{s+1}$ and therefore the chain rule implies

$$\left[\nabla^q f_U(a^{(s)})\right]_{j_1,\ldots,j_q} = \sum_{i_1,\ldots,i_q\in S_{s+1}} \left[\nabla^q f(z^{(s)})\right]_{i_1,\ldots,i_q} u^{(i_1)}_{j_1}\cdots u^{(i_q)}_{j_q}.$$

Since $\mathsf{A} \in \mathcal{A}^{(p)}_{\mathsf{det}}$ is deterministic, $a^{(t)}$ is a function of $\nabla^q f(z^{(s)})$ for $q \in [p]$ and $s \in [t-1]$, and thus $\mathsf{Z_A}$ can simulate and compute it. To satisfy the support condition $\mathrm{supp}\{z^{(t)}\} \subseteq S_t$ we require that $\langle u^{(j)}, a^{(t)}\rangle = 0$ for every $j \notin S_t$. This also means that to compute $z^{(t)} = U^\top a^{(t)}$ we require only the columns of U indexed by the support S_t.

Finally, we need to show that after computing S_{t+1} we can find the vectors $\{u^{(i)}\}_{i\in S_{t+1}\setminus S_t}$ satisfying $\langle u^{(j)}, a^{(s)}\rangle = 0$ for every $s \le t$ and $j \in S_{t+1}\setminus S_t$, and additionally that U be orthogonal. Thus, we need to choose $\{u^{(i)}\}_{i\in S_{t+1}\setminus S_t}$ in the orthogonal complement of span $\left\{a^{(1)}, \ldots, a^{(t)}, \{u^{(i)}\}_{i\in S_t}\right\}$. This orthogonal complement has dimension at least $d' - t - |S_t| = |S_t^c| + T_0 - t \ge |S_t^c|$. Since $|S_{t+1}\setminus S_t| \le |S_t^c|$, there exist orthonormal vectors $\{u^{(i)}\}_{i\in S_{t+1}\setminus S_t}$ that meet the requirements. This completes the induction.

Finally, note that the arguments above hold unchanged for $p = \infty$. □

With Lemma 7 in hand, the propositions follow easily.

Proposition 1 *Let $p \in \mathbb{N} \cup \{\infty\}$, $\mathcal{F}$ be an orthogonally invariant function class and $\epsilon > 0$. Then*

$$\mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{det}}, \mathcal{F}\big) \geq \mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{zr}}, \mathcal{F}\big).$$

Proof We may assume that $\mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{det}}, \mathcal{F}\big) < T_0$ for some integer $T_0 < \infty$, as otherwise we have $\mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{det}}, \mathcal{F}\big) = \infty$ and the result holds trivially. For any $\mathsf{A} \in \mathcal{A}^{(p)}_{\mathsf{det}}$ and the value T_0, we invoke Lemma 7 to construct $\mathsf{Z}_\mathsf{A} \in \mathcal{A}^{(p)}_{\mathsf{zr}}$ such that $\mathsf{T}_\epsilon\big(\mathsf{A}, f_U\big) \geq \min\{T_0, \mathsf{T}_\epsilon\big(\mathsf{Z}_\mathsf{A}, f\big)\}$ for every $f \in \mathcal{F}$ and some orthogonal matrix U that depends on f and A. Consequently, we have

$$\begin{aligned}
&\mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{det}}, \mathcal{F}\big)\\
&\quad= \inf_{\mathsf{A}\in\mathcal{A}^{(p)}_{\mathsf{det}}} \sup_{f\in\mathcal{F}} \mathsf{T}_\epsilon\big(\mathsf{A}, f\big) \overset{(i)}{\geq} \inf_{\mathsf{A}\in\mathcal{A}^{(p)}_{\mathsf{det}}} \sup_{f\in\mathcal{F}} \mathsf{T}_\epsilon\big(\mathsf{A}, f_U\big) \overset{(ii)}{\geq} \min\left\{T_0, \inf_{\mathsf{A}\in\mathcal{A}^{(p)}_{\mathsf{det}}} \sup_{f\in\mathcal{F}} \mathsf{T}_\epsilon\big(\mathsf{Z}_\mathsf{A}, f\big)\right\}\\
&\quad\overset{(iii)}{\geq} \min\left\{T_0, \inf_{\mathsf{B}\in\mathcal{A}^{(p)}_{\mathsf{zr}}} \sup_{f\in\mathcal{F}} \mathsf{T}_\epsilon\big(\mathsf{B}, f\big)\right\} = \min\left\{T_0, \mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{zr}}, \mathcal{F}\big)\right\},
\end{aligned}$$

where inequality (i) uses that $f_U \in \mathcal{F}$ because $\mathcal{F}$ is orthogonally invariant, step (ii) uses $\mathsf{T}_\epsilon\big(\mathsf{A}, f_U\big) \geq \min\{T_0, \mathsf{T}_\epsilon\big(\mathsf{Z}_\mathsf{A}, f\big)\}$ and step (iii) is due to $\mathsf{Z}_\mathsf{A} \in \mathcal{A}^{(p)}_{\mathsf{zr}}$ by construction. As we chose T_0 for which $\mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{det}}, \mathcal{F}\big) < T_0$, the chain of inequalities implies $\mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{det}}, \mathcal{F}\big) \geq \mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{zr}}, \mathcal{F}\big)$, concluding the proof. □

Proposition 2 *Let $p \in \mathbb{N}\cup\{\infty\}$, $\mathcal{F}$ be an orthogonally invariant function class, $f \in \mathcal{F}$ with domain of dimension d, and $\epsilon > 0$. If $\mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{zr}}, \{f\}\big) \geq T$, then*

$$\mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{det}}, \mathcal{F}\big) \geq \mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{det}}, \{f_U \mid U \in \mathsf{O}(d+T, d)\}\big) \geq T,$$

where $f_U := f(U^\top z)$ and $\mathsf{O}(d+T, d)$ is the set of $(d+T)\times d$ orthogonal matrices, so that $\{f_U \mid U \in \mathsf{O}(d+T, d)\}$ contains only function with domain of dimension $d+T$.

Proof For any $\mathsf{A} \in \mathcal{A}^{(p)}_{\mathsf{det}}$, we invoke Lemma 7 with $T_0 = T$ to obtain $\mathsf{Z}_\mathsf{A} \in \mathcal{A}^{(p)}_{\mathsf{zr}}$ and orthogonal matrix U' (dependent on f and A) for which

$$\mathsf{T}_\epsilon\big(\mathsf{A}, f_{U'}\big) \geq \min\{T, \mathsf{T}_\epsilon\big(\mathsf{Z}_\mathsf{A}, f\big)\} = T,$$

where the last equality is due to $\inf_{\mathsf{B}\in\mathcal{A}^{(p)}_{\mathsf{zr}}} \mathsf{T}_\epsilon\big(\mathsf{B}, f\big) = \mathcal{T}_\epsilon\big(\mathcal{A}^{(p)}_{\mathsf{zr}}, \{f\}\big) \geq T$. Since $f_{U'} \in \{f_U \mid U \in \mathsf{O}(d+T, d)\}$, we have

$$\sup_{f'\in\{f_U \mid U\in\mathsf{O}(d+T,d)\}} \mathsf{T}_\epsilon\big(\mathsf{A}, f'\big) \geq T,$$

and taking the infimum over $\mathsf{A} \in \mathcal{A}^{(p)}_{\mathsf{det}}$ concludes the proof. □

B Technical results

B.1 Proof of Lemma 1

Lemma 1 *The functions Ψ and Φ satisfy the following.*

i. *For all $x \le \frac{1}{2}$ and all $k \in \mathbb{N}$, $\Psi^{(k)}(x) = 0$.*
ii. *For all $x \ge 1$ and $|y| < 1$, $\Psi(x)\Phi'(y) > 1$.*
iii. *Both Ψ and Φ are infinitely differentiable, and for all $k \in \mathbb{N}$ we have*

$$\sup_x |\Psi^{(k)}(x)| \le \exp\left(\frac{5k}{2}\log(4k)\right) \text{ and } \sup_x |\Phi^{(k)}(x)| \le \exp\left(\frac{3k}{2}\log\frac{3k}{2}\right).$$

iv. *The functions and derivatives Ψ, Ψ', Φ and Φ' are non-negative and bounded, with*

$$0 \le \Psi < e, \;\; 0 \le \Psi' \le \sqrt{54/e}, \;\; 0 < \Phi < \sqrt{2\pi e}, \text{ and } 0 < \Phi' \le \sqrt{e}.$$

Each of the statements in the lemma is immediate except for part iii. To see this part, we require a few further calculations. We begin by providing bounds on the derivatives of $\Phi(x) = e^{\frac{1}{2}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt$. To avoid annoyances with scaling factors, we define $\phi(t) = e^{-\frac{1}{2}t^2}$.

Lemma 8 *For all $k \in \mathbb{N}$, there exist constants $c_i^{(k)}$ satisfying $|c_i^{(k)}| \le (2\max\{i, 1\})^k$, and*

$$\phi^{(k)}(t) = \left(\sum_{i=0}^{k} c_i^{(k)} t^i\right)\phi(t).$$

Proof We prove the result by induction. We have $\phi'(t) = -te^{-\frac{1}{2}t^2}$, so that the base case of the induction is satisfied. Now, assume for our induction that

$$\phi^{(k)}(t) = \sum_{i=0}^{k} c_i^{(k)} t^i e^{-\frac{1}{2}t^2} = \sum_{i=0}^{k} c_i^{(k)} t^i \phi(t).$$

where $|c_i^{(k)}| \le 2^k(\max\{i, 1\})^k$. Then taking derivatives, we have

$$\phi^{(k+1)}(t) = \sum_{i=1}^{k}\left[i \cdot c_i^{(k)} t^{i-1} - c_i^{(k)} t^{i+1}\right]\phi(t) - c_0^{(k)} t\phi(t) = \sum_{i=0}^{k+1} c_i^{(k+1)} t^i \phi(t)$$

where $c_i^{(k+1)} = (i+1)c_{i+1}^{(k)} - c_{i-1}^{(k)}$ (and we treat $c_{k+1}^{(k)} = 0$) and $|c_{k+1}^{(k+1)}| = 1$. With the induction hypothesis that $c_i^{(k)} \le (2\max\{i, 1\})^k$, we obtain

$$|c_i^{(k+1)}| \le 2^k(i+1)(i+1)^k + 2^k(\max\{i, 1\})^k \le 2^{k+1}(i+1)^{k+1}.$$

This gives the result. □

With this result, we find that for any $k \geq 1$,

$$\Phi^{(k)}(x) = \sqrt{e}\left(\sum_{i=0}^{k-1} c_i^{(k-1)} x^i\right)\phi(x).$$

The function $\log(x^i\phi(x)) = i\log x - \frac{1}{2}x^2$ is maximized at $x = \sqrt{i}$, so that $x^i\phi(x) \leq \exp(\frac{i}{2}\log\frac{i}{e})$. We thus obtain the numerically verifiable upper bound

$$|\Phi^{(k)}(x)| \leq \sqrt{e}\sum_{i=0}^{k-1}(2\max\{i,1\})^{k-1}\exp\left(\frac{i}{2}\log\frac{i}{e}\right) \leq \exp\left(1.5k\log(1.5k)\right).$$

Now, we turn to considering the function $\Psi(x)$. We assume w.l.o.g. that $x > \frac{1}{2}$, as otherwise $\Psi^{(k)}(x) = 0$ for all k. Recall $\Psi(x) = \exp\left(1 - \frac{1}{(2x-1)^2}\right)$ for $x > \frac{1}{2}$. We have the following lemma regarding its derivatives.

Lemma 9 *For all $k \in \mathbb{N}$, there exist constants $c_i^{(k)}$ satisfying $|c_i^{(k)}| \leq 6^k(2i+k)^k$ such that*

$$\Psi^{(k)}(x) = \left(\sum_{i=1}^{k}\frac{c_i^{(k)}}{(2x-1)^{k+2i}}\right)\Psi(x).$$

Proof We provide the proof by induction over k. For $k = 1$, we have that

$$\Psi'(x) = \frac{4}{(2x-1)^3}\exp\left(1 - \frac{1}{(2x-1)^2}\right) = \frac{4}{(2x-1)^3}\Psi(x),$$

which yields the base case of the induction. Now, assume that for some k, we have

$$\Psi^{(k)}(x) = \left(\sum_{i=1}^{k}\frac{c_i^{(k)}}{(2x-1)^{k+2i}}\right)\Psi(x).$$

Then

$$\begin{aligned}\Psi^{(k+1)}(x) &= \left(-\sum_{i=1}^{k}\frac{2(k+2i)c_i^{(k)}}{(2x-1)^{k+1+2i}} + \sum_{i=1}^{k}\frac{4c_i^{(k)}}{(2x-1)^{k+3+2i}}\right)\Psi(x)\\ &= \left(\sum_{i=1}^{k+1}\frac{4c_{i-1}^{(k)} - 2(k+2i)c_i^{(k)}}{(2x-1)^{k+1+2i}}\right)\Psi(x),\end{aligned}$$

where $c_{k+1}^{(k)} = 0$ and $c_0^{(k)} = 0$. Defining $c_1^1 = 4$ and $c_i^{(k+1)} = 4c_{i-1}^{(k)} - 2(k+2i)c_i^{(k)}$ for $i > 1$, then, under the inductive hypothesis that $|c_i^{(k)}| \le 6^k(2i+k)^k$, we have

$$\begin{aligned} |c_i^{(k+1)}| &\le 4 \cdot 6^k(k-2+2i)^k + 2 \cdot 6^k(k+2i)(k+2i)^k \\ &\le 6^{k+1}(k+2i)^{k+1} \le 6^{k+1}(k+1+2i)^{k+1} \end{aligned}$$

which gives the result. □

As in the derivation immediately following Lemma 8, by replacing $t = \frac{1}{2x-1}$, we have that $t^{k+2i}e^{-t^2}$ is maximized by $t = \sqrt{(k+2i)/2}$, so that

$$\frac{1}{(2x-1)^{k+2i}}\Psi(x) \le \exp\left(1 + \frac{k+2i}{2}\log\frac{k+2i}{2e}\right),$$

which yields the numerically verifiable upper bound

$$|\Psi^{(k)}(x)| \le \sum_{i=1}^{k} \exp\left(1 + k\log(6k+12i) + \frac{k+2i}{2}\log\frac{k+2i}{2e}\right) \le \exp\left(2.5k\log(4k)\right).$$

B.2 Proof of Lemma 3

Lemma 3 *The function $\bar{f}_T$ satisfies the following.*

i. *We have $\bar{f}_T(0) - \inf_x \bar{f}_T(x) \le 12T$.*
ii. *For all $x \in \mathbb{R}^d$, $\left\|\nabla \bar{f}_T(x)\right\| \le 23\sqrt{T}$.*
iii. *For every $p \ge 1$, the p-th order derivatives of $\bar{f}_T$ are ℓ_p-Lipschitz continuous, where $\ell_p \le \exp(\frac{5}{2}p\log p + cp)$ for a numerical constant $c < \infty$.*

Proof Part i follows because $\bar{f}_T(0) < 0$ and, since $0 \le \Psi(x) \le e$ and $0 \le \Phi(x) \le \sqrt{2\pi e}$,

$$\bar{f}_T(x) \ge -\Psi(1)\,\Phi(x_1) - \sum_{i=2}^{T}\Psi(x_{i-1})\,\Phi(x_i) > -T \cdot e \cdot \sqrt{2\pi e} \ge -12T.$$

Part ii follows additionally from $\Psi(x) = 0$ on $x < 1/2$, $0 \le \Psi'(x) \le \sqrt{54e^{-1}}$ and $0 \le \Phi'(x) \le \sqrt{e}$, which when substituted into

$$\begin{aligned} \frac{\partial \bar{f}_T}{\partial x_j}(x) = &-\Psi\left(-x_{j-1}\right)\Phi'\left(-x_j\right) - \Psi\left(x_{j-1}\right)\Phi'\left(x_j\right) \\ &-\Psi'\left(-x_j\right)\Phi\left(-x_{j+1}\right) - \Psi'\left(x_j\right)\Phi\left(x_{j+1}\right) \end{aligned}$$

yields

$$\left|\frac{\partial \bar{f}_T}{\partial x_j}(x)\right| \le e \cdot \sqrt{e} + \sqrt{54e^{-1}} \cdot \sqrt{2\pi e} \le 23$$

for every x and j. Consequently, $\left\|\nabla \bar{f}_T(x)\right\| \le \sqrt{T} \le 23\sqrt{T}$.

To establish part iii, fix a point $x \in \mathbb{R}^T$ and a unit vector $v \in \mathbb{R}^T$. Define the real function $h_{x,v} : \mathbb{R} \to \mathbb{R}$ by the directional projection of $\bar{f}_T$, $h_{x,v}(\theta) := \bar{f}_T(x + \theta v)$. The function $\theta \mapsto h_{x,v}(\theta)$ is infinitely differentiable for every x and v. Therefore, $\bar{f}_T$ has ℓ_p-Lipschitz p-th order derivatives if and only if $|h_{x,v}^{(p+1)}(0)| \le \ell_p$ for every x, v. Using the shorthand notation $\partial_{i_1} \cdots \partial_{i_k}$ for $\frac{\partial^k}{\partial x_{i_1} \cdots \partial x_{i_k}}$, we have

$$h_{x,v}^{(p+1)}(0) = \sum_{i_1,\ldots,i_{p+1}=1}^{T} \partial_{i_1} \cdots \partial_{i_{p+1}} \bar{f}_T(x)\, v_{i_1} \cdots v_{i_{p+1}}.$$

Examining $\bar{f}_T$, we see that $\partial_{i_1} \cdots \partial_{i_{p+1}} \bar{f}_T$ is non-zero if and only if $\left|i_j - i_k\right| \le 1$ for every $j, k \in [p+1]$. Consequently, we can rearrange the above summation as

$$h_{x,v}^{(p+1)}(0) = \sum_{\delta_1,\delta_2,\ldots,\delta_p \in \{0,1\}^p \cup \{0,-1\}^p} \sum_{i=1}^{T} \partial_{i+\delta_1} \cdots \partial_{i+\delta_p} \partial_i \bar{f}_T(x)\, v_{i+\delta_1} \cdots v_{i+\delta_p} v_i,$$

where we take $v_0 := 0$ and $v_{T+1} := 0$. Brief calculation show that

$$\begin{aligned}
&\sup_{x \in \mathbb{R}^T} \max_{i \in [T]} \max_{\delta \in \{0,1\}^p \cup \{0,-1\}^p} \left|\partial_{i+\delta_1} \cdots \partial_{i+\delta_p} \partial_i \bar{f}_T(x)\right| \\
&\quad \le \max_{k \in [p+1]} \left\{ 2 \sup_{x \in \mathbb{R}} \left|\Psi^{(k)}(x)\right| \sup_{x' \in \mathbb{R}} \left|\Phi^{(p+1-k)}(x')\right| \right\} \\
&\quad \le 2\sqrt{2\pi e} \cdot e^{2.5(p+1)\log(4(p+1))} \le \exp(2.5p\log p + 4p + 9).
\end{aligned}$$

where the second inequality uses Lemma 1.iii, and $\Phi(x') \le \sqrt{2\pi e}$ for the case $k = p+1$. Defining $\ell_p = 2^{p+1} e^{2.5p\log p + 4p + 9} \le e^{2.5p\log p + 5p + 10}$, we thus have

$$\begin{aligned}
\left|h_{x,v}^{(p+1)}(0)\right| &\le \sum_{\delta \in \{0,1\}^p \cup \{0,-1\}^p} 2^{-(p+1)} \ell_p \left|\sum_{i=1}^{T} v_{i+\delta_1} \cdots v_{i+\delta_p} v_i\right| \\
&\le \left(2^{p+1} - 1\right) 2^{-(p+1)} \ell_p \le \ell_p,
\end{aligned}$$

where we have used $|\sum_{i=1}^{T} v_{i+\delta_1} \cdots v_{i+\delta_p} v_i| \le 1$ for every $\delta \in \{0,1\}^p \cup \{0,-1\}^p$. To see this last claim is true, recall that v is a unit vector and note that

$$\sum_{i=1}^{T} v_{i+\delta_1} \cdots v_{i+\delta_p} v_i = \sum_{i=1}^{T} v_i^{p+1-\sum_{j=1}^{p}\delta_j} v_{i\pm1}^{\sum_{j=1}^{p}\delta_j}.$$

If $\delta = 0$ then $|\sum_{i=1}^{T} v_{i+\delta_1} \cdots v_{i+\delta_p} v_i| = |\sum_{i=1}^{T} v_i^{p+1}| \le \sum_{i=1}^{T} v_i^2 = 1$. Otherwise, letting $1 \le \sum_{j=1}^{p} |\delta_j| = n \le p$, the Cauchy-Swartz inequality implies

$$\left|\sum_{i=1}^{T} v_{i+\delta_1} \cdots v_{i+\delta_p} v_i\right| = \left|\sum_{i=1}^{T} v_i^{p+1-n} v_{i+s}^{n}\right|$$
$$\le \sqrt{\sum_{i=1}^{T} v_i^{2(p+1-n)}} \sqrt{\sum_{i=1}^{T} v_{i+s}^{2n}} \le \sum_{i=1}^{T} v_i^2 = 1,$$

where $s = -1$ or 1. This gives the result. □

B.3 Proof of Lemma 4

The proof of Lemma 4 uses a number of auxiliary arguments, marked as Lemmas 4a, 4b and 4c . Readers looking to gain a high-level view of the proof of Lemma 4 can safely skip the proofs of these sub-lemmas. In the following, recall that $U \in \mathbb{R}^{d\times T}$ is drawn from the uniform distribution over $d \times T$ orthogonal matrices (satisfying $U^T U = I$, as $d > T$), that the columns of U are denoted $u^{(1)}, \ldots, u^{(T)}$, and that $\tilde{f}_{T;U}(x) = \bar{f}_T(U^\top x)$.

Lemma 4 *Let $\delta > 0$ and $R \ge \sqrt{T}$, and let $x^{(1)}, \ldots, x^{(T)}$ be informed by $\tilde{f}_{T;U}$ and bounded, so that $\|x^{(t)}\| \le R$ for each T. If $d \ge 52TR^2 \log \frac{2T^2}{\delta}$ then with probability at least $1 - \delta$, for all $t \le T$ and each $j \in \{t, \ldots, T\}$, we have*

$$|\langle u^{(j)}, x^{(t)}\rangle| < 1/2.$$

For $t \in \mathbb{N}$, let $P_t \in \mathbb{R}^{d\times d}$ denote the projection operator to the span of $x^{(1)}, u^{(1)}, \ldots, x^{(t)}, u^{(t)}$, and let $P_t^\perp = I - P_t$ denote its orthogonal complement. We define the event G_t as

$$G_t = \left\{ \max_{j\in\{t,\ldots,T\}} \left|\left\langle u^{(j)}, P_{t-1}^\perp x^{(t)}\right\rangle\right| \le \alpha \left\| P_{t-1}^\perp x^{(t)}\right\| \right\} \text{ where } \alpha = \frac{1}{5R\sqrt{T}}. \quad (21)$$

For every t, define

$$G_{\le t} = \cap_{i\le t} G_i \text{ and } G_{<t} = \cap_{i<t} G_i \ .$$

The following linear-algebraic result justifies the definition (21) of G_t.

Lemma 4a *For all $t \le T$, $G_{\le t}$ implies $|\langle u^{(j)}, x^{(s)}\rangle| < 1/2$ for every $s \in \{1, \ldots, t\}$ and every $j \in \{s, \ldots, T\}$.*

Proof First, notice that since $G_{\le t}$ implies $G_{\le s}$ for every $s \le t$, it suffices to show that $G_{\le t}$ implies $|\langle u^{(j)}, x^{(t)}\rangle| < 1/2$ for every $j \in \{t, \ldots, T\}$. We will in fact prove a stronger statement:

$$\text{For every } t,\ G_{<t} \text{ implies } \left\| P_{t-1}u^{(j)} \right\|^2 \le 2\alpha^2 (t-1) \text{ for every } j \in \{t, \ldots, T\}, \tag{22}$$

where we recall that $P_t \in \mathbb{R}^{d\times d}$ is the projection operator to the span of $x^{(1)}, u^{(1)}, \ldots, x^{(t)}, u^{(t)}$, $P_t^{\perp} = I_d - P_t$ and $\alpha = 1/(5R\sqrt{T})$. Before proving (22), let us show that it implies our result. Fixing $j \in \{t, \ldots, T\}$, we have

$$\left|\left\langle u^{(j)}, x^{(t)} \right\rangle\right| \le \left|\left\langle u^{(j)}, P_{t-1}^{\perp}x^{(t)} \right\rangle\right| + \left|\left\langle u^{(j)}, P_{t-1}x^{(t)} \right\rangle\right|.$$

Since G_t holds, its definition (21) implies $|\langle u^{(j)}, P_{t-1}^{\perp}x^{(t)}\rangle| \le \alpha \left\| P_{t-1}^{\perp}x^{(t)} \right\| \le \alpha \left\| x^{(t)} \right\|$. Moreover, by Cauchy-Schwarz and the implication (22), we have $|\langle u^{(j)}, P_{t-1}x^{(t)}\rangle| \le \left\| P_{t-1}u^{(j)} \right\| \left\| x^{(t)} \right\| \le \sqrt{2\alpha^2(t-1)} \left\| x^{(t)} \right\|$. Combining the two bounds, we obtain the result of the lemma,

$$\left|\left\langle u^{(j)}, x^{(t)} \right\rangle\right| \le \left\| x^{(t)} \right\| (\alpha + \sqrt{2\alpha^2(t-1)}) < \frac{5}{2}\sqrt{t}R\alpha \le \frac{1}{2},$$

where we have used $\left\| x^{(t)} \right\| \le R$ and $\alpha = 1/(5R\sqrt{T})$.

We prove bound (22) by induction. The basis of the induction, $t = 1$, is trivial, as $P_0 = 0$. We shall assume (22) holds for $s \in \{1, \ldots, t-1\}$ and show that it consequently holds for $s = t$ as well. We may apply the Graham-Schmidt procedure on the sequence $x^{(1)}, u^{(1)}, \ldots, x^{(t-1)}, u^{(t-1)}$ to write

$$\left\| P_{t-1}u^{(j)} \right\|^2 = \sum_{i=1}^{t-1} \left| \left\langle \frac{P_{i-1}^{\perp}x^{(i)}}{\left\| P_{i-1}^{\perp}x^{(i)} \right\|}, u^{(j)} \right\rangle \right|^2 + \sum_{i=1}^{t-1} \left| \left\langle \frac{\hat{P}_{i-1}^{\perp}u^{(i)}}{\left\| \hat{P}_{i-1}^{\perp}u^{(i)} \right\|}, u^{(j)} \right\rangle \right|^2 \tag{23}$$

where $\hat{P}_k$ is the projection to the span of $\{x^{(1)}, u^{(1)}, \ldots, x^{(k)}, u^{(k)}, x^{(k+1)}\}$,

$$\hat{P}_k = P_k + \frac{1}{\left\| P_k^{\perp}x^{(k+1)} \right\|^2} \left(P_k^{\perp}x^{(k+1)}\right)\left(P_k^{\perp}x^{(k+1)}\right)^{\top}.$$

Then for every $j > i$ we have

$$\begin{aligned}\left\langle \hat{P}_{i-1}^{\perp}u^{(i)}, u^{(j)} \right\rangle = -\left\langle \hat{P}_{i-1}u^{(i)}, u^{(j)} \right\rangle &= -\left\langle P_{i-1}u^{(i)}, u^{(j)} \right\rangle \\ &\quad -\frac{\left\langle u^{(i)}, P_{i-1}^{\perp}x^{(i)} \right\rangle \left\langle u^{(j)}, P_{i-1}^{\perp}x^{(i)} \right\rangle}{\left\| P_{i-1}^{\perp}x^{(i)} \right\|^2},\end{aligned}$$

where the equalities hold by $\left\langle u^{(i)}, u^{(j)} \right\rangle = 0$, $\hat{P}_{i-1}^{\perp} = I - \hat{P}_{i-1}$, and the definition of $\hat{P}_{i-1}$.

The P_i matrices are projections, so $P_{i-1}^2 = P_{i-1}$, and Cauchy-Swartz and the induction hypothesis imply

$$\left|\left\langle P_{i-1}u^{(i)}, u^{(j)}\right\rangle\right| = \left|\left\langle P_{i-1}u^{(i)}, P_{i-1}u^{(j)}\right\rangle\right| \le \left\| P_{i-1}u^{(i)}\right\| \left\| P_{i-1}u^{(j)}\right\| \le 2\alpha^2 \cdot (i-1)\,.$$

Moreover, the event G_i implies $\left|\langle u^{(i)}, P_{i-1}^{\perp}x^{(i)}\rangle\langle u^{(j)}, P_{i-1}^{\perp}x^{(i)}\rangle\right| \le \alpha^2 \left\| P_{i-1}^{\perp}x^{(i)}\right\|^2$, so

$$\begin{aligned}\left|\left\langle \hat{P}_{i-1}^{\perp}u^{(i)}, u^{(j)}\right\rangle\right| &\le \left|\left\langle P_{i-1}u^{(i)}, u^{(j)}\right\rangle\right| + \left|\frac{\left\langle u^{(i)}, P_{i-1}^{\perp}x^{(i)}\right\rangle\left\langle u^{(j)}, P_{i-1}^{\perp}x^{(i)}\right\rangle}{\left\| P_{i-1}^{\perp}x^{(i)}\right\|^2}\right| \\ &\le \alpha^2\,(2i-1) \le \frac{\alpha}{2},\end{aligned} \tag{24a}$$

where the last transition uses $\alpha = \frac{1}{5R\sqrt{T}} \le \frac{1}{4i}$ because $R \ge \sqrt{T} \ge i$. We also have the lower bound

$$\begin{aligned}\left\| \hat{P}_{i-1}^{\perp}u^{(i)}\right\|^2 = \left|\left\langle \hat{P}_{i-1}^{\perp}u^{(i)}, u^{(i)}\right\rangle\right| &= 1 - \left\| P_{i-1}u^{(i)}\right\|^2 - \frac{\left(\left\langle u^{(i)}, P_{i-1}^{\perp}x^{(i)}\right\rangle\right)^2}{\left\| P_{i-1}^{\perp}x^{(i)}\right\|^2} \\ &\ge 1 - \alpha^2\,(2i-1) \ge \frac{1}{2},\end{aligned} \tag{24b}$$

where the first equality uses $(P_{i-1}^{\perp})^2 = P_{i-1}^{\perp}$, the second the definition of $\hat{P}_{i-1}$, and the inequality uses $\langle u^{(j)}, P_{i-1}^{\perp}x^{(i)}\rangle \le \alpha\| P_{i-1}^{\perp}x^{(i)}\|$ and $\| P_{i-1}u^{(j)}\|^2 \le 2\alpha^2\,(i-1)$.

Combining the observations (24a) and (24b), we can bound each summand in the second summation in (23). Since the summands in the first summation are bounded by α^2 by definition (21) of G_i, we obtain

$$\left\| P_{t-1}u^{(j)}\right\|^2 \le \sum_{i=1}^{t-1}\alpha^2 + \sum_{i=1}^{t-1}\frac{(\alpha/2)^2}{1/2} = \alpha^2\left(t-1+\frac{t-1}{2}\right) \le 2\alpha^2\,(t-1)\,,$$

which completes the induction. □

By Lemma 4a the event $G_{\le T}$ implies our result, so using $\mathbb{P}(G_{\le T}^c) \le \sum_{t=1}^{T}\mathbb{P}(G_t^c \mid G_{<t})$, it suffices to show that

$$\mathbb{P}\left(G_{\le T}^c\right) \le \sum_{t=1}^{T}\mathbb{P}(G_t^c \mid G_{<t}) \le \delta. \tag{25}$$

Let us therefore consider $\mathbb{P}\left(G_t^c \mid G_{<t}\right)$. By the union bound and fact that $\left\| P_{t-1}^{\perp}u^{(j)}\right\| \le 1$ for every t and j,

$$
\begin{aligned}
&\mathbb{P}(G_t^c \mid G_{<t}) \\
&\quad\le \sum_{j\in\{t,\ldots,T\}} \mathbb{P}\left(\left|\left\langle u^{(j)}, \frac{P_{t-1}^{\perp}x^{(t)}}{\left\| P_{t-1}^{\perp}x^{(t)}\right\|}\right\rangle\right| > \alpha \mid G_{<t}\right) \\
&\quad= \sum_{j\in\{t,\ldots,T\}} \mathbb{E}_{\xi,U_{(<t)}}\mathbb{P}\left(\left|\left\langle u^{(j)}, \frac{P_{t-1}^{\perp}x^{(t)}}{\left\| P_{t-1}^{\perp}x^{(t)}\right\|}\right\rangle\right| > \alpha \mid \xi, U_{(<t)}, G_{<t}\right) \\
&\quad\le \sum_{j\in\{t,\ldots,T\}} \mathbb{E}_{\xi,U_{(<t)}}\mathbb{P}\left(\left|\left\langle \frac{P_{t-1}^{\perp}u^{(j)}}{\left\| P_{t-1}^{\perp}u^{(j)}\right\|}, \frac{P_{t-1}^{\perp}x^{(t)}}{\left\| P_{t-1}^{\perp}x^{(t)}\right\|}\right\rangle\right| > \alpha \mid \xi, U_{(<t)}, G_{<t}\right),
\end{aligned}
\tag{26}
$$

where $U_{(<t)}$ is shorthand for $u^{(1)}, \ldots, u^{(t-1)}$ and ξ is the random variable generating $x^{(1)}, \ldots, x^{(T)}$.

In the following lemma, we state formally that conditioned on $G_{<i}$, the iterate $x^{(i)}$ depends on U only through its first $(i-1)$ columns.

Lemma 4b *For every $i \le T$, there exist measurable functions $\mathsf{A}_+^{(i)}$ and $\mathsf{A}_-^{(i)}$ such that*

$$
x^{(i)} = \mathsf{A}_+^{(i)}\left(\xi, U_{(<i)}\right) 1_{(G_{<i})} + \mathsf{A}_-^{(i)}\left(\xi, U\right) 1_{(G_{<i}^c)}. \tag{27}
$$

Proof Since the iterates are informed by $\tilde{f}_{T;U}$, we may write each one as (recall definition (4))

$$
x^{(i)} = \mathsf{A}^{(i)}\left(\xi, \nabla^{(0,\ldots,p)}\tilde{f}_{T;U}(x^{(1)}), \ldots, \nabla^{(0,\ldots,p)}\tilde{f}_{T;U}(x^{(i-1)})\right) = \mathsf{A}_-^{(i)}\left(\xi, U\right),
$$

for measurable functions $\mathsf{A}^{(i)}$, $\mathsf{A}_-^{(i)}$, where we recall the shorthand $\nabla^{(0,\ldots,p)}h(x)$ for the derivatives of h at x to order p. Crucially, by Lemma 4a, $G_{<i}$ implies $|\langle u^{(j)}, x^{(s)}\rangle| < \frac{1}{2}$ for every $s < i$ and every $j \ge s$. As $\bar{f}_T$ is a fixed robust zero-chain (Definition 4), for any $s < i$, the derivatives of $\tilde{f}_{T;U}$ at $x^{(s)}$ can therefore be expressed as functions of $x^{(s)}$ and $u^{(1)}, \ldots, u^{(s-1)}$, and—applying this argument recursively—we see that $x^{(i)}$ is of the form (27) for every $i \le T$. □

Consequently (as $G_{<t}$ implies $G_{<i}$ for every $i \le t$), conditioned on $\xi, U_{(<t)}$ and $G_{<t}$, the iterates $x^{(1)}, \ldots, x^{(t)}$ are deterministic, and so is $P_{t-1}^{\perp}x^{(t)}$. If $P_{t-1}^{\perp}x^{(t)} = 0$ then G_t holds and $\mathbb{P}(G_t^c \mid G_{<t}) = 0$, so we may assume without loss of generality that $P_{t-1}^{\perp}x^{(t)} \ne 0$. We may therefore regard $P_{t-1}^{\perp}x^{(t)}/\left\| P_{t-1}^{\perp}x^{(t)}\right\|$ in (26) as a deterministic unit vector in the subspace $P_{t-1}^{\perp}$ projects to. We now characterize the conditional distribution of $P_{t-1}^{\perp}u^{(j)}/\left\| P_{t-1}^{\perp}u^{(j)}\right\|$.

Lemma 4c *Let $t \le T$, and $j \in \{t, \ldots, T\}$. Then conditioned on ξ, $U_{(<t)}$ and $G_{<t}$, the vector $\frac{P_{t-1}^{\perp}u^{(j)}}{\left\| P_{t-1}^{\perp}u^{(j)}\right\|}$ is uniformly distributed on the unit sphere in the subspace to which $P_{t-1}^{\perp}$ projects.*

Proof This lemma is subtle. The vectors $u^{(j)}$, $j \geq t$, conditioned on $U_{(<t)}$, are certainly uniformly distributed on the unit sphere in the subspace orthogonal to $U_{(<t)}$. However, the additional conditioning on $G_{<t}$ requires careful handling. Throughout the proof we fix $t \leq T$ and $j \in \{t, \ldots, T\}$. We begin by noting that by (22), $G_{<t}$ implies

$$\left\| P_{t-1}^{\perp} u^{(j)} \right\|^2 = 1 - \left\| P_{t-1} u^{(j)} \right\|^2 \geq 1 - 2\alpha^2 (t-1) > 0.$$

Therefore, when $G_{<t}$ holds we have $P_{t-1}^{\perp} u^{(j)} \neq 0$ so $P_{t-1}^{\perp} u^{(j)} / \left\| P_{t-1}^{\perp} u^{(j)} \right\|$ is well-defined.

To establish our result, we will show that the density of $U_{(\geq t)} = [u^{(t)}, \ldots, u^{(T)}]$ conditioned on $\xi, U_{(<t)}, G_{<t}$ is invariant to rotations that preserve the span of $x^{(1)}, u^{(1)}, \ldots, x^{(t-1)}, u^{(t-1)}$. More formally, let $p_{\geq t}$ denote the density of $U_{(\geq t)}$ conditional on ξ, $U_{(<t)}$ and $G_{<t}$. We wish to show that

$$p_{\geq t}\left(U_{(\geq t)} \mid \xi, U_{(<t)}, G_{<t}\right) = p_{\geq t}\left(Z U_{(\geq t)} \mid \xi, U_{(<t)}, G_{<t}\right) \tag{28}$$

for every rotation $Z \in \mathbb{R}^{d \times d}$, $Z^{\top} Z = I_d$, satisfying

$$Z v = v = Z^{\top} v \ \text{ for all } \ v \in \left\{ x^{(1)}, u^{(1)}, \ldots, x^{(t-1)}, u^{(t-1)} \right\}.$$

Throughout, we let Z denote such a rotation. Letting $p_{\xi,U}$ and p_U denote the densities of (ξ, U) and U, respectively, we have

$$\begin{aligned} p_{\geq t}\left(U_{(\geq t)} \mid \xi, U_{(<t)}, G_{<t}\right) &= \frac{\mathbb{P}\left(G_{<t} \mid \xi, U\right) p_{\xi,U}\left(\xi, U\right)}{\mathbb{P}\left(G_{<t} \mid \xi, U_{(<t)}\right) p_{\xi, U_{(<t)}}\left(\xi, U_{(<t)}\right)} \\ &= \frac{\mathbb{P}\left(G_{<t} \mid \xi, U\right) p_U\left(U\right)}{\mathbb{P}\left(G_{<t} \mid \xi, U_{(<t)}\right) p_{U_{(<t)}}\left(U_{(<t)}\right)} \end{aligned}$$

where the first equality holds by the definition of conditional probability and second by the independence of ξ and U. We have $Z U_{(<t)} = U_{(<t)}$ and therefore, by the invariance of U to rotations, $p_U([U_{(<t)}, Z U_{(\geq t)}]) = p_U(ZU) = p_U(U)$. Hence, replacing U with ZU in the above display yields

$$p_{\geq t}\left(Z U_{(\geq t)} \mid \xi, U_{(<t)}, G_{<t}\right) = \frac{\mathbb{P}\left(G_{<t} \mid \xi, ZU\right) p_U\left(U\right)}{\mathbb{P}\left(G_{<t} \mid \xi, U_{(<t)}\right) p_{U_{(<t)}}\left(U_{(<t)}\right)}.$$

Therefore if we prove $\mathbb{P}(G_{<t} \mid \xi, U) = \mathbb{P}(G_{<t} \mid \xi, ZU)$—as we proceed to do—then we can conclude the equality (28) holds.

First, note that $\mathbb{P}\left(G_{<t} \mid \xi, U\right)$ is supported on $\{0, 1\}$ for every ξ, U, as they completely determine $x^{(1)}, \ldots, x^{(T)}$. It therefore suffices to show that $\mathbb{P}(G_{<t} \mid \xi, U) = 1$ if and only if $\mathbb{P}\left(G_{<t} \mid \xi, ZU\right) = 1$. Set $U' = ZU$, observing that $u'^{(i)} = Z u^{(i)} = u^{(i)}$ for any $i < t$, and let $x'^{(1)}, \ldots, x'^{(T)}$ be the sequence generated from ξ and U'. We will prove by induction on i that $\mathbb{P}(G_{<t} \mid \xi, U) = 1$ implies $\mathbb{P}(G_{<i} \mid \xi, U') = 1$ for every $i \leq t$. The basis of the induction is trivial as $G_{<1}$ always holds. Suppose now that $\mathbb{P}(G_{<i} \mid \xi, U') = 1$ for $i < t$, and therefore $x'^{(1)}, \ldots, x'^{(i)}$ can be written as functions of ξ and $u'^{(1)}, \ldots, u'^{(i-1)} = u^{(1)}, \ldots, u^{(i-1)}$ by Lemma 4b. Consequently, $x'^{(l)} = x^{(l)}$ for any $l \le i$ and also $P'^{\perp}_{i-1} x'^{(i)} = P^{\perp}_{i-1} x^{(i)}$. Therefore, for any $l \ge i$,

$$\left|\left\langle u'^{(l)}, \frac{P'^{\perp}_{i-1} x'^{(i)}}{\left\| P'^{\perp}_{i-1} x'^{(i)} \right\|} \right\rangle\right| \overset{(i)}{=} \left|\left\langle u^{(l)}, Z^{\top} \frac{P^{\perp}_{i-1} x^{(i)}}{\| P^{\perp}_{i-1} x^{(i)} \|} \right\rangle\right| \overset{(ii)}{=} \left|\left\langle u^{(l)}, \frac{P^{\perp}_{i-1} x^{(i)}}{\| P^{\perp}_{i-1} x^{(i)} \|} \right\rangle\right| \overset{(iii)}{\le} \alpha,$$

where in (i) we substituted $u'^{(l)} = Zu^{(l)}$ and $P'^{\perp}_{i-1} x'^{(i)} = P^{\perp}_{i-1} x^{(i)}$, (ii) is because $P^{\perp}_{i-1} x^{(i)} = x^{(i)} - P_{i-1} x^{(i)}$ is in the span of vectors $\left\{x^{(1)}, u^{(1)}, \ldots, x^{(i-1)}, u^{(i-1)}, x^{(i)}\right\}$ and therefore not modified by $Z^{\top}$, and (iii) is by our assumption that $G_{<t}$ holds, and so G_i holds. Therefore $\mathbb{P}\left(G_i \mid \xi, U'\right) = 1$ and $\mathbb{P}\left(G_{<i+1} \mid \xi, U'\right) = 1$, concluding the induction. An analogous argument shows that $\mathbb{P}\left(G_{<t} \mid \xi, U'\right) = 1$ implies $\mathbb{P}\left(G_{<t} \mid \xi, U\right) = \mathbb{P}\left(G_{<t} \mid \xi, Z^{\top} U'\right) = 1$ and thus $\mathbb{P}\left(G_{<t} \mid \xi, U\right) = \mathbb{P}\left(G_{<t} \mid \xi, ZU\right)$ as required.

Marginalizing the density (28) to obtain a density for $u^{(j)}$ and recalling that $P^{\perp}_{t-1}$ is measurable $\xi, U_{(<t)}, G_{<t}$, we conclude that, conditioned on $\xi, U_{(<t)}, G_{<t}$ the random variable $\frac{P^{\perp}_{t-1} u^{(j)}}{\| P^{\perp}_{t-1} u^{(j)} \|}$ has the same density as $\frac{P^{\perp}_{t-1} Z u^{(j)}}{\| P^{\perp}_{t-1} Z u^{(j)} \|}$. However, $P^{\perp}_{t-1} Z = Z P^{\perp}_{t-1}$ by assumption on Z, and therefore

$$\frac{P^{\perp}_{t-1} Z u^{(j)}}{\| P^{\perp}_{t-1} Z u^{(j)} \|} = Z \frac{P^{\perp}_{t-1} u^{(j)}}{\| P^{\perp}_{t-1} u^{(j)} \|}.$$

We conclude that the conditional distribution of the unit vector $\frac{P^{\perp}_{t-1} u^{(j)}}{\| P^{\perp}_{t-1} u^{(j)} \|}$ is invariant to rotations in the subspace to which $P^{\perp}_{t-1}$ projects. □

Summarizing the discussion above, the conditional probability in (26) measures the inner product of two unit vectors in a subspace of $\mathbb{R}^d$ of dimension $d' = \operatorname{tr}\left(P^{\perp}_{t-1}\right) \ge d - 2\,(t-1)$, with one of the vectors deterministic and the other uniformly distributed. We may write this as

$$\mathbb{P}\left(\left|\left\langle \frac{P^{\perp}_{t-1} u^{(j)}}{\| P^{\perp}_{t-1} u^{(j)} \|}, \frac{P^{\perp}_{t-1} x^{(t)}}{\| P^{\perp}_{t-1} x^{(t)} \|} \right\rangle\right| > \alpha \mid \xi, U_{(<t)}, G_{<t} \right) = \mathbb{P}(|v_1| > \alpha),$$

where v is uniformly distributed on the unit sphere in $\mathbb{R}^{d'}$. By a standard concentration of measure bound on the sphere [5, Lecture 8],

$$\mathbb{P}(|v_1| > \alpha) \le 2e^{-d'\alpha^2/2} \le 2e^{-\frac{\alpha^2}{2}(d-2t)}.$$

Substituting this bound back into the probability (26) gives

$$\mathbb{P}\left(G^c_t \mid G_{<t}\right) \le 2\,(T - t + 1)\, e^{-\frac{\alpha^2}{2}(d-2t)} \le 2Te^{-\frac{\alpha^2}{2}(d-2T)}.$$

Substituting this in turn into the bound (25), we have $\mathbb{P}(G^c_{\le T}) \le \sum_{t=1}^{T} \mathbb{P}(G^c_t \mid G_{<t}) \le 2T^2 e^{-\frac{\alpha^2}{2}(d-2T)}$. Setting $d \ge 52TR^2 \log \frac{2T^2}{\delta} \ge \frac{2}{\alpha^2} \log \frac{2T^2}{\delta} + 2T$ establishes $\mathbb{P}(G^c_{\le T}) \le \delta$, concluding Lemma 4. □

B.4 Proof of Lemma 6

Lemma 6 *The function $\hat{f}_{T;U}$ satisfies the following.*

i. *We have $\hat{f}_{T;U}(0) - \inf_x \hat{f}_{T;U}(x) \le 12T$.*
ii. *For every $p \ge 1$, the pth order derivatives of $\hat{f}_{T;U}$ are $\hat{\ell}_p$-Lipschitz continuous, where $\hat{\ell}_p \le \exp(cp \log p + c)$ for a numerical constant $c < \infty$.*

Proof Part i holds because $\hat{f}_{T;U}\,(0) = \bar{f}_T\,(0)$ and $\hat{f}_{T;U}\,(x) \ge \tilde{f}_{T;U}\,(\rho(x))$ for every x, so

$$\inf_{x\in\mathbb{R}^d} \hat{f}_{T;U}\,(x) \ge \inf_{x\in\mathbb{R}^d} \tilde{f}_{T;U}\,(\rho(x)) = \inf_{\|x\|\le R} \bar{f}_T\,(x) \ge \inf_{x\in\mathbb{R}^d} \bar{f}_T\,(x)\,,$$

and therefore by Lemma 3.i, we have $\hat{f}_{T;U}(0) - \inf_x \hat{f}_{T;U}(x) \le \bar{f}_T(0) - \inf_x \bar{f}_T(x) \le 12T$.

Establishing part ii requires substantially more work. Since smoothness with respect to Euclidean distances is invariant under orthogonal transformations, we take U to be the first T columns of the d-dimensional identity matrix, denoted $U = I_{d,T}$. Recall the scaling $\rho(x) = Rx/\sqrt{R^2 + \|x\|^2}$ with "radius" $R = 230\sqrt{T}$ and the definition $\hat{f}_{T;U}(x) = \bar{f}_T(U^\top \rho(x)) + \frac{1}{10}\|x\|^2$.

The quadratic $\frac{1}{10}\|x\|^2$ term in $\hat{f}_{T;U}$ has $\frac{1}{5}$-Lipschitz first derivative and 0-Lipschitz higher order derivatives (as they are all constant or zero), and we take $U = I_{d,T}$ without loss of generality, so we consider the function

$$\hat{f}_{T;I}(x) := \bar{f}_T(\rho(x)) = \bar{f}_T\left([\rho\,(x)]_1\,, \ldots, [\rho\,(x)]_T\right).$$

We now compute the partial derivatives of $\hat{f}_{T;I}$. Defining $y = \rho(x)$, let $\widetilde{\nabla}^k_{j_1,\ldots,j_k} := \frac{\partial^k}{\partial y_{j_1}\cdots\partial y_{j_k}}$ denote derivatives with respect to y. In addition, define $\mathcal{P}_k$ to be the set of all partitions of $[k] = \{1, \ldots, k\}$, i.e. $(S_1, \ldots, S_L) \in \mathcal{P}_k$ if and only if the S_i are disjoint and $\cup_l S_l = [k]$. Using the chain rule, we have for any k and set of indices $i_1, \ldots, i_k \le T$ that

$$\begin{aligned}&\nabla^k_{i_1,\ldots,i_k} \hat{f}_{T;I}(x)\\ &\quad= \sum_{(S_1,\ldots,S_L)\in\mathcal{P}_k} \sum_{j_1,\ldots,j_L=1}^{T} \left(\prod_{l=1}^{L} \nabla^{|S_l|}_{i_{S_l}} \rho_{j_l}(x)\right) \widetilde{\nabla}^L_{j_1,\ldots,j_L} \bar{f}_T(y), \quad y = \rho(x), \quad (29)\end{aligned}$$

where we have used the shorthand $\nabla_{i_S}^{|S|}$ to denote the partial derivatives with respect to each of x_{i_j} for $j \in S$. We use the equality (29) to argue that (recall the identity (2))

$$\left\| \nabla^{p+1} \hat{f}_{T;I}(x) \right\|_{\mathrm{op}} = \sup_{\|v\|=1} \langle \nabla^{p+1} \hat{f}_{T;I}(x), v^{\otimes(p+1)} \rangle := \hat{\ell}_p - \frac{1}{5} 1_{(p=1)} \le e^{cp \log p + c},$$

for some numerical constant[5], $0 < c < \infty$ and every $p \ge 1$. As explained in Sect. 2.1, this implies $\hat{f}_{T;U}$ has $e^{cp \log p + c}$-Lipschitz pth order derivative, giving part ii of the lemma.

To do this, we begin by considering the partitioned sum (29). Let $v \in \mathbb{R}^d$ be an arbitrary direction with $\|v\| = 1$. Then for $j \in [d]$ and $k \in \mathbb{N}$ we define the quantity

$$\widetilde{v}_j^k = \widetilde{v}_j^k(x) := \langle \nabla^k \rho_j(x), v^{\otimes k} \rangle,$$

algebraic manipulations and rearrangement of the sum (29) yield

$$\begin{aligned}
&\langle \nabla^k \hat{f}_{T;I}(x), v^{\otimes k} \rangle \\
&\quad = \sum_{(S_1,\ldots,S_L)\in\mathcal{P}_k} \sum_{i_1,\ldots,i_k=1}^{d} v_{i_1} v_{i_2} \cdots v_{i_k} \sum_{j_1,\ldots,j_L=1}^{T} \left(\prod_{l=1}^{L} \nabla_{i_{S_l}}^{|S_l|} \rho_{j_l}(x) \right) \widetilde{\nabla}_{j_1,\ldots,j_L}^{L} \bar{f}_T(y) \\
&\quad = \sum_{(S_1,\ldots,S_L)\in\mathcal{P}_k} \sum_{j_1,\ldots,j_L=1}^{T} \widetilde{v}_{j_1}^{|S_1|} \cdots \widetilde{v}_{j_L}^{|S_L|} \widetilde{\nabla}_{j_1,\ldots,j_L}^{L} \bar{f}_T(y) \\
&\quad = \sum_{(S_1,\ldots,S_L)\in\mathcal{P}_k} \left\langle \widetilde{\nabla}^L \bar{f}_T(y), \widetilde{v}^{|S_1|} \otimes \cdots \otimes \widetilde{v}^{|S_L|} \right\rangle.
\end{aligned}$$

We claim that there exists a numerical constant $c < \infty$ such that for all $k \in \mathbb{N}$,

$$\sup_x \|\widetilde{v}^k(x)\| \le \exp(ck \log k + c) R^{1-k}. \tag{30}$$

Before proving inequality (30), we show how it implies the desired lemma. By the preceding display, we have

$$|\langle \nabla^{p+1} \hat{f}_{T;I}(x), v^{\otimes(p+1)} \rangle| \le \sum_{(S_1,\ldots,S_L)\in\mathcal{P}_{p+1}} \left\| \widetilde{\nabla}^L \bar{f}_T(y) \right\|_{\mathrm{op}} \prod_{l=1}^{L} \|\widetilde{v}^{|S_l|}\|.$$

Lemma 3 shows that there exists a numerical constant $c < \infty$ such that

$$\left\| \nabla^{(L)} \bar{f}_T(y) \right\|_{\mathrm{op}} \le \ell_{L-1} \le \exp(cL \log L + c) \text{ for all } L \ge 2.$$

[5] To simplify notation we allow c to change from equation to equation throughout the proof, always representing a finite numerical constant independent of d, T, k or p.

When the number of partitions $L = 1$, we have $|S_1| = p + 1 \geq 2$, and so Lemma 3.ii yields

$$\left\| \nabla \bar{f}_T(y) \right\|_{\text{op}} \|\widetilde{v}^{|S_1|}\| = \left\| \nabla \bar{f}_T(y) \right\| \|\widetilde{v}^{|S_1|}\| \leq 23\sqrt{T} \cdot R^{-p} \exp(cp \log p + c) \\ \leq \exp(cp \log p + c),$$

where we have used $R = 230\sqrt{T}$. Using $|S_1| + \cdots + |S_L| = p + 1$ and the fact that $q(x) = (x + 1) \log(x + 1)$ satisfies $q(x) + q(y) \leq q(x + y)$ for every $x, y > 0$, we have

$$\left\| \widetilde{\nabla}^L \bar{f}_T(y) \right\|_{\text{op}} \prod_{l=1}^{L} \|\widetilde{v}^{|S_l|}\| \leq \exp(cp \log p + c)$$

for some $c < \infty$ and every $(S_1, \ldots, S_L) \in \mathcal{P}_{p+1}$. Bounds on Bell numbers [6, Thm. 2.1] give that there are at most $\exp(k \log k)$ partitions in $\mathcal{P}_k$, which combined with the bound above gives desired result.

Let us return to the derivation of inequality (30). We begin by recalling Faà di Bruno's formula for the chain rule. Let $f, g : \mathbb{R} \to \mathbb{R}$ be appropriately smooth functions. Then

$$\frac{d^k}{dt^k} f(g(t)) = \sum_{P \in \mathcal{P}_k} f^{(|P|)}(g(t)) \cdot \prod_{S \in P} g^{(|S|)}(t), \tag{31}$$

where $|P|$ denotes the number of disjoint elements of partition $P \in \mathcal{P}_k$. Define the function $\overline{\rho}(\xi) = \xi/\sqrt{1 + \|\xi\|^2}$, and let $\lambda(\xi) = \sqrt{1 + \|\xi\|^2}$ so that $\overline{\rho}(\xi) = \nabla\lambda(\xi)$ and $\rho(\xi) = R\overline{\rho}(\xi/R)$. Let $\overline{v}_j^k(\xi) = \langle \nabla^k \overline{\rho}_j(\xi), v^{\otimes k} \rangle$, so that

$$\overline{v}^k(\xi) = \nabla\langle \nabla^k \lambda(\xi), v^{\otimes k} \rangle \ \text{ and } \ \widetilde{v}^k = R^{1-k} \overline{v}^k(x/R). \tag{32}$$

With this in mind, we consider the quantity $\langle \nabla^k \lambda(\xi), v^{\otimes k} \rangle$. Defining temporarily the functions $\alpha(r) = \sqrt{1 + 2r}$ and $\beta(t) = \frac{1}{2}\|\xi + tv\|^2$, and their composition $h(t) = \alpha(\beta(t))$, we evidently have

$$h^{(k)}(0) = \langle \nabla^k \lambda(\xi), v^{\otimes k} \rangle = \sum_{P \in \mathcal{P}_k} \alpha^{(|P|)}(\beta(0)) \cdot \prod_{S \in P} \beta^{(|S|)}(0),$$

where the second equality used Faá di Bruno's formula (31). Now, we note the following immediate facts:

$$\alpha^{(l)}(r) = (-1)^l \frac{(2l - 1)!!}{(1 + 2r)^{l - 1/2}} \ \text{ and } \ \beta^{(l)}(t) = \begin{cases} \langle v, \xi \rangle + t \|v\|^2 & l = 1 \\ \|v\|^2 & l = 2 \\ 0 & l > 2. \end{cases}$$

Thus, if we let $\mathcal{P}_{k,2}$ denote the partitions of $[k]$ consisting only of subsets with one or two elements, we have

$$h^{(k)}(0) = \sum_{P \in \mathcal{P}_{k,2}} (-1)^{|P|} \frac{(2|P|-1)!!}{(1+\|\xi\|^2)^{|P|-1/2}} \langle \xi, v \rangle^{\mathsf{C}_1(P)} \|v\|^{2\mathsf{C}_2(P)}$$

where $\mathsf{C}_i(P)$ denotes the number of sets in P with precisely i elements. Noting that $\|v\| = 1$, we may rewrite this as

$$\langle \nabla^k \lambda(\xi), v^{\otimes k} \rangle = \sum_{l=1}^{k} \sum_{P \in \mathcal{P}_{k,2}, \mathsf{C}_1(P)=l} (-1)^{|P|} \frac{(2|P|-1)!!}{(1+\|\xi\|^2)^{|P|-1/2}} \langle \xi, v \rangle^l.$$

Taking derivatives we obtain

$$\widetilde{v}^k = \nabla \langle \nabla^k \lambda(\xi), v^{\otimes k} \rangle = \left(\sum_{l=1}^{k} a_l(\xi) \langle \xi, v \rangle^{l-1} \right) v + \left(\sum_{l=1}^{k} b_l(\xi) \langle \xi, v \rangle^{l} \right) \xi$$

where

$$a_l(\xi) = l \cdot \sum_{P \in \mathcal{P}_{k,2}, \mathsf{C}_1(P)=l} \frac{(-1)^{|P|}(2|P|-1)!!}{(1+\|\xi\|^2)^{|P|-1/2}}$$
$$\text{and } \; b_l(\xi) = \sum_{P \in \mathcal{P}_{k,2}, \mathsf{C}_1(P)=l} \frac{(-1)^{|P|+1}(2|P|+1)!!}{(1+\|\xi\|^2)^{|P|+1/2}}.$$

We would like to bound $a_l(\xi)\langle \xi, v \rangle^{l-1}$ and $b_l(\xi)\langle \xi, v \rangle^{l} \xi$. Note that $|P| \ge \mathsf{C}_1(P)$ for every $P \in \mathcal{P}_k$, so $|P| \ge l$ in the sums above. Moreover, bounds for Bell numbers [6, Thm. 2.1] show that there are at most $\exp(k \log k)$ partitions of $[k]$, and $(2k-1)!! \le \exp(k \log k)$ as well. As a consequence, we obtain

$$\sup_{\xi} |a_l(\xi) \langle \xi, v \rangle^{l-1}| \le \exp(cl \log l) \sup_{\xi} \frac{|\langle \xi, v \rangle|^{l-1}}{(1+\|\xi\|^2)^{(l-1)/2}} < \exp(cl \log l),$$

where we have used $|\langle \xi, v \rangle| \le \|\xi\|$ due to $\|v\| = 1$. We similarly bound $\sup_\xi |b_l(\xi)| |\langle \xi, v \rangle|^l \, \|\xi\|$. Returning to expression (32), we have

$$\sup_{x} \|\widetilde{v}^k(x)\| \le \exp\left(ck \log k + c\right) R^{1-k},$$

for a numerical constant $c < \infty$. This is the desired bound (30), completing the proof. □

C Proof of Theorem 3

Theorem 3 *There exist numerical constants $0 < c_0, c_1 < \infty$ such that the following lower bound holds. For any $p \geq 1$, let D, L_p, and ϵ be positive. Then*

$$\mathcal{T}_\epsilon\left(\mathcal{A}_{\mathsf{rand}}, \mathcal{F}_p^{\mathrm{dist}}(D, L_p)\right) \geq c_0 \cdot D^{1+p}\left(\frac{L_p}{\ell'_p}\right)^{\frac{1+p}{p}} \epsilon^{-\frac{1+p}{p}},$$

where $\ell'_p \leq e^{c_1 p \log p + c_1}$. The lower bound holds even if we restrict $\mathcal{F}_p^{\mathrm{dist}}(D, L_p)$ to functions with domain of dimension $1 + c_2 q\left(D^{1+p}\left(L_p/\ell'_p\right)^{\frac{1+p}{p}} \epsilon^{-\frac{1+p}{p}}\right)$, for a some numerical constant $c_2 < \infty$ and $q(x) = x^2 \log(2x)$.

We divide the proof of the theorem into two parts, as in our previous results, first providing a few building blocks, then giving the theorem. The basic idea is to introduce a negative "bump" that is challenging to find, but which is close to the origin.

To make this precise, let $e^{(j)}$ denote the jth standard basis vector. Then we define the bump function $\bar{h}_T : \mathbb{R}^T \to \mathbb{R}$ by

$$\begin{aligned}\bar{h}_T(x) &= \Psi\left(1 - \frac{25}{2}\left\|x - \frac{4}{5}e^{(T)}\right\|^2\right)\\ &= \begin{cases} 0 & \left\|x - \frac{4}{5}e^{(T)}\right\| \geq \frac{1}{5} \\ \exp\left(1 - \frac{1}{\left(1 - 25\left\|x - \frac{4}{5}e^{(T)}\right\|^2\right)^2}\right) & \text{otherwise.}\end{cases}\end{aligned} \tag{33}$$

As Fig. 2 shows, $\bar{h}_T$ features a unit-height peak centered at $\frac{4}{5}e^{(T)}$, and it is identically zero when the distance from that peak exceeds $\frac{1}{5}$. The volume of the peak vanishes exponentially with T, making it hard to find by querying $\bar{h}_T$ locally. We list the properties of $\bar{h}_T$ necessary for our analysis.

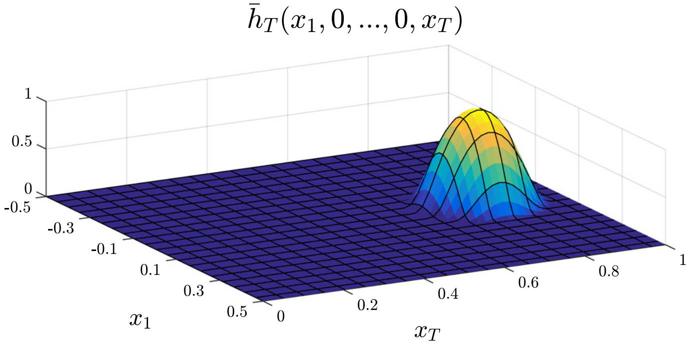

Fig. 2 Two-dimensional cross-section of the bump function $\bar{h}_T$

Lemma 10 *The function $\bar{h}_T$ satisfies the following.*

i. $\bar{h}_T\left(0.8e^{(T)}\right) = 1$ *and* $\bar{h}_T(x) \in [0, 1]$ *for all* $x \in \mathbb{R}^T$.
ii. $\bar{h}_T(x) = 0$ *on the set* $\{x \in \mathbb{R}^d \mid x_T \le \frac{3}{5} \text{ or } \|x\| \ge 1\}$.
iii. *For* $p \ge 1$, *the pth order derivatives of* $\bar{h}_T$ *are* $\tilde{\ell}_p$*-Lipschitz continuous, where* $\tilde{\ell}_p < e^{3p\log p + cp}$ *for some numerical constant* $c < \infty$.

We prove the lemma in Sect. C.1; the proof is similar to that of Lemma 6. With these properties in hand, we can prove Theorem 3.

C.1 Proof of Lemma 10

Properties i and ii are evident from the definition (33) of $\bar{h}_T$. To show property iii, consider $h(x) = \bar{h}_T(\frac{x+0.8e^{(T)}}{5}) = \Psi(1 - \frac{1}{2}\|x\|^2)$, which is a translation and scaling of $\bar{h}_T$, so if we show h has $(\tilde{\ell}_p/5^{p+1})$-Lipschitz pth order derivatives, for every $p \ge 1$, we obtain the required results. For any $x, v \in \mathbb{R}^T$ with $\|v\| \le 1$ we define the directional projection $h_{x,v}(t) = h(x + t \cdot v)$. The required smoothness bound is equivalent to

$$\left|h_{x,v}^{(p+1)}(0)\right| \le \tilde{\ell}_p/5^{p+1} \le e^{cp\log p + c}$$

for every $x, v \in \mathbb{R}^d$ with $\|v\| \le 1$, every $p \ge 1$ and some numerical constant $c < \infty$ (which we allow to change from equation to equation, caring only that it is finite and independent of T and p).

As in the proof of Lemma 6, we write $h_{x,v}(t) = \Psi(\beta(t))$ where $\beta(t) = 1 - \frac{1}{2}\|x + tv\|^2$, and use Faá di Bruno's formula (31) to write, for any $k \ge 1$,

$$h_{x,v}^{(k)}(0) = \sum_{P \in \mathcal{P}_k} \Psi^{(|P|)}(\beta(0)) \cdot \prod_{S \in P} \beta^{(|S|)}(0),$$

where $\mathcal{P}_k$ is the set of partitions of $[k]$ and $|P|$ denotes the number of set in partition P. Noting that $\beta'(0) = -\langle x, v\rangle$, $\beta''(0) = -\|v\|^2$ and $\beta^{(n)}(0) = 0$ for any $n > 2$, we have

$$h_{x,v}^{(k)}(0) = \sum_{P \in \mathcal{P}_{k,2}} (-1)^{|P|}\Psi^{(|P|)}\left(1 - \frac{1}{2}\|x\|^2\right)\langle x, v\rangle^{\mathsf{C}_1(P)}\|v\|^{2\mathsf{C}_2(P)}$$

where $\mathcal{P}_{k,2}$ denote the partitions of $[k]$ consisting only of subsets with one or two elements and $\mathsf{C}_i(P)$ denotes the number of sets in P with precisely i elements.

Noting that $\Psi^{(k)}(1 - \frac{1}{2}\|x\|^2) = 0$ for any $k \ge 0$ and $\|x\| > 1$, we may assume $\|x\| \le 1$. Since $\|v\| \le 1$, we may bound $|h_{x,v}^{(p+1)}(0)|$ by

$$\begin{aligned}\left|h_{x,v}^{(p+1)}(0)\right| &\le \left|\mathcal{P}_{p+1,2}\right| \cdot \max_{k \in [p+1]} \sup_{x \in \mathbb{R}} |\Psi^{(k)}(x)| \\ &\overset{(i)}{\le} e^{\frac{p+1}{2}\log(p+1)} \cdot e^{\frac{5(p+1)}{2}\log(\frac{5}{2}(p+1))} \le e^{3p\log p + cp}\end{aligned}$$

for some absolute constant $c < \infty$, where inequality (i) follows from Lemma 1.iv and that the number of matchings in the complete graph (or the kth telephone number [21, Lem. 2]) has bound $|\mathcal{P}_{k,2}| \le e^{\frac{k}{2}\log k}$. This gives the result.

C.2 Proof of Theorem 3

For some $T \in \mathbb{N}$ and $\sigma > 0$ to be specified, and $d = \lceil 52 \cdot 230^2 \cdot T^2 \log(4T^2) \rceil$, consider the function $f_U : \mathbb{R}^d \to \mathbb{R}$ indexed by orthogonal matrix $U \in \mathbb{R}^{d\times T}$ and defined as

$$f_U(x) = \frac{L_p \sigma^{p+1}}{\ell'_p} \hat{f}_{T;U}(x/\sigma) - \frac{L_p D^{p+1}}{\ell'_p} \bar{h}_T(U^\top x/D),$$

where $\hat{f}_{T;U}(x) = \tilde{f}_{T;U}(\rho(x)) + \frac{1}{10}\|x\|^2$ is the randomized hard instance construction (13) with $\rho(x) = x/\sqrt{1+\|x/R\|^2}$, $\bar{h}_T$ is the bump function (33) and $\ell'_p = \hat{\ell}_p + \tilde{\ell}_p$, for $\hat{\ell}_p$ and $\tilde{\ell}_p$ as in Lemmas 6.ii and 10.iii, respectively. By the lemmas, f_U has L_p-Lipschitz pth order derivatives and $\ell'_p \le e^{c_1 p \log p + c_1}$ for some $c_1 < \infty$. We assume that $\sigma \le D$; our subsequent choice of σ will obey this constraint.

Following our general proof strategy, we first demonstrate that $f_U \in \mathcal{F}_p^{\text{dist}}(D, L_p)$, for which all we need do is guarantee that the global minimizers of f_U have norm at most D. By the constructions (13) and (10) of $\hat{f}_{T;U}$ and $\tilde{f}_{T;U}$, Lemma 10.i implies

$$\begin{aligned} &f_U\left((0.8D)u^{(T)}\right) \\ &\quad= \frac{L_p\sigma^{p+1}}{\ell'_p}\bar{f}_T(\rho(e^{(T)})) + \frac{L_p\sigma^{p+1}}{10\ell'_p}\left\|\frac{4Du^{(T)}}{5\sigma}\right\|^2 - \frac{L_pD^{p+1}}{\ell'_p}\bar{h}_T(0.8e^{(T)}) \\ &\quad= \frac{L_p\sigma^{p+1}}{\ell'_p}\bar{f}_T(0) + \frac{8L_p\sigma^{p-1}D^2}{125\ell'_p} + \frac{-L_pD^{p+1}}{\ell'_p} < -\frac{117}{125}\frac{L_pD^{p+1}}{\ell'_p} \\ &\qquad+ \frac{L_p\sigma^{p+1}}{\ell'_p}\bar{f}_T(0) \end{aligned}$$

with the final inequality using our assumption $\sigma \le D$. On the other hand, for any x such that $\bar{h}_T(U^\top x/D) = 0$, we have by Lemma 6.i (along with $\hat{f}_{T;U}(0) = 0$) that

$$f_U(x) \ge \frac{L_p\sigma^{p+1}}{\ell'_p}\inf_x \hat{f}_{T;U}(x) \ge -12\frac{L_p\sigma^{p+1}}{\ell'_p}T + \frac{L_p\sigma^{p+1}}{\ell'_p}\bar{f}_T(0).$$

Combining the two displays above, we conclude that if

$$12\frac{L_p\sigma^{p+1}}{\ell'_p}T \le \frac{117}{125}\frac{L_pD^{p+1}}{\ell'_p},$$

then all global minima $x^\star$ of f_U must satisfy $\bar{h}_T(U^\top x^\star/D) > 0$. Inspecting the definition (18) of $\bar{h}_T$, this implies $\left\|x^\star/D - 0.8u^{(T)}\right\| < \frac{1}{5}$, and therefore $\|x^\star\| \le D$. Thus, by setting

$$T = \left\lfloor \frac{D^{p+1}}{13\sigma^{p+1}} \right\rfloor, \tag{34}$$

we guarantee that $f_U \in \mathcal{F}_p^{\text{dist}}(D, L_p)$ as long as $\sigma \le D$.

It remains to show that, for an appropriately chosen σ, any randomized algorithm requires (with high probability) more than T iterations to find x such that $\|\nabla f_U(x)\| < \epsilon$. We claim that when $\sigma \le D$, for any $x \in \mathbb{R}^d$,

$$|\langle u^{(T)}, \rho(x/\sigma)\rangle| < \frac{1}{2} \text{ implies } \bar{h}_T(U^\top y/D) = 0 \text{ for } y \text{ in a neighborhood of } x. \tag{35}$$

We defer the proof of claim (35) to the end of this section.

Now, let $U \in \mathbb{R}^{d\times T}$ be an orthogonal matrix chosen uniformly at random from $\mathsf{O}(d, T)$. Let $x^{(1)}, \ldots, x^{(t)}$ be a sequence of iterates generated by algorithm $\mathsf{A} \in \mathcal{A}_{\mathsf{rand}}$ applied on f_U. We argue that $|\langle u^{(T)}, \rho(x^{(t)}/\sigma)\rangle| < 1/2$ for all $t \le T$, with high probability. To do so, we briefly revisit the proof of Lemma 4 (Sect. B.3) where we replace $\tilde{f}_{T;U}$ with f_U and $x^{(t)}$ with $\rho(x^{(t)}/\sigma)$. By Lemma 4a we have that for every $t \le T$ the event $G_{\le t}$ implies $|\langle u^{(T)}, \rho(x^{(s)}/\sigma)\rangle| < 1/2$ for all $s \le t$, and therefore by the claim (35) we have that Lemma 4b holds (as we may replace the terms $\bar{h}_T(U^\top x^{(s)}/D)$, $s < t$, with 0 whenever $G_{<t}$ holds). The rest of the proof of Lemma 4a proceeds unchanged and gives us that with probability greater than 1/2 (over any randomness in A and the uniform choice of U),

$$|\langle u^{(T)}, \rho(x^{(t)}/\sigma)\rangle| < \frac{1}{2} \text{ for all } t \le T.$$

By claim (35), this implies $\nabla\bar{h}_T(U^\top x^{(t)}/D) = 0$, and by Lemma 5, $\|\nabla \hat{f}_{T;U}(x^{(t)}/\sigma)\| > 1/2$. Thus, after scaling,

$$\left\|\nabla f_U(x^{(t)})\right\| > \frac{L_p\sigma^p}{2\ell'_p}$$

for all $t \le T$, with probability greater that 1/2. As in the proof of Theorem 2, By taking $\sigma = (2\ell'_p\epsilon/L_p)^{1/p}$ we guarantee

$$\inf_{\mathsf{A}\in\mathcal{A}_{\mathsf{det}}} \sup_{U\in\mathsf{O}(d,T)} \mathsf{T}_\epsilon\big(\mathsf{A}, f_U\big) \ge 1 + T.$$

where $T = \left\lfloor D^{p+1}/13\sigma^{p+1}\right\rfloor$ is defined in Eq. (34). Thus, as $f_U \in \mathcal{F}_p^{\text{dist}}(D, L_p)$ for our choice of T, we immediately obtain

$$\mathcal{T}_\epsilon\big(\mathcal{A}_{\mathsf{rand}}, \mathcal{F}_p^{\mathrm{dist}}(D, L_p)\big) \geq T + 1 \geq \frac{D^{1+p}}{52}\left(\frac{L_p}{\ell'_p}\right)^{\frac{1+p}{p}} \epsilon^{-\frac{1+p}{p}},$$

as long as our initial assumption $\sigma \leq D$ holds. When $\sigma > D$, we have that $\frac{2\ell'_p}{L_p}\epsilon > D^p$, or $1 > D^{p+1}(\frac{L_p}{2\ell'_p})^{\frac{1+p}{p}}\epsilon^{-\frac{1+p}{p}}$, so that the bound is vacuous in this case regardless: every method must take at least 1 step.

Finally, we return to demonstrate claim (35). Note that $|\langle u^{(T)}, \rho(x/\sigma)\rangle| < 1/2$ is equivalent to $|\langle u^{(T)}, x\rangle| < \frac{\sigma}{2}\sqrt{1 + \|\frac{x}{\sigma R}\|^2}$, and consider separately the cases $\|x/\sigma\| \leq R/2$ and $\|x/\sigma\| > R/2 = 115\sqrt{T}$. In the first case, we have $|\langle u^{(T)}, x\rangle| < (\sqrt{5}/4)\sigma < (3/5)D$, by our assumption $\sigma \leq D$. Therefore, by Lemma 10.ii we have that $\bar{h}_T(U^\top y/D) = 0$ for y near x. In the second case, we have $\|x\| > (4R/\sqrt{5})|\langle u^{(T)}, x\rangle| > 230|\langle u^{(T)}, x\rangle|$. If in addition $|\langle u^{(T)}, x\rangle| < (3/5)D$ then our conclusion follows as before. Otherwise, $\|x\|/D > 230 \cdot (3/5) > 1$, so again the conclusion follows by Lemma 10.ii.

References

1. Agarwal, A., Bartlett, P.L., Ravikumar, P., Wainwright, M.J.: Information-theoretic lower bounds on the oracle complexity of convex optimization. IEEE Trans. Inf. Theory **58**(5), 3235–3249 (2012)
2. Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., Ma, T.: Finding approximate local minima faster than gradient descent. In: Proceedings of the Forty-Ninth Annual ACM Symposium on the Theory of Computing (2017)
3. Arjevani, Y., Shalev-Shwartz, S., Shamir, O.: On lower and upper bounds in smooth and strongly convex optimization. J. Mach. Learn. Res. **17**(126), 1–51 (2016)
4. Arjevani, Y., Shamir, O., Shiff, R.: Oracle complexity of second-order methods for smooth convex optimization (2017). arXiv:1705.07260 [math.OC]
5. Ball, K.: An elementary introduction to modern convex geometry. In: Levy, S. (ed.) Flavors of Geometry, pp. 1–58. MSRI Publications, Cambridge (1997)
6. Berend, D., Tassa, T.: Improved bounds on Bell numbers and on moments of sums of random variables. Prob. Math. Stat. **30**(2), 185–205 (2010)
7. Birgin, E.G., Gardenghi, J.L., Martínez, J.M., Santos, S.A., Toint, P.L.: Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. Math. Program. **163**(1–2), 359–368 (2017)
8. Boumal, N., Voroninski, V., Bandeira, A.: The non-convex Burer–Monteiro approach works on smooth semidefinite programs. Adv. Neural Inf. Process. Syst. **30**, 2757–2765 (2016)
9. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
10. Braun, G., Guzmán, C., Pokutta, S.: Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory. IEEE Trans. Inf. Theory **63**(7), 4709–4724 (2017)
11. Burer, S., Monteiro, R.D.: A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. Math. Program. **95**(2), 329–357 (2003)
12. Candès, E.J., Li, X., Soltanolkotabi, M.: Phase retrieval via Wirtinger flow: theory and algorithms. IEEE Trans. Inf. Theory **61**(4), 1985–2007 (2015)
13. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Convex until proven guilty: dimension-free acceleration of gradient descent on non-convex functions. In: Proceedings of the 34th International Conference on Machine Learning (2017)
14. Carmon, Y., Duchi, J. C., Hinder, O., Sidford, A.: Lower bounds for finding stationary points II: first-order methods (2017). arXiv: 1711.00841 [math.OC]. URL https://arxiv.org/pdf/1711.00841.pdf
15. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Accelerated methods for non-convex optimization. SIAM J. Optim. **28**(2), 1751–1772 (2018)

16. Cartis, C., Gould, N.I., Toint, P.L.: On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. SIAM J. Optim. **20**(6), 2833–2852 (2010)
17. Cartis, C., Gould, N.I., Toint, P.L.: Complexity bounds for second-order optimality in unconstrained optimization. J. Complex. **28**(1), 93–108 (2012)
18. Cartis, C., Gould, N.I.M., Toint, P.L.: How much patience do you have? A worst-case perspective on smooth nonconvex optimization. Optima **88**, 1–10 (2012)
19. Cartis, C., Gould, N.I., Toint, P.L.: A note about the complexity of minimizing nesterov's smooth Chebyshev–Rosenbrock function. Optim. Methods Softw. **28**, 451–457 (2013)
20. Cartis, C., Gould, N.I.M., Toint, P.L.: Worst-case evaluation complexity and optimality of second-order methods for nonconvex smooth optimization (2017). arXiv:1709.07180 [math.OC]
21. Chowla, S., Herstein, I.N., Moore, W.K.: On recursions connected with symmetric groups I. Can. J. Math. **3**, 328–334 (1951)
22. Conn, A.R., Gould, N.I.M., Toint, P.L.: Trust Region Methods. MPS-SIAM Series on Optimization. SIAM, Bangkok (2000)
23. Hager, W.W., Zhang, H.: A survey of nonlinear conjugate gradient methods. Pac. J. Optim. **2**(1), 35–58 (2006)
24. Hinder, O.: Cutting plane methods can be extended into nonconvex optimization. In: Proceedings of the Thirty First Annual Conference on Computational Learning Theory (2018)
25. Jarre, F.: On Nesterov's smooth Chebyshev–Rosenbrock function. Optim. Methods Softw. **28**(3), 478–500 (2013)
26. Jin, C., Ge, R., Netrapalli, P., Kakade, S.M., Jordan, M.I.: How to escape saddle points efficiently. In: Proceedings of the 34th International Conference on Machine Learning (2017)
27. Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from noisy entries. J. Mach. Learn. Res. **11**, 2057–2078 (2010)
28. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
29. Liu, D., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Math. Program. **45**(1), 503–528 (1989)
30. Loh, P.-L., Wainwright, M.J.: High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. Ann. Stat. **40**(3), 1637–1664 (2012)
31. Loh, P.-L., Wainwright, M.J.: Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. J. Mach. Learn. Res. **16**, 559–616 (2013)
32. Monteiro, R.D., Svaiter, B.F.: An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. SIAM J. Optim. **23**(2), 1092–1125 (2013)
33. Murty, K., Kabadi, S.: Some NP-complete problems in quadratic and nonlinear programming. Math. Program. **39**, 117–129 (1987)
34. Nemirovski, A.: Efficient methods in convex programming. The Israel Institute of Technology, Technion (1994)
35. Nemirovski, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley, Hoboken (1983)
36. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Sov. Math. Dokl. **27**(2), 372–376 (1983)
37. Nesterov, Y.: Introductory Lectures on Convex Optimization. Kluwer Academic Publishers, Cambridge (2004)
38. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim. **22**(2), 341–362 (2012)
39. Nesterov, Y.: How to make the gradients small. Optima **88**, 10–11 (2012)
40. Nesterov, Y., Polyak, B.: Cubic regularization of Newton method and its global performance. Math. Program. Ser. A **108**, 177–205 (2006)
41. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, Berlin (2006)
42. Sun, J., Qu, Q., Wright, J.: A geometric analysis of phase retrieval. Found. Comput. Math. **18**(5), 1131–1198 (2018)
43. Traub, J., Wasilkowski, H., Wozniakowski, H.: Information-Based Complexity. Academic Press, Cambridge (1988)
44. Vavasis, S.A.: Black-box complexity of local minimization. SIAM J. Optim. **3**(1), 60–80 (1993)
45. Woodworth, B.E., Srebro, N.: Tight complexity bounds for optimizing composite objectives. Adv. Neural Inf. Process. Syst. **30**, 3639–3647 (2016)

46. Woodworth, B. E., Srebro, N.: Lower bound for randomized first order convex optimization (2017). arXiv:1709.03594 [math.OC]
47. Zhang, X., Ling, C., Qi, L.: The best rank-1 approximation of a symmetric tensor and related spherical optimization problems. SIAM J. Matrix Anal. Appl. **33**(3), 806–821 (2012)