Lower Bounds for Linear Minimization Oracle Methods Optimizing over Strongly Convex Sets

Benjamin Grimmer* Ning Liu†

Abstract

We consider the oracle complexity of constrained convex optimization given access to a Linear Minimization Oracle (LMO) for the constraint set and a gradient oracle for the L-smooth, strongly convex objective. This model includes Frank-Wolfe methods and their many variants. Over the problem class of strongly convex constraint sets S, our main result proves that no such deterministic method can guarantee a final objective gap less than ε in fewer than $\Omega(\sqrt{L\,\mathrm{diam}(S)^2/\varepsilon})$ iterations. Our lower bound matches, up to constants, the accelerated Frank-Wolfe theory of Garber and Hazan [1]. Together, these establish this as the optimal complexity for deterministic LMO methods over strongly convex constraint sets. Second, we consider optimization over β-smooth sets, finding that in the modestly smooth regime of $\beta = \Omega(1/\sqrt{\varepsilon})$, no complexity improvement for span-based LMO methods is possible against either compact convex sets or strongly convex sets.

1 Introduction

In this work, we consider convex constrained optimization problems of the form

$$\begin{cases} \min & f(x) \\ \text{s.t.} & x \in S \subseteq \mathbb{R}^d. \end{cases}$$

In particular, we consider high-dimensional problems where d may be arbitrarily large. Frank-Wolfe methods and the broader family of "projection-free" algorithms using a linear minimization subroutine have found renewed interest due to their scalability. See the survey [2] and references therein. In a basic form, these methods iterate from an initialization x_0, producing for $k = 0, 1, \ldots$

$$\begin{aligned} p_k &= \nabla f(x_k), \\ z_{k+1} &\in \mathrm{argmin}_{x\in S}\langle p_k, x\rangle, \\ x_{k+1} &\in \mathrm{conv}\{x_k, z_{k+1}\}. \end{aligned}$$

As examples, an exact line search implementation of Frank-Wolfe would set x_{k+1} as the minimizer of f on the segment $[x_k, z_{k+1}]$. Similarly, a fixed "open-loop" stepsize implementation would fix $x_{k+1} = \theta_k x_k + (1-\theta_k) z_{k+1}$ for predetermined $\theta_k \in [0, 1]$. The two key computational oracles assumed here are access to gradients of the objective f to compute p_k and access to a Linear Minimization Oracle (LMO) to produce z_{k+1}. As a shorthand, we denote $\mathtt{LMO}_S(p) \in \mathrm{argmin}_{x\in S}\langle p, x\rangle$ as an oracle providing a selection of this minimizer (potentially selected adversarially).

*Johns Hopkins University, Department of Applied Mathematics and Statistics, grimmer@jhu.edu
†Johns Hopkins University, Department of Applied Mathematics and Statistics, nliu15@jhu.edu

In this work, we provide complexity lower bounds for the family of all deterministic first-order methods using a linear minimization oracle, containing the above Frank-Wolfe methods. Our focus is on lower bounds for problem classes where the constraint set S possesses structural properties like strong convexity or smoothness. Below, we formalize these classes of problems and algorithms.

The Families of Smooth and Strongly Convex Constraint Sets and Functions. A problem instance is defined by a function $f\colon \mathbb{R}^d \to \mathbb{R}$ and a set $S \subseteq \mathbb{R}^d$ as well as an initialization x_0, which we will set as $x_0 = 0$ without loss of generality. A problem class is defined by a set of allowable f and S values. We consider the standard family of convex objective functions parameterized by $0 \leq \mu \leq L$, defined as

$$\begin{aligned} f \text{ is } \mu\text{-strongly convex if} \quad & f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) - \frac{\lambda(1-\lambda)\mu}{2}\|y-x\|_2^2 \\ & \qquad \forall x, y \in \mathbb{R}^d, \lambda \in [0,1], \\ f \text{ is } L\text{-smooth if} \quad & \|\nabla f(y) - \nabla f(x)\|_2 \leq L\|y-x\|_2 \quad \forall x, y \in \mathbb{R}^d. \end{aligned}$$

If $L = \mu$, then f must be a quadratic function of the form $\frac{L}{2}\|x - x_\star\|_2^2$ up to an additive constant. Such simple quadratics will suffice for our theoretical development.

We consider compact convex sets S with diameter $\operatorname{diam}(S) = \max_{x,y\in S}\|x-y\|_2$ with at least one of the natural parallel notions of strong convexity and smoothness for constraint sets. We say

$$S \text{ is } \alpha\text{-strongly convex if} \quad \lambda x + (1-\lambda)y + B\left(0,\ \frac{\lambda(1-\lambda)\alpha}{2}\|y-x\|_2^2\right) \subseteq S$$

for all $x, y \in S, \lambda \in [0,1]$ where $B(0,r) = \{x \in \mathbb{R}^d \mid \|x\|_2 \leq r\}$ denotes the closed ball of radius r. Normal vectors $n \in N_S(x) := \{n \mid \langle n, y - x\rangle \leq 0\ \forall y \in S\}$ at each $x \in \operatorname{bdry} S$ serve as the analogues of gradients. Considering any unit length normal vectors, we define

$$\begin{aligned} S \text{ is } \beta\text{-smooth if} \quad \|n_y - n_x\|_2 \leq \beta\|y-x\|_2 \quad & \forall x, y \in \operatorname{bdry} S, \\ & n_y \in \{n \in N_S(y) \mid \|n\|_2 = 1\}, \\ & n_x \in \{n \in N_S(x) \mid \|n\|_2 = 1\}. \end{aligned}$$

Note that the above definition requires that smooth sets have a unique unit normal vector at each boundary point. For a classical reference providing equivalent characterizations of the strong convexity of a set, see [3]. A modernized treatment of smoothness and strong convexity of sets was given by [4]. Therein, parallels to the smoothness and strong convexity of functions are explored.

The Family of First-Order Linear Minimization Oracle Methods (`FO-LMO`). For a given problem instance (f, S) and initialization $x_0 \in S$, a `FO-LMO` method generates sequences of search directions $p_k \in \mathbb{R}^d$, linear minimization solutions $z_{k+1} = \mathtt{LMO}_S(p_k) \in \operatorname{argmin}_{x\in S}\langle p_k, x\rangle$, and iterates x_{k+1}. We require that p_k is a deterministic function of the oracle responses $\{(f(x_i), \nabla f(x_i))\}_{i=0}^k$ and $\{z_i\}_{i=1}^k$ so far and that x_{k+1} is a deterministic function of the responses $\{(f(x_i), \nabla f(x_i))\}_{i=0}^k$ and $\{z_i\}_{i=1}^{k+1}$. Note that we do not require that x_{k+1} lie in the convex hull of x_0 and $\{z_i\}_{i=1}^{k+1}$. However, when outside this convex hull, the iterates may fail to be feasible. This `FO-LMO` family of methods includes, for example, the Away-step Frank–Wolfe methods [5,6], Pairwise Frank–Wolfe methods [5,7], Fully Corrective Frank–Wolfe methods [5,8], and Blended Pairwise Frank–Wolfe methods [9].

In this model, each iteration of the algorithm can make one call to a first-order oracle, returning $(f(x_k), \nabla f(x_k))$, and one call to the LMO. Our theory then bounds iteration complexity to measure the minimum number of such pairs of oracle calls needed to reach a target suboptimality.

For the minimization of an L-smooth convex function f over a compact convex set S, Frank-Wolfe methods [10, 11] are known to provide convergence rates of the following form

$$f(x_T) - \min_{x\in S} f(x) \le \mathcal{O}\left(\frac{L\,\mathrm{diam}(S)^2}{T}\right).$$

This provides an upper bound on the iteration complexity of computing a point with ε-suboptimality (i.e., $f(x_T) - \min_{x\in S} f(x) \le \varepsilon$) of $\mathcal{O}(L\,\mathrm{diam}(S)^2/\varepsilon)$. A matching complexity lower bound was provided by Lan [12], establishing this as the order of the optimal LMO complexity. Garber and Hazan [1] showed that given the additional structure that f and S are strongly convex, this rate can be accelerated to

$$f(x_T) - \min_{x\in S} f(x) \le \mathcal{O}\left(\frac{L\,\mathrm{diam}(S)^2}{T^2}\right)$$

giving an iteration complexity of $\mathcal{O}(\sqrt{L\,\mathrm{diam}(S)^2/\varepsilon})$. This accelerated result poses the natural question of whether further acceleration is possible or if this complexity is the optimal strongly convex order.

Providing recent progress on this question, Halbey et al. [13] showed that two particular variants of Frank-Wolfe (exact line search and the short stepsize procedure) cannot improve upon this $\mathcal{O}(1/T^2)$ rate. However, these results do not preclude the possibility of other `FO-LMO` methods exceeding this rate. Similar limitations in unconstrained minimization were overcome by the foundational work [14], establishing information/oracle complexity lower bounds against the whole family of gradient methods. This was accomplished by the design of hard functions whose gradient reveals only one new coordinate of information per step. This property, known as a "zero-chain" property, can prevent *any gradient-span method* from having made substantial progress until dimension-many steps have been taken. By combining this with an adversarial "resisting oracle", lower bounds against *all deterministic gradient methods* were achieved.

Our Contributions. We extend the classical zero-chain approach to LMOs to derive complexity lower bounds on any `FO-LMO` method over the problem classes corresponding to minimizing a quadratic $f(x) = \frac{L}{2}\|x - x_\star\|_2^2$ over structured sets, possessing either strong convexity or smoothness. Our primary result shows that over α-strongly convex constraints, no `FO-LMO` method can guarantee ε-suboptimality in fewer than $\Omega(\sqrt{L\,\mathrm{diam}(S)^2/\varepsilon})$ iterations. By matching the upper bounding guarantee of [1], this establishes $\Theta(\sqrt{L\,\mathrm{diam}(S)^2/\varepsilon})$ as the optimal order of complexity for this class. Formally, we prove the following in Section 3.

Theorem 1.1. *Consider any $T \ge 1$ and $L, \alpha > 0$. Then there exist an α-strongly convex set S and an L-smooth, L-strongly convex function f in dimension $d = 2(T+1)$ such that every `FO-LMO` method applied to (f, S) with $x_0 = 0$ has x_T infeasible or*

$$f(x_T) - \min_{x\in S} f(x) \ge \frac{1}{528}\frac{L\,\mathrm{diam}(S)^2}{(T+1)^2}.$$

Consequently, for any $\varepsilon > 0$, there exist problem instances where $T \ge \sqrt{\frac{L\,\mathrm{diam}(S)^2}{528\varepsilon}} - 1$ iterations are required for any `FO-LMO` method to reach a suboptimality of $f(x_T) - \min_{x\in S} f(x) \le \varepsilon$.

While Theorem 1.1 applies for any $L, \alpha > 0$, it does not allow a free selection of $\mathrm{diam}(S)$. Rather, our constructed hard "zero-chain" set S has $\mathrm{diam}(S) = \Theta(1/\alpha d)$. One cannot arbitrarily select α and $\mathrm{diam}(S)$ as they must satisfy, for example, that $\mathrm{diam}(S) \le 2/\alpha$. In the limit where $\mathrm{diam}(S) = 2/\alpha$, the set S is forced to be (up to translation) the ball $B(0, 1/\alpha)$. Such forced structure enables faster algorithms[1]. The development of hard instances for any selection of $\mathrm{diam}(S) \le 2/\alpha$ and resulting more nuanced complexity bounds is left as an important future direction.

As a secondary result, we provide a partial generalization to β-smooth sets. These bounds are meaningful in the regime of only modestly smooth sets, having $\beta = \Omega(1/\sqrt{\varepsilon})$. In this regime, no `LMO-span` method (see Section 2 for a formal definition) can improve past the optimal compact convex set complexity of $\Theta(L\,\mathrm{diam}(S)^2/\varepsilon)$ or over the optimal α-strongly convex set complexity of $\Theta(\sqrt{L\,\mathrm{diam}(S)^2/\varepsilon})$. Theorems 4.1 and 4.2 formalize these limits on acceleration due to smoothness.

Since our constructions throughout use an L-smooth, L-strongly convex objective, all of our lower bounds apply against the wider class of problems with smooth, strongly convex objectives having any $0 \le \mu \le L$. Hence, our theory highlights a fundamental difficulty of constrained optimization via LMOs: *Even on perfectly conditioned objective functions, hard adversarial constraint sets constitute a barrier to linear convergence.* This stands in contrast to gradient methods with an orthogonal projection oracle where convergence is dominated by objective function conditioning.

Outline. Section 2 first derives a lower bound via a novel construction of a strongly convex feasible region that is hard for all `FO-LMO` methods with an additional span restriction. Section 3 removes this restriction, proving Theorem 1.1. This construction is the main technical innovation of our work. Section 4 then provides results extending lower bounds to the specialized setting of sets with modest levels of smoothness.

1.1 Related Work

Our Theorem 1.1 differs from the previously mentioned lower bounding result of [13] in two aspects, making the result complementary. In terms of algorithmic scope, our bound provides a wider guarantee, establishing a universal lower bound against all `FO-LMO` methods. In contrast, [13] provides a hard instance for two standard implementations of Frank-Wolfe, namely those fixing $p_k = \nabla f(x_k)$ and using an exact line search or the short stepsize procedure to select x_{k+1}. In terms of problem class scope, our construction requires a nonsmooth feasible region and a large problem dimension d (linear in the number of iterations to be run). Such a high dimensionality assumption (i.e., $d > T$) is classical and widespread in the optimization complexity literature [14]. In contrast, [13] is able to provide a hard instance using a smooth ball in $d = 2$ dimensions. As a result, they provide a stronger illustration of the limitations of the two methods that their theory covers.

As mentioned above, most classical lower bounding results in unconstrained first-order optimization rely on setting the problem dimension larger than the number of iterations to be conducted. This enables "zero-chain" arguments where the objective function is designed to reveal only one new coordinate to the given gradient-span algorithm at each iteration. The main technical innovation of our work is the design of a hard strongly convex set where the LMO possesses a similar zero-chain property to these classical hard objective constructions. Such constructions open the possibility to extend proof techniques and insights from existing unconstrained optimization lower bounds to constrained LMO/projection-free settings.

For example, although our strongly convex convergence rate lower bound matches the upper bound of [1] in order, the two differ by constant factors: our lower bound has a universal constant of $1/528$ and their upper bound has $9/2$. This leaves open the question of determining exactly minimax optimal algorithms and hard problem instances. In settings of unconstrained first-order minimization, the Performance Estimation Problem (PEP) techniques pioneered by [15–17] have provided such

[1]For example, when $\mathrm{diam}(S) = 2/\alpha$ and $S = B(0, 1/\alpha)$, an LMO can be used to explicitly compute orthogonal projections onto the feasible region S. Then the class of `FO-LMO` methods includes projected gradient methods, which are known to converge linearly for smooth, strongly convex objectives.

theory. For example, in smooth convex optimization, these facilitated the identification of the Optimized Gradient Method [18] and an exactly matching lower bound [19]. PEP was extended to cover structured smooth and strongly convex sets by Luner and Grimmer [20]. As a result, future work may leverage PEP to similarly close the constant-factor gap left for `FO-LMO` methods.

Another important property of most Frank-Wolfe methods is that their trajectory is independent of the choice of inner product used. This differs from projected-gradient methods, where the choice of inner product and notion of orthogonality for projections affects the algorithm trajectory, making good preconditioning important for practical success. The line of work [21–23] provided Frank-Wolfe with "affine-covariant" convergence theory, matching the method's affine covariant nature by avoiding notions like smoothness and strong convexity defined in terms of a fixed inner product. Developing lower bounding theory in affine-covariant terms is an interesting future direction.

One can view the oracle model of an LMO as assuming first-order access to the support function $p \mapsto \sup\{\langle p, x\rangle \mid x \in S\}$. For sets with $0 \in S$, a polar transformation establishes this as dual to assuming first-order access to the Minkowski gauge $y \mapsto \inf\{\gamma > 0 \mid y/\gamma \in S\}$. This gauge model has been studied by the works [24–28] as an alternative projection-free framework. The works [4] and [29] developed accelerated $\mathcal{O}(1/T^2)$ convergence guarantees for gauge methods over smooth sets, complementing the accelerated LMO rates for strongly convex sets. Identifying any structural relationships between the complexity of these dual oracle models and problem settings is another interesting direction.

2 Lower Bounds for Strongly Convex Sets and Span Methods

In this section, we develop the core ideas and constructions underlying our lower bounding theory. We do this against a weakened family of `FO-LMO` methods, called `LMO-span` methods, restricted to select search directions and iterates from fixed spans and convex hulls. Using a "resisting oracle", in the next section, our constructions here will extend to lower bounds for any `FO-LMO` method.

A Family of First-Order Linear Minimization Oracle Span Methods (`LMO-span`). For a given problem instance (f, S) and initialization $x_0 \in S$, a `LMO-span` method generates sequences of search directions p_k, linear minimization solutions z_{k+1}, and iterates x_{k+1} as follows for $k = 0, 1, \dots$

$$p_k \in \operatorname{span}\left\{\{x_i - x_0\}_{i=1}^k \cup \{z_i - x_0\}_{i=1}^k \cup \{\nabla f(x_i)\}_{i=0}^k\right\} \setminus \{0\}, \tag{2.1}$$

$$z_{k+1} = \mathtt{LMO}_S(p_k) \in \operatorname{argmin}_{x\in S}\langle p_k, x\rangle, \tag{2.2}$$

$$x_{k+1} \in \operatorname{conv}\{x_0, z_1, \dots, z_{k+1}\}. \tag{2.3}$$

No computational restrictions are placed on how p_k and x_{k+1} are computed. So operations like exact line searches and any usage of a memory/bundle of past gradients are allowed within this model. We also remark that the restriction that $p_k \neq 0$ is for ease of our development and without loss of generality. An adversarial LMO given $p_k = 0$ can return any $z_{k+1} \in S$ as a minimizer. In particular, it could return the prior iterate x_k, providing the algorithm with no new information.

The following theorem provides universal lower bounds against any such span method applied to a strongly convex problem in the high-dimensional regime where $d > T$.

Theorem 2.1. *For any $d \geq 1$ and $L, \alpha > 0$, there exist an α-strongly convex set S and an L-smooth, L-strongly convex function f such that every `LMO-span` method applied to (f, S) with $x_0 = 0$ has*

$$f(x_t) - \min_{x\in S} f(x) \geq \frac{2}{5}\frac{(d-t)L\operatorname{diam}(S)^2}{(d+2)^3} \qquad \forall t \in \{0, \dots, d-1\}. \tag{2.4}$$

In particular, for any fixed budget $T \geq 1$, there exist S and f in dimension $d = 2(T+1)$ such that

$$f(x_T) - \min_{x \in S} f(x) \geq \frac{1}{20} \frac{L \operatorname{diam}(S)^2}{(T+2)^2}.$$

Consequently, for any $\varepsilon > 0$, there exist problem instances where $T \geq \sqrt{\frac{L \operatorname{diam}(S)^2}{20\varepsilon}} - 2$ iterations are required for any `LMO-span` method to reach a suboptimality of $f(x_T) - \min_{x\in S} f(x) \leq \varepsilon$.

Our proof of this result is developed in three parts. To reduce notation, without loss of generality, we assume $L = 1$ and $\alpha = 1$ throughout. These two values can be ensured by rescaling in objective values and in variable space: Given a 1-smooth function f and a 1-strongly convex set S guaranteeing for any `LMO-span` that

$$f(x_T) - \min_{x \in S} f(x) \geq \frac{1}{20} \frac{\operatorname{diam}(S)^2}{(T+2)^2},$$

one could consider the equivalent rescaled L-smooth and α-strongly convex problem on

$$\tilde{f}(x) := \frac{L}{\alpha^2} f(\alpha x), \qquad \tilde{S} := \frac{1}{\alpha} S = \{x/\alpha \mid x \in S\}. \tag{2.5}$$

Under these rescalings, any `LMO-span` method has $\tilde{f}(x_T) - \min_{x\in\tilde{S}} \tilde{f}(x) \geq \frac{1}{20} \frac{L \operatorname{diam}(\tilde{S})^2}{(T+2)^2}$.

First, Section 2.1 constructs our candidate hard problem instance for each dimension and verifies its validity (computing its strong convexity constant and diameter). Then Section 2.2 establishes a key "zero-chain" property of these hard instances, showing that any `LMO-span` method applied will only discover one new coordinate per iteration. Finally, by leveraging this property, Section 2.3 proves this section's main result, Theorem 2.1, showing that no method in T steps can guarantee a suboptimality less than $\Omega(L \operatorname{diam}(S)^2/T^2)$. Combined with the upper bound of [1], this establishes the optimal iteration complexity for strongly convex constrained `LMO-span` optimization as $\Theta(\sqrt{L \operatorname{diam}(S)^2/\varepsilon})$.

2.1 Construction of a Hard Problem Instance

For any given problem dimension d, we construct our hard problem instance as follows: The feasible region is defined by

$$S := \left\{ x \in \mathbb{R}^d \mid \frac{1}{2}\|x\|_2^2 + \sum_{i=1}^d w_i |x_i| \leq C^2 \right\} \tag{2.6}$$

where the normalization constant C and a sequence of weights w_i are defined as

$$C := \frac{1}{\sqrt{d^2 + d + 2}}, \qquad w_i := C\sqrt{2i} \quad \text{for } i = 1, \ldots, d. \tag{2.7}$$

Note that $x_0 = 0$ is feasible, lying in the interior of S. The objective function is constructed to have its minimizer $x_\star = \nu \mathbf{1}$ where ν is the unique positive root of

$$\frac{1}{2} d\nu^2 + \left(\sum_{i=1}^d w_i \right) \nu = C^2. \tag{2.8}$$

This choice guarantees that $x_\star$ lies on the boundary of S and hence is feasible. We then set $f(x) := \frac{1}{2}\|x - x_\star\|_2^2$, ensuring that the minimum of f over S is zero.

In the remainder of this subsection, we verify that S has strong convexity constant $\alpha = 1$ in Lemma 2.1 and compute its diameter in Lemma 2.2. Together, these establish the validity of the considered problem instance defined by f and S.

Lemma 2.1 (Strong Convexity of S)**.** *The constructed set S in* (2.6) *is $\alpha = 1$-strongly convex.*

Proof. We prove this by using the fact that strong convexity is preserved under intersections [3, Proposition 2]. Hence, it suffices to show that S is equal to an intersection of 1-strongly convex sets. We do this below by reformulating S into the intersection of 2^d shifted unit balls $\{x \in \mathbb{R}^d \mid \|x + s \circ w\|_2 \le 1\}$ where $s \circ w = (s_1 w_1, \ldots, s_d w_d)$ for a given sign vector $s \in \{\pm 1\}^d$. Namely, one has that

$$\begin{aligned}
S &= \left\{ x \in \mathbb{R}^d \mid \max_{s \in \{\pm 1\}^d} \left\{ \frac{1}{2}\|x\|_2^2 + \sum_{i=1}^d w_i s_i x_i \right\} \le C^2 \right\} \\
&= \bigcap_{s \in \{\pm 1\}^d} \left\{ x \in \mathbb{R}^d \mid \frac{1}{2}\|x\|_2^2 + \sum_{i=1}^d s_i w_i x_i \le C^2 \right\} \\
&= \bigcap_{s \in \{\pm 1\}^d} \left\{ x \in \mathbb{R}^d \mid \|x\|_2^2 + 2\sum_{i=1}^d s_i w_i x_i + \|w\|_2^2 \le 2C^2 + \|w\|_2^2 \right\} \\
&= \bigcap_{s \in \{\pm 1\}^d} \left\{ x \in \mathbb{R}^d \mid \|x + s \circ w\|_2^2 \le 1 \right\}
\end{aligned}$$

where the first equality uses the identity $|u| = \max_{s \in \{\pm 1\}} su$ component-wise, the second rewrites this as an intersection, the third multiplies by two and adds the constant $\|w\|_2^2$ to each inequality, and the fourth completes the square and notes that the definition of C and the fact that $\|w\|_2^2 = C^2(d^2 + d)$ ensure that $1 = 2C^2 + \|w\|_2^2$. Since each shifted unit ball is 1-strongly convex, so is S. ☐

Lemma 2.2 (Exact Diameter of S)**.** *The diameter of S is* $\mathrm{diam}(S) = 2C(2 - \sqrt{2})$.

Proof. Since S is centrally symmetric, its diameter is $\mathrm{diam}(S) = 2\max_{x \in S} \|x\|_2$. Hence we maximize $\|x\|_2$ subject to the defining constraint (2.6). Since the weights are increasing, one can lower bound the weighted sum $\sum_{i=1}^d w_i |x_i|$ in terms of this two-norm as

$$\sum_{i=1}^d w_i |x_i| \ge w_1 \sum_{i=1}^d |x_i| = C\sqrt{2}\|x\|_1 \ge C\sqrt{2}\|x\|_2.$$

Both inequalities above hold with equality if and only if $x_i = 0$ for all $i \ge 2$. Applying this bound to the defining constraint of (2.6), every $x \in S$ must have $\frac{1}{2}\|x\|_2^2 + C\sqrt{2}\|x\|_2 \le C^2$. Completing the square gives $\frac{1}{2}(\|x\|_2 + C\sqrt{2})^2 - C^2 \le C^2$, which simplifies directly to $(\|x\|_2 + C\sqrt{2})^2 \le 4C^2$. Hence $\|x\|_2 \le C(2 - \sqrt{2})$. This upper bound is attained when the support of x is restricted to the first coordinate. The boundary points $(\pm C(2 - \sqrt{2}), 0, \ldots, 0) \in S$ then attain $\mathrm{diam}(S) = 2C(2 - \sqrt{2})$. ☐

2.2 A Zero-Chain Property

Lemma 2.3 establishes that linear minimization over the constructed S corresponds to a soft-thresholding operation. This form is essential for making S a hard instance for all LMO-based algorithms. In the subsequent Lemma 2.4, we present our key technical result: this LMO has a "zero-chain" property that prevents an `LMO-span` algorithm from discovering more than one new nonzero coordinate per iteration.

Lemma 2.3 (Exact LMO)**.** *For any search direction $p \in \mathbb{R}^d \setminus \{0\}$, the linear minimization oracle $z = \mathtt{LMO}_S(p) \in \mathrm{argmin}_{x \in S}\langle p, x\rangle$ is given elementwise by the soft-thresholding operator*

$$z_i = -\mathrm{sign}(p_i) \max\left(0, \frac{|p_i|}{\lambda} - w_i\right), \tag{2.9}$$

where $\lambda > 0$ is the unique KKT multiplier ensuring $\frac{1}{2}\|z\|_2^2 + \sum_{i=1}^d w_i|z_i| = C^2$.

Proof. Let $h(x) = \frac{1}{2}\|x\|_2^2 + \sum_{i=1}^d w_i|x_i|$. The LMO solves $\min_x \langle p, x\rangle$ subject to $h(x) \le C^2$. Since the feasible region is compact, this is attained by some z. Since the constraint set is strongly convex and the objective is a non-constant linear function (i.e., $p \neq 0$), this z is the unique minimizer. Further, it must lie on the boundary of the set, making the constraint $h(z) = C^2$ active. Slater's condition holds here as the origin is in the interior ($h(0) = 0 < C^2$). Hence, the KKT conditions provide a multiplier $\lambda \ge 0$ with $0 \in p + \lambda\partial h(z)$. Since $p \neq 0$, λ must be positive. The Lagrangian is

$$\mathcal{L}(x, \lambda) = \langle p, x\rangle + \lambda\left(h(x) - C^2\right).$$

We consider the optimality condition for each z_i with coordinate $i \in \{1, \dots, d\}$ separately. These require that $0 \in p_i + \lambda(z_i + w_i\partial|z_i|)$, giving the key inclusion $-\frac{p_i}{\lambda} \in z_i + w_i\partial|z_i|$. We consider the three possible cases on the value of z_i: positive, negative, or equal to zero.

- **Case $z_i > 0$:** Here $\partial|z_i| = \{1\}$, requiring $-\frac{p_i}{\lambda} = z_i + w_i$. Since $z_i > 0$ and $w_i > 0$, this implies $p_i = -\lambda(w_i + z_i) < 0$. Thus $\operatorname{sign}(p_i) = -1$ and so $z_i = \frac{|p_i|}{\lambda} - w_i$.
- **Case $z_i < 0$:** Here $\partial|z_i| = \{-1\}$, requiring $-\frac{p_i}{\lambda} = z_i - w_i$. Since $z_i < 0$ and $w_i > 0$, this implies $p_i = \lambda(w_i - z_i) > 0$. Thus $\operatorname{sign}(p_i) = 1$ and so $z_i = -\left(\frac{|p_i|}{\lambda} - w_i\right)$.
- **Case $z_i = 0$:** Here $\partial|z_i| = [-1, 1]$, requiring $-\frac{p_i}{\lambda} \in [-w_i, w_i]$. This directly evaluates to the condition $\frac{|p_i|}{\lambda} \le w_i$.

Combining these three cases gives the elementwise soft-thresholding form in (2.9). Since $H(\lambda) = \frac{1}{2}\|z(\lambda)\|_2^2 + \sum_{i=1}^d w_i|z_i(\lambda)|$ is continuous, decreasing in λ whenever $H(\lambda)$ is positive, and ranges from ∞ to 0, there is a unique $\lambda > 0$ with the defining constraint active $H(\lambda) = C^2$. □

Lemma 2.4 (Zero-Chain Property)**.** *Suppose at iteration $t \le d - 2$, all past iterations $k \le t$ satisfy $(x_k)_i = 0$ and $(z_k)_i = 0$ for all $i \ge t + 1$. Then for any search direction p_t satisfying* (2.1)*, the LMO point $z_{t+1} = \mathtt{LMO}_S(p_t) \in \operatorname{argmin}_{z\in S}\langle p_t, z\rangle$ satisfies*

$$(z_{t+1})_i = 0 \qquad \forall i \in \{t + 2, \dots, d\}.$$

Proof. Noting $x_0 = 0$, any valid search direction p_t must take the form $p_t = \sum_{k=0}^t \alpha_k\nabla f(x_k) + \sum_{k=1}^t \beta_k x_k + \sum_{k=1}^t \gamma_k z_k$ for some multipliers $\alpha_k, \beta_k, \gamma_k \in \mathbb{R}$. Since $(x_k)_i = 0$ for all $i \ge t + 1$ and $x_\star = \nu\mathbf{1}$, each past gradient $\nabla f(x_k) = x_k - x_\star$ has $(\nabla f(x_k))_i = -\nu$ for all $i \ge t + 1$.

$$(p_t)_i = \sum_{k=0}^t \alpha_k(-\nu) + \sum_{k=1}^t \beta_k(0) + \sum_{k=1}^t \gamma_k(0) = -\nu\sum_{k=0}^t \alpha_k =: \sigma.$$

If $\sigma = 0$, the LMO (2.9) evaluates to 0 for all remaining coordinates, satisfying the lemma.

Now assume $\sigma \neq 0$, and suppose for contradiction that $(z_{t+1})_j \neq 0$ for some $j \ge t + 2$. Since $|(p_t)_j| = |\sigma|$, the soft-thresholding operator formula (2.9) requires this coordinate to satisfy $\frac{|\sigma|}{\lambda} > w_j$. Since the weights w_i are increasing, $\frac{|\sigma|}{\lambda} > w_{t+2} > w_{t+1}$. It follows that for coordinate $t + 1$, which also has $|(p_t)_{t+1}| = |\sigma|$, the LMO formula (2.9) must return a strictly positive absolute value

$$|(z_{t+1})_{t+1}| = \frac{|\sigma|}{\lambda} - w_{t+1} > w_{t+2} - w_{t+1} > 0.$$

This strict inequality and the monotonicity of $u \mapsto \frac{1}{2}u^2 + w_{t+1}u$ on $u > 0$ imply that

$$\begin{aligned}\frac{1}{2}|(z_{t+1})_{t+1}|^2 + w_{t+1}|(z_{t+1})_{t+1}| &> \frac{1}{2}(w_{t+2} - w_{t+1})^2 + w_{t+1}(w_{t+2} - w_{t+1}) \\ &= \frac{1}{2}(w_{t+2}^2 - 2w_{t+2}w_{t+1} + w_{t+1}^2) + w_{t+1}w_{t+2} - w_{t+1}^2 \\ &= \frac{1}{2}(w_{t+2}^2 - w_{t+1}^2) = C^2\end{aligned}$$

where the final equality uses that by definition, adjacent squared weights satisfy $w_{t+2}^2 - w_{t+1}^2 = 2C^2$. However, this strict inequality contradicts the feasibility of the LMO output z_{t+1} as the sum $\sum_{i=1}^d \left(\frac{1}{2}z_i^2 + w_i|z_i|\right)$ must also strictly exceed C^2 since all other summands are nonnegative. Hence, the LMO output must have $(z_{t+1})_j = 0$ for all $j \geq t+2$. $\square$

2.3 Proof of Theorem 2.1

We first inductively show that for any considered `LMO-span` method, at each iteration t, $(x_t)_i = 0$ for all coordinates $i \geq t+1$ and $(z_{t+1})_i = 0$ for all coordinates $i \geq t+2$. When $t = 0$, $x_0 = 0$ and so $(x_0)_i = 0$ for $i \geq 1$. Applying Lemma 2.4 then gives the needed property for z_1. Now, suppose that for all $k \leq t-1$, x_k is supported only on its first k coordinates and z_{k+1} is supported only on its first $k+1$ coordinates. Since $x_t \in \operatorname{conv}\{0, z_1, \dots, z_t\}$ and each of these LMO solutions is zero on all coordinates greater than t, it follows that $(x_t)_i = 0$ for $i \geq t+1$. Then applying Lemma 2.4 restricts the support of z_{t+1}, guaranteeing that this returned LMO point has $(z_{t+1})_i = 0$ for all $i \geq t+2$, completing the induction.

From this, we can conclude that for any iteration $t \leq d-1$, the suboptimality is lower bounded because the remaining $d-t$ coordinates of x_t are all forced to equal zero:

$$f(x_t) - f(x_\star) = \frac{1}{2}\|x_t - x_\star\|_2^2 \geq \frac{1}{2}\sum_{i=t+1}^d (0-\nu)^2 = \frac{d-t}{2}\nu^2. \tag{2.10}$$

Therefore, to establish the theorem, it suffices to provide a lower bound on ν^2.

Let $W = \sum_{i=1}^d w_i$. Applying the quadratic formula to (2.8) and rationalizing the numerator provides the following formula for ν of

$$\nu = \frac{-W + \sqrt{W^2 + 2dC^2}}{d} = \frac{2C^2}{W + \sqrt{W^2 + 2dC^2}}. \tag{2.11}$$

The summation W is upper bounded by $\frac{2\sqrt{2}}{3}C(d+1)^{3/2}$ via the following integral upper bound

$$\sum_{i=1}^d \sqrt{i} \leq \int_0^d \sqrt{x+1}\,dx = \left[\frac{2}{3}(x+1)^{3/2}\right]_0^d = \frac{2}{3}\left((d+1)^{3/2} - 1\right) < \frac{2}{3}(d+1)^{3/2}.$$

Since the formula (2.11) for ν is decreasing in W, we can lower bound ν^2 by

$$\nu^2 \geq \frac{(2C^2)^2}{C^2\left(\frac{2\sqrt{2}}{3}(d+1)^{3/2} + \sqrt{\frac{8}{9}(d+1)^3 + 2d}\right)^2} \geq \frac{9C^2}{8(d+2)^3}$$

where the final step bounds $\left(\frac{2\sqrt{2}}{3}(d+1)^{3/2}+\sqrt{\frac{8}{9}(d+1)^3+2d}\right)^2 \leq \frac{32}{9}(d+2)^3$, which is valid[2] for all $d \geq 1$. Finally, we recall from Lemma 2.2 that the diameter of S is $2C(2-\sqrt{2})$. Rearranging, we have that $C^2 = \frac{3+2\sqrt{2}}{8}\mathrm{diam}(S)^2$. Hence by (2.10), the suboptimality after t iterations of any `LMO-span` is bounded by

$$f(x_t) - f(x_\star) \geq \frac{d-t}{2}\,\frac{9(3+2\sqrt{2})\mathrm{diam}(S)^2}{64(d+2)^3} \geq \frac{2}{5}\frac{(d-t)\mathrm{diam}(S)^2}{(d+2)^3}$$

with the last inequality reducing the absolute constant to a simpler lower bound, $\frac{9(3+2\sqrt{2})}{128} \geq \frac{2}{5}$.

For a fixed budget of iterations $T \geq 1$, fix $d = 2(T+1)$ in the above construction. Then our suboptimality lower bound at $t = T$ becomes

$$f(x_T) - f(x_\star) \geq \frac{2}{5}\frac{(T+2)\ \mathrm{diam}(S)^2}{(2T+4)^3} = \frac{1}{20}\frac{\mathrm{diam}(S)^2}{(T+2)^2}.$$

3 A Resisting Oracle Extension to Deterministic FO-LMO Methods

The lower bound provided in Theorem 2.1 establishes that no `LMO-span` method can achieve an optimization rate faster than $\Omega(L\,\mathrm{diam}(S)^2/T^2)$. However, if the span restriction is relaxed, one can design a method capable of exactly solving any single fixed hard problem. Trivially, one could consider the algorithm that always returns $\nu\mathbf{1}$ for every problem instance. Although often not an effective algorithm, this would exactly solve our previously proposed hard instance. To resolve this, Section 3.1 constructs an adversarial variant of our previous hard problem. Section 3.2 derives a "zero-chain" property for this adversarial LMO, resisting any `FO-LMO` method. Finally, Section 3.3 proves Theorem 1.1.

3.1 A Permuted Family of Hard Problem Instances

Let $\mathcal{P}_d$ denote the symmetric group of all permutations on $\{1,\dots,d\}$. Given a permutation $\pi \in \mathcal{P}_d$, we define the permuted constraint set

$$S_\pi := \left\{ x \in \mathbb{R}^d \mid \frac{1}{2}\|x\|_2^2 + \sum_{i=1}^d w_{\pi(i)}|x_i| \leq C^2 \right\}, \tag{3.1}$$

where C and $w_i = C\sqrt{2i}$ are defined identically to (2.7). Since S_π is a coordinate permutation of the original set S, it retains the exact same diameter and $\alpha = 1$ strong convexity as S.

For any $d \geq 3$, we define the 1-smooth, 1-strongly convex objective function as

$$f(x) := \frac{1}{2}\|x - M\mathbf{1}\|_2^2, \tag{3.2}$$

where $M := \frac{\rho}{1-\rho}w_d$ with $\rho \in (0,1)$ as the unique scalar satisfying

$$\sum_{i=1}^d \left[\frac{1}{2}\rho^2(w_d - w_i)^2 + \rho w_i(w_d - w_i)\right] = C^2. \tag{3.3}$$

[2]To verify this algebraic bound, let $x = \frac{8}{9}(d+1)^3$. Then the left-hand side equals $(\sqrt{x}+\sqrt{x+2d})^2 = 2x + 2d + 2\sqrt{x^2+2dx}$. Since $\sqrt{x^2+2dx} \leq x + d$, this is bounded above by $\frac{32}{9}(d+1)^3 + 4d$. For $d \geq 1$, this can be further bounded above by $\frac{32}{9}(d+2)^3$, giving the final simplified bound.

Existence and uniqueness follow since the above polynomial is increasing for $\rho > 0$, evaluating to 0 at $\rho = 0$, and $\sum_{i=1}^{d} \frac{1}{2}(w_d^2 - w_i^2) = C^2 \frac{d(d-1)}{2} > C^2$ at $\rho = 1$ since $d \geq 3$. The following lemma computes the true constrained optimum $x_\star^{(\pi)} = \operatorname{argmin}_{x \in S_\pi} f(x)$ and provides a lower bound on ρ.

Lemma 3.1 (The Permuted Problem Minimizer)**.** *For any $\pi \in \mathcal{P}_d$, the optimal solution $x_\star^{(\pi)}$ to $\min_{x \in S_\pi} f(x)$ is unique and given elementwise by*

$$(x_\star^{(\pi)})_i = \rho\left(w_d - w_{\pi(i)}\right) \quad \forall i \in \{1, \dots, d\}.$$

Furthermore, the scalar $\rho \in (0,1)$ is bounded below by $\rho \geq 2/d^2$.

Proof. We verify optimality of the proposed solution, defined elementwise by $(x_\star^{(\pi)})_i = \rho(w_d - w_{\pi(i)})$, by showing that $x_\star^{(\pi)} - M\mathbf{1} + \gamma(x_\star^{(\pi)} + \zeta) = 0$ for some subgradient $\zeta \in \partial(\sum_{i=1}^{d} w_{\pi(i)}|u_i|)(x_\star^{(\pi)})$ and multiplier $\gamma \geq 0$. In particular, we set $\gamma = \frac{\rho}{1-\rho}$. Since $w_d \geq w_{\pi(i)}$, we have $(x_\star^{(\pi)})_i \geq 0$. Further, when $w_{\pi(i)} < w_d$, we have $(x_\star^{(\pi)})_i > 0$ and we can set the required subgradient as $\zeta_i = w_{\pi(i)}$. Then the needed KKT condition holds by substituting $\rho + \frac{\rho^2}{1-\rho} = \gamma = \frac{\rho}{1-\rho}$ and $M = \gamma w_d$ into

$$\rho(w_d - w_{\pi(i)}) - M + \frac{\rho}{1-\rho}(\rho(w_d - w_{\pi(i)}) + w_{\pi(i)}) = 0.$$

For the unique component where $w_{\pi(i)} = w_d$, we have $(x_\star^{(\pi)})_i = 0$, and the KKT condition requires $-M + \gamma\zeta_i = 0$, implying $\zeta_i = M/\gamma = w_d$. Since $w_d \in [-w_d, w_d]$, this is a valid subgradient. Finally, substituting $(x_\star^{(\pi)})_i$ into the defining boundary constraint of S_π yields the polynomial equation (3.3), confirming feasibility.

Finally we lower bound ρ. Observe that for each coordinate i, one has

$$\rho(w_d - w_i)\left[\frac{1}{2}\rho(w_d - w_i) + w_i\right] \leq \rho(w_d - w_i)\left[\frac{1}{2}(w_d - w_i) + w_i\right] = \frac{1}{2}\rho(w_d^2 - w_i^2).$$

Summing over i yields $C^2 \leq \frac{1}{2}\rho \sum_{i=1}^{d}(w_d^2 - w_i^2) = \frac{1}{2}\rho C^2 d(d-1)$, and so $\rho \geq \frac{2}{d(d-1)} \geq \frac{2}{d^2}$. ∎

3.2 A Generalized Zero-Chain Property

Any first-order query at an iterate x_k yields a gradient $\nabla f(x_k) = x_k - M\mathbf{1}$. Since M is independent of the choice of permutation π, this gradient for any x_k reveals no information about π. Thus `FO-LMO` algorithms must use the LMO to reveal information about $x_\star^{(\pi)}$. We formalize the information revealed by the LMO via an adversarial "resisting oracle" that selects the permutation π at runtime. The oracle maintains a set $K_t \subseteq \{1, \dots, d\}$ of assigned coordinates, initially empty. At each iteration, when the algorithm queries the LMO with an arbitrary vector $p_t \neq 0$, the oracle selects an unassigned coordinate attaining the largest query magnitude

$$i_\star \in \operatorname{argmax}_{i \notin K_t} |(p_t)_i|.$$

The adversary assigns this coordinate the smallest available weight, setting $\pi(i_\star) = |K_t| + 1$ and $K_{t+1} = K_t \cup \{i_\star\}$. The oracle will then be able to compute the exact LMO output using only this partial assignment, leaving the permutation on $\{1, \dots, d\} \setminus K_{t+1}$ undefined. Alternatively, if given $p_t = 0$, we define the oracle output as $z_{t+1} = 0$, which is feasible for all S_π and reveals no new information.

Lemma 3.2 (Generalized Zero-Chain Property). *Under the resisting oracle's assignment, for any p_t, the returned LMO point $z_{t+1} = \mathtt{LMO}_{S_\pi}(p_t)$ has $(z_{t+1})_j = 0$ for all unassigned coordinates $j \notin K_{t+1}$, regardless of how the remainder of π is eventually assigned.*

Proof. This result is immediate if $p_t = 0$ as $z_{t+1} = 0$. As a result, we can assume $p_t \neq 0$ and let $\lambda > 0$ be the KKT multiplier for the LMO query. Assume there exists some $j \notin K_{t+1}$ with $(z_{t+1})_j \neq 0$. From the elementwise soft-thresholding formula derived in (2.9), this requires $\frac{|(p_t)_j|}{\lambda} > w_{\pi(j)}$. Since $i_\star$ was chosen as a maximizer of the search direction's magnitude among all unassigned coordinates, we have $|(p_t)_{i_\star}| \geq |(p_t)_j|$. Furthermore, the oracle assigns weights monotonically. So any unassigned coordinate j will eventually receive a weight $w_{\pi(j)} \geq w_{|K_t|+2}$. Thus, we have

$$\frac{|(p_t)_{i_\star}|}{\lambda} \geq \frac{|(p_t)_j|}{\lambda} > w_{\pi(j)} \geq w_{|K_t|+2}.$$

It follows that the LMO output z_{t+1} at coordinate $i_\star$ is nonzero and lower bounded absolutely by

$$|(z_{t+1})_{i_\star}| = \frac{|(p_t)_{i_\star}|}{\lambda} - w_{|K_t|+1} > w_{|K_t|+2} - w_{|K_t|+1} > 0.$$

Since the function $u \mapsto \frac{1}{2}u^2 + w_{|K_t|+1}u$ is increasing for $u > 0$, we have that

$$\frac{1}{2}(z_{t+1})_{i_\star}^2 + w_{|K_t|+1}|(z_{t+1})_{i_\star}| > \frac{1}{2}(w_{|K_t|+2}^2 - w_{|K_t|+1}^2) = C^2,$$

where the last equality follows from the definition $w_k^2 = 2C^2k$. However, this violates the constraint $\sum_{i=1}^d (\frac{1}{2}z_i^2 + w_{\pi(i)}|z_i|) \leq C^2$ since all other terms in the sum are nonnegative. From this contradiction, we conclude that z_{t+1} is zero on all remaining coordinates, regardless of how the permutation π is completed. □

3.3 Proof of Theorem 1.1

By Lemma 3.2, our resisting LMO reveals at most one coordinate of the permutation π at each iteration. After T iterations, the set of assigned coordinates K_T has size at most T. Let $U = \{1, \ldots, d\} \setminus K_T$ denote the remaining coordinates, with $m = |U| \geq d - T$.

Any considered `FO-LMO` method must set x_T deterministically as a function of the history of oracle responses. Crucially, while the algorithm must fix $(x_T)_i$ for $i \in U$, the resisting oracle remains free to select any completion of the permutation π over U. Let $\mathcal{P}_U \subseteq \mathcal{P}_d$ denote all such permutations. If some completion $\pi \in \mathcal{P}_U$ has $x_T \notin S_\pi$, then the adversarial oracle can select that π, making x_T infeasible and the theorem hold trivially. Hence, we can assume for all $\pi \in \mathcal{P}_U$ that $x_T \in S_\pi$. From the strong convexity of f and optimality of $x_\star^{(\pi)}$ guaranteeing $\langle \nabla f(x_\star^{(\pi)}), x_T - x_\star^{(\pi)} \rangle \geq 0$ since $x_T \in S_\pi$, the suboptimality for any $\pi \in \mathcal{P}_U$ is bounded by

$$f(x_T) - f(x_\star^{(\pi)}) \geq \frac{1}{2}\|x_T - x_\star^{(\pi)}\|_2^2 \geq \frac{1}{2}\sum_{j \in U}\left((x_T)_j - (x_\star^{(\pi)})_j\right)^2. \tag{3.4}$$

Let $\pi_\star \in \mathcal{P}_U$ denote the permutation choice maximizing this lower bound. Then, we have

$$\begin{aligned} f(x_T) - f(x_\star^{(\pi_\star)}) &\geq \frac{1}{2}\sum_{j \in U}\left((x_T)_j - (x_\star^{(\pi_\star)})_j\right)^2 \\ &\geq \frac{1}{|\mathcal{P}_U|}\sum_{\pi \in \mathcal{P}_U}\left(\frac{1}{2}\sum_{j \in U}\left((x_T)_j - (x_\star^{(\pi)})_j\right)^2\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\sum_{j\in U}\frac{1}{|\mathcal{P}_U|}\sum_{\pi\in\mathcal{P}_U}\left((x_T)_j - (x_\star^{(\pi)})_j\right)^2 \\
&\geq \frac{m}{2}\mathrm{Var}\left\{\rho(w_d - w_k)\right\}_{k=d-m+1}^{d} \\
&= \frac{\rho^2 m}{2}\mathrm{Var}\{w_k\}_{k=d-m+1}^{d} \\
&= \rho^2 m C^2 \mathrm{Var}\{\sqrt{k}\}_{k=d-m+1}^{d}
\end{aligned}
$$

where the first inequality is by (3.4), the second inequality lower bounds this maximal $\pi_\star$ by the average lower bound over $\mathcal{P}_U$, and the third inequality notes that each inner sum is minimized when $(x_T)_j$ is the average of $\{\rho(w_d - w_{d-m+1}), \dots, \rho(w_d - w_d)\}$ yielding the stated variance lower bound.

To derive a lower bound on the above variance, recall that the variance over a uniform distribution of m elements is $\mathrm{Var}(X) = \frac{1}{2m^2}\sum_{i,j}(X_i - X_j)^2$. For $X_i = \sqrt{i}$, we can lower bound $(\sqrt{i}-\sqrt{j})^2 = (i-j)^2/(\sqrt{i}+\sqrt{j})^2 \geq (i-j)^2/4d$. So $\mathrm{Var}(X) \geq \mathrm{Var}(i)/4d = (m^2-1)/48d$, using that the variance of m contiguous integers is $(m^2-1)/12$. Further, note that for $d = 2T+2$, we have $m \geq T+2$. Then applying that $\rho \geq 2/d^2$ and $C^2 = \frac{3+2\sqrt{2}}{8}\mathrm{diam}(S_\pi)^2$, we conclude that there exists a problem instance resisting any given `FO-LMO` method, having

$$
f(x_T) - f(x_\star^{(\pi_\star)}) \geq \rho^2 m C^2 \frac{m^2-1}{48d} \geq \frac{3+2\sqrt{2}}{8}\frac{\mathrm{diam}(S_\pi)^2}{384(T+1)^2} > \frac{1}{528}\frac{\mathrm{diam}(S_\pi)^2}{(T+1)^2}
$$

where the last inequality above just reduces to a simpler fractional coefficient. Hence, f and $S_{\pi^\star}$ together constitute a hard problem instance for the given first-order method, adversarially constructed. The case of general $L, \alpha > 0$ follows by considering the rescaling formulas (2.5).

4 An Extension of Lower Bounds to Modestly Smooth Sets

Finally, we provide a direct approach to extend our span-based lower bounds to guarantees against methods applied to β-smooth sets. We do this by considering the smoothing of hard instances given by taking Minkowski sums with a ball $B(0, 1/\beta)$. For sufficiently large values of $\beta = \Omega(1/\sqrt{\varepsilon})$, we show that this perturbation of the problem instance cannot notably improve worst-case performance. As a result, we find that no acceleration is possible for `LMO-span` methods on only modestly smooth sets. We leave open whether acceleration is possible when sets possess constant levels of smoothness. The numerical survey of [20] using performance estimation techniques suggested that for small values of T, no clear big-O acceleration could be numerically identified for more general ranges of β.

Below, we provide lower bounds for optimization over smooth convex sets and over smooth, strongly convex sets. In the $\beta = \Omega(1/\sqrt{\varepsilon})$ regime, our smooth convex set lower bound matches the general optimal complexity for optimization over convex sets of [11,12]. Similarly, in this regime, our smooth, strongly convex set lower bound matches the optimal strongly convex complexity established in the previous section of $\Omega(\sqrt{L\,\mathrm{diam}(S)^2/\varepsilon})$.

Theorem 4.1. *For any $d \geq 1$ and $\beta > 0$, there exist a convex β-smooth set S_β and a 1-smooth, 1-strongly convex function f_β such that every `LMO-span` method applied to (f_β, S_β) starting from $x_0 = 0$ has for all $t \leq d-1$*

$$
f_\beta(x_t) - \min_{x\in S_\beta} f_\beta(x) \geq \max\left(0, \sqrt{\frac{d-t}{2}}\left(\frac{\mathrm{diam}(S_\beta) - 2/\beta}{\sqrt{2}d} + \frac{1}{\beta\sqrt{d}}\right) - \frac{1}{\sqrt{2}\beta}\right)^2. \tag{4.1}
$$

In particular, for any fixed budget $T \geq 1$, there exist S_β and f_β in dimension $d = 2T$ such that

$$f_\beta(x_T) - \min_{x \in S_\beta} f_\beta(x) \geq \max\left(0, \frac{\mathrm{diam}(S_\beta) - 2/\beta}{4\sqrt{T}} - \frac{\sqrt{2}-1}{2\beta}\right)^2.$$

Theorem 4.2. *For any $d \geq 1$ and $\beta > 0$, there exist a β-smooth, $1/(1+1/\beta)$-strongly convex set S_β and a 1-smooth, 1-strongly convex function f_β such that every* `LMO-span` *method applied to (f_β, S_β) starting from $x_0 = 0$ has for all $t \leq d-1$*

$$f_\beta(x_t) - \min_{x \in S_\beta} f_\beta(x) \geq \max\left(0, \sqrt{\frac{2}{5}\frac{(d-t)(\mathrm{diam}(S_\beta) - 2/\beta)^2}{(d+2)^3}} - \frac{1}{\sqrt{2}\beta}\right)^2. \tag{4.2}$$

In particular, for any fixed budget $T \geq 1$, there exist S_β and f_β in dimension $d = 2(T+1)$ such that

$$f_\beta(x_T) - \min_{x \in S_\beta} f_\beta(x) \geq \max\left(0, \frac{\mathrm{diam}(S_\beta) - 2/\beta}{\sqrt{20}(T+2)} - \frac{1}{\sqrt{2}\beta}\right)^2.$$

In both of these theorems, the lower bounds are only meaningful in the regime of $\beta = \Omega(1/\sqrt{\varepsilon})$ as otherwise, the two-term maximums above will take value zero, making the lower bound vacuous.

To prove these results, we use the following lemma which allows us to relate a set S with a suitable zero-chain property to an approximate zero-chain property holding on the set $S_\beta = S + B(0, 1/\beta)$ given by a Minkowski sum with a ball. Note that S_β is β-smooth for any closed convex set S by [4, Lemma 9]. Further, the LMO for S_β is given by summing the LMOs for S and $B(0, 1/\beta)$: by the additivity of linear minimization over Minkowski sums, it follows that

$$\mathtt{LMO}_{S_\beta}(p) = \mathtt{LMO}_S(p) - \frac{1}{\beta}\frac{p}{\|p\|_2}.$$

We find that $\mathtt{LMO}_{S_\beta}(p)$ can inherit an approximate zero-chain property from $\mathtt{LMO}_S(p)$ as follows.

Lemma 4.1 (Approximate Zero-Chain Property)**.** *Consider a compact convex set S and $f_\beta(x) = \frac{1}{2}\|x - \nu\mathbf{1}\|_2^2$ for some $\nu > 0$. Let $S_\beta = S + B(0, 1/\beta)$. Suppose that whenever, for all $k \leq t$, each of x_k and z_k has identical coordinates for all indices $i \geq t+1$, one can guarantee for any p_t satisfying* (2.1) *that the base-set LMO point $\hat{z}_{t+1} = \mathtt{LMO}_S(p_t)$ satisfies $(\hat{z}_{t+1})_i = 0$ for all $i \geq t+2$. Then, in this case, the smoothed-set LMO point $z_{t+1} = \mathtt{LMO}_{S_\beta}(p_t)$ satisfies*

$$(z_{t+1})_i = \delta_{t+1} \qquad \forall i \in \{t+2, \dots, d\}$$

where the common tail value δ_{t+1} is bounded by $|\delta_{t+1}| \leq \frac{1}{\beta\sqrt{d-t}}$.

Proof. By the hypothesis, for each $k \leq t$, the vectors x_k and z_k each have all coordinates $i \geq t+1$ equal (within each vector). Since $x_\star = \nu\mathbf{1}$, the gradients $\nabla f_\beta(x_k) = x_k - x_\star$ also have $(\nabla f_\beta(x_k))_i$ constant among all $i \geq t+1$. As a result, any valid search direction p_t satisfying (2.1) is constant among its coordinates $i \geq t+1$. Denote this shared scalar value by σ.

Let $\hat{z}_{t+1} = \mathtt{LMO}_S(p_t)$. By assumption, $(\hat{z}_{t+1})_i = 0$ for all $i \geq t+2$. Since $p_t \neq 0$ under (2.1), linear minimization over Minkowski sums gives

$$z_{t+1} = \mathtt{LMO}_{S_\beta}(p_t) = \mathtt{LMO}_S(p_t) - \frac{1}{\beta}\frac{p_t}{\|p_t\|_2} = \hat{z}_{t+1} - \frac{1}{\beta}\frac{p_t}{\|p_t\|_2}.$$

Therefore, for every $i \geq t+2$, the coordinate $(z_{t+1})_i$ is determined solely by the ball term, having $(z_{t+1})_i = -\frac{\sigma}{\beta\|p_t\|_2} =: \delta_{t+1}$. Hence, all such tail coordinates are equal. If $\sigma = 0$, then $\delta_{t+1} = 0$, satisfying the lemma's claim. If $\sigma \neq 0$, then since p_t contains at least $d - t$ copies of σ (at indices $t+1, \dots, d$), its two-norm is at least $\|p_t\|_2 \geq \sqrt{d-t}\,|\sigma|$. Rearrangement gives the claimed bound as $|\delta_{t+1}| = \frac{|\sigma|}{\beta\|p_t\|_2} \leq \frac{|\sigma|}{\beta\sqrt{d-t}\,|\sigma|} = \frac{1}{\beta\sqrt{d-t}}$. □

4.1 Proof of Theorem 4.1

Given any $d > 0$, consider the simplex $S = \left\{x \in \mathbb{R}^d \mid \sum_{i=1}^d x_i \leq 1, x_i \geq 0\right\}$. This choice follows directly from that of Lan [12] where a simplex was shown to possess a zero-chain property—each extreme point is a 1-sparse basis vector, from which an adversarial LMO can return the attaining basis vector with minimal index. From this, Lan provided a hard instance for any LMO method applied to general convex constrained optimization. Fixing any $\beta > 0$, we define

$$S_\beta := S + B(0, 1/\beta).$$

The diameter of S_β is $\mathrm{diam}(S_\beta) = \mathrm{diam}(S) + 2/\beta = \sqrt{2} + 2/\beta$. To keep the target minimizer on the boundary of this set, we define the optimal solution $x_\star = (1/d + 1/(\beta\sqrt{d}))\mathbf{1}$. We continue as in previous constructions by setting $f_\beta(x) := \frac{1}{2}\|x - x_\star\|_2^2$.

Inductively applying Lemma 4.1, the iterate $x_t \in \mathrm{conv}\{x_0, z_1, \ldots, z_t\}$ must have final $d - t$ coordinates $i \geq t + 1$ equal with value bounded by $|(x_t)_i| \leq \max_{1 \leq k \leq t} |\delta_k| \leq \frac{1}{\beta\sqrt{d-t+1}}$. Then the suboptimality $f_\beta(x_t) - f_\beta(x_\star) = \frac{1}{2}\|x_t - x_\star\|_2^2$ is lower bounded by

$$\begin{aligned}
f_\beta(x_t) - f_\beta(x_\star) &\geq \frac{1}{2}\sum_{i=t+1}^{d}\left((x_t)_i - \left(\frac{1}{d} + \frac{1}{\beta\sqrt{d}}\right)\right)^2 \\
&\geq \frac{d-t}{2}\max\left(0, \frac{1}{d} + \frac{1}{\beta\sqrt{d}} - \frac{1}{\beta\sqrt{d-t+1}}\right)^2 \\
&\geq \max\left(0, \sqrt{\frac{d-t}{2}}\left(\frac{\mathrm{diam}(S_\beta) - 2/\beta}{\sqrt{2}d} + \frac{1}{\beta\sqrt{d}}\right) - \frac{1}{\sqrt{2}\beta}\right)^2.
\end{aligned}$$

Specializing to $t = T$ and $d = 2T$, this provides a lower bound of

$$f_\beta(x_T) - f_\beta(x_\star) \geq \max\left(0, \frac{\mathrm{diam}(S_\beta) - 2/\beta}{4\sqrt{T}} - \frac{\sqrt{2}-1}{2\beta}\right)^2.$$

4.2 Proof of Theorem 4.2

Given any $d > 0$, consider the set S previously defined in (2.6) and the parameter ν defined by (2.8). Fix any $\beta > 0$ and define

$$S_\beta := S + B(0, 1/\beta).$$

From Lemma 2.2, it is immediate that the diameter of S_β is $\mathrm{diam}(S_\beta) = \mathrm{diam}(S) + 2/\beta = 2C(2 - \sqrt{2}) + 2/\beta$. Further, S_β is β-smooth and $\alpha = 1/(1 + 1/\beta)$-strongly convex by the calculus rules for Minkowski sums [4, Lemma 9].

To keep the target minimizer on the boundary of this set, we define $x_\star = \nu_\beta \mathbf{1}$ with ν_β as the unique positive root of

$$\min_{x \in S}\|\nu_\beta \mathbf{1} - x\|_2 = \frac{1}{\beta}.$$

Since $0 \in \mathrm{int}(S)$, the distance from $\eta\mathbf{1}$ to the set S is zero for $0 \leq \eta \leq \nu$ and strictly increasing for $\eta > \nu$, mapping $[\nu, \infty)$ to $[0, \infty)$. Thus, this equation uniquely defines a scalar $\nu_\beta > \nu$. Given this, we continue as in previous constructions by setting $f_\beta(x) := \frac{1}{2}\|x - x_\star\|_2^2$.

Inductively applying the approximate zero-chain property of S_β, the iterate $x_t \in \mathrm{conv}\{x_0, z_1, \dots, z_t\}$ must have final $d-t$ coordinates $i \geq t+1$ equal with value bounded by $|(x_t)_i| \leq \max_{1\leq k\leq t} |\delta_k| \leq \frac{1}{\beta\sqrt{d-t+1}}$. Then the suboptimality is lower bounded by

$$f_\beta(x_t) - f_\beta(x_\star) \geq \frac{1}{2}\sum_{i=t+1}^{d} ((x_t)_i - \nu_\beta)^2 \geq \frac{d-t}{2}\max\left(0, \nu_\beta - \frac{1}{\beta\sqrt{d-t+1}}\right)^2.$$

By the bounds developed in Section 2.3, $\nu_\beta \geq \nu \geq \frac{2}{\sqrt{5}(d+2)^{3/2}}\mathrm{diam}(S)$. Hence

$$\begin{aligned} f_\beta(x_t) - f_\beta(x_\star) &\geq \frac{d-t}{2}\max\left(0, \frac{2(\mathrm{diam}(S_\beta) - 2/\beta)}{\sqrt{5}(d+2)^{3/2}} - \frac{1}{\beta\sqrt{d-t+1}}\right)^2 \\ &\geq \max\left(0, \sqrt{\frac{2}{5}\frac{(d-t)(\mathrm{diam}(S_\beta) - 2/\beta)^2}{(d+2)^3}} - \frac{1}{\sqrt{2}\beta}\right)^2. \end{aligned}$$

Specializing to $t = T$ and $d = 2(T+1)$, this provides a lower bound of

$$f_\beta(x_T) - f_\beta(x_\star) \geq \max\left(0, \frac{\mathrm{diam}(S_\beta) - 2/\beta}{\sqrt{20}(T+2)} - \frac{1}{\sqrt{2}\beta}\right)^2.$$

Acknowledgments. Benjamin Grimmer was supported as an Alfred P. Sloan Foundation fellow.

References

[1] Dan Garber and Elad Hazan. Faster rates for the Frank-Wolfe method over strongly-convex sets. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 541–549. PMLR, 2015.

[2] Gábor Braun, Alejandro Carderera, Cyrille W. Combettes, Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Sebastian Pokutta. *Conditional gradient methods: From core principles to AI applications*. MOS-SIAM Series on Optimization. SIAM, 2025.

[3] Jean-Philippe Vial. Strong convexity of sets and functions. *Journal of Mathematical Economics*, 9(1-2):187–205, 1982.

[4] Ning Liu and Benjamin Grimmer. Gauges and accelerated optimization over smooth and/or strongly convex sets, 2023. arXiv:2303.05037.

[5] Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. *Advances in neural information processing systems*, 28, 2015.

[6] Amir Beck and Shimrit Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 164(1):1–27, 2017.

[7] Kazuma K Tsuji, Ken'ichiro Tanaka, and Sebastian Pokutta. Pairwise conditional gradients without swap steps and sparser kernel herding. In *International Conference on Machine Learning*, pages 21864–21883. PMLR, 2022.

[8] Charles A Holloway. An extension of the Frank and Wolfe method of feasible directions. *Mathematical Programming*, 6(1):14–27, 1974.

[9] Gábor Braun, Sebastian Pokutta, Dan Tu, and Stephen Wright. Blended conditional gradients. In *International conference on machine learning*, pages 735–743. PMLR, 2019.

[10] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[11] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pages 427–435. PMLR, 2013.

[12] Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle, 2013. arXiv:1309.5550.

[13] Jannis Halbey, Daniel Deza, Max Zimmer, Christophe Roux, Bartolomeo Stellato, and Sebastian Pokutta. Lower bounds for Frank-Wolfe on strongly convex sets, 2026. arXiv:2602.04378.

[14] Arkadij Semenovič Nemirovski and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Chichester, 1983.

[15] Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482, 2014.

[16] Adrien Taylor, Julien Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161:307–345, 2017.

[17] Adrien Taylor, Julien Hendrickx, and François Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27(3):1283–1313, 2017.

[18] Donghwan Kim and Jeffrey A Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical programming*, 159(1):81–107, 2016.

[19] Yoel Drori. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39:1–16, 2017.

[20] Alan Luner and Benjamin Grimmer. Performance estimation for smooth and strongly convex sets, 2024. arXiv:2410.14811.

[21] Javier F. Peña. Affine invariant convergence rates of the conditional gradient method. *SIAM Journal on Optimization*, 33(4):2654–2674, 2023.

[22] Elias Wirth, Javier Peña, and Sebastian Pokutta. Accelerated affine-invariant convergence rates of the Frank–Wolfe algorithm with open-loop step-sizes. *Mathematical Programming*, 214(1–2):201–245, 2025.

[23] Elias Wirth, Javier Peña, and Sebastian Pokutta. Fast convergence of Frank-Wolfe algorithms on polytopes. *Mathematics of Operations Research*, 2025.

[24] James Renegar. "Efficient" Subgradient Methods for General Convex Optimization. *SIAM Journal on Optimization*, 26(4):2649–2676, 2016.

[25] Benjamin Grimmer. Radial duality part i: foundations. *Mathematical Programming*, 205(1–2):33–68, 2024.

[26] Benjamin Grimmer. Radial duality part ii: applications and algorithms. *Mathematical Programming*, 205(1–2):69–105, 2023.

[27] Zakaria Mhammedi. Efficient projection-free online convex optimization with membership oracle. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5314–5390, London, UK, 2022. PMLR.

[28] Zhou Lu, Nataly Brukhim, Paula Gradu, and Elad Hazan. Projection-free adaptive regret with membership oracles. In Shipra Agrawal and Francesco Orabona, editors, *International Conference on Algorithmic Learning Theory, February 20-23, 2023, Singapore*, volume 201 of *Proceedings of Machine Learning Research*, pages 1055–1073, Singapore, 2023. PMLR.

[29] Thabo Samakhoana and Benjamin Grimmer. Scalable projection-free optimization methods via multiradial duality theory, 2024. arXiv:2403.13688.