Lower Bounds for Finding Stationary Points I

Yair Carmon  John C. Duchi  Oliver Hinder  Aaron Sidford
{yairc,jduchi,ohinder,sidford}@stanford.edu

Abstract

We prove lower bounds on the complexity of finding $\epsilon$-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 $p$th order derivatives such that $A$ requires at least $\epsilon^{-(p+1)/p}$ queries to find an $\epsilon$-stationary point. Our lower bounds are sharp to within constants, and they show that gradient descent, cubic-regularized Newton’s method, and generalized $p$th order regularization are worst-case optimal within their natural function classes.

1 Introduction

Consider the optimization problem

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

where $f : \mathbb{R}^d \to \mathbb{R}$ is smooth, but possibly non-convex. In general, it is intractable to even approximately minimize such $f$ [32, 30], so—following an established line of research—we consider the problem of finding an $\epsilon$-stationary point of $f$, meaning some $x \in \mathbb{R}^d$ such that

$$\|\nabla f(x)\| \leq \epsilon.$$  \hspace{1cm} (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 [27, 28], phase retrieval [12, 39], and non-convex (low-rank) reformulations of semidefinite programs and matrix completion [11, 24, 8], 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 [25], motivates a growing body of work on algorithms with strong complexity guarantees for finding stationary points [37, 7, 13, 2, 14]. In contrast to this algorithmic progress, algorithm-independent lower bounds (fundamental limits on complexity) for finding stationary points are largely unexplored.

Even for non-convex functions $f$, it is possible to find $\epsilon$-stationary points in time polynomial in $1/\epsilon$. Of particular interest are methods for which the number of function and derivative evaluations does not depend on $d$, the dimension of $\text{dom} f$, 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 $f(x^{(0)}) - \inf_x f(x) \leq \Delta$ at the initial point $x^{(0)}$, gradient descent finds an $\epsilon$-stationary point in at
most $2L_1\Delta \epsilon^{-2}$ iterations [34]. Under the additional assumption that $f$ has Lipschitz continuous Hessian, our work [13] and Agarwal et al. [2] exhibit randomized first-order methods that find an $\epsilon$-stationary point in time scaling as $\epsilon^{-7/4}\log \frac{d}{\epsilon}$ (ignoring other problem-dependent constants). In subsequent work [14], 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 $\epsilon$-stationary point.

By evaluation of higher order derivatives, such as the Hessian, it is possible to achieve better $\epsilon$ dependence. Nesterov and Polyak’s cubic regularization of Newton’s method [37, 16] guarantees $\epsilon$-stationarity (1) in $\epsilon^{-3/2}$ iterations, but each iteration may be expensive when the dimension $d$ is large. More generally, $p$th-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 approximately minimizing 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 [15], we focus on the converse problem: providing dimension-free complexity lower bounds for finding $\epsilon$-stationary points. We show fundamental limits on the best achievable $\epsilon$ 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 $\epsilon$-suboptimal points, that is, $x$ satisfying $f(x) \leq f(x^*) + \epsilon$ for some $\epsilon > 0$, where $x^* \in \arg\min_x f(x)$. Here we review only the dimension-free optimal rates, as those are most relevant for our results. Given a point $x^{(0)}$ satisfying $\|x^{(0)} - x^*\| \leq D < \infty$, if $f$ is convex with $L_1$-Lipschitz gradient, Nesterov’s accelerated gradient method finds an $\epsilon$-suboptimal point in $O(\sqrt{L_1}D\epsilon^{-1/2})$ gradient evaluations, which is optimal even among randomized, higher-order algorithms [33, 32, 34, 41].

For non-smooth problems, that is, when $f$ is $L_0$-Lipschitz, subgradient methods achieve the optimal rate of $O(L_0^2D^2/\epsilon^2)$ subgradient evaluations (cf. [10, 32, 34]). In Part II of this paper [15], 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 [32, §1.6] show that for functions $f: \mathbb{R}^d \to \mathbb{R}$ with Lipschitz $1, \ldots, p$th derivatives, and algorithms receiving all derivatives of $f$ at the query point $x$, the worst case complexity of finding $\epsilon$-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 in $\mathbb{R}^d$. Nesterov [36] 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. Cartis et al. [16, 17] show that for important but specific algorithms, namely gradient descent and cubic regularization of Newton’s method, the respective performance guarantees are tight in the worst case. They also extend these results to certain structured classes of methods [18, 19]. We provide the first algorithm-independent lower bounds for finding stationary points in the unconstrained setting.

---

1Higher order methods can yield improvements under additional smoothness: if in addition $f$ has $L_2$-Lipschitz Hessian and $\epsilon \leq L_1^{-1/3}L_2^{-4/3}D^{2/3}$, an accelerated Newton method achieves the (optimal) rate $\Theta((L_2D^3/\epsilon)^{2/3})$ [4, 29].
1.2 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 Section 5), is as follows. Let $p \in \mathbb{N}$ and $\Delta, L_p$, 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 $p$th derivative and satisfies $f(x^{(0)}) - f(x^*) \leq \Delta$, such that, with high probability, $A$ requires at least

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

oracle queries to find an $\epsilon$-stationary point of $f$. The constructed function $f$ has domain of 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 [34, 36] 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 [37, 16] achieves the optimal rate $\epsilon^{-3/2}$, and for general $p$ we show that $p$th 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 $\epsilon^{-2}$ [14]. In Part II of this series [15], 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 [14].

1.3 Our approach and paper organization

In Section 2 we introduce the classes of functions and algorithms we consider as well as our notion of complexity. Then, in Section 3, we present the generic technique we use to prove lower bound for deterministic algorithms in both this paper and Part II [15]. While essentially present in previous work, our technique abstracts away and generalizes the central arguments in many lower bounds [32, 31, 41, 4]. 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 [32]. The key building blocks of the technique are Nesterov’s notion of a “chain-like” function [34], which is difficult for a certain subclass of algorithms, and a “resisting oracle” [32, 34] reduction that turns a lower bound for this subclass into a lower bound for all deterministic algorithms.

In Section 4 we apply this generic method to produce lower bounds for deterministic 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 Section 5. Following Woodworth and Srebro [41], 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 Section 6 we provide lower bounds on finding stationary points of functions where $\|x^{(0)} - x^*\|$ is bounded, rather than the function value gap $f(x^{(0)}) - f(x^*)$; these bounds have the same $\epsilon$ 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 \( p \geq 1 \) and \( p \) times continuously differentiable \( f : \mathbb{R}^d \to \mathbb{R} \), we let \( \nabla^p f(x) \) denote the tensor of \( p \)th order partial derivatives of \( f \) at point \( x \), so \( \nabla^p f(x) \) is an order \( p \) symmetric tensor with entries

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

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

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

where \( \langle \cdot, \cdot \rangle \) is the Euclidean inner product on tensors, defined for order \( k \) tensors \( T \) and \( M \) by

\[
\langle T, M \rangle = \sum_{i_1, \ldots, i_k} T_{i_1, \ldots, i_k} M_{i_1, \ldots, i_k}, \quad \text{and} \quad \otimes \text{ denotes the Kronecker product.}
\]

We let \( \otimes^k \) denote \( d \times \cdots \times d \), \( k \) times, so that \( T \in \mathbb{R}^{\otimes^k} \) 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 \( T \in \mathbb{R}^{\otimes^k} \), the Euclidean operator norm of \( T \) is

\[
\|T\|_{\text{op}} := \sup_{\|v\|=1} \left\{ \langle T, v^{(1)} \otimes \cdots \otimes v^{(k)} \rangle = \sum_{i_1, \ldots, i_k} T_{i_1, \ldots, i_k} v^{(1)}_{i_1} \cdots v^{(k)}_{i_k} \mid \|v^{(i)}\| \leq 1, i = 1, \ldots, k \right\}.
\]

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. [43, Thm. 2.1] show that

\[
\|T\|_{\text{op}} = \sup_{\|v\|=1} \left| \langle T, v^{\otimes k} \rangle \right|, \quad \text{where} \quad v^{\otimes k} = v \otimes v \otimes \cdots \otimes v.
\]

(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 \( g \preceq h \) or \( g = O(h) \) if there exists \( c > 0 \) such that \( g(s) \leq c \cdot h(s) \) for every \( s \in \mathcal{S} \). We write \( g = \tilde{O}(h) \) if \( g \preceq h \log(h+2) \).

## 2 Preliminaries

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

### 2.1 Function classes

Measures of function regularity are crucial for the design and analysis of optimization algorithms [34, 9, 32]. 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 \( p \)th order derivatives if it is \( p \) times continuously differentiable, and for every \( x \in \mathbb{R}^d \)
and direction $v \in \mathbb{R}^d$, $\|v\| \leq 1$, the directional projection $f_{x,v}(t) := f(x + t \cdot v)$ of $f$, defined for $t \in \mathbb{R}$, satisfies
\[
\left| f_{x,v}^{(p)}(t) - f_{x,v}^{(p)}(t') \right| \leq L_p |t - t'|
\]
for all $t, t' \in \mathbb{R}$, where $f_{x,v}^{(p)}(\cdot)$ is the $p$th derivative of $t \mapsto f_{x,v}(t)$. If $f$ is $p + 1$ times continuously differentiable, this is equivalent to requiring
\[
\left| f_{x,v}^{(p+1)}(0) \right| \leq L_p \quad \text{or} \quad \|\nabla^{p+1} f(x)\|_{\text{op}} \leq L_p
\]
for all $x, v \in \mathbb{R}^d$, $\|v\| \leq 1$. We occasionally refer to a function with Lipschitz $p$th order derivatives as $p$th 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 $p \geq 1$, $\Delta > 0$ 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 $C^\infty$ functions $f : \mathbb{R}^d \to \mathbb{R}$ with $L_p$-Lipschitz $p$th derivative and $f(0) - \inf_x f(x) \leq \Delta$.

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” problems [32, 34]. This definition allows clean presentation of our results: 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. That $\mathcal{F}_p(\Delta, L_p)$ contains only $C^\infty$ functions is no real restriction, because our lower bounds become only stronger if we additionally allow less smooth functions.

For our results, we also require the following important invariance notion, proposed (in the context of optimization) by Nemirovski and Yudin [32, 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 procedures we consider. For any dimension $d \in \mathbb{N}$, an algorithm $A$ (also referred to as method or procedure) 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
\[
A[f] = \{x(t)\}_{t=1}^\infty
\]
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 [32] develop (see also [40, 1, 10]), and view every every
iterate $x^{(t)}$ as a query to an \textit{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 assumptions on how the algorithm uses this information (e.g. deterministically). Our approach is

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

as shorthand for the response of a $p$th 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.

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

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

where $A^{(i)}$ is a measurable mapping to $\mathbb{R}^d$ (the dependence on dimension $d$ is implicit). We denote the class of $p$th-order deterministic algorithms by $A^{(p)}_\det$ and let $A^{(\infty)}_\det$ 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 $\text{REG}_{p,L} \in A^{(p)}_\det$ that produces iterates by minimizing the sum of a $p$th order Taylor expansion and an order $p+1$ proximal term:

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

For $p = 1$, $\text{REG}_{p,L}$ is gradient descent with step-size $1/L$, for $p = 2$ it is cubic-regularized Newton’s method [37], and for general $p$ it is a simplified form of the scheme that Birgin et al. [7] propose.

\textbf{Randomized algorithms (and function-informed processes)} A $p$th-order randomized algorithm $A$ is a distribution on $p$th-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 $\xi \sim \text{Uni}[0, 1]$ (independently of $f$), then producing iterates of the form

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

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

\textbf{Zero-respecting sequences and algorithms} While deterministic and randomized algorithm 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 (Sec. 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 $\text{supp} \{v\} := \{ i \in [d] \mid v_i \neq 0 \}$ 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
$T_i \in \mathbb{R}^{k-1,d}$ be the order $(k - 1)$ tensor defined by $[T_i]_{j_1,...,j_{k-1}} = T_{i,j_1,...,j_{k-1}}$. With this notation, we define

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

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

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

The definition (5) says that $x^{(1)}_i = 0$ if all partial derivatives involving the $i$th coordinate of $f$ (up to the $p$th 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 $x^{(1)}_j = 0$. The requirement (5) implies that $x^{(1)} = 0$.

An algorithm $A \in \mathcal{A}_\text{rand}$ is $p$th order zero-respecting if for any $f : \mathbb{R}^d \rightarrow \mathbb{R}$, the (potentially random) iterate sequence $A[f]$ is $p$th 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 [22], BFGS and L-BFGS [26, 38],\footnote{if the initial Hessian approximation is a diagonal matrix, as is typical} Newton’s method (with and without cubic regularization [37]) and trust-region methods [21]. We denote the class of $p$th order zero-respecting algorithms by $\mathcal{A}^{(p)}$, and let $\mathcal{A}^{(\infty)} := \mathcal{A}^{(\infty)}_\text{rand}$.

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 [34, 3]. 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 [35] 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 $\|\nabla f(x)\| \leq \epsilon$. Thus for a deterministic sequence $\{x^{(t)}\}_{t \in \mathbb{N}}$ we define

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

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 distribution $\mathbb{P}$, we define

$$T_\epsilon(\mathbb{P}, f) := \inf \left\{ t \in \mathbb{N} \mid \mathbb{P} \left(\|\nabla f(x^{(s)})\| > \epsilon \text{ for all } s \leq t \right) \leq \frac{1}{2} \right\}, \quad (6)$$

where the randomness is over $x^{(t)}$, according to $\mathbb{P}$. The complexity $T_\epsilon(\mathbb{P}, f)$ is also the median of the random variable $T_\epsilon(\{x^{(t)}\}_{t \in \mathbb{N}}, f)$. By Markov’s inequality, definition (6) lower bounds expectation-based alternatives, as

$$\inf \left\{ t \in \mathbb{N} \mid \mathbb{E}\|\nabla f(x^{(t)})\| \leq \epsilon \right\} \geq T_{2\epsilon}(\mathbb{P}, f) \quad \text{and} \quad \mathbb{E} T_\epsilon(\{x^{(t)}\}_{t \in \mathbb{N}}, f) \geq \frac{1}{2} T_\epsilon(\mathbb{P}, f).$$
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

$$T_\epsilon(A, f) := T_\epsilon(A[f], f)$$

as the complexity of $A$ on $f$. With this setup, we define the complexity of algorithm class $A$ on function class $F$ as

$$T_\epsilon(A, F) := \inf_{A \in A} \sup_{f \in F} T_\epsilon(A, f). \tag{7}$$

Many algorithms guarantee “dimension independent” convergence [34] 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 $\text{REG}_{p, L}$ defined by the recursion (3) guarantees

$$T_\epsilon(A^{(p)}_{\text{det}} \cap A^{(p)}_{2\pi}, \mathcal{F}_p(\Delta, L_p)) \leq \sup_{f \in \mathcal{F}_p(\Delta, L_p)} T_\epsilon(\text{REG}_{p, L_p}, f) \lesssim \Delta L_p^{1/p} e^{-(1+p)/p} \tag{8}$$

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 $\inf_{A \in A} T_\epsilon(A, f)$ 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 $\inf_{A \in A_{\text{rand}}} \int T_\epsilon(A, f) dP(f)$, bounding the distributional complexity [32, 10], 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 [32, 34, 41, 4]. 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. $\epsilon$-stationarity, $\epsilon$-optimality).

We begin by defining zero-chain functions, that generalize a construction due to Nesterov [34]. Zero-chains limit the rate with which zero-respecting algorithms (Sec. 2.2) gather information (Observation 1). This immediately suggests a strategy for proving lower bounds on the complexity of zero-respecting algorithms. By using the orthogonal invariance of our function classes, we construct a resisting oracle [32, Ch. 7.2] (Proposition 1) that reduces any deterministic algorithm to a zero-respecting one. While this approach fails for randomized algorithms, we discuss how to make the basic ideas “robust” so that they apply to any randomized algorithm and local information oracle.

#### 3.1 Zero-chains

Nesterov [34, 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 \( x_t = x_{t+1} = \ldots = x_d = 0 \), 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 \( p \)-th order zero-chain if for every \( x \in \mathbb{R}^d \),

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

We say \( f \) is a zero-chain if it is a \( p \)-th 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 \( \text{supp}\{\nabla^2 f(0)\} = [d] \). Informally, at a point for which \( x_{i-1} = x_i = \ldots = x_d = 0 \), 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.

**Observation 1.** Let \( f : \mathbb{R}^d \to \mathbb{R} \) be a \( p \)-th order zero-chain and let \( x^{(1)} = 0, x^{(2)}, \ldots \) be a \( p \)-th order 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 \( \text{supp}\{x^{(t)}\} \subseteq [t - 1] \) 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 \( \bigcup_{q \in [k]} \text{supp}\{\nabla^q f(x^{(t)})\} \subseteq \{1, \ldots, t\} \) for every \( t \leq k \). Therefore, by the zero-respecting property (5), we have \( \text{supp}\{x^{(k+1)}\} \subseteq \bigcup_{q \in [k]} \bigcup_{t < k+1} \text{supp}\{\nabla^q f(x^{(t)})\} \subseteq [k] \), 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. Therefore, for any function class \( \mathcal{F} \) and \( p, T \in \mathbb{N} \), to show that \( \mathcal{T}_e(\mathcal{A}^{(p)}_{\mathcal{F}}) > T \) it suffices to find \( f_\epsilon : \mathbb{R}^T \to \mathbb{R} \) such that

(i) \( f_\epsilon \) is a \( p \)-th order zero-chain,

(ii) \( f_\epsilon \) belongs to the function class, i.e. \( f_\epsilon \in \mathcal{F} \), and

(iii) \( \|\nabla f_\epsilon(x)\| > \epsilon \) for every \( x \) such that \( x_T = 0 \).

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

\[
\mathcal{T}_e(\mathcal{A}^{(p)}_{\mathcal{F}}) = \inf_{A \in \mathcal{A}^{(p)}_{\mathcal{F}}} \sup_{f \in \mathcal{F}} \mathcal{T}_e(A, f) \geq \sup_{f \in \mathcal{F}} \inf_{A \in \mathcal{A}^{(p)}_{\mathcal{F}}} \mathcal{T}_e(A, f) \geq \inf_{A \in \mathcal{A}^{(p)}_{\mathcal{F}}} \mathcal{T}_e(A, f_\epsilon) > T.
\]

If \( f \) is a zero-chain, then so is the function \( x \mapsto \mu f(x/\sigma) \) for any multiplier \( \mu \) and scale parameter \( \sigma \). This is useful for our development, as we construct zero-chains \( \{g_T\}_{T \in \mathbb{N}} \) such that \( \|\nabla g_T(x)\| > \epsilon \) for every \( x \) such that \( x_T = 0 \).

---

[^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 \( x_T = 0 \).
for every \( x \) with \( x_T = 0 \) and some constant \( c > 0 \). By setting \( f_\epsilon(x) = \mu g_T(x/\sigma) \), then choosing \( T, \mu, \) and \( \sigma \) 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 \( \epsilon \) tends to zero. Thus, the hard functions we construct are fundamentally high-dimensional, making this strategy suitable only for dimension-free lower bounds.

### 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 \( p \geq 1 \) and \( \mathcal{F} \) be an orthogonally invariant function class. Then

\[
\mathcal{T}_\epsilon(\mathcal{A}_\text{det}, \mathcal{F}) \geq \mathcal{T}_\epsilon(\mathcal{A}_\text{det}^{(p)}, \mathcal{F})
\]

and

\[
\mathcal{T}_\epsilon(\mathcal{A}_\text{det}, \mathcal{F}) \geq \mathcal{T}_\epsilon(\mathcal{A}_\text{det}, \mathcal{F})
\]

The proof of Proposition 1, which we detail in Appendix A, builds on the classical notion of a “resisting oracle” [32, 34], which we briefly sketch here. Let \( \mathcal{A} \in \mathcal{A}_\text{det} \), and let \( f \in \mathcal{F}, f : \mathbb{R}^d \rightarrow \mathbb{R} \). We may sequentially construct an orthogonal matrix \( U \in \mathbb{R}^{d \times d} \) (for some finite \( d' > d \)), such that for the function \( f_U(z) := f(U^Tz) \in \mathcal{F} \), the sequence \( U^T \mathcal{A}[f_U] \subset \mathbb{R}^d \) is zero-respecting with respect to \( f \). We do this by choosing the columns of \( U \) to be orthogonal to components in \( \mathcal{A}[f] \) that would otherwise violate the zero-respecting property of \( U^T \mathcal{A}[f] \). Thus, there exists an algorithm \( Z_{\mathcal{A}} \in \mathcal{A}_\text{det} \cap \mathcal{A}_\text{det} \) such that \( Z_{\mathcal{A}}[f] = U^T \mathcal{A}[f_U] \), implying \( \mathcal{T}_\epsilon(\mathcal{A}, f_U) = \mathcal{T}_\epsilon(Z_{\mathcal{A}}, f) \). Therefore,

\[
\inf_{\mathcal{A} \in \mathcal{A}_\text{det}} \sup_{f \in \mathcal{F}} \mathcal{T}_\epsilon(\mathcal{A}, f) \geq \inf_{\mathcal{A} \in \mathcal{A}_\text{det}} \sup_{f \in \mathcal{F}, U} \mathcal{T}_\epsilon(\mathcal{A}, f_U) = \inf_{\mathcal{A} \in \mathcal{A}_\text{det}} \sup_{f \in \mathcal{F}} \mathcal{T}_\epsilon(\mathcal{A}, f_U) \geq \inf_{\mathcal{A} \in \mathcal{A}_\text{det}} \sup_{f \in \mathcal{F}} \mathcal{T}_\epsilon(\mathcal{A}, f),
\]

giving Proposition 1.

The adversarial rotation argument that yields Proposition 1 is more or less apparent in the proofs of previous lower bounds in convex optimization [32, 41, 4] for deterministic algorithms. We believe it is instructive to separate the proof of lower bounds on \( \mathcal{T}_\epsilon(\mathcal{A}_\text{det}, \mathcal{F}) \) and the reduction from \( \mathcal{A}_\text{det} \) to \( \mathcal{A}_\text{det} \), as the latter holds in great generality. Indeed, Proposition 1 holds for any complexity measure \( \mathcal{T}_\epsilon(\cdot, \cdot) \) that satisfies

1. Orthogonal invariance: for every \( f : \mathbb{R}^d \rightarrow \mathbb{R} \), every \( U \in \mathbb{R}^{d \times d} \) such that \( U^T U = I_d \) and every sequence \( \{z(t)\}_{t \in \mathbb{N}} \subset \mathbb{R}^d \), we have

\[
\mathcal{T}_\epsilon(\{z(t)\}_{t \in \mathbb{N}}, f(U^T \cdot)) = \mathcal{T}_\epsilon(\{U^T z(t)\}_{t \in \mathbb{N}}, f).
\]

2. “Stopping time” invariance: for any \( T_0 \in \mathbb{N} \), if \( \mathcal{T}_\epsilon(\{x(t)\}_{t \in \mathbb{N}}, f) \leq T_0 \) then \( \mathcal{T}_\epsilon(\{\hat{x}(t)\}_{t \in \mathbb{N}}, f) = \mathcal{T}_\epsilon(\{\hat{x}(t)\}_{t \in \mathbb{N}}, f) \) for any sequence \( \{\hat{x}(t)\}_{t \in \mathbb{N}} \) such that \( \hat{x}(t) = x(t) \) for \( t \leq T_0 \).

These properties hold for the typical performance measures used in optimization. Examples include time to \( \epsilon \)-optimality, in which case \( \mathcal{T}_\epsilon(\{x(t)\}_{t \in \mathbb{N}}, f) = \inf\{t \in \mathbb{N} | f(x(t)) - f(x(t)) \leq \epsilon\} \), and the second-order stationarity desired in many non-convex optimization problems [13, 23], where for \( \epsilon_1, \epsilon_2 > 0 \) we define \( \mathcal{T}_\epsilon(\{x(t)\}_{t \in \mathbb{N}}, f) = \inf\{t \in \mathbb{N} | \|\nabla f(x(t))\| \leq \epsilon_1 \) and \( \nabla^2 f(x(t)) \geq -\epsilon_2 I\).
3.4 Randomized algorithms

Proposition 1 does not apply to randomized algorithms, as it requires 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 \(x \in \mathbb{R}^d\),

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

A robust zero-chain is also an “ordinary” zero-chain. In Section 5 we replace the adversarial rotation \(U\) of § 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. Adapting a lemma by Woodworth and Srebro [41] to our setting, we show that for every \(A \in A_{\text{rand}}\), \(A[f_U]\) satisfies an approximate form of Observation 1 (w.h.p.) whenever the iterates \(A[f_U]\) 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 \(A_{\text{rand}}\).

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 § 3.2 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

\[
\bar{f}_T(x) = -\Psi(1) \Phi(x) + \sum_{i=2}^{T} \left[ \Psi(-x_{i-1}) \Phi(-x_i) - \Psi(x_{i-1}) \Phi(x_i) \right],
\]

where the component functions are

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

We illustrate the construction in Figure 1. The choices of \(\Psi\) and \(\Phi\) are not particularly special; any pair of increasing, bounded, \(C^\infty\) functions for which \(\Psi(x) = 0\) for \(x\) below some constant yield similar results.

Our construction has two key properties. First is that \(f\) is a zero-chain (Observation 2 in the sequel). Second, as we show in Lemma 2, \(\|\nabla \bar{f}_T(x)\|\) is large unless \(|x_i| \geq 1\) 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 \(\Psi\), \(\Phi\) and \(\bar{f}_T\).

**Lemma 1.** The functions \(\Psi\) and \(\Phi\) satisfy the following.

i. For all \(x \leq \frac{1}{2}\) and all \(k \in \mathbb{N}\), \(\Psi^{(k)}(x) = 0\).
ii. For all $x \geq 1$ and $|y| < 1$, $\Psi(x)\Phi'(y) > 1$.

iii. Both $\Psi$ and $\Phi$ are infinitely differentiable, and for all $k \in \mathbb{N}$ we have
\[
\sup_x |\Psi^{(k)}(x)| \leq \exp \left( \frac{5k}{2} \log(4k) \right) \quad \text{and} \quad \sup_x |\Phi^{(k)}(x)| \leq \exp \left( \frac{3k}{2} \log \frac{3k}{2} \right).
\]

iv. The functions and derivatives $\Psi, \Psi', \Phi$ and $\Phi'$ are non-negative and bounded, with
\[
0 \leq \Psi < e, \quad 0 \leq \Psi' \leq \sqrt{\frac{54}{e}}, \quad 0 < \Phi < \sqrt{2\pi e}, \quad \text{and} \quad 0 < \Phi' \leq \sqrt{e}.
\]

We provide a proof of Lemma 1 in Appendix B.1; it is essentially mechanical.

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 2.** 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 2 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 then 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 $i \leq T$, then there exists $j \leq i$ such that $|x_j| < 1$ and
\[
\|\nabla \bar{f}_T(x)\| \geq \left| \frac{\partial}{\partial x_j} \bar{f}_T(x) \right| > 1.
\]

**Proof.** We take $j \leq i$ to be the smallest $j$ for which $|x_j| < 1$, so that $|x_{j-1}| \geq 1$ (where we use the shorthand $x_0 \equiv 1$). Therefore, we have
\[
\frac{\partial \bar{f}_T}{\partial x_j}(x) = -\Psi(-x_{j-1}) \Phi'(-x_j) - \Psi(x_{j-1}) \Phi'(x_j) - \Psi'(-x_j) \Phi(-x_{j+1}) - \Psi'(x_j) \Phi(x_{j+1}) \quad (11)
\]
\[
\overset{(i)}{\leq} -\Psi(-x_{j-1}) \Phi'(-x_j) - \Psi(x_{j-1}) \Phi'(x_j) \overset{(ii)}{=} -\Psi(|x_{j-1}|)\Phi'(x_j \text{ sign}(x_{j-1})) \overset{(iii)}{<} -1.
\]
In the chain of inequalities, inequality (i) follows because $\Psi(x)\Phi(y) \geq 0$ for every $x, y$; inequality (ii) follows because $\Psi(x) = 0$ for $x \leq 1/2$, while equality (iii) follows from Lemma 1.ii and the pairing of $|x_j| < 1$ and $|x_{j-1}| \geq 1$.

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

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

i. We have $\tilde{f}_T(0) - \inf_x \tilde{f}_T(x) \leq 12T$.

ii. For all $x \in \mathbb{R}^d$, $\|\nabla \tilde{f}_T(x)\| \leq 23\sqrt{T}$.

iii. For every $p \geq 1$, the $p$-th order derivatives of $\tilde{f}_T$ are $\ell_p$-Lipschitz continuous, where $\ell_p \leq \exp(\frac{5}{2}p\log p + cp)$ 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 $\tilde{f}_T$ are in $\mathcal{F}_p(\Delta, L_p)$. Property ii is necessary for analysis of the randomized construction in Section 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 $p$th order smooth functions using full derivative information and zero-respecting algorithms (the class $\mathcal{A}_{Zr}$). Proposition 1 transforms this bound into one on all deterministic algorithms (the class $\mathcal{A}_{det}$).

**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 $\Delta$, $L_p$, and $\epsilon$ be positive. Then

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

where $\ell_p \leq e^{\frac{5}{2}p\log p + c_1 p}$.

Before we prove the theorem, a few remarks are in order. First, our lower bound matches the upper bound (8) that $p$th-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 $\Delta, L_p$, and $\epsilon$.

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

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

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

Finally, the scaling of $\ell_p$ 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 $C^\infty$ 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, $\ell_p$ is the Lipschitz constant of the $p$th derivative of $\bar{f}_T$. Second, it seems that attaining optimal rate of convergence for a given $p$ requires evaluation of the full $p$th order derivative; in the companion [15] we prove that the rate $\epsilon^{-3/2} = \epsilon^{-(p+1)/p}$, for $p = 2$, is unachievable with first-order methods, and we believe that similarly $\epsilon^{-(p+1)/p}$ is unachievable with only the first $p-1$ derivatives. The cases of practical interest are thus $p \in \{1, 2\}$, in which case $\ell_p^{1/p} \lesssim p^2$ is a numerical constant.

4.3 Proof of Theorem 1

To prove Theorem 1, we set up the hard instance $f : \mathbb{R}^T \rightarrow \mathbb{R}$ for some integer $T$ by appropriately scaling $\bar{f}_T$ 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. Fix $A \in \mathcal{A}_x$ and let $x^{(1)} = 0, x^{(2)}, \ldots, x^{(T)}$ be the iterates produced by $A$ applied on $f$. Since $f$ is also a zero-chain, by Observation 1 we have $x^{(t)} = 0$ for all $t \leq T$. Applying Lemma 2 guarantees that $\left\| \nabla \bar{f}_T(x^{(t)}/\sigma) \right\| > 1$, and therefore

$$\left\| \nabla f(x^{(t)}) \right\| = \frac{L_p \sigma^p}{\ell_p} \left\| \nabla \bar{f}_T(x^{(t)}/\sigma) \right\| > \frac{L_p \sigma^p}{\ell_p}. \quad (12)$$

It remains to choose $T$ and $\sigma$ based on $\epsilon$ such that $\left\| \nabla f(x^{(t)}) \right\| > \epsilon$ and $f \in F_p(\Delta, L_p)$. By the lower bound (12), the choice $\sigma = (\ell_p \epsilon / L_p)^{1/p}$ guarantees $\left\| \nabla f(x^{(t)}) \right\| > \epsilon$ and hence $T_\epsilon(f, A) \geq T + 1$. 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 $f \in F_p(\Delta, L_p)$ it suffices to show that $f(0) - \inf_x f(x) \leq \Delta$. By the first part of Lemma 3 we have

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

where in the last transition we substituted $\sigma = (\ell_p \epsilon / L_p)^{1/p}$. We conclude that $f \in F_p(\Delta, L_p)$ for

$$T = \left[ \frac{\Delta L_p^{1/p}}{12 \ell_p^{1/p} \epsilon^{1+p}} \right] \quad \text{so} \quad T_\epsilon(A_x, F_p(\Delta, L_p)) \geq \inf_{A \in A_x} T_\epsilon(A, f) \geq 1 + T \geq \frac{\Delta L_p^{1/p}}{12 \ell_p^{1/p} \epsilon^{1+p}},$$

with $\ell_p$ bounded from above as in Lemma 3.iii.

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 Section 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 $\tilde{f}_T(U^\top \cdot)$ must (with high probability) satisfy that $|\langle x(t), u(j) \rangle| \leq \frac{1}{2}$ for all $t$ and $j \geq t$, so that by Lemma 2, $\|\nabla \tilde{f}_T(U^\top x(t))\|$ must be large (step (b)). Second, in Section 5.2 we further augment our construction to force boundedness of the iterates by composing $\tilde{f}_T(U^\top \cdot)$ with a soft projection, so that an algorithm cannot “cheat” with unbounded iterates. Finally, we present our general lower bounds in Section 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

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

(13)

We assume throughout that $U$ is chosen uniformly at random from the space of orthogonal matrices $O(d, T) = \{V \in \mathbb{R}^{d \times T} \mid V^\top V = I_T\}$; 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 $\tilde{f}_T$ 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 $A[f]$ for some randomized algorithm $f$ (with iteration (4)).

**Lemma 4.** Let $\delta > 0$ and $R \geq \sqrt{T}$, and let $x(1), \ldots, x(T)$ be informed by $\tilde{f}_{T;U}$ and bounded, so that $\|x(t)\| \leq R$ for each $T$. If $d \geq 52TR^2 \log \frac{2T^2}{\delta}$ then with probability at least $1 - \delta$, for all $t \leq T$ 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 [41, 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 Section 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 $s \leq t$ and $j \in \{s, \ldots, T\}$; we argue that this (roughly) implies that $|\langle u(j), x(t+1) \rangle| < 1/2$ for every $j \in \{t+1, \ldots, T\}$ 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 $f_T$ (Definition 4 and Observation 2) implies that for every $s \leq t$ we have

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

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 $u(1), \ldots, u(s)$, as $\tilde{f}_T$ is known. Therefore, given $u(1), x(1), \ldots, u(t), x(t)$ it is possible to reconstruct all the information the algorithm has collected up to iteration $t$. This means that beyond possibly revealing $u(1), \ldots, u(t)$, these derivatives contain no additional information on

\(^4\) In a recent note Woodworth and Srebro [42] independently provide a revision of their proof that is similar, but not identical, to the one we propose here.
u^{(t+1)}, ..., u^{(T)}. Consequently, any component of x^{(t+1)} outside the span of u^{(1)}, x^{(1)}, ..., 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 span\{u^{(1)}, x^{(1)}, ..., u^{(t)}, x^{(t)}\}. We show that conditioned on u^{(1)}, ..., 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 |⟨\( \hat{u}^{(j)}, x^{(t+1)} \rangle⟩| \( \lesssim \|x^{(t+1)}\|/\sqrt{d} \leq R/\sqrt{d} \) for any individual \( j \geq t + 1 \), with high probability.

Using an appropriate induction hypothesis, this is sufficient to guarantee that for every \( t + 1 \leq j \leq T \), |⟨u^{(j)}, x^{(t+1)}⟩| \( \lesssim \sqrt{T/\log T}/d \), 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 (13) thus fails for unbounded random iterates, since as long as \( U \) is informed by \( \tilde{u}^{(j)} \), set \( \hat{u}^{(j)} = \tilde{u}^{(j)} \), orthogonal complement of span \{\( \tilde{u}^{(j)} \)\}.

Set \( \rho(x) = \frac{x}{\sqrt{1 + \|x\|^2/R^2}} \) and \( R = 230\sqrt{T} \).

The quadratic term in \( \hat{f}_{T,U}(x) \) 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 \( x^{(1)}, ..., x^{(T)} \) 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 \( 1 - \delta \),

\[
\|\nabla \hat{f}_{T,U}(x^{(t)})\| > 1/2 \quad \text{for all } t \leq T.
\]

**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 \( \rho \) is fixed). Therefore, the process \( y^{(1)}, ..., y^{(T)} \) 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 \( 1 - \delta \),

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

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

\[
|\langle u^{(j)}, y^{(t)} \rangle| < 1 \quad \text{and} \quad |\langle u^{(j)}, \nabla \tilde{f}_{T,U}(y^{(t)}) \rangle| > 1.
\]

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 \( \frac{\partial \hat{f}}{\partial x}(x) = \frac{1 - \rho(x)\rho(x)^\top / R^2}{\sqrt{1 + \|x\|^2 / R^2}} \) to write
\[
\left\langle u^{(j)}, \nabla \hat{f}_{T;U}(x^{(t)}) \right\rangle = \left\langle u^{(j)}, \frac{\partial \hat{f}}{\partial x}(x^{(t)}) \nabla \hat{f}_{T;U}(y^{(t)}) \right\rangle + \frac{1}{5} \left\langle u^{(j)}, x^{(t)} \right\rangle
\]
\[
= \frac{(u^{(j)}, \nabla \hat{f}_{T;U}(y^{(t)})) - (u^{(j)}, y^{(t)})(\nabla \hat{f}_{T;U}(y^{(t)}))}{\sqrt{1 + \|x^{(t)}\|^2 / R^2}} + \frac{1}{5} \left\langle u^{(j)}, y^{(t)} \right\rangle \sqrt{1 + \|x^{(t)}\|^2 / R^2}.
\]
Therefore, for \( \|y^{(t)}\| \leq \|x^{(t)}\| \leq R/2 \) we have
\[
\left| \left\langle u^{(j)}, \nabla \hat{f}_{T;U}(x^{(t)}) \right\rangle \right| \geq \frac{2}{\sqrt{5}} \left| \left\langle u^{(j)}, \nabla \hat{f}_{T;U}(y^{(t)}) \right\rangle \right| - \left| \left\langle u^{(j)}, y^{(t)} \right\rangle \right| \left( \frac{\|\nabla \hat{f}_{T;U}(y^{(t)})\|}{2R} + \frac{1}{2\sqrt{5}} \right).
\]
By Lemma 3.ii we have \( \|\nabla \hat{f}_{T;U}(y^{(t)})\| \leq 23\sqrt{T} = R/10 \), which combined with (15) and the above display yields \( \|\nabla \hat{f}_{T;U}(x^{(T)})\| \geq \|u^{(j)}, \nabla \hat{f}_{T;U}(x^{(T)})\| \geq \frac{2}{\sqrt{5}} - \frac{1}{20} - \frac{1}{2\sqrt{5}} > \frac{1}{2} \).

In the second case, \( \|x^{(t)}\| \geq R/2 \), we have for any \( x \) satisfying \( \|x\| \geq R/2 \) and \( y = \rho(x) \) that
\[
\left\| \nabla \hat{f}_{T;U}(x) \right\| \geq \frac{1}{5} \|x\| - \left\| \frac{\partial \hat{f}}{\partial x}(x) \right\|_{\text{op}} \left\| \nabla \hat{f}_{T;U}(y) \right\| \geq \frac{R}{10} - \frac{2}{\sqrt{5}} \frac{R}{10} > \sqrt{T} \geq 1,
\]
where we used \( \left\| \frac{\partial \hat{f}}{\partial x}(x) \right\|_{\text{op}} \leq \frac{1}{\sqrt{1 + \|x\|^2 / R^2}} \leq 2/\sqrt{5} \) and that \( \|\nabla \hat{f}_{T;U}(y)\| \leq 23\sqrt{T} = R/10 \).

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 \( p \)th order derivatives of \( \hat{f}_{T;U} \) are \( \hat{L}_p \)-Lipschitz continuous, where \( \hat{L}_p \leq \exp(c p \log(p + c)) \) for a numerical constant \( c < \infty \).

We defer the (computationally involved) proof of this lemma to Section B.5.

### 5.3 Final lower bounds

With Lemmas 5 and 6 in hand, we can state our a lower bound for all algorithms, randomized or otherwise, given access to all derivatives of a \( C^\infty \) function. Note that our construction also implies an identical lower bound for (slightly) more general algorithms that use any local oracle \([32, 10]\), 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 \( \Delta, L_p, \) and \( \epsilon \) be positive. Then
\[
\mathcal{F}_\epsilon(A_{\text{rand}};, F_p(\Delta, L_p)) \geq c_0 \cdot \Delta \left( \frac{\hat{L}_p}{\hat{L}_p} \right)^{1/p} \epsilon^{- \frac{1+c}{p}},
\]
where \( \hat{L}_p \leq e^{c_1 p \log p + c_1} \).
We return to the proof of Theorem 2 in Sec. 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 $d \geq \lceil 52 \cdot (230)^2 \cdot T^2 \log(2T^2/\delta) \rceil$ where $T = \lfloor c_0 \Delta(L_p/\ell_p)^{1/p} \epsilon^{-1 + \frac{1}{p}} \rfloor$ as in the theorem statement. In the proof we construct a probability measure $\mu$ on functions in $\mathcal{F}_p(\Delta, L_p)$, of fixed dimension $d$, such that
\[
\inf_{A \in \mathcal{A}_{\text{rand}}} \int \mathbb{P}_A \left( \| \nabla f(x^{(t)}) \| > \epsilon \text{ for all } t \leq T \mid f \right) d\mu(f) > 1 - \delta,
\]
where the randomness in $\mathbb{P}_A$ depends only on $A$. Therefore, by definition (6), for any $A \in \mathcal{A}_{\text{rand}}$ a function $f$ drawn from $\mu$ satisfies
\[
T_\epsilon(A, f) > T \text{ with probability greater than } 1 - 2\delta,
\]
implying Theorem 2 for any $\delta \geq 1/2$. 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, as it applies for a broader class of algorithms. However, our probabilistic analysis requires that the functions constructed to prove Theorem 2 be of dimension at least $T^2$. In contrast to Theorem 1, which requires dimension $2T$, this is a drawback; we do not speculate whether such an increase in dimension is necessary.

## 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 $U \in \mathcal{O}(d, T)$, by appropriately scaling $\hat{f}_{T,U}$ defined in (14),
\[
f_U(x) := \frac{L_p\sigma^{p+1}}{\ell_p} \hat{f}_{T,U}(x/\sigma),
\]
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 $\ell_p \leq \exp(c_1p \log p + c_1)$ is defined in Lemma 6.ii.

Fix $A \in \mathcal{A}_{\text{rand}}$ and let $x^{(1)}(\hat{A}), x^{(2)}(\hat{A}), \ldots, x^{(T)}(\hat{A})$ 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 Sec. 2.2), and therefore we may apply Lemma 5 with $\delta = 1/2$ and our large enough choice of dimension $d$ to conclude that
\[
\mathbb{P}_{A,U} \left( \| \nabla \hat{f}_{T,U} \left( x^{(t)}/\sigma \right) \| > \frac{1}{2} \text{ for all } t \leq T \right) > \frac{1}{2},
\]
where the probability is taken over both the random orthogonal $U$ and any randomness in $A$. As $A$ is arbitrary, taking $\sigma = (2\ell_p^2/L_p)^{1/p}$, this inequality becomes the desired strong inequality (17) with $\delta = 1/2$ and $\mu$ induced by the distribution of $U$. Thus, by (18), for every $A \in \mathcal{A}_{\text{rand}}$ there exists $U_A \in \mathcal{O}(d, T)$ such that $T_\epsilon(A, f_{U_A}) \geq 1 + T$, so
\[
\inf_{A \in \mathcal{A}_{\text{det}}} \sup_{U \in \mathcal{O}(d, T)} T_\epsilon(A, f_U) \geq 1 + T.
\]

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 $p$th order derivatives. By Lemma 6.i, we have
\[
f_U(0) - \inf_x f_U(x) \leq \frac{L_p\sigma^{p+1}}{\ell_p} \left( \hat{f}_T(0) - \inf_x \hat{f}_T(x) \right) \leq \frac{12L_p\sigma^{p+1}}{\ell_p} T = \frac{24(2\ell_p)^{1/p} \epsilon^{\frac{p+1}{p}}}{L_p^{1/p}} T,
\]
...
where in the last transition we have substituted \( \sigma = (2 \ell_p \epsilon / L_p)^{1/p} \). Setting \( T = \left[ \frac{\Delta}{\log(L_p/\hat{L}_p)} \right]^{1/p} \) 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

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

be the class of \( C^\infty \) functions with \( L_p \)-Lipschitz \( p \)th order derivatives satisfying

\[
\sup_x \{ \| x \| \mid x \in \arg\min f \} \leq D,
\]

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 \( \epsilon \) 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 \( \epsilon \) dependence than their value-bounded counterparts (see Section PII.3). 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 [32].

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 \( \epsilon \) be positive. Then

\[
T_\epsilon(A_{\text{rand}}, \mathcal{F}^{\text{dist}}_p(D, L_p)) \geq c_0 \cdot D^{1+p} \left( \frac{L_p}{\ell_p^p} \right)^{1+p} \epsilon^{-\frac{1+p}{p}},
\]

where \( \ell_p^p \leq e c_1 p \log p + c_1 \).

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

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

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

\[
T_\epsilon(A^{(p)}_\text{det} \cap A^{(p)}_{zr}, \mathcal{F}^{\text{dist}}_{1,p}(D, L_1, L_p)) \lesssim D^2 L_1^{1/p} \epsilon^{-\frac{p+1}{p}}.
\]

Acknowledgments

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

[1] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[2] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the Forty-Ninth Annual ACM Symposium on the Theory of Computing*, 2017. URL https://arxiv.org/abs/1611.01146.

[3] Y. Arjevani, S. Shalev-Shwartz, and O. Shamir. On lower and upper bounds in smooth and strongly convex optimization. *Journal of Machine Learning Research*, 17(126):1–51, 2016. URL http://jmlr.org/papers/v17/15-106.html.

[4] Y. Arjevani, O. Shamir, and R. Shiff. Oracle complexity of second-order methods for smooth convex optimization. arXiv:1705.07260 [math.OC], 2017.

[5] K. Ball. An elementary introduction to modern convex geometry. In S. Levy, editor, *Flavors of Geometry*, pages 1–58. MSRI Publications, 1997.

[6] D. Berend and T. Tassa. Improved bounds on Bell numbers and on moments of sums of random variables. *Probability and Mathematical Statistics*, 30(2):185–205, 2010.

[7] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and P. L. Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, 163(1–2):359–368, 2017.

[8] N. Boumal, V. Voroninski, and A. Bandeira. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems 29*, pages 2757–2765, 2016.

[9] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[10] G. Braun, C. Guzmán, and S. Pokutta. Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory. *IEEE Transactions on Information Theory*, 63(7), 2017.

[11] S. Burer and R. D. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[12] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[13] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Accelerated methods for non-convex optimization. arXiv:1611.00756 [math.OC], 2016.

[14] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Convex until proven guilty: dimension-free acceleration of gradient descent on non-convex functions. In *Proceedings of the 33rd International Conference on Machine Learning*, 2017.

[15] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points II: First-order methods. arXiv: [math.OC], 2017.
[16] C. Cartis, N. I. Gould, and P. L. Toint. On the complexity of steepest descent, Newton’s and regularized Newton’s methods for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010.

[17] C. Cartis, N. I. Gould, and P. L. Toint. Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, 28(1):93–108, 2012.

[18] C. Cartis, N. I. M. Gould, and P. L. Toint. How much patience do you have? A worst-case perspective on smooth nonconvex optimization. *Optima*, 88, 2012.

[19] C. Cartis, N. I. M. Gould, and P. L. Toint. Worst-case evaluation complexity and optimality of second-order methods for nonconvex smooth optimization. *arXiv:1709.07180 [math.OC]*, 2017.

[20] S. Chowla, I. N. Herstein, and W. K. Moore. On recursions connected with symmetric groups I. *Canadian Journal of Mathematics*, 3:328–334, 1951.

[21] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust Region Methods*. MPS-SIAM Series on Optimization. SIAM, 2000.

[22] W. W. Hager and H. Zhang. A survey of nonlinear conjugate gradient methods. *Pacific Journal of Optimization*, 2(1):35–58, 2006.

[23] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 33rd International Conference on Machine Learning*, 2017.

[24] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 2010.

[25] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[26] D. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

[27] P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *Annals of Statistics*, 40(3):1637–1664, 2012.

[28] P.-L. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616, 2013.

[29] R. D. Monteiro and B. F. Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

[30] K. Murty and S. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39:117–129, 1987.

[31] A. Nemirovski. Efficient methods in convex programming. Technion: The Israel Institute of Technology, 1994.

[32] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.
[33] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady, 27(2):372–376, 1983.

[34] Y. Nesterov. Introductory Lectures on Convex Optimization. Kluwer Academic Publishers, 2004.

[35] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization, 22(2):341–362, 2012.

[36] Y. Nesterov. How to make the gradients small. Optima, 88, 2012.

[37] Y. Nesterov and B. Polyak. Cubic regularization of Newton method and its global performance. Mathematical Programming, Series A, 108:177–205, 2006.

[38] J. Nocedal and S. J. Wright. Numerical Optimization. Springer, 2006.

[39] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. Foundations of Computational Mathematics, To appear, 2017.

[40] J. Traub, H. Wasilkowski, and H. Wozniakowski. Information-Based Complexity. Academic Press, 1988.

[41] B. E. Woodworth and N. Srebro. Tight complexity bounds for optimizing composite objectives. In Advances in Neural Information Processing Systems 29, pages 3639–3647, 2016.

[42] B. E. Woodworth and N. Srebro. Lower bound for randomized first order convex optimization. arXiv:1709.03594 [math.OC], 2017. URL https://arxiv.org/pdf/1709.03594.

[43] X. Zhang, C. Ling, and L. Qi. The best rank-1 approximation of a symmetric tensor and related spherical optimization problems. SIAM Journal on Matrix Analysis and Applications, 33(3):806–821, 2012.
A Proof of Proposition 1

**Proposition 1.** Let $p \geq 1$ and $\mathcal{F}$ be an orthogonally invariant function class. Then

$$\mathcal{T}_\epsilon(A^{(p)}_{\text{det}}, \mathcal{F}) \geq \mathcal{T}_\epsilon(A^{(p)}_{\text{det}}, \mathcal{F})$$

and

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

**Proof.** We may assume that $\mathcal{T}_\epsilon(A^{(p)}_{\text{det}}, \mathcal{F}) < T_0$ for some integer $T_0 < \infty$, as otherwise we have $\mathcal{T}_\epsilon(A^{(p)}_{\text{det}}, \mathcal{F}) = \infty$ and the result holds trivially. It is therefore sufficient to consider only algorithms with worst-case complexity better than $T_0$, i.e. $A \in A^{(p)}_{\text{det}}$ such that $\sup_{f \in \mathcal{F}} \mathcal{T}_\epsilon(A, f) < T_0 < \infty$.

Fixing such $A$, we construct an algorithm $Z_A \in A^{(p)}_{\text{det}}$ with the following property: for every every $f : \mathbb{R}^d \to \mathbb{R}$ in $\mathcal{F}$, there exists an orthogonal $U \in \mathbb{R}^{(d+T_0)\times d}$, $U^T U = I_d$, such that $f_U(x) := f(U^T x)$ satisfies that the first $T_0$ iterates in sequences $Z_A[f]$ and $U^T A[f_U]$ are identical. (Recall the notation $A[f] = \{a(t)\}_{t \in \mathbb{N}}$ where $a(t)$ are the iterates of $A$ on $f$, and we use the obvious shorthand $U^T A \{a(t)\}_{t \in \mathbb{N}} = \{U^T a(t)\}_{t \in \mathbb{N}}$.) Before explaining the construction of the zero-respecting algorithm $Z_A$, let us see how its defining property implies Proposition 1.

First, note that if $Z_A[f]$ and $U^T A[f_U]$ are identical over their first $T_0$ iterates,

$$\mathcal{T}_\epsilon(A, f_U) := \mathcal{T}_\epsilon(A[f_U], f_U) = \mathcal{T}_\epsilon(U^T A[f_U], f) = \mathcal{T}_\epsilon(Z_A, f).$$

(20)

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

$$\mathcal{T}_\epsilon(\{a(t)\}_{t \in \mathbb{N}}, f_U) = \inf \{t \in \mathbb{N} \mid \|\nabla f_U(a(t))\| \leq \epsilon\} = \inf \{t \in \mathbb{N} \mid \|\nabla f(U^T a(t))\| \leq \epsilon\} = \mathcal{T}_\epsilon(\{U^T a(t)\}_{t \in \mathbb{N}}, f)$$

and in equality (i) we let $\{a(t)\}_{t \in \mathbb{N}} = A[f_U]$. The equality (ii) holds because $\mathcal{T}_\epsilon(\cdot, \cdot)$ is a “stopping time”: if $\mathcal{T}_\epsilon(U^T A[f_U], f) \leq T_0$ then the first $T_0$ iterates of $U^T A[f_U]$ determine $\mathcal{T}_\epsilon(U^T A[f_U], f)$, and these $T_0$ iterates are identical to the first $T_0$ iterates of $Z_A[f]$ by assumption. With the identity (20) in hand, we obtain our result by straightforward manipulation as

$$\mathcal{T}_\epsilon(A^{(p)}_{\text{det}}, \mathcal{F}) = \inf_{A \in A^{(p)}_{\text{det}}} \sup_{f \in \mathcal{F}} \mathcal{T}_\epsilon(A, f) \geq \inf_{A \in A^{(p)}_{\text{det}}} \sup_{f \in \mathcal{F}} \mathcal{T}_\epsilon(A, f_U) \geq \inf_{B \in A^{(p)}_{\text{det}}} \sup_{f \in \mathcal{F}} \mathcal{T}_\epsilon(B, f)$$

where inequality (i) uses that $f_U \in \mathcal{F}$ because $\mathcal{F}$ is orthogonally invariant, step (ii) uses equality (20) and step (iii) is due to $Z_A \in A^{(p)}_{\text{det}}$ by construction.

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 $\{z(t)\}_{t \in \mathbb{N}} = Z_A[f]$. 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 $\{u(t)\}_{t \in S_t} \subset \mathbb{R}^{d'}$ identified with this support. The support condition (5) defines the set $S_t$,

$$S_t = \bigcup_{q \in [p]} \bigcup_{s < t} \text{supp} \left\{ \nabla^q f(z(s)) \right\},$$
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 $i$th column is $u^{(i)}$—even though $U$ may not be completely determined throughout the runtime of $Z_A$, our partial knowledge of it will suffice to simulate the operation of $A$ on $f_U(a) = f(U^T a)$. Letting $\{a^{(i)}\}_{i \in \mathbb{N}} = A[f_U]$, our requirements $Z_A[f] = U^T A[f_U]$ and $Z_A \in A_{2r}$ are equivalent to

$$z^{(t)} = U^T a^{(t)} \quad \text{and} \quad \text{supp}\{z^{(t)}\} \subseteq S_t$$

for every $t \leq T_0$ (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 (21) at $t = 1$ by requiring that $(u^{(j)}, a^{(1)}) = 0$ for every $j \in [d]$, whence the first iterate of $Z_A$ satisfies $z^{(1)} = 0 \in \mathbb{R}^d$. Assume now the equality and containment (21) holds for every $s < t$, where $t \leq T_0$ (implying that $Z_A$ has emulated the iterates $a^{(2)}, \ldots, a^{(t-1)}$ of $A$); we show how $Z_A$ can emulate $a^{(t)}$, the $t$th iterate of $A$, and from it can construct $z^{(t)}$ that satisfies (21). To obtain $a^{(t)}$, note that for every $q \leq p$, 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 $\text{supp}\{\nabla^q f(z^{(s)})\} \subseteq S_{s+1}$ 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 $A \in A_{\det}^{(p)}$ is deterministic, $a^{(t)}$ is a function of $\nabla^q f(z^{(s)})$ for $q \in [p]$ and $s \in [t-1]$, and thus $Z_A$ can simulate and compute it. To satisfy the support condition $\text{supp}\{z^{(t)}\} \subseteq S_t$ we require that $(u^{(j)}, a^{(t)}) = 0$ for every $j \notin S_t$. This also means that to compute $z^{(t)} = U^T a^{(t)}$ 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 $(u^{(j)}, a^{(s)}) = 0$ for every $s \leq t$ 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 $\{a^{(1)}, \ldots, a^{(t)}, \{u^{(i)}\}_{i \in S_t}\}$. This orthogonal complement has dimension at least $d' - t - |S_t| = |S_{t+1}^c| + T_0 - t \geq |S_{t+1}^c|$. Since $|S_{t+1} \setminus S_t| \leq |S_{t+1}^c|$, there exist orthonormal vectors $\{u^{(i)}\}_{i \in S_{t+1} \setminus S_t}$ that meet the requirements. This completes the induction.

Note that the arguments above hold unchanged for $p = \infty$, so we have the second result $T_c(A_{\det}, F) = T_c(A_{\det}^{(\infty)}, F) \geq T_c(A_{2r}^{(\infty)}, F) = T_c(A_{2r}, F)$. □

B Technical Results

B.1 Proof of Lemma 1

Lemma 1. The functions $\Psi$ and $\Phi$ satisfy the following.

i. For all $x \leq \frac{1}{2}$ and all $k \in \mathbb{N}$, $\Psi^{(k)}(x) = 0$.

ii. For all $x \geq 1$ and $|y| < 1$, $\Psi(x)\Phi'(y) > 1$.

iii. Both $\Psi$ and $\Phi$ are infinitely differentiable, and for all $k \in \mathbb{N}$ we have

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

iv. The functions and derivatives $\Psi, \Psi', \Phi$ and $\Phi'$ are non-negative and bounded, with

$$0 \leq \Psi < e, \quad 0 \leq \Psi' \leq \sqrt{54/e}, \quad 0 < \Phi < \sqrt{2\pi e}, \quad \text{and} \quad 0 < \Phi' \leq \sqrt{e}.$$
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 $\phi(t) = e^{-\frac{1}{2} t^2}$. We have the following lemma, whose proof we defer temporarily.

**Lemma 7.** For all $k \in \mathbb{N}$, there exist constants $c_i^{(k)}$ satisfying $|c_i^{(k)}| \leq (2 \max\{i, 1\})^k$, and

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

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 $x^i \phi(x) \leq \exp(\frac{1}{2} \log \frac{1}{e})$. We thus obtain the numerically verifiable upper bound

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

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

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

As in the derivation immediately following Lemma 7, 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

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

which yields the numerically verifiable upper bound

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

**Proof of Lemma 7.** 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

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


Lemma 3. The function $f_T$ satisfies the following.

i. We have $f_T(0) - \inf_x f_T(x) \leq 12T$.

ii. For all $x \in \mathbb{R}^d$, $\|\nabla f_T(x)\| \leq 23\sqrt{T}$.

iii. For every $p \geq 1$, the $p$-th order derivatives of $f_T$ are $\ell_p$-Lipschitz continuous, where $\ell_p \leq \exp\left(\frac{5}{2} p \log p + cp \right)$ for a numerical constant $c < \infty$. 

\[ \phi^{(k+1)}(t) = \sum_{i=0}^{k+1} c_i^{(k+1)} t^i \phi(t) \]

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

\[ |c_i^{(k+1)}| \leq 2^k(i + 1)(i + 1)^k + 2^k(\max\{i, 1\})^k \leq 2^{k+1}(i + 1)^{k+1} \]

This gives the result.

Proof of Lemma 8. 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

\[ \Psi^{(k+1)}(x) = \left(\sum_{i=1}^{k} \frac{2(k + 2i)c_i^{(k)}}{(2x - 1)^{k+2i+2}} + \sum_{i=1}^{k} \frac{4c_i^{(k)}}{(2x - 1)^{k+3+2i}}\right) \Psi(x) \]

\[ = \left(\sum_{i=1}^{k+1} \frac{4c_{i-1}^{(k)} - 2(k + 2i)c_i^{(k)}}{(2x - 1)^{k+1+2i}}\right) \Psi(x), \]

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

\[ |c_i^{(k+1)}| \leq 4 \cdot 6^k(k - 2 + 2i)^k + 2 \cdot 6^k(k + 2i)(k + 2i)^k \leq 6^{k+1}(k + 2i)^{k+1} \leq 6^{k+1}(k + 1 + 2i)^{k+1} \]

which gives the result. 

\[ \square \]
Proof. Part i follows because \( \tilde{f}_T(0) < 0 \) and, since \( 0 \leq \Psi(x) \leq e \) and \( 0 \leq \Phi(x) \leq \sqrt{2\pi e} \),
\[
\tilde{f}_T(x) = -\Psi(x) - \sum_{i=2}^T \Psi(x_{i-1}) \Phi(x_i) > -T \cdot e \cdot \sqrt{2\pi e} \geq -12T.
\]

Part ii follows additionally from \( \Psi(x) = 0 \) on \( x < 1/2, 0 \leq \Psi'(x) \leq \sqrt{54e^{-1}} \) and \( 0 \leq \Phi'(x) \leq \sqrt{e} \), which when substituted into equality (11) yields
\[
\left| \frac{\partial \tilde{f}_T}{\partial x_j}(x) \right| \leq e \cdot \sqrt{e} + \sqrt{54e^{-1}} \cdot \sqrt{2\pi e} \leq 23
\]
for every \( x \) and \( j \). Consequently, \( \| \nabla \tilde{f}_T(x) \| \leq \sqrt{T} \leq 23\sqrt{T} \).

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 \( h_{x,v} : \mathbb{R} \rightarrow \mathbb{R} \) by the directional projection of \( \tilde{f}_T \), \( h_{x,v}(\theta) := \tilde{f}_T(x + \theta v) \). The function \( \theta \mapsto h_{x,v}(\theta) \) is infinitely differentiable for every \( v \). Therefore, \( \tilde{f}_T \) has \( \ell_p \)-Lipschitz \( p \)-th order derivatives if and only if \( |h_{x,v}^{(p+1)}(0)| \leq \ell_p \) for every \( x, v \). Using the shorthand notation \( \partial_{i_1} \cdots \partial_{i_k} \) for \( \frac{\partial^k}{\partial x_{i_1} \cdots x_{i_k}} \), we have
\[
h_{x,v}^{(p+1)}(0) = \sum_{i_1, \ldots, i_{p+1}=1}^T \partial_{i_1} \cdots \partial_{i_{p+1}} \tilde{f}_T(x) v_{i_1} \cdots v_{i_{p+1}}.
\]
Examing \( \tilde{f}_T \), we see that \( \partial_{i_1} \cdots \partial_{i_{p+1}} \tilde{f}_T \) is non-zero if and only if \( |i_j - i_k| \leq 1 \) for every \( j, k \in [p+1] \). Consequently, we can rearrange the above summation as
\[
h_{x,v}^{(p+1)}(0) = \sum_{\delta_1, \delta_2, \ldots, \delta_p \in \{0,1\}^p \cup \{0,1\}^p} \sum_{i=1}^T \partial_{i+\delta_1} \cdots \partial_{i+\delta_p} \partial_i \tilde{f}_T(x) v_{i+\delta_1} \cdots v_{i+\delta_p} v_i,
\]
where we take \( v_0 := 0 \) and \( v_{T+1} := 0 \). Brief calculation show that
\[
\begin{align*}
\sup_{x \in \mathbb{R}^T} \max_{i \in [T]} \max_{\delta \in \{0,1\}^p \cup \{0,1\}^p} \left| \partial_{i+\delta_1} \cdots \partial_{i+\delta_p} \partial_i \tilde{f}_T(x) \right| &\leq \max_{k \in [p+1]} \left\{ 2\sup_{x \in \mathbb{R}} \Psi^{(k)}(x) \sup_{x' \in \mathbb{R}} \Phi^{(p+1-k)}(x') \right\} \\
&\leq 2\sqrt{2\pi e} \cdot e^{2.5(p+1)\log(4(p+1))} \leq \exp (2.5p \log p + 4p + 9).
\end{align*}
\]
where the second inequality uses Lemma 1.iii, and \( \Phi(x') \leq \sqrt{2\pi e} \) for the case \( k = p + 1 \). Defining \( \ell_p = 2^{p+1}e^{2.5(p+1)\log p+4+9} \leq e^{2.5p \log p+5+9+10} \), we thus have
\[
\left| h_{x,v}^{(p+1)}(0) \right| \leq \sum_{\delta \in \{0,1\}^p \cup \{0,1\}^p} 2^{-(p+1)} \ell_p \left| \sum_{i=1}^T v_{i+\delta_1} \cdots v_{i+\delta_p} v_i \right| \leq (2^{p+1} - 1) 2^{-(p+1)} \ell_p \leq \ell_p,
\]
where we have used \( \left| \sum_{i=1}^T v_{i+\delta_1} \cdots v_{i+\delta_p} v_i \right| \leq 1 \) 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
\[
\sum_{i=1}^T v_{i+\delta_1} \cdots v_{i+\delta_p} v_i = \sum_{i=1}^T v_{i}^{p+1-\sum_{j=1}^p \delta_j} \sum_{i_{p+1}=1}^{v_{i}} v_{i_{p+1}}.
\]
If \( \delta = 0 \) then \( \left| \sum_{i=1}^T v_{i+\delta_1} \cdots v_{i+\delta_p} v_i \right| = \left| \sum_{i=1}^T v_{i}^{p+1} \right| \leq \sum_{i=1}^T v_{i}^2 = 1 \). Otherwise, letting \( 1 \leq \sum_{j=1}^p |\delta_j| = n \leq p \), the Cauchy-Swartz inequality implies
\[
\left| \sum_{i=1}^T v_{i+\delta_1} \cdots v_{i+\delta_p} v_i \right| \leq \left( \sum_{i=1}^T v_{i}^{2(p+1-n)} \right)^{1/2} \left( \sum_{i=1}^T v_{i}^{2n} \right)^{1/2} \leq \sum_{i=1}^T v_{i}^2 = 1,
\]
where \( s = -1 \) or 1. This gives the result. \( \square \)
B.3 Proof of Lemma 4: main argument

The proof of Lemma 4 contains multiple sub-lemmas, with some proofs in Section B.4. We use notation defined in the proof of Lemma 4 within these sub-lemmas.

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

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

**Proof.** 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

$$G_t = \left\{ \max_{j \in \{t, \ldots, T\}} \left| \langle u^{(j)}, P_{t-1}^\perp x^{(t)} \rangle \right| \leq \alpha \left\| P_{t-1}^\perp x^{(t)} \right\| \right\} \text{ where } \alpha = \frac{1}{5R\sqrt{T}}. \tag{22}$$

For every $t$, define

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

The following linear-algebraic result justifies the definition (22) of $G_t$; we defer the proof to Section B.4.1.

**Lemma 9.** For all $t \leq T$, $G_{\leq t}$ implies $\|\langle u^{(j)}, x^{(s)} \rangle\| < \frac{1}{2}$ for every $s \in \{1, \ldots, t\}$ and every $j \in \{s, \ldots, T\}$.

By Lemma 9 the event $G_{\leq t}$ implies our result, so using that $\mathbb{P}(G_{<T}^c) \leq \sum_{t=1}^T \mathbb{P}(G_t^c \mid G_{<t})$, it suffices to show that

$$\mathbb{P}(G_{<T}^c) \leq \sum_{t=1}^T \mathbb{P}(G_t^c \mid G_{<t}) \leq \delta. \tag{23}$$

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

$$\mathbb{P}(G_t^c \mid G_{<t}) \leq \sum_{j \in \{t, \ldots, T\}} \mathbb{P} \left( \left| \frac{\langle u^{(j)}, P_{t-1}^\perp x^{(t)} \rangle}{\left\| P_{t-1}^\perp x^{(t)} \right\|} \right| > \alpha \mid G_{<t} \right)$$

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

$$\leq \sum_{j \in \{t, \ldots, T\}} \mathbb{E}_{\xi, U_{(<t)}} \mathbb{P} \left( \left| \frac{P_{t-1}^\perp u^{(j)}}{\left\| P_{t-1}^\perp u^{(j)} \right\|}, \frac{P_{t-1}^\perp x^{(t)}}{\left\| P_{t-1}^\perp x^{(t)} \right\|} \right| > \alpha \mid \xi, U_{(\leq t)}, G_{<t} \right), \tag{24}$$

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

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

**Lemma 10.** For every $i \leq T$, there exist measurable functions $A^{(i)}_+$ and $A^{(i)}_-$ such that

$$x^{(i)} = A^{(i)}_+ (\xi, U_{(\leq i)}) 1_{G_{\leq i}} + A^{(i)}_- (\xi, U) 1_{G_{< i}}. \tag{25}$$
Proof. Since the iterates are informed by $\tilde{f}_{T; U}$, we may write each one as (recall definition (4))

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

for measurable functions $A^{(i)}, A^{(i)}$, where we recall the shorthand $\nabla^{(0, \ldots, p)}h(x)$ for the derivatives of $h$ at $x$ to order $p$. Crucially, by Lemma 9, $G_{< i}$ implies $|\langle u^{(j)}, x^{(s)} \rangle| < \frac{1}{2}$ for every $s < i$ and every $j \geq s$. As $\tilde{f}_T$ 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 (25) for every $i \leq T$.

Consequently (as $G_{< t}$ implies $G_{< i}$ for every $i \leq t$), 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 $\mathbb{P}(G^c_t \mid G_{< t}) = 0$, so we may assume without loss of generality that $P_{t-1}^{\perp}x^{(t)} \neq 0$. We may therefore regard $P_{t-1}^{\perp}x^{(t)}/\|P_{t-1}^{\perp}x^{(t)}\|$ in (24) as a deterministic unit vector in the subspace $P_{t-1}^{\perp}$ projects to. We now characterize the conditional distribution of $P_{t-1}^{\perp}u^{(j)}/\|P_{t-1}^{\perp}u^{(j)}\|$.

Lemma 11. Let $t \leq T$, and $j \in \{t, \ldots, T\}$. Then conditioned on $\xi, U_{(< t)}$ and $G_{< t}$, the vector\

$$\frac{P_{t-1}^{\perp}u^{(j)}}{\|P_{t-1}^{\perp}u^{(j)}\|}$$

is uniformly distributed on the unit sphere in the subspace to which $P_{t-1}^{\perp}$ projects.

Lemma 11 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; we prove the lemma in Section B.4.2.

Summarizing the discussion above, the conditional probability in (24) measures the inner product of two unit vectors in a subspace of $\mathbb{R}^d$ of dimension $d' = \text{tr}(P_{t-1}^{\perp}) \geq d - 2(t - 1)$, with one of the vectors deterministic and the other uniformly distributed. We may write this as

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

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],

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

Substituting this bound back into the probability (24) gives

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

Substituting this in turn into the bound (23), we have $\mathbb{P}(G_{< T}^c) \leq \sum_{t=1}^T \mathbb{P}(G^c_t \mid G_{< t}) \leq 2T^2e^{-\frac{\alpha^2}{2(d-2T)}}$. Setting $d \geq 52TR^2 \log \frac{2T^2}{\delta} > \frac{\alpha}{2} \log \frac{2T^2}{\delta} + 2T$ establishes $\mathbb{P}(G_{< T}^c) \leq \delta$, concluding Lemma 4.

B.4 Proof of Lemma 4: auxiliary arguments

B.4.1 Proof of Lemma 9

Lemma 9. For all $t \leq T$, $G_{< t}$ implies $|\langle u^{(j)}, x^{(s)} \rangle| < 1/2$ for every $s \in \{1, \ldots, t\}$ and every $j \in \{s, \ldots, T\}$.
Moreover, the event $G_{t}$ implies $G_{≤t}$ for every $s ≤ t$, it suffices to show that $G_{≤t}$ implies $|\langle u^{(j)}, x^{(t)} \rangle| < 1/2$ for every $j ∈ \{t, \ldots, T\}$. We will in fact prove a stronger statement:

For every $t$, $G_{≤t}$ implies $\|P_{t-1}u^{(j)}\|^2 ≤ 2α^2(t - 1)$ for every $j ∈ \{t, \ldots, T\}$,  

(26) 

where we recall that $P_{t} ∈ \mathbb{R}^{d×d}$ is the projection operator to the span of $x^{(1)}, u^{(1)}, \ldots, x^{(t)}, u^{(t)}$, $P_{t}^⊥ = I_d - P_{t}$ and $α = 1/(5R\sqrt{T})$. Before proving (26), let us show that it implies our result. Fixing $j ∈ \{t, \ldots, T\}$, we have

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

Since $G_{t}$ holds, its definition (22) implies $|\langle u^{(j)}, P_{t-1}x^{(t)} \rangle| ≤ α\|P_{t-1}x^{(t)}\| ≤ α\|x^{(t)}\|$. Moreover, by Cauchy-Schwarz and the implication (26), we have $|\langle u^{(j)}, P_{t-1}x^{(t)} \rangle| ≤ \|P_{t-1}u^{(j)}\|\|x^{(t)}\| ≤ \sqrt{2α^2(t - 1)}\|x^{(t)}\|$. Combining the two bounds, we obtain the result of the lemma,

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

where we have used $\|x^{(t)}\| ≤ R$ and $α = 1/(5R\sqrt{T})$.

We prove bound (26) by induction. The basis of the induction, $t = 1$, is trivial, as $P_{0} = 0$. We shall assume (26) holds for $s ∈ \{1, \ldots, t - 1\}$ 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

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

(27)

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

$$\hat{P}_{k} = P_{k} + \frac{1}{\|P_{k}x^{(k+1)}\|} \left(\left|\left| P_{k}x^{(k+1)} \right|\right| \right)^2 \left(\left|\left| P_{k}x^{(k+1)} \right|\right| \right)^+.$$

Then for every $j > i$ we have

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

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

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

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

Moreover, the event $G_{t}$ implies $|\langle u^{(i)}, P_{t-1}x^{(i)} \rangle| ≤ α^2 \|P_{t-1}x^{(i)}\|^2$, so

$$\left\langle \hat{P}_{t-1}u^{(i)}, u^{(j)} \right\rangle ≤ \left\langle P_{t-1}u^{(i)}, u^{(j)} \right\rangle + \frac{\left\langle u^{(i)} , P_{t-1}x^{(i)} \right\rangle \left\langle u^{(j)} , P_{t-1}x^{(i)} \right\rangle}{\|P_{t-1}x^{(i)}\|^2} ≤ α^2(2i - 1) ≤ \frac{α}{2},$$
where the last transition uses $\alpha = \frac{1}{5R \sqrt{T}} \leq \frac{1}{4t}$ because $R \geq \sqrt{T} \geq i$. We also have the lower bound
\[
\|P_{t-1}^\perp u^{(j)}\|^2 = \left\langle \hat{P}_{t-1}^\perp u^{(i)}, u^{(i)} \right\rangle = 1 - \|P_{t-1} u^{(i)}\|^2 - \frac{\left(\langle u^{(i)}, P_{t-1}^\perp x^{(i)} \rangle\right)^2}{\|P_{t-1}^\perp x^{(i)}\|^2} \geq 1 - \alpha^2 (2i - 1) \geq \frac{1}{2}, \tag{28b}
\]
where the first equality uses $(P_{t-1}^\perp)^2 = P_{t-1}^\perp$, the second the definition of $\hat{P}_{t-1}$, and the inequality uses $\langle u^{(j)}, P_{t-1}^\perp x^{(i)} \rangle \leq \alpha \|P_{t-1}^\perp x^{(i)}\|$ and $\|P_{t-1} u^{(j)}\|^2 \leq 2\alpha^2 (i - 1)$.

Combining the observations (28a) and (28b), we can bound each summand in the second summation in (27). Since the summands in the first summation are bounded by $\alpha^2$ by definition (22) of $G_t$, we obtain
\[
\|P_{t-1} u^{(j)}\|^2 \leq \sum_{i=1}^{t-1} \alpha^2 + \sum_{i=1}^{t-1} \frac{(\alpha/2)^2}{1/2} = \alpha^2 \left( t - 1 + \frac{t - 1}{2} \right) \leq 2\alpha^2 (t - 1),
\]
which completes the induction. \hfill \Box

### B.4.2 Proof of Lemma 11

**Lemma 11.** Let $t \leq T$, and $j \in \{t, \ldots, T\}$. Then conditioned on $\xi, U_{(< t)}$ and $G_{< t}$, the vector $\frac{P_{t-1}^\perp u^{(j)}}{\|P_{t-1}^\perp u^{(j)}\|}$ is uniformly distributed on the unit sphere in the subspace to which $P_{t-1}^\perp$ projects.

**Proof.** Throughout the proof we fix $t \leq T$ and $j \in \{t, \ldots, T\}$. We begin by noting that by (26), $G_{< t}$ implies
\[
\|P_{t-1}^\perp u^{(j)}\|^2 = 1 - \|P_{t-1} u^{(j)}\|^2 \geq 1 - 2\alpha^2 (t - 1) > 0.
\]
Therefore, when $G_{< t}$ holds we have $\frac{P_{t-1}^\perp u^{(j)}}{\|P_{t-1}^\perp u^{(j)}\|} \neq 0$ so $\frac{P_{t-1}^\perp u^{(j)}}{\|P_{t-1}^\perp u^{(j)}\|}$ 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 $\xi, U_{(< t)}$ and $G_{< t}$. We wish to show that
\[
p_{\geq t}(U_{(\geq t)} \mid \xi, U_{(< t)}, G_{< t}) = p_{\geq t}(ZU_{(\geq t)} \mid \xi, U_{(< t)}, G_{< t}) \tag{29}
\]
for every rotation $Z \in \mathbb{R}^{d \times d}$, $Z^\top Z = I_d$, satisfying
\[Z v = v \text{ for all } v \in \{x^{(1)}, u^{(1)}, \ldots, x^{(t-1)}, u^{(t-1)}\}.
\]
Throughout, we let $Z$ denote such a rotation. Letting $p_{\xi, U}$ and $p_U$ denote the densities of $(\xi, U)$ and $U$, respectively, we have
\[
p_{\geq t}(U_{(\geq t)} \mid \xi, U_{(< t)}, G_{< t}) = \frac{\mathbb{P}(G_{< t} \mid \xi, U)p_{\xi, U}(\xi, U)}{\mathbb{P}(G_{< t} \mid \xi, U_{(< t)})p_{\xi, U_{(< t)}}(\xi, U_{(< t)})} = \frac{\mathbb{P}(G_{< t} \mid \xi, U)p_U(U)}{\mathbb{P}(G_{< t} \mid \xi, U_{(< t)})p_{U_{(< t)}}(U_{(< t)})}
\]
where the first equality holds by the definition of conditional probability and second by the independence of $\xi$ and $U$. We have $ZU_{(< t)} = U_{(< t)}$ and therefore, by the invariance of $U$ to rotations, $p_{U_{(< t)}}(U_{(< t)}) = p_U(U)$. Hence, replacing $U$ with $ZU$ in the above display yields
\[
p_{\geq t}(ZU_{(\geq t)} \mid \xi, U_{(< t)}, G_{< t}) = \frac{\mathbb{P}(G_{< t} \mid \xi, ZU)p_U(U)}{\mathbb{P}(G_{< t} \mid \xi, U_{(< t)})p_{U_{(< t)}}(U_{(< t)})}.
\]
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 (29) holds.

First, note that $\mathbb{P}(G_{<t} \mid \xi, U)$ is supported on $\{0, 1\}$ for every $\xi, U$, 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 $\mathbb{P}(G_{<t} \mid \xi, ZU) = 1$. Set $U' = ZU$, observing that $u^{(i)} = Zu^{(i)} = u^{(i)}$ for any $i < t$, and let $x^{(1)}, \ldots, x^{(T)}$ be the sequence generated from $\xi$ and $U'$. We will prove by induction on $i$ that $\mathbb{P}(G_{<t} \mid \xi, U') = 1$ 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 $\xi$ and $u^{(1)}, \ldots, u^{(i-1)} = u^{(1)}, \ldots, u^{(i-1)}$ by Lemma 10. Consequently, $x^{(l)} = x^{(l)}$ for any $l \leq i$ and also $P_{i-1}u^{(i)} = P_{i-1}x^{(i)}$. Therefore, for any $l \geq i$,

$$\left\langle u^{(l)} \left( \frac{P_{i-1}u^{(i)}}{\|P_{i-1}u^{(i)}\|} \right) \right\rangle = \left\langle u^{(l)} \left( Z^\top \frac{P_{i-1}x^{(i)}}{\|P_{i-1}x^{(i)}\|} \right) \right\rangle = \left\langle u^{(l)} \left( \frac{P_{i-1}x^{(i)}}{\|P_{i-1}x^{(i)}\|} \right) \right\rangle \leq \alpha,$$

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

Marginalizing the density (29) to obtain a density for $u^{(i)}$ and recalling that $P_{i-1}$ is measurable $\xi, U_{(<t)}, G_{<t}$, we conclude that, conditioned on $\xi, U_{(<t)}, G_{<t}$ the random variable $\frac{P_{i-1}u^{(i)}}{\|P_{i-1}u^{(i)}\|}$ has the same density as $\frac{P_{i-1}Zu^{(i)}}{\|P_{i-1}Zu^{(i)}\|}$. However, $P_{i-1}Z = ZP_{i-1}$ by assumption on $Z$, and therefore

$$\frac{P_{i-1}Zu^{(i)}}{\|P_{i-1}Zu^{(i)}\|} = \frac{Z}{\|Z\|} \frac{P_{i-1}u^{(i)}}{\|P_{i-1}u^{(i)}\|}.$$

We conclude that the conditional distribution of the unit vector $\frac{P_{i-1}u^{(i)}}{\|P_{i-1}u^{(i)}\|}$ is invariant to rotations in the subspace to which $P_{i-1}$ projects. \qed

### B.5 Proof of Lemma 6

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

i. We have $\hat{f}_{T,U}(0) - \inf_x \hat{f}_{T,U}(x) \leq 12T$.

ii. For every $p \geq 1$, the $p$th order derivatives of $\hat{f}_{T,U}$ are $\hat{e}_p$-Lipschitz continuous, where $\hat{e}_p \leq \exp(cp \log p + c)$ for a numerical constant $c < \infty$.

**Proof.** Part i holds because $\hat{f}_{T,U}(0) = \hat{f}_T(0)$ and $\hat{f}_{T,U}(x) \geq \hat{f}_{T,U}(\rho(x))$ for every $x$, so

$$\inf_{x \in \mathbb{R}^d} \hat{f}_{T,U}(x) \geq \inf_{x \in \mathbb{R}^d} \hat{f}_{T,U}(\rho(x)) = \inf_{\|x\| \leq R} \hat{f}_{T}(x) \geq \inf_{x \in \mathbb{R}^d} \hat{f}_{T}(x),$$

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

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.

32
of the $d$-dimensional identity matrix, denoted $U = I_d$. Recall the scaling $\rho(x) = R x / \sqrt{R^2 + \|x\|^2}$ with “radius” $R = 230\sqrt{T}$ and the definition $\hat{f}_{T;U}(x) = \tilde{f}_T(U^T \rho(x)) + \frac{1}{10} \|x\|^2$. 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 $U = I_d$ without loss of generality, so we consider the function

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

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

$$\tilde{\nabla}_{i_1, \ldots, i_k} \hat{f}_{T;I}(x) = \sum_{(S_1, \ldots, S_L) \in P_k} \sum_{j_1, \ldots, j_L = 1}^T \left( \prod_{l = 1}^L \nabla_{i_{S_l}}^{S_l}\rho_{j_l}(x) \right) \tilde{\nabla}_{j_1, \ldots, j_L} \hat{f}_T(y), \quad y = \rho(x), \quad (30)$$

where we have used the shorthand $\nabla_{i_{S_l}}^{S_l}$ to denote the partial derivatives with respect to each of $x_{i_j}$ for $j \in S$. We use the equality (30) to argue that (recall the identity (2))

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

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

To do this, we begin by considering the partitioned sum (30). 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

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

algebraic manipulations and rearrangement of the sum (30) yield

$$\langle \nabla^k \hat{f}_{T;I}(x), v^\otimes k \rangle = \sum_{(S_1, \ldots, S_L) \in P_k} \sum_{i_1, \ldots, i_k = 1}^d v_{i_1} v_{i_2} \cdots v_{i_k} \sum_{j_1, \ldots, j_L = 1}^T \left( \prod_{l = 1}^L \nabla_{i_{S_l}}^{S_l}\rho_{j_l}(x) \right) \tilde{\nabla}_{j_1, \ldots, j_L} \hat{f}_T(y)$$

$$= \sum_{(S_1, \ldots, S_L) \in P_k} \sum_{j_1, \ldots, j_L = 1}^T \tilde{v}^{|S_1|}_{j_1} \cdots \tilde{v}^{|S_L|}_{j_L} \tilde{\nabla}_{j_1, \ldots, j_L} \hat{f}_T(y)$$

$$= \sum_{(S_1, \ldots, S_L) \in P_k} \left( \tilde{\nabla}^L \hat{f}_T(y), \tilde{v}^{|S_1|} \otimes \cdots \otimes \tilde{v}^{|S_L|} \right).$$

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

$$\sup_x \|\tilde{v}^k(x)\| \leq \exp(ck \log k + c)R^{1-k}. \quad (31)$$

Before proving inequality (31), we show how it implies the desired lemma. By the preceding display, we have

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

$^5$ 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 $p$. 

33
Lemma 3 shows that there exists a numerical constant $c < \infty$ such that
\[
\left\| \nabla^{(L)} f_T(y) \right\|_{\text{op}} \leq \ell_{L-1} \leq \exp(cl \log L + c) \text{ for all } L \geq 2.
\]
When the number of partitions $L = 1$, we have $|S_1| = p + 1 \geq 2$, and so Lemma 3.ii yields
\[
\left\| \nabla f_T(y) \right\|_{\text{op}} \left\| \nabla^{(S_1)} \right\| = \left\| \nabla f_T(y) \right\| \left\| \nabla^{(S_1)} \right\| \leq 23\sqrt{T} \cdot R^{-p} \exp(c \log p + c) \leq \exp(c \log p + c),
\]
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
\[
\left\| \nabla f_T(y) \right\|_{\text{op}} \prod_{l=1}^{L} \left\| \nabla^{(S_l)} \right\| \leq \exp(c \log p + c)
\]
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 (31). 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
\[
\frac{d^k}{dt^k} f(g(t)) = \sum_{P \in \mathcal{P}_k} f^{(|P|)}(g(t)) \cdot \prod_{S \in P} g^{(|S|)}(t),
\]
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 $\overline{\theta}_j^{(k)}(\xi) = (\nabla^k \overline{\theta}_j(\xi), v_{\otimes k})$, so that
\[
\overline{\theta}_j^{(k)}(\xi) = \nabla^k \lambda(\xi), v_{\otimes k} \text{ and } \overline{\theta}^{(k)} = R^{1-k}\overline{\theta}^{(k)}(x/R).
\]
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 Faà di Bruno’s formula (32). 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
\[
h^{(k)}(0) = \sum_{P \in \mathcal{P}_{k,2}} (-1)^{|P|} \frac{(2|P| - 1)!!}{(1 + \|\xi\|^2)^{|P| - 1/2}} \langle \xi, v \rangle \mathcal{C}_1(P) \|v\|^{2\mathcal{C}_2(P)}
\]
where $C_i(P)$ denotes the number of sets in $P$ with precisely $i$ elements. Noting that $\|v\| = 1$, we may rewrite this as
\[
\langle \nabla^k \lambda(\xi), v_{\otimes k} \rangle = \sum_{l=1}^{k} \sum_{P \in \mathcal{P}_{k,2}, C_1(P) = l} (-1)^{|P|} \frac{(2|P| - 1)!!}{(1 + \|\xi\|^2)^{|P| - 1/2}} \langle \xi, v \rangle^l.
\]
34
Taking derivatives we obtain
\[
\bar{v}^k = \nabla \langle \nabla^k \lambda(\xi), v^{\otimes k} \rangle = \left( \sum_{l=1}^{k} a_l(\xi) \langle \xi, v \rangle^{l-1} \right) v + \left( \sum_{l=1}^{k} b_l(\xi) \langle \xi, v \rangle^l \right) \xi
\]
where
\[
a_l(\xi) = \sum_{P \in \mathcal{P}_{k,2,\mathcal{C}_1(P)=l}} \frac{(-1)^{|P|}(2|P| - 1)!!}{(1 + \|\xi\|^2)^{|P| - 1/2}} \quad \text{and} \quad b_l(\xi) = \sum_{P \in \mathcal{P}_{k,2,\mathcal{C}_1(P)=l}} \frac{(-1)^{|P|+1}(2|P| + 1)!!}{(1 + \|\xi\|^2)^{|P|+1/2}}.
\]

We would like to bound \(a_l(\xi)\langle \xi, v \rangle^{l-1}\) and \(b_l(\xi)\langle \xi, v \rangle^l\). Note that \(|P| \geq \mathcal{C}_1(P)\) for every \(P \in \mathcal{P}_k\), so \(|P| \geq l\) 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 \((2k - 1)!! \leq \exp(k \log k)\) as well. As a consequence, we obtain
\[
\sup_\xi |a_l(\xi)\langle \xi, v \rangle^{l-1}| \leq \exp(cl \log l) \sup_\xi \frac{\|\langle \xi, v \rangle\|^{l-1}}{(1 + \|\xi\|^2)^{(l-1)/2}} \leq \exp(cl \log l),
\]
where we have used \(\|\langle \xi, v \rangle\| \leq \|\xi\|\) due to \(\|v\| = 1\). We similarly bound \(\sup_\xi |b_l(\xi)\|\langle \xi, v \rangle\| \|\xi\|\).

Returning to expression (33), we have
\[
\sup_x \|v^k(x)\| \leq \exp(ck \log k + c) R^{1-k},
\]
for a numerical constant \(c < \infty\). This is the desired bound (31), completing the proof. \(\Box\)

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 \(D, D_p, \) and \(\epsilon\) be positive. Then
\[
\mathcal{T}_\epsilon(A_{\text{rand}}, \mathcal{T}_p^{\text{dist}}(D, D_p)) \geq c_0 \cdot D^{1+\epsilon} \left( \frac{L_p}{\ell_p} \right)^{1+\epsilon} \epsilon^{-\frac{1+\epsilon}{p}},
\]
where \(\ell_p \leq \epsilon e^{c_1 p \log p + c_1} \).

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 \(j\)th standard basis vector. Then we define the bump function \(\bar{h}_T : \mathbb{R}^T \rightarrow \mathbb{R}\) by
\[
\bar{h}_T(x) = \Psi \left( 1 - \frac{25}{2} \| x - \frac{4}{5} e^{(T)} \|^2 \right) = \begin{cases} 0 & \| x - \frac{4}{5} e^{(T)} \| \geq \frac{1}{5} \\ \exp \left( 1 - \frac{1}{(1 - 25 \| x - \frac{4}{5} e^{(T)} \|^2)^2} \right) & \text{otherwise.} \end{cases} \tag{34}
\]
As Figure 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.
Lemma 12. The function $\tilde{h}_T$ satisfies the following.

i. $\tilde{h}_T(0.8e(T)) = 1$ and $\tilde{h}_T(x) \in [0, 1]$ for all $x \in \mathbb{R}^T$.

ii. $\tilde{h}_T(x) = 0$ on the set $\{x \in \mathbb{R}^d | x_T \leq \frac{3}{5} \text{ or } ||x|| \geq 1\}$.

iii. For $p \geq 1$, the $p$th order derivatives of $\tilde{h}_T$ are $\tilde{\ell}_p$-Lipschitz continuous, where $\tilde{\ell}_p < e^{3p\log p + c}$ for some numerical constant $c < \infty$.

We prove the lemma in Section C.2; the proof is similar to that of Lemma 6. With these properties in hand, we can prove Theorem 3.

C.1 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

$$f_U(x) = \frac{L_p \sigma^{p+1}}{\ell'_p} \hat{f}_{U;U}(x/\sigma) - \frac{L_p D^{p+1}}{\ell'_p} \tilde{h}_T(U^\top x/D),$$

where $\hat{f}_{U;U}(x) = \hat{f}_{T;U}(\rho(x)) + \frac{1}{10} ||x||^2$ is the randomized hard instance construction (14) with $\rho(x) = x/\sqrt{1 + ||x/R||^2}$, $\tilde{h}_T$ is the bump function (34) and $\ell'_p = \hat{\ell}_p + \tilde{\ell}_p$, for $\hat{\ell}_p$ and $\tilde{\ell}_p$ as in Lemmas 6.ii and 12.iii, respectively. By the lemmas, $f_U$ has $L_p$-Lipschitz $p$th order derivatives and $\ell'_p \leq e^{c_1 p \log p + c}$ for some $c_1 < \infty$. We assume that $\sigma \leq D$; our subsequent choice of $\sigma$ will obey this constraint.

Following our general proof strategy, we first demonstrate that $f_U \in F_{p}^\text{dist}(D, L_p)$, for which all we need do is guarantee that the global minimizers of $f_U$ have norm at most $D$. By the constructions (14) and (10) of $\hat{f}_{T;U}$ and $\tilde{f}_{T;U}$, Lemma 12.i implies

$$f_U \left((0.8D)u^{(T)}\right) = \frac{L_p \sigma^{p+1}}{\ell'_p} \hat{f}_T(\rho(u^{(T)})) + \frac{L_p \sigma^{p+1}}{10 \ell'_p} \left\| \frac{4DU^{(T)}}{5\sigma} \right\|^2 - \frac{L_p D^{p+1}}{\ell'_p} \tilde{h}_T(0.8e(T))$$

$$= \frac{L_p \sigma^{p+1}}{\ell'_p} \hat{f}_T(0) + \frac{8L_p \sigma^{-p} D^2}{125 \ell'_p} + \frac{-L_p D^{p+1}}{\ell'_p} < -\frac{117}{125} \frac{L_p D^{p+1}}{\ell'_p} + \frac{L_p \sigma^{p+1}}{\ell'_p} \hat{f}_T(0)$$
with the final inequality using our assumption \( \sigma \leq D \). On the other hand, for any \( x \) such that \( \tilde{h}_T(U^\top x/D) = 0 \), we have by Lemma 6.i (along with \( \tilde{f}_{T,U}(0) = 0 \)) that

\[
f_U(x) \geq \frac{L_p\sigma^{p+1}}{\ell_p^p} \inf_x \tilde{f}_{T,U}(x) \geq -12 \frac{L_p\sigma^{p+1}}{\ell_p^p} T + \frac{L_p\sigma^{p+1}}{\ell_p^p} \tilde{f}_T(0).
\]

Combining the two displays above, we conclude that if

\[
12 \frac{L_p\sigma^{p+1}}{\ell_p^p} T \leq \frac{117}{125} \frac{L_p D^{p+1}}{\ell_p^p},
\]

then all global minima \( x^* \) of \( f_U \) must satisfy \( \tilde{h}_T(U^\top x^*/D) > 0 \). Inspecting the definition (19) of \( \tilde{h}_T \), this implies \( \|x^*/D - 0.8u(T)\| < \frac{1}{\ell} \), and therefore \( \|x^*\| \leq D \). Thus, by setting

\[
T = \left\lfloor \frac{D^{p+1}}{13\sigma^{p+1}} \right\rfloor,
\]

we guarantee that \( f_U \in F_p^{\text{dist}}(D, L_p) \) as long as \( \sigma \leq D \).

It remains to show that, for an appropriately chosen \( \sigma \), 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 \( \sigma \leq D \), for any \( x \in \mathbb{R}^d \),

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

We defer the proof of claim (36) to the end of this section.

Now, let \( U \in \mathbb{R}^{d \times T} \) be an orthogonal matrix chosen uniformly at random from \( O(d, T) \). Let \( x^{(1)}, \ldots, x^{(t)} \) be a sequence of iterates generated by algorithm \( A \in \mathcal{A}_{\text{rand}} \) applied on \( f_U \). We argue that \( |\langle u(T), \rho(x^{(t)}/\sigma) \rangle| < 1/2 \) for all \( t \leq T \), with high probability. To do so, we briefly revisit the proof of Lemma 4 (Sec. B.3) where we replace \( \tilde{f}_{T,U} \) with \( f_U \) and \( x^{(t)} \) with \( \rho(x^{(t)}/\sigma) \). By Lemma 9 we have that for every \( t \leq T \) the event \( G_{\leq t} \) implies \( |\langle u(T), \rho(x^{(s)}/\sigma) \rangle| < 1/2 \) for all \( s \leq t \), and therefore by the claim (36) we have that Lemma 10 holds (as we may replace the terms \( \tilde{h}_T(U^\top x^{(s)}/D) \), \( s < t \), with 0 whenever \( G_{< t} \) holds). The rest of the proof of Lemma 9 proceeds unchanged and gives us that with probability greater than 1/2 (over any randomness in \( A \) and the uniform choice of \( U \)),

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

By claim (36), this implies \( \nabla h_T(U^\top x^{(t)}/D) = 0 \), and by Lemma 5, \( \|\nabla \tilde{f}_{T,U}(x^{(t)}/\sigma)\| > 1/2 \). Thus, after scaling,

\[
\|\nabla f_U(x^{(t)})\| > \frac{L_p\sigma^p}{2\ell_p^p}
\]

for all \( t \leq T \), with probability greater that 1/2. As in the proof of Theorem 2, By taking \( \sigma = (2\ell_p^p/L_p)^{1/p} \) we guarantee

\[
\inf_{A \in \mathcal{A}_{\text{rand}}} \sup_{U \in O(d, T)} T_\epsilon(A, f_U) \geq 1 + T.
\]

where \( T = \left\lfloor D^{p+1}/13\sigma^{p+1} \right\rfloor \) is defined in Eq. (35). Thus, as \( f_U \in F_p^{\text{dist}}(D, L_p) \) for our choice of \( T \), we immediately obtain

\[
T_\epsilon(\mathcal{A}_{\text{rand}}, F_p^{\text{dist}}(D, L_p)) \geq T + 1 \geq \frac{D^{1+p}}{52} \left( \frac{L_p}{\ell_p^p} \right)^{1+p} \epsilon^{-\frac{1+p}{p}},
\]

37
as long as our initial assumption \( \sigma \leq D \) holds. When \( \sigma > D \), we have that \( \frac{2p}{L_p} \epsilon > D^p \), or
\[ 1 > D^{p+1} \left( \frac{L_p}{2L_p} \right)^{\frac{1+p}{p}} - e^{-\frac{1+p}{p}}, \]
so that the bound is vacuous in this case regardless: every method must take at least 1 step.

Finally, we return to demonstrate claim (36). Note that \(|\langle u(T), \rho(x/\sigma) \rangle| < 1/2\) is equivalent to
\[ |\langle u(T), x \rangle| < \frac{\sigma}{2} \| x \|^2 + \frac{1}{2} \| x \|^2, \]
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 12.ii we have that \( \hat{h}_T(U^\top y/D) = 0 \) for \( y \) near \( x \). In the second case, we have \( |\langle u(T), x \rangle| > (4R/\sqrt{5})|\langle u(T), x \rangle| > 230|\langle u(T), x \rangle| \). 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 12.ii.

### C.2 Proof of Lemma 12

Properties i and ii are evident from the definition (34) of \( \tilde{h}_T \). To show property iii, consider
\[ h(x) = \tilde{h}_T \left( \frac{x + 5 \epsilon_p x}{5} \right) = \Psi \left( 1 - \frac{1}{2} \| x \|^2 \right), \]
which is a translation and scaling of \( \tilde{h}_T \), so if we show \( h \) has \( (\tilde{c}_p/5^{p+1}) \)-Lipschitz \( p \)th order derivatives, for every \( p \geq 1 \), we obtain the required results. For any \( x, v \in \mathbb{R}^T \) with \( \| v \| \leq 1 \) we define the directional projection \( h_{x,v}(t) = h(x + t \cdot v) \). The required smoothness bound is equivalent to
\[ \left| h_{x,v}^{(p+1)}(0) \right| \leq \frac{\tilde{c}_p}{5^{p+1}} \leq e^{c_p \log p + c} \]
for every \( x, v \in \mathbb{R}^d \) with \( \| v \| \leq 1 \), every \( p \geq 1 \) 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 Faà di Bruno’s formula (32) to write, for any \( k \geq 1 \),
\[ h_{x,v}^{(k)}(0) = \sum_{P \in \mathcal{P}_k} \Psi(|P|)(\beta(0)) \cdot \prod_{S \in P} \beta(|S|)(0), \]
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
\[ h_{x,v}^{(k)}(0) = \sum_{P \in \mathcal{P}_{k,2}} (-1)^{|P|} \Psi(|P|) \left( 1 - \frac{1}{2} \| x \|^2 \right) \langle x, v \rangle C_1(P) \| v \|^{2C_2(P)} \]
where \( \mathcal{P}_{k,2} \) denote the partitions of \( [k] \) consisting only of subsets with one or two elements and \( C_i(P) \) denotes the number of sets in \( P \) with precisely \( i \) elements.

Noting that \( \Psi^{(k)}(1 - \frac{1}{2} \| x \|^2) = 0 \) for any \( k \geq 0 \) and \( \| x \| > 1 \), we may assume \( \| x \| \leq 1 \). Since \( \| v \| \leq 1 \), we may bound \( |h_{x,v}^{(p+1)}(0)| \) by
\[ \left| h_{x,v}^{(p+1)}(0) \right| \leq |\mathcal{P}_{p+1,2}| \cdot \max_{k \in [p+1]} \sup_{x, v = \mathbb{R}} |\Psi^{(k)}(x)| \leq e^{\frac{p+1}{2} \log(p+1)} \cdot e^{\frac{2}{5(p+1)} \log(\frac{2}{5}(p+1))} \leq e^{3p \log p + cp} \]
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 \( k \)th telephone number [20, Lem. 2]) has bound \( |\mathcal{P}_{k,2}| \leq e^{\frac{k}{2} \log k} \). This gives the result.