

Performance Estimation for Smooth and Strongly Convex Sets

Alan Luner* Benjamin Grimmer†

Abstract

We extend recent computer-assisted design and analysis techniques for first-order optimization over structured functions—known as performance estimation—to apply to structured sets. We prove “interpolation theorems” for smooth and strongly convex sets with interior point conditions and bounded diameter, showing a wide range of extremal questions amount to structured mathematical programs. Prior function interpolation theorems are recovered as a limit of our set interpolation theory. Our theory provides finite-dimensional formulations of performance estimation problems for algorithms utilizing separating hyperplane oracles and linear optimization oracles of smooth/strongly convex sets. As applications of this computer-assisted machinery, we identify a minimax optimal separating hyperplane method and areas for improvement in the theory of Frank-Wolfe and non-Lipschitz Smooth Optimization.

1 Introduction

Given a proposed algorithm, the study of its worst-case convergence guarantees over some family of problem instances can be framed as a meta-optimization problem. In this work, we will consider examples of first-order optimization methods applied to problems from a structured family of objective functions and constraint sets. Given some measure of performance (e.g., final objective gap), the exact worst-case performance of the method corresponds to computing the maximum of this measure at the algorithm’s output over all problem instances. This is known as the Performance Estimation Problem (PEP). In general, optimizing over a family of functions and sets is an infinite-dimensional problem, typically beyond direct approach. Surprisingly, recent advances have shown that PEPs can often be computer-solved and have led to several state-of-the-art results in the design and analysis of optimization algorithms.

Performance estimation was first proposed by Drori and Teboulle in [1], where tractable relaxations of these worst-case PEPs were considered. Shortly afterward, the “interpolation theorems” of Taylor, Hendrickx, and Glineur [2, 3] proved that for many gradient methods and structured function classes, the PEP exactly corresponds to a tractable finite-dimensional problem. Specifically, their function interpolation results showed that the infinite-dimensional optimization problem of finding a worst-case problem instance for N steps of a given gradient method could be equivalently reformulated as a nonconvex Quadratically Constrained Quadratic Program (QCQP), which can often then be reformulated as a Semidefinite Program (SDP). As a result, for a modest fixed number of algorithm steps N , computers can be used to determine an algorithm’s worst-case problem instances and, dually, to produce optimal convergence proofs. Similar techniques from the perspective of control were developed by [4]. These computer-assisted techniques provide an invaluable starting point for formally proving improved theory for general N .

* Johns Hopkins University, Department of Applied Mathematics and Statistics, aluner1@jhu.edu

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

Since this initial application, performance estimation theory has been applied to many different settings and algorithms, including conditional gradient [3], proximal gradient [3, 5, 6], alternating projections [3, 7], stochastic gradient descent [8], line-search [9, 10], strong convexity [2, 10, 11], fixed point iteration [12, 13], and much more. See the growing collection of examples documented in the PEPit toolbox [14]¹.

Computer PEP solutions directly produce tight numerical convergence guarantees and, in many cases, have led to tight analytical convergence rates. They also have a record of directly improving the state of the art in algorithm design by enabling the optimization of parameter selections. In general, the task of finding algorithmic parameters (e.g., stepsizes and momentum sequences) minimizing the worst-case PEP guarantee is a nonconvex problem². Despite this apparent hardness, many recent works have found success at this task, often by observing structure in numerical results and devising corresponding analytical solutions. Some such advances include:

- For smooth convex optimization, Kim and Fessler [16] derived an Optimized Gradient Method, improving on Nesterov’s accelerated method by constants and attaining the minimax optimal rate by Drori [17]. In [18], they proposed an optimized method (OGM-G) for minimizing gradient norm.
- For smooth strongly convex optimization, Taylor and Drori [11] derived an exactly optimal method (ITEM) improving upon the asymptotically optimized triple momentum method of [19].
- Altschuler and Parrilo [20, 21] and Grimmer, Shu, and Wang [22, 23] showed that gradient descent with fractal stepsizes provably accelerates in big-O.

These advances on long-standing, best-known results are in large part due to the insights generated from performance estimation and computer assistance.

The primary aim of this work is the development of new interpolation theorems for smooth and strongly convex sets, potentially with interior point conditions and bounded diameter, paralleling the function-oriented theorems of Taylor et al. [2, 3]. Prior function interpolation theorems in [3] can be used to model simply convex constraint sets but do not capture additional structural properties of set smoothness and strong convexity. Our theorems enable the formulation of computer-solvable PEPs over these structured families of constraint sets. Formal definitions of smoothness and strong convexity of closed convex sets and functions (and several equivalent conditions) are given in Section 2. In brief, a closed convex function f is smooth and strongly convex if each subgradient provides certain quadratic upper and lower bounds on f , respectively. Similarly, a closed convex set C is smooth and strongly convex if each normal vector provides certain balls inner and outer approximating C .

Our Contributions. Our primary contribution is a collection of interpolation theorems for structured sets, extending the prior theory of [3] for convex sets. Our theorems provide tractable, finite representations for a variety of settings, including smooth and strongly convex sets, sets with bounded diameters, and sets with interior point conditions. See Theorems 3.1-3.3. These interpolations allow us to express the worst-case performance of many constrained optimization algorithms as numerically solvable performance estimation problems. Theorem 3.7 relates our theory to prior function interpolation theory.

Outline. Section 2 defines basic notations and presents the function interpolation theorems of [2]. Section 3 then presents our main result: interpolation theorems for structured sets. The remainder of this paper presents three applications, showing computational³ and modeling potential of our

¹Available at <https://pepit.readthedocs.io/en/latest/examples.html>.

²A branch-and-bound solver for this hard task was developed by Das Gupta et al. [15].

³All numerical results were obtained with Mosek [24] or Gurobi [25] via JuMP [26]. Code for reproducibility is

theory:

- Section 4 shows, given a separating hyperplane oracle for a set C , our interpolation theorems allow one to determine any proposed method's convergence to feasibility as a semidefinite program. Matching resulting numerics, we derive a simple, minimax optimal, separating hyperplane algorithm.
- Section 5 shows, for constrained optimization, that our interpolation theorems enable the performance of the Frank-Wolfe method to be modeled as a semidefinite program with separable nonconvex equality constraints. By solving such nonconvex problems (globally and locally), we survey the potential for the Frank-Wolfe method to accelerate given structured sets.
- Section 6 shows, for unconstrained optimization, that our interpolation theorems allow us to analyze optimization of functions with a smooth epigraph. Preliminary results, solving semidefinite programs with a rank constraint, show potential to generalize current accelerated theory.

2 Definitions and Preliminaries

Section 2.1 first introduces standard definitions of smooth and strongly convex functions and the function interpolation of Taylor et al. [2]. Section 2.2 then introduces the analogous concepts for sets in preparation for our main results.

2.1 Preliminaries on Smooth and Strongly Convex Functions

Consider a closed convex function $f: \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$. We denote $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ for the Euclidean inner product and associated norm. We denote the domain of f by $\text{dom } f$ and its subdifferential at some $x \in \text{dom } f$ by $\partial f(x) = \{g \mid f(y) \geq f(x) + \langle g, y - x \rangle \quad \forall y \in \mathbb{R}^d\}$. Each element $g \in \partial f(x)$ is referred to as a subgradient and must be the gradient $\nabla f(x)$ when f is differentiable.

Typically, L -smoothness of a function is defined as $\nabla f(x)$ being L -Lipschitz. For the sake of developing other symmetries, here we instead (equivalently) define a closed convex f as being **L -smooth** if for all $x, y \in \mathbb{R}^d$ and $g \in \partial f(x)$,

$$f(y) \leq f(x) + \langle g, y - x \rangle + \frac{L}{2} \|y - x\|^2. \quad (2.1)$$

That is, each subgradient provides a quadratic upper bound on f . Mirroring this, we say f is **μ -strongly convex** if for all $x, y \in \mathbb{R}^d$ and $g \in \partial f(x)$,

$$f(y) \geq f(x) + \langle g, y - x \rangle + \frac{\mu}{2} \|y - x\|^2. \quad (2.2)$$

That is, each subgradient provides a quadratic lower bound on f . We allow the extreme cases of $L, \mu \in \{0, \infty\}$, taking limits above. Every convex f is vacuously ∞ -smooth and 0-strongly convex. The only 0-smooth functions are linear and the only ∞ -strongly convex functions are indicators of points. Denote the set of all μ -strongly convex and L -smooth functions f by $\mathcal{F}_{\mu, L}$.

As a motivating example of the performance estimation problem, consider minimizing some $f \in \mathcal{F}_{0, L}$ via N steps of a gradient method, defined as

$$x_{i+1} = x_i - \frac{1}{L} \sum_{j=0}^i H_{i,j} \nabla f(x_j) \quad (\text{GM})$$

available at github.com/alanlunier/PEPStructuredSets.

with initial point x_0 and a lower triangular matrix H . Note this model captures simple methods like gradient descent via a diagonal H and more complex momentum methods like Nesterov's accelerated method [27] and the optimal gradient method of [16]. Suppose as an initial condition, we are guaranteed the initial distance to a minimizer $\|x_0 - x_\star\|$ is bounded above by R . Then the infinite-dimensional PEP of seeking a problem instance with objective in $\mathcal{F}_{0,L}$ achieving the largest final objective gap can be formulated as

$$p_S(L, R) := \begin{cases} \max_{x_i, f} & f(x_N) - f(x_\star) \\ \text{s.t.} & f \in \mathcal{F}_{0,L} \\ & x_{i+1} = x_i - \frac{1}{L} \sum_{j=0}^i H_{i,j} \nabla f(x_j) \\ & \|x_0 - x_\star\| \leq R, \quad \nabla f(x_\star) = 0. \end{cases} \quad (2.3)$$

Note that aside from the constraint $f \in \mathcal{F}_{0,L}$, this optimization problem only depends on f via its function value and gradient at finitely many points, namely $\{x_0, \dots, x_N, x_\star\}$. Denote these values by $f_i = f(x_i)$ and $g_i = \nabla f(x_i)$ for each $i \in \{0, \dots, N, \star\}$. The following definition of function interpolation from [2] allows us to reformulate into these more limited quantities:

Definition 2.1. Consider a set of observations $S = \{(x_i, g_i, f_i)\}_{i \in \mathcal{I}}$ where $x_i, g_i \in \mathbb{R}^d$ and $f_i \in \mathbb{R}$ for all $i \in \mathcal{I}$. The set S is $\mathcal{F}_{\mu,L}$ -**interpolable** if there exists a function $f \in \mathcal{F}_{\mu,L}$ such that $g_i \in \partial f(x_i)$ and $f(x_i) = f_i$ for all $i \in \mathcal{I}$.

Function interpolability describes whether there exists an appropriate function that matches the given finite set of function values and subgradients and interpolates between them (See Figure 2). In terms of the working example (2.3), this can then be expressed finitely as

$$p_S(L, R) = \begin{cases} \max_{x_i, g_i, f_i} & f_N - f_\star \\ \text{s.t.} & \{(x_i, g_i, f_i)\}_{i \in \{0, \dots, N, \star\}} \text{ is } \mathcal{F}_{0,L}\text{-interpolable} \\ & x_{i+1} = x_i - \frac{1}{L} \sum_{j=0}^i H_{i,j} g_j \\ & \|x_0 - x_\star\| \leq R, \quad g_\star = 0. \end{cases} \quad (2.4)$$

Although finite, this formulation is still not particularly tractable as the interpolability constraint is, at first glance, not approachable. The critical insight of Taylor et al. [2, Theorem 4] was establishing explicit necessary and sufficient conditions for function interpolability.

Theorem 2.2. ([2, Theorem 4]) *A set $S = \{(x_i, g_i, f_i)\}_{i \in \mathcal{I}}$ is $\mathcal{F}_{\mu,L}$ -interpolable if and only if the following quadratic condition holds for all $i, j \in \mathcal{I}$:*

$$f_i - f_j + \langle g_i, x_j - x_i \rangle + \frac{1}{2(1 - \mu/L)} \left(\frac{1}{L} \|g_i - g_j\|^2 + \mu \|x_i - x_j\|^2 - \frac{2\mu}{L} \langle g_i - g_j, x_i - x_j \rangle \right) \leq 0. \quad (2.5)$$

Plugging this into (2.4) yields an explicit QCQP exactly describing gradient descent's worst-case behavior over the considered family of structured functions. Semidefinite programming lifting techniques (and duality) can be further used to gain insights from this new formulation.

Remark 1. Throughout the literature and in this work, the *equivalent* reformulation of the QCQP into a semidefinite program depends on an assumption of high dimension (e.g., $d \geq N + 2$). This assumption is only required for the SDP to be an equivalent problem, rather than a relaxation. For any d , the resulting SDP provides a valid, dimension-independent upper bound.

It is worth noting that there are many collections of necessary constraints on the set of observation data S . Any standard inequality for μ -strongly convex, L -smooth functions must necessarily hold: for example, the definitions (2.1) and (2.2). However, as Taylor et al. [2] highlight, these other inequalities prove not to be sufficient for interpolation. Hence, identification of the right inequalities, e.g., (2.5), is key for the “if and only if” nature of interpolation theory and, consequently, its power in enabling computer assistance. Similarly, in Section 3, we find that while there are many necessary constraints for set interpolability, a careful selection is needed for sufficiency.

2.2 Preliminaries on Smooth and Strongly Convex Sets

We now switch our focus to smoothness and strong convexity of sets. These notions mirror their function counterparts; see [28–30] as classic references. These properties are, however, quite underexplored in the contemporary first-order method literature compared to their function counterparts, which are the backbone of much modern convergence theory. In part, this work aims to provide tools to eventually alleviate this discrepancy.

Consider a nonempty closed convex set $C \subseteq \mathbb{R}^d$. We denote the interior of a set C by $\text{int } C$ and its boundary by $\text{bdry } C$. We denote the normal cone of C at some $z \in \text{bdry } C$ by $N_C(z) = \{v \mid \langle v, x - z \rangle \leq 0 \ \forall x \in C\}$ and refer to individual elements as normal vectors. Let $B(x, r)$ be the closed ball of radius r centered at x . We define $\text{diam}(C) = \sup\{\|x - y\| \mid x, y \in C\}$ and the δ -interior of C by $\text{int}_\delta C = \{x \mid B(x, \delta) \subseteq C\}$. Note $\text{int}_0 C = C$. Lastly, we define the Minkowski sum of two sets C_1 and C_2 as

$$C_1 + C_2 = \{x + y \mid x \in C_1, y \in C_2\} .$$

We now define smoothness and strong convexity with respect to sets. For a more full discussion of the equivalent definitions of these properties and their analogs with smooth and strongly convex functions, we refer readers to [31, Section 2]. Overall, one can recover the analogous definitions for functions by replacing unit normal vectors with gradients, bounding balls with bounding quadratics, and sets with epigraphs. A set C is β -**smooth** if for any $z \in \text{bdry } C$ and unit vector $n \in N_C(z)$,

$$B(z - \frac{1}{\beta}n, \frac{1}{\beta}) \subseteq C .$$

A set C is α -**strongly convex** if for any $z \in \text{bdry } C$ and unit vector $n \in N_C(z)$,

$$B(z - \frac{1}{\alpha}n, \frac{1}{\alpha}) \supseteq C .$$

Again, we allow the limiting cases of $\alpha, \beta \in \{0, \infty\}$. Every convex set is 0-strongly convex and ∞ -smooth. The only ∞ -strongly convex sets are singletons and the only 0-smooth sets are halfspaces. Let $\mathcal{C}_{\alpha, \beta, D}$ denote the set of all closed convex sets C that are α -strongly convex, β -smooth, and have $\text{diam}(C) \leq D$. See an illustration of smooth and strongly convex sets in Figure 1.

Our discussion of set smoothness and strong convexity throughout will focus on these properties at full dimension. One could consider smoothness/strong convexity with respect to a fixed subspace. Through a simple change of variables, our definitions and theory could be applied.

Below are equivalent characterizations of set smoothness and strong convexity that are useful in our analysis, paralleling classic results for functions.

Proposition 2.3. ([31, Proposition 3]) *For any nonempty closed convex set C , the following are equivalent:*

1. C is β -smooth.

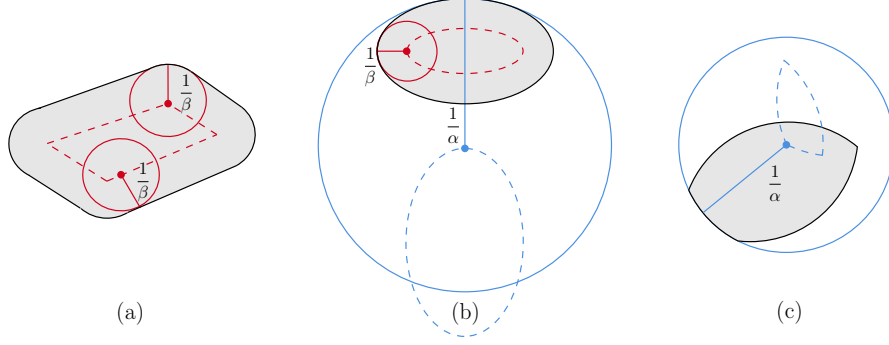


Figure 1: Example closed convex sets with red inner approximating balls from smoothness and blue outer approximating balls from strong convexity. Namely, (a) is β -smooth but not strongly convex, (b) is β -smooth and α -strongly convex, and (c) is α -strongly convex but not smooth. The red dotted lines indicate the set C_0 from Proposition 2.3 and the blue dotted lines indicate the set C_0 from Proposition 2.4. The blue dot marks the origin where applicable.

2. For any $z_1, z_2 \in \text{bdry } C$, with unit normals $n_1 \in N_C(z_1)$ and $n_2 \in N_C(z_2)$,

$$\langle z_1 - z_2, n_1 - n_2 \rangle \geq \frac{1}{\beta} \|n_1 - n_2\|^2.$$

3. There exists a closed convex set C_0 such that $C_0 + B(0, \frac{1}{\beta}) = C$.

Proposition 2.4. ([30, Theorem 2.1]) *For any nonempty closed convex set C , the following are equivalent:*

1. C is α -strongly convex.

2. For any $z_1, z_2 \in \text{bdry } C$, with unit normals $n_1 \in N_C(z_1)$ and $n_2 \in N_C(z_2)$,

$$\langle z_1 - z_2, n_1 - n_2 \rangle \leq \frac{1}{\alpha} \|n_1 - n_2\|^2.$$

3. There exists a closed convex set C_0 such that $C_0 + C = B(0, \frac{1}{\alpha})$.

Many more equivalent characterizations of strongly convex sets are given in [30]. The following pair of lemmas will also help develop our theory.

Lemma 2.5. *A set C is β -smooth with $\text{diam}(C) \leq D$ if and only if $C = C_0 + B(0, \frac{1}{\beta})$ for some convex set C_0 with $\text{diam}(C_0) \leq D - \frac{2}{\beta}$.*

Proof. We know from Proposition 2.3 that C is β -smooth if and only if there exists a convex set C_0 such that $C = C_0 + B(0, \frac{1}{\beta})$. So the result follows as

$$\begin{aligned} \text{diam}(C) &= \sup\{\|x + \frac{1}{\beta}n_x - (y + \frac{1}{\beta}n_y)\| \mid x, y \in C_0, \|n_x\| \leq 1, \|n_y\| \leq 1\} \\ &= \sup\{\|x - y\| \mid x, y \in C_0\} + \frac{2}{\beta} = \text{diam}(C_0) + \frac{2}{\beta}. \end{aligned} \quad \square$$

Lemma 2.6. *A set C is α -strongly convex and β -smooth if and only if $C = C_0 + B(0, \frac{1}{\beta})$ for some γ -strongly convex set C_0 , with $\gamma := (\frac{1}{\alpha} - \frac{1}{\beta})^{-1} = \frac{\alpha\beta}{\beta - \alpha}$.*

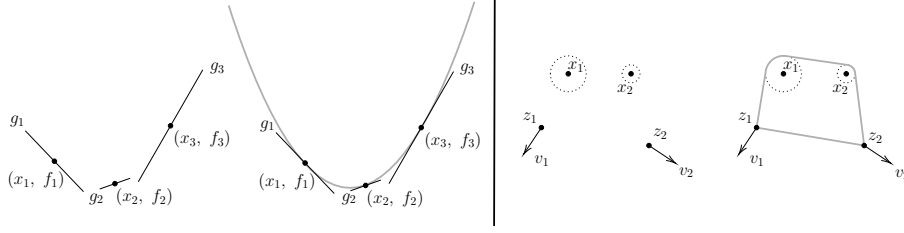


Figure 2: Left: Function interpolation of function values f_i and subgradients g_i evaluated at x_i . Right: Set interpolation of boundary points z_i with normals v_i and δ_k -interior points x_k .

Proof. This follows from property 3 in both Propositions 2.3 and 2.4. Namely, C is α -strongly convex and β -smooth if and only if there exist convex sets C_0^{sc} and C_0^{sm} such that $C + C_0^{\text{sc}} = B(0, \frac{1}{\alpha})$ and $C = C_0^{\text{sm}} + B(0, \frac{1}{\beta})$. Together, these properties imply $C_0^{\text{sm}} + B(0, \frac{1}{\beta}) + C_0^{\text{sc}} = B(0, \frac{1}{\alpha})$, and so $C_0^{\text{sm}} + C_0^{\text{sc}} = B(0, \frac{1}{\gamma})$, i.e., C_0^{sm} is γ -strongly convex. Conversely, if $C = C_0 + B(0, \frac{1}{\beta})$ for some γ -strongly convex C_0 , it immediately follows that C is β -smooth. Further some convex C'_0 must exist with $C_0 + C'_0 = B(0, \frac{1}{\gamma})$. Hence $C + C'_0 = B(0, \frac{1}{\alpha})$, i.e., C is α -strongly convex. $\square$

Finally, we define set interpolability, paralleling the function interpolability of Definition 2.1. In addition to smoothness and strong convexity, we model two additional structures: diameter bounds and various interior point conditions, which occur widely in constrained optimization theory.

We consider a set of observations S defined by two types of data: points $z_i \in \mathbb{R}^d$ on the boundary of the constraint set with nonzero normal vectors $v_i \in \mathbb{R}^d$ and points $x_k \in \mathbb{R}^d$ in the δ_k -interior of the set. Note the index sets $i \in \mathcal{I}$ and $k \in \mathcal{K}$ need not be of the same size or even both be nonempty. Given target strong convexity α , smoothness β , and diameter D , we can formalize whether given observation data can be interpolated by some such set.

Definition 2.7. Consider a set of observations $S = (\{(z_i, v_i)\}_{i \in \mathcal{I}}, \{(x_k, \delta_k)\}_{k \in \mathcal{K}})$, with $z_i, v_i, x_k \in \mathbb{R}^d$ and $\delta_k \in \mathbb{R}_{\geq 0}$ for all $i \in \mathcal{I}$ and all $k \in \mathcal{K}$. The set S is $\mathcal{C}_{\alpha, \beta, D}$ -**interpolable** if there exists $C \in \mathcal{C}_{\alpha, \beta, D}$ such that $z_i \in C$ and $v_i \in N_C(z_i)$ for all $i \in \mathcal{I}$ and $x_k \in \text{int}_{\delta_k} C$ for all $k \in \mathcal{K}$.

Note that this definition allows $\delta_k = 0$ (denoting the corresponding x_k simply must be a member of C) and all limiting values $\alpha, \beta, D \in \{0, \infty\}$.

This work's primary contribution is identifying computationally tractable, verifiable equivalent conditions for set interpolability. Our theory builds upon the interpolation theory over convex sets introduced in [3, Theorem 3.6].

Before presenting our theory in Section 3, we consider a simplified setting where $\alpha = 0$ and $\beta = \infty$. In this case, interpolability corresponds to determining if a bounded diameter convex set exists with the required normals $v_i \in N_C(z_i)$ and δ_k -interior points x_k . When the interpolation conditions introduced in Section 3 hold, a simple, explicit construction of the interpolating convex, bounded set exists: $C = \text{conv}(\{z_i\}_{i \in \mathcal{I}}, \{B(x_k, \delta_k)\}_{k \in \mathcal{K}})$. Figure 2 shows an example of this simplified case's construction. Designing and analyzing such a construction is the primary step in proving the sufficiency of given interpolation conditions. Our constructions grow substantially in complexity from this easy case, but remain explicit. As a result, one can always recover a (smooth, strongly convex, bounded diameter) set that exactly interpolates any observed normal vectors and interior points satisfying our theory's conditions.

3 Main Result: Set Interpolation Theory

This section presents and proves our main results on set interpolability.

(Interp1)	(Interp2)	(Interp3)	(Interp4)	(Interp5)	(Interp6)
All settings	$\mathcal{K} \neq \emptyset$	$\beta < \infty$ and $\mathcal{K} \neq \emptyset$	$D < \infty$	$D < \infty$ and $\mathcal{K} \neq \emptyset$	$D < \infty$ and $\mathcal{K} \neq \emptyset$

Table 1: Summary of settings in which each interpolation condition is applicable.

We first define a set of conditions that will be necessary throughout our theory. Consider a set of observations $S = (\{(z_i, v_i)\}_{i \in \mathcal{I}}, \{(x_k, \delta_k)\}_{k \in \mathcal{K}})$ and denote the associated unit normal vectors by $n_i = \frac{v_i}{\|v_i\|}$ for all $i \in \mathcal{I}$. Due to the added complexity of our set interpolation model, our conditions involve the introduction of auxiliary parameters $\{w_k\}_{k \in \mathcal{K}} \subset \mathbb{R}^d$ along with a constant factor $\lambda \in [\frac{1}{\sqrt{2}}, 1]$, typically equal to one, enabling strengthened versions of these conditions. We find that for any $\mathcal{C}_{\alpha, \beta, D}$ -interpolable observation data $S = (\{(z_i, v_i)\}_{i \in \mathcal{I}}, \{(x_k, \delta_k)\}_{k \in \mathcal{K}})$, there must exist auxiliary parameters $\{w_k\}_{k \in \mathcal{K}}$ and a choice of λ such that for all $i, j \in \mathcal{I}$ and $k, l \in \mathcal{K}$,

$$\|z_i - \frac{1}{\alpha}n_i - (z_j - \frac{1}{\beta}n_j)\| \leq \frac{1}{\gamma} \quad (\text{Interp1})$$

$$\|z_i - \frac{1}{\alpha}n_i - w_k\| \leq \frac{1}{\gamma} - s_k \quad (\text{Interp2})$$

$$\|x_k - w_k\| \leq \frac{1}{\beta} - \delta_k + s_k \quad (\text{Interp3})$$

$$\|z_i - \frac{1}{\beta}n_i - (z_j - \frac{1}{\beta}n_j)\| \leq \lambda(D - \frac{2}{\beta}) \quad (\text{Interp4})$$

$$\|z_i - \frac{1}{\beta}n_i - w_k\| \leq \lambda(D - \frac{2}{\beta}) - s_k \quad (\text{Interp5})$$

$$\|w_k - w_l\| \leq \lambda(D - \frac{2}{\beta}) - s_k - s_l \quad (\text{Interp6})$$

where $\gamma = (\frac{1}{\alpha} - \frac{1}{\beta})^{-1}$ and $s_k = \max\{0, \delta_k - \frac{1}{\beta}\}$ for all $k \in \mathcal{K}$. In the case $\alpha = \beta$, we will use the convention that $\frac{1}{\gamma} = 0$. As a shorthand, we write $S \in \text{Interp}(\alpha, \beta, D; \lambda)$ if $\{w_k\}_{k \in \mathcal{K}} \subset \mathbb{R}^d$ exist satisfying all six conditions.

We can briefly motivate our auxiliary parameters w_k and λ . The values w_k arise from interpolating a smooth set, which can be decomposed into $C = C_0 + B(0, \frac{1}{\gamma})$. Effectively, for each interior point $x_k \in C$, the point w_k represents a corresponding point in C_0 . The auxiliary parameter λ arises from possible slack in bounding the diameter of a strongly convex set (See Remark 4). Note that if a condition holds for some λ , then it holds for all $\hat{\lambda} > \lambda$. Lastly, we emphasize the normalized vectors n_i . In Section 5, we will see that having simultaneous constraints on v_i and $\frac{v_i}{\|v_i\|}$ results in more computationally difficult PEP formulations.

If we simplify our set requirements, the conditions above simplify in turn. If one only enforces smoothness and strong convexity (i.e., $\mathcal{K} = \emptyset, D = \infty$), only a single condition (Interp1) is needed, mirroring the single interpolation condition (2.5) of Taylor et al. [2]. The other conditions are only applicable depending on the constraints on the set structure. We summarize this dependency in Table 1. In particular, the auxiliary variables w_k are only necessary for smooth sets ($\beta < \infty$) with interior point conditions ($\mathcal{K} \neq \emptyset$); otherwise, the condition reduces to $x_k = w_k$.

Our three main theorems below show that these conditions are always necessary for interpolation (Theorem 3.1); they become sufficient if the diameter bound is tightened by a factor of at most $\sqrt{2}$ (Theorem 3.2); they are necessary and sufficient if either the strong convexity or diameter bound is omitted (Theorem 3.3). Immediately afterward, we discuss the necessity of this $\sqrt{2}$ gap between our

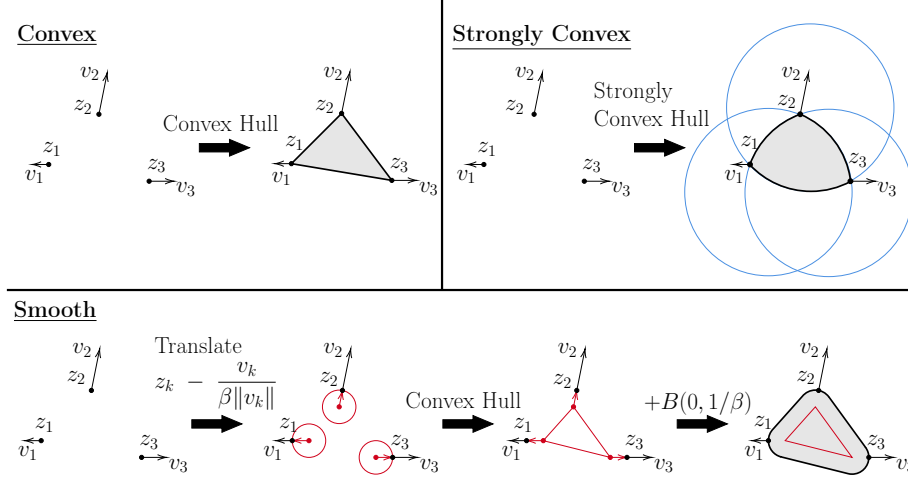


Figure 3: Examples of constructed interpolations with no required interior points ($\mathcal{K} = \emptyset$).

first two theorems and provide several convenient corollaries. Figure 3 presents sample constructions for convex, smooth, and strongly convex interpolating sets for the case where $\mathcal{K}$ is empty. These examples are special cases of the constructions used in our sufficiency proofs.

The following theorems, proven in the following subsections, consider a set of observations $S = (\{(z_i, v_i)\}_{i \in \mathcal{I}}, \{(x_k, \delta_k)\}_{k \in \mathcal{K}})$. Recall d denotes the dimension of space being considered, i.e., $z_i, v_i, x_k \in \mathbb{R}^d$.

Theorem 3.1. *If S is $\mathcal{C}_{\alpha, \beta, D}$ -interpolable, then $S \in \text{Interp}(\alpha, \beta, D; 1)$.*

Theorem 3.2. *If $S \in \text{Interp}(\alpha, \beta, D; \sqrt{\frac{d+1}{2d}})$, then S is $\mathcal{C}_{\alpha, \beta, D}$ -interpolable. In particular, if $S \in \text{Interp}(\alpha, \beta, D; \frac{1}{\sqrt{2}})$, then S is $\mathcal{C}_{\alpha, \beta, D}$ -interpolable.*

Theorem 3.3. *If $\alpha = 0$ or $D \geq \frac{2}{\alpha}$, then S is $\mathcal{C}_{\alpha, \beta, D}$ -interpolable if and only if $S \in \text{Interp}(\alpha, \beta, D; 1)$.*

Remark 2 (On Indicator Function Formulation). Instead of using boundary points and normal vectors, one can equivalently define set interpolation in terms of indicator functions. For any set C and indicator $\iota_C = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{otherwise,} \end{cases}$ one has that $v_i \in N_C(z_i)$ if and only if $v_i \in \partial \iota_C(z_i)$. Similarly, one can define a δ -interior point as some x such that $\iota_C(x + \delta \zeta) = 0$ for all $\|\zeta\| \leq 1$. We focus on the normal-vector formulation to align with our later applications.

Remark 3 (On the Computational Cost of Auxiliary Variables). The function interpolation result of Theorem 2.2 provides a direct means to algebraically check if a collection of observation data $\{(x_i, f_i, g_i)\}_{i \in \mathcal{I}}$ is interpolable. One must check $O(|\mathcal{I}|^2)$ quadratic inequalities. Our set interpolation theory does not present such an easy algebraic check. One must determine whether $\{w_k\}_{k \in \mathcal{K}}$ exist satisfying the needed conditions. This corresponds to solving a second-order cone program with $O(|\mathcal{K}|)$ variables and $O(|\mathcal{I}|^2 + |\mathcal{K}|^2)$ constraints. However, in cases with no or very few interior point conditions ($|\mathcal{K}| = 0$ or small), or where $\beta = \infty$, this cost reduces back to checking $O(|\mathcal{I}|^2)$ inequalities.

Remark 4 (On the Tightness of the Gap Between Theorems 3.1 and 3.2). The gap between $\lambda = 1$ and $\lambda = \sqrt{\frac{d+1}{2d}} \geq 1/\sqrt{2}$ in Theorems 3.1 and 3.2 is fundamental: Consider the regular

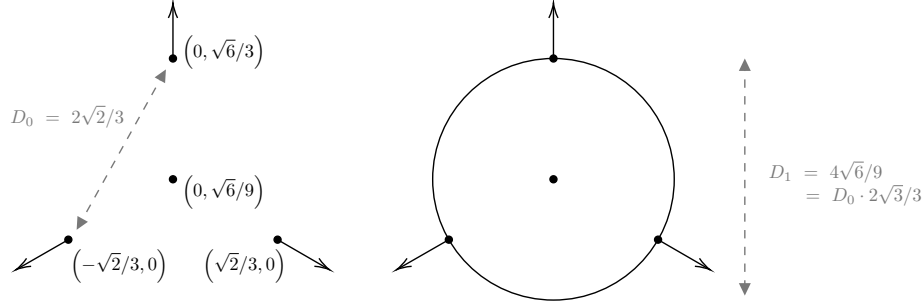


Figure 4: Example of slackness in diameter constraint for the 2-simplex (after projecting into $\mathbb{R}^2$ and translating).

d -simplex in $\mathbb{R}^{d+1}$, with $d > 1$. Placing the centroid at 0, our vertices become $(\frac{d}{d+1}, \frac{-1}{d+1}, \dots, \frac{-1}{d+1})$, $(\frac{-1}{d+1}, \frac{d}{d+1}, \frac{-1}{d+1}, \dots, \frac{-1}{d+1})$, etc. Taking $\mathcal{I} = [0:d]$, we choose $\{z_i\}_{i \in \mathcal{I}}$ to be these $d+1$ vertices and $v_i = \frac{z_i}{\|z_i\|}$ for all $i \in \mathcal{I}$. Further, set $\mathcal{K} = \emptyset$, $S = (\{(z_i, v_i)\}_{i \in \mathcal{I}}, \emptyset)$, and $\alpha = \sqrt{\frac{d+1}{d}}$, $\beta = \infty$, $D = \sqrt{2}$. Observing that

$$\|z_i - \frac{1}{\alpha}n_i - (z_j - \frac{1}{\beta}n_j)\| = \|z_j\| = \sqrt{\frac{d}{d+1}}, \quad \|z_i - z_j\| = \sqrt{2}$$

for all $i \neq j$, (Interp1) and (Interp4) are satisfied with $\lambda = 1$, so $S \in \text{Interp}(\alpha, \beta, \sqrt{2}; 1)$. However, since all z_i satisfy $\|z_i\| = \frac{1}{\alpha}$, $B(0, \frac{1}{\alpha})$ is the unique α -strongly convex (and vacuously β -smooth) set that interpolates S . So S is not $\mathcal{C}_{\alpha, \beta, \sqrt{2}}$ -interpolable as any interpolating set must have diameter $2\sqrt{\frac{d}{d+1}} > \sqrt{2}$, exactly matching the gap in our theorems. See Figure 4 demonstrating a similar example in $\mathbb{R}^2$.

Remark 5 (On the Special Case of Interpolating with Strongly Convex Sets). If we take β to ∞ , that is, entirely relaxing the requirement of smoothness, then our interpolation conditions reduce to

$$\|z_i - \frac{1}{\alpha}n_i - z_j\| \leq \frac{1}{\alpha} \quad (3.1)$$

$$\|z_i - \frac{1}{\alpha}n_i - x_k\| \leq \frac{1}{\alpha} - \delta_k \quad (3.2)$$

$$\|z_i - z_j\| \leq \lambda D \quad (3.3)$$

$$\|z_i - x_k\| \leq \lambda D - \delta_k \quad (3.4)$$

$$\|x_k - x_l\| \leq \lambda D - \delta_k - \delta_l \quad (3.5)$$

with no more dependence on auxiliary parameters $\{w_k\}_{k \in \mathcal{K}}$. As an immediate corollary, we have the following result for strongly convex sets.

Corollary 3.4. *Let $S = (\{(z_i, v_i)\}_{i \in \mathcal{I}}, \{(x_k, \delta_k)\}_{k \in \mathcal{K}})$. If S is $\mathcal{C}_{\alpha, \infty, D}$ -interpolable, then for all $i, j \in \mathcal{I}$ and $k, l \in \mathcal{K}$, (3.1-3.5) are satisfied for $\lambda = 1$. Conversely, if for all $i, j \in \mathcal{I}$ and $k, l \in \mathcal{K}$, (3.1-3.5) are satisfied for $\lambda = 1/\sqrt{2}$, then S is $\mathcal{C}_{\alpha, \infty, D}$ -interpolable.*

Remark 6 (On the Special Case of Interpolating with Smooth, Convex Sets). We can similarly relax the strong convexity condition, taking α to zero and deriving simpler conditions for smooth convex interpolation from Theorem 3.3. Suppose that $\|z_i - \frac{1}{\alpha}n_i - (z_j - \frac{1}{\beta}n_j)\| \leq \frac{1}{\gamma}$. Then

$$\langle z_i - \frac{1}{\alpha}n_i - z_j + \frac{1}{\beta}n_j, z_i - \frac{1}{\alpha}n_i - z_j + \frac{1}{\beta}n_j \rangle \leq \frac{1}{\gamma^2}$$

$$\begin{aligned}
&\Leftrightarrow \|z_i - \frac{1}{\beta}n_i - z_j + \frac{1}{\beta}n_j\|^2 - 2\langle \frac{1}{\gamma}n_i, z_i - \frac{1}{\beta}n_i - z_j + \frac{1}{\beta}n_j \rangle + \frac{1}{\gamma^2} \leq \frac{1}{\gamma^2} \\
&\Leftrightarrow \langle n_i, z_j - \frac{1}{\beta}n_j - z_i + \frac{1}{\beta}n_i \rangle \leq -\frac{\gamma}{2}\|z_i - \frac{1}{\beta}n_i - z_j + \frac{1}{\beta}n_j\|^2.
\end{aligned}$$

Taking the limit as $\alpha \rightarrow 0$ and correspondingly $\gamma \rightarrow 0$, this inequality becomes

$$\langle n_i, z_j - \frac{1}{\beta}n_j - z_i + \frac{1}{\beta}n_i \rangle \leq 0 \quad (3.6)$$

in place of (Interp1). Similarly, for w_k , we obtain

$$\langle n_i, w_k + s_k n_i - z_i + \frac{1}{\beta}n_i \rangle \leq 0 \quad (3.7)$$

in place of (Interp2). We can then state a corollary for smooth sets.

Corollary 3.5. *Let $S = (\{(z_i, v_i)\}_{i \in \mathcal{I}}, \{(x_k, \delta_k)\}_{k \in \mathcal{K}})$. S is $\mathcal{C}_{0,\beta,D}$ -interpolable if and only if there exist $\{w_k\}_{k \in \mathcal{K}} \subset \mathbb{R}^d$ such that for all $i, j \in \mathcal{I}$ and $k, l \in \mathcal{K}$, (3.6) and (3.7) along with (Interp3)-(Interp6) are satisfied for $\lambda = 1$.*

Taking the limit as β tends to ∞ yields a final result for nonsmooth, non-strongly convex interpolation.

Corollary 3.6. *A set $S = (\{(z_i, v_i)\}_{i \in \mathcal{I}}, \{(x_k, \delta_k)\}_{k \in \mathcal{K}})$ is $\mathcal{C}_{0,\infty,D}$ -interpolable if and only if for all $i, j \in \mathcal{I}$ and $k, l \in \mathcal{K}$,*

$$\begin{aligned}
&\langle n_i, z_j - z_i \rangle \leq 0 \\
&\langle n_i, x_k + \delta_k n_i - z_i \rangle \leq 0 \\
&\|x_k - x_l\| \leq D - \delta_k - \delta_l \\
&\|z_i - x_k\| \leq D - \delta_k \\
&\|z_i - z_j\| \leq D
\end{aligned}$$

where $n_i = \frac{v_i}{\|v_i\|}$ for all $i \in \mathcal{I}$.

Finally, we note the strong connections between our set interpolation and the function interpolation of [3]. In the convex case of Corollary 3.6 above, if $\delta_k = 0$ for all k (or if $\mathcal{K} = \emptyset$), we can replace all uses of the normalized vectors n_i with v_i and recover an equivalent reformulation of [3, Theorem 3.6]. More generally, we show below that smooth function interpolation [2] is recovered as a limit of our set interpolation theory. We defer the proof to Appendix A.1.

Theorem 3.7. *The set of observations $\{(x_i, g_i, f_i)\}_{i \in \mathcal{I}}$ is $\mathcal{F}_{0,L}$ -interpolable if and only if $(\{((\frac{x_i}{\eta}, f_i), (\eta g_i, -1))\}_{i \in \mathcal{I}}, \emptyset)$ is $\mathcal{C}_{0,\eta^2 L, \infty}$ -interpolable for all $\eta > 0$.*

3.1 Proof of Theorem 3.1

Suppose that S is $\mathcal{C}_{\alpha,\beta,D}$ -interpolable. Then there exists some $C \in \mathcal{C}_{\alpha,\beta,D}$ such that $v_i \in N_C(z_i)$, $z_i \in C$, and $x_k \in \text{int}_{\delta_k} C$ for all $i \in \mathcal{I}, k \in \mathcal{K}$. From Lemma 2.6, we know that $C = C_0 + B(0, \frac{1}{\beta})$ for some γ -strongly convex set C_0 , where $\frac{1}{\gamma} + \frac{1}{\beta} = \frac{1}{\alpha}$. In particular, as shown in [31], we can write C_0 explicitly with the Minkowski difference $C_0 = C - B(0, \frac{1}{\beta}) := \{x \mid x + B(0, \frac{1}{\beta}) \subseteq C\}$.

As a first intermediate result, we claim for each $i \in \mathcal{I}$,

$$v_i \in N_{C_0}(z_i - \frac{v_i}{\beta\|v_i\|}). \quad (3.8)$$

This follows since $v_i \in N_C(z_i)$ implies for all $y \in C$, $\langle v_i, y - z_i \rangle \leq 0$. So noting that every $p \in C_0$ has $p + \frac{v_i}{\beta\|v_i\|} \in C$, we have $\langle v_i, p - (z_i - \frac{v_i}{\beta\|v_i\|}) \rangle \leq 0$ for all $p \in C_0$. From the definition of C_0 , $z_i - \frac{v_i}{\beta\|v_i\|} \in C_0$. Together these yield (3.8).

Since C_0 is γ -strongly convex, (3.8) implies that $C_0 \subseteq B(z_i - \frac{v_i}{\beta\|v_i\|} - \frac{v_i}{\gamma\|v_i\|}, \frac{1}{\gamma})$. Since $z_j - \frac{v_j}{\beta\|v_j\|} \in C_0$ for all $j \in \mathcal{I}$, we have $z_j - \frac{v_j}{\beta\|v_j\|} \in B(z_i - \frac{v_i}{\beta\|v_i\|} - \frac{v_i}{\gamma\|v_i\|}, \frac{1}{\gamma})$. So (Interp1) holds as

$$\|z_i - \frac{v_i}{\alpha\|v_i\|} - (z_j - \frac{v_j}{\beta\|v_j\|})\| = \|z_i - \frac{v_i}{\beta\|v_i\|} - \frac{v_i}{\gamma\|v_i\|} - (z_j - \frac{v_j}{\beta\|v_j\|})\| \leq \frac{1}{\gamma}.$$

Now, we verify (Interp2) and (Interp3) for each $k \in \mathcal{K}$ through two cases:

Case 1: $x_k \in C_0$. In this case, we let $w_k = x_k$. Suppose that $\delta_k > \frac{1}{\beta}$, so $s_k := \max\{0, \delta_k - \frac{1}{\beta}\} = \delta_k - \frac{1}{\beta}$. We claim that $x_k \in \text{int}_{s_k} C_0$. Suppose that for some ζ , $x_k + (\delta_k - \frac{1}{\beta}) \frac{\zeta}{\|\zeta\|} \notin C_0$. Then $x_k + (\delta_k - \frac{1}{\beta}) \frac{\zeta}{\|\zeta\|} + \frac{1}{\beta} \frac{\zeta}{\|\zeta\|} = x_k + \delta_k \frac{\zeta}{\|\zeta\|} \notin C$. But since $x_k \in \text{int}_{\delta_k} C$, this is a contradiction, so we must have $x_k \in \text{int}_{s_k} C_0$. Now suppose $\delta_k \leq \frac{1}{\beta}$. Then $s_k = 0$, so $x_k \in \text{int}_{s_k} C_0$ is true by assumption. Since $x_k = w_k$, we have shown that $w_k + B(0, s_k) \subseteq C_0$. When combined with the previous fact that $C_0 \subseteq B(z_i - \frac{v_i}{\beta\|v_i\|} - \frac{v_i}{\gamma\|v_i\|}, \frac{1}{\gamma})$, it follows that

$$\|z_i - \frac{v_i}{\alpha\|v_i\|} - w_k\| \leq \frac{1}{\gamma} - s_k.$$

This is equivalent to (Interp2). Finally, noting $\|w_k - x_k\| = 0 \leq \max\{0, \frac{1}{\beta} - \delta_k\} = \frac{1}{\beta} - \delta_k + s_k$, we conclude (Interp3) is satisfied.

Case 2: $x_k \notin C_0$. Since $C = C_0 + B(0, \frac{1}{\beta})$ and $x_k \notin C_0$, we must have $\delta_k < \frac{1}{\beta}$. We therefore have $s_k = 0$, and we set w_k as the orthogonal projection of x_k onto C_0 and $\zeta = \frac{x_k - w_k}{\|x_k - w_k\|}$ as the corresponding normal vector to C_0 . Since $w_k \in C_0$, the same reasoning as above leveraging (3.8) implies (Interp2). We know $x_k \in \text{int}_{\delta_k} C$, so it follows that $w_k + \zeta(\|w_k - x_k\| + \delta_k) \in C$. However, since $\zeta \in N_{C_0}(w_k)$, it follows that $\|w_k - x_k\| + \delta_k \leq \frac{1}{\beta}$. Recalling $s_k = 0$, that is exactly (Interp3).

Finally, we verify the diameter conditions (Interp4), (Interp5), and (Interp6). For all $i \in \mathcal{I}, k \in \mathcal{K}$, we have shown $z_i - \frac{v_i}{\beta\|v_i\|} \in C_0$ and in either case above, we have that $w_k \in \text{int}_{s_k} C_0$. Hence the diameter bound of $\text{diam}(C_0) \leq D - \frac{2}{\beta}$ from Lemma 2.5 yields the remaining three conditions for $\lambda = 1$.

3.2 Proof of Theorem 3.2

Suppose our six conditions hold for $\lambda = \sqrt{\frac{d+1}{2d}} \geq \frac{1}{\sqrt{2}}$. We begin by constructing our interpolating set C . First, we construct C_0 as the γ -strongly convex hull of $\{B(w_k, s_k)\}_{k \in \mathcal{K}}$ and $\{z_i - \frac{v_i}{\beta\|v_i\|}\}_{i \in \mathcal{I}}$. More formally, by [29, Proposition 2.5], we set $C_0 = \bigcap_{y \in Y} B(y, \frac{1}{\gamma})$, where

$$Y = \{y \mid \|z_i - \frac{v_i}{\beta\|v_i\|} - y\| \leq \frac{1}{\gamma} \quad \forall i \in \mathcal{I}, \quad \|w_k - y\| \leq \frac{1}{\gamma} - s_k \quad \forall k \in \mathcal{K}\}.$$

Observe that each ball $B(y, \frac{1}{\gamma})$ is γ -strongly convex, and this property is preserved under intersections [28, Proposition 2], so C_0 is γ -strongly convex. We then define $C = C_0 + B(0, \frac{1}{\beta})$. By Lemma 2.6, we see that C is α -strongly convex and β -smooth. All that remains is to verify this set C correctly interpolates the given observation data S and has the desired diameter bound D .

Verification that $z_i \in C$ and $v_i \in N_C(z_i)$ for each $i \in \mathcal{I}$. From the definition of Y , $z_i - \frac{v_i}{\beta\|v_i\|} \in B(y, \frac{1}{\gamma})$ for all $y \in Y$. Hence $z_i - \frac{v_i}{\beta\|v_i\|} \in C_0$. Since $C = C_0 + B(0, \frac{1}{\beta})$, it follows that

$z_i \in C$. Further, by (Interp1), $\|z_i - \frac{v_i}{\alpha\|v_i\|} - (z_j - \frac{v_j}{\beta\|v_j\|})\| \leq \frac{1}{\gamma}$ for all $j \in \mathcal{I}$, and by (Interp2), $\|z_i - \frac{v_i}{\alpha\|v_i\|} - w_k\| \leq \frac{1}{\gamma} - s_k$ for all $k \in \mathcal{K}$. Hence $z_i - \frac{v_i}{\alpha\|v_i\|} \in Y$. Letting $B_i = B(z_i - \frac{v_i}{\alpha\|v_i\|}, \frac{1}{\gamma})$, observe that $v_i \in N_{B_i}(z_i - \frac{v_i}{\alpha\|v_i\|} + \frac{v_i}{\gamma\|v_i\|}) = N_{B_i}(z_i - \frac{v_i}{\beta\|v_i\|})$. Combining this with the fact that $C_0 \subseteq B_i$ since $z_i - \frac{v_i}{\alpha\|v_i\|} \in Y$, one has that $v_i \in N_{C_0}(z_i - \frac{v_i}{\beta\|v_i\|})$. Finally, since $C = C_0 + B(0, \frac{1}{\beta})$, it follows that $v_i \in N_C(z_i)$ as well.

Verification that $x_k \in \text{int}_{\delta_k} C$ for each $k \in \mathcal{K}$. Next, we consider w_k and x_k . By definition of Y , we see that for all $y \in Y$, $w_k \in \text{int}_{s_k} B(y, \frac{1}{\gamma})$. Consequently, by construction, $w_k \in \text{int}_{s_k} C_0$. Then using the fact that $\|x_k - w_k\| \leq \frac{1}{\beta} + s_k - \delta_k$ by (Interp3), for any ζ , it follows that

$$\begin{aligned} \|x_k + \delta_k \frac{\zeta}{\|\zeta\|} - (w_k + s_k \frac{\zeta}{\|\zeta\|})\| &\leq \|x_k - w_k\| + |\delta_k - s_k| = \|x_k - w_k\| + \delta_k - s_k \\ &\leq \frac{1}{\beta} + s_k - \delta_k + (\delta_k - s_k) = \frac{1}{\beta}. \end{aligned}$$

Since $w_k \in \text{int}_{s_k} C_0$, for any nonzero ζ , one has $\hat{w}_k := w_k + s_k \frac{\zeta}{\|\zeta\|} \in C_0$. Hence $\|x_k + \delta_k \frac{\zeta}{\|\zeta\|} - \hat{w}_k\| \leq \frac{1}{\beta}$ from which one can conclude for any ζ , $x_k + \delta_k \frac{\zeta}{\|\zeta\|} \in C$ and consequently $x_k \in \text{int}_{\delta_k} C$.

Verification of diameter bound $\text{diam}(C) \leq D$. By Lemma 2.5, it is sufficient to show that $\text{diam}(C_0) \leq D - \frac{2}{\beta}$. Since C_0 is γ -strongly convex, we already know $\text{diam}(C_0) \leq \frac{2}{\gamma}$. So, all that remains is to prove the needed bound when $D - \frac{2}{\beta} < \frac{2}{\gamma}$. Jung's theorem [32] states that given a compact set $X \subseteq \mathbb{R}^d$, there exists a closed ball B with radius $R = \text{diam}(X) \sqrt{\frac{d}{2(d+1)}}$ containing X .

Applied to the set $X = \{z_i - \frac{v_i}{\beta\|v_i\|}\}_{i \in \mathcal{I}} \cup \{B(w_k, s_k)\}_{k \in \mathcal{K}}$, which has diameter at most $\sqrt{\frac{d+1}{2d}}(D - \frac{2}{\beta})$ by (Interp4), (Interp5), and (Interp6), there must exist $q \in \mathbb{R}^d$ such that $X \subseteq B(q, R)$ where $R = \left(\sqrt{\frac{d+1}{2d}}(D - \frac{2}{\beta})\right) \sqrt{\frac{d}{2(d+1)}} = \frac{1}{2}(D - \frac{2}{\beta})$. Hence the distance from q to any point in X is at most $\frac{1}{2}(D - \frac{2}{\beta}) < 1/\gamma$. As a result, a neighborhood of q lies in Y . Namely, $B(q, \frac{1}{\gamma} - \frac{1}{2}(D - \frac{2}{\beta})) \subseteq Y$. So,

$$C_0 = \bigcap_{y \in Y} B\left(y, \frac{1}{\gamma}\right) \subseteq \bigcap_{y \in B(q, \frac{1}{\gamma} - \frac{1}{2}(D - \frac{2}{\beta}))} B\left(y, \frac{1}{\gamma}\right) = B\left(q, \frac{1}{2}(D - \frac{2}{\beta})\right),$$

proving the needed bound $\text{diam}(C_0) \leq D - \frac{2}{\beta}$.

3.3 Proof of Theorem 3.3

The forward direction is already proven by Theorem 3.1, so we only need to prove the reverse. First consider the case of $D \geq \frac{2}{\alpha}$. By our argument in Theorem 3.2, if our conditions hold for $\lambda = 1$, then there exists an α -strongly convex, β -smooth set C (of unspecified diameter) that interpolates S . However, since any α -strongly convex set C must satisfy $C \subseteq B(z - \frac{1}{\alpha}n, \frac{1}{\alpha})$ for $z \in \text{bdry } C$ with unit normal vector n , we see that $\text{diam}(C) \leq \text{diam}(B(z - \frac{1}{\alpha}n, \frac{1}{\alpha})) = \frac{2}{\alpha} \leq D$. Therefore, S is $C_{\alpha, \beta, D}$ -interpolable.

Now consider the case of $\alpha = 0$. Suppose our six conditions hold for $\lambda = 1$. As shown in Remark 6, our conditions (Interp1) and (Interp2) reduce to (3.6) and (3.7). Define $C_0 = \text{conv}(\{z_i - \frac{v_i}{\beta\|v_i\|}\}_{i \in \mathcal{I}}, \{B(w_k, s_k)\}_{k \in \mathcal{K}})$. Clearly C_0 is convex with $w_k \in \text{int}_{s_k} C_0$ and $z_i - \frac{v_i}{\beta\|v_i\|} \in C_0$. We then construct $C = C_0 + B(0, \frac{1}{\beta})$. From Proposition 2.3, C is β -smooth.

Next, we show that $z_i \in C$ and $v_i \in N_C(z_i)$. Since $z_i - \frac{v_i}{\beta\|v_i\|} \in C_0$, we immediately have that $z_i \in C$. Consider any $y \in C$ and let $p \in C_0$ be such that $y = p + \frac{1}{\beta}\zeta$ for some $\|\zeta\| \leq 1$. Since C_0 is a

convex hull, we write p as the convex combination

$$p = \sum_{j \in \mathcal{I}} \sigma_j \left(z_j - \frac{v_j}{\beta \|v_j\|} \right) + \sum_{k \in \mathcal{K}} \phi_k (w_k + s_k \xi_k)$$

where $\sigma_j, \phi_k \geq 0$ and $\sum_j \sigma_j + \sum_k \phi_k = 1$, and $\|\xi_k\| \leq 1$. We then have

$$\begin{aligned} \langle v_i, y - z_i \rangle &= \langle v_i, p + \frac{1}{\beta} \zeta - z_i \rangle \leq \langle v_i, p - (z_i - \frac{v_i}{\beta \|v_i\|}) \rangle \\ &= \sum_{j \in \mathcal{I}} \sigma_j \langle v_i, z_j - \frac{v_j}{\beta \|v_j\|} - (z_i - \frac{v_i}{\beta \|v_i\|}) \rangle + \sum_{k \in \mathcal{K}} \phi_k \langle v_i, w_k + s_k \xi_k - (z_i - \frac{v_i}{\beta \|v_i\|}) \rangle \\ &\leq \sum_{j \in \mathcal{I}} \sigma_j \langle v_i, z_j - \frac{v_j}{\beta \|v_j\|} - (z_i - \frac{v_i}{\beta \|v_i\|}) \rangle + \sum_{k \in \mathcal{K}} \phi_k \langle v_i, w_k + s_k \frac{v_i}{\|v_i\|} - (z_i - \frac{v_i}{\beta \|v_i\|}) \rangle \\ &\leq 0 \end{aligned}$$

where the last inequality follows from (3.6) and (3.7). Hence $v_i \in N_C(z_i)$.

By the same argument as in Theorem 3.2, we have that $x_k \in \text{int}_{\delta_k} C$. Lastly, since C_0 is a convex hull, (Interp4), (Interp5), and (Interp6) establish that

$$\text{diam}(C_0) = \text{diam} \left(\{z_i - \frac{v_i}{\beta \|v_i\|}\}_{i \in \mathcal{I}} \cup \{B(w_k, s_k)\}_{k \in \mathcal{K}} \right) \leq D - \frac{2}{\beta}.$$

Then, by Lemma 2.5, we conclude that $\text{diam}(C) \leq D$.

4 Separating Hyperplane Algorithms as an SDP

In the remainder of this paper, we provide three applications of our interpolation theorems. In this section, we consider a family of simple algorithms computing an element of a convex set by iteratively using a separating hyperplane oracle. Our interpolation theorems allow us to quantify the worst-case stopping time of such a method over any smooth and/or strongly convex set with an interior point via semidefinite programming. Motivated by and confirming our numerical results, we identify a simple minimax optimal separating hyperplane algorithm.

For any closed convex set $C \subseteq \mathbb{R}^d$, we denote the set of (unit) separating hyperplanes of C at some $x \in \mathbb{R}^d$ by

$$SH_C(x) = \{n \in \mathbb{R}^d \mid \|n\| = 1, \langle n, y - x \rangle \leq 0 \quad \forall y \in C\}.$$

Note this set is nonempty exactly when $x \notin \text{int } C$. Here, our primary interest is in designing algorithms constructing a member of $\text{int } C$ using a sequence of separating hyperplane oracle queries. For example, this generalizes convex minimization, seeking an ϵ -minimizer where each subgradient provides such a separating hyperplane to the ϵ -level set. As a general form of method with N queries, consider any iteration producing points x_i via

$$x_{i+1} = x_i - \sum_{j=0}^i H_{i,j} n_j, \quad n_i \in SH_C(x_i), \quad \forall i = 0, \dots, N-1 \quad (4.1)$$

parameterized by the lower triangular matrix of stepsizes H . Note this method must halt once $x_i \in \text{int } C$ as $SH_C(x_i)$ is empty.

Given fixed α, β, δ, R , we consider the problem finding a feasible point to any set $C \in \mathcal{C}_{\alpha, \beta, \infty}$ containing some $q \in \text{int}_{\delta} C$ and any initialization x_0 with $\|x_0 - q\| \leq R$. Since no diameter bound is enforced on C , our interpolation conditions are necessary and sufficient (i.e., Theorem 3.3 applies).

The question of whether a problem instance exists where an algorithm H can fail to construct some $x_i \in \text{int } C$ with $i = 0, \dots, N$ corresponds to

$$\exists x_i, n_i, q, C \quad \text{s.t.} \quad \begin{cases} C \in \mathcal{C}_{\alpha, \beta, \infty} \\ B(q, \delta) \subseteq C \\ n_i \in SH_C(x_i) \\ \|x_0 - q\| \leq R \\ x_{i+1} = x_i - \sum_{j=0}^i H_{i,j} n_j \end{cases}$$

A successful algorithm design would be a selection of H such that no solution to the above system exists, as a failure to have $n_k \in SH_C(x_k)$ exist for some k implies $x_k \in \text{int } C$. Noting that every separating hyperplane n_i must be normal to the set C at some z_i , this system can be rewritten in terms of interpolation with $\mathcal{I} = \{0, \dots, N\}$. Theorem 3.3 enables this decision problem to be described as a system of quadratic inequalities using our first three interpolation conditions (Interp1)-(Interp3):

$$\exists x_i, z_i, n_i, q, w \quad \text{s.t.} \quad \begin{cases} \|z_i - \frac{1}{\alpha} n_i - (z_j - \frac{1}{\beta} n_j)\|^2 \leq \frac{1}{\gamma^2} \\ \|z_i - \frac{1}{\alpha} n_i - w\|^2 \leq (\frac{1}{\gamma} - s)^2 \\ \|q - w\|^2 \leq (\frac{1}{\beta} - \delta + s)^2 \\ \langle n_i, z_i - x_i \rangle \leq 0 \\ \|n_i\|^2 = 1 \\ \|x_0 - q\|^2 \leq R^2 \\ x_{i+1} = x_i - \sum_{j=0}^i H_{i,j} n_j \end{cases} \quad (4.2)$$

where $\gamma = (\frac{1}{\alpha} - \frac{1}{\beta})^{-1}$ and $s = \max\{0, \delta - \frac{1}{\beta}\}$. Without loss of generality, we fix $q = 0$. Assuming $d \geq 2N+4$, a Gram matrix reformulation of this quadratic system yields an equivalent SDP feasibility problem with variable G capturing every quadratic term in the original variables x_0, z_i, n_i, w , i.e.,

$$\Lambda = [x_0 | z_0 | z_1 | \dots | z_N | n_0 | n_1 | \dots | n_N | w] \in \mathbb{R}^{d \times (2N+4)}, \\ G = \Lambda^T \Lambda \in \mathbb{S}_+^{2N+4}.$$

In the case $d < 2N+4$, this reformulation still represents a valid upper bound on our performance estimation problem, but it is no longer guaranteed to be tight (See Remark 1). Appendix B.1 presents this reformulation in full for the sake of completeness. Such reformulations are widespread in the existing PEP literature. Hence, we can efficiently certify whether a proposed separating hyperplane algorithm is guaranteed to construct a feasible point.

Consider the simple separating hyperplane method fixing H to be diagonal with constant value:

$$x_{i+1} = x_i - h n_i, \quad n_i \in SH_C(x_i), \quad \forall i = 0, \dots, N-1. \quad (4.3)$$

For a given family of problem instances fixing α, β, δ, R , we can compute the number of iterations needed to guarantee a strictly feasible point by growing N until the corresponding SDP becomes infeasible. Denote this maximal number of steps needed to ensure an interior point is found by $N_{\max}$. Fixing $h = \max\{\delta, \frac{1}{\beta}\}$, Figure 5 shows $N_{\max}$ as δ and R vary under this stepsize. From these numerical results, one can readily identify a formula $N_{\max} = \lfloor \frac{(R+h-\delta)^2}{h^2} \rfloor$.

This simple method and numerically observed rate turn out to be the minimax optimal method and guarantee for any separating hyperplane method of the general form (4.1). Appendix B.2 presents an elementary proof of this fact.

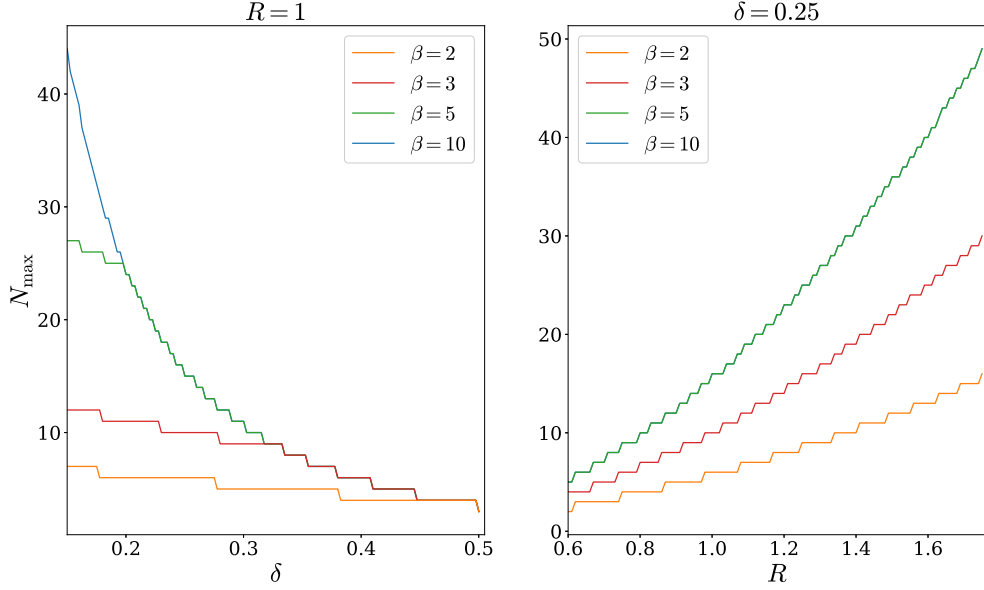


Figure 5: Numerical PEP result for $N_{\max}$ given a constant stepsize $h = \max\{\delta, \frac{1}{\beta}\}$ as δ varies on the left and R on the right. In the second plot, the line for $\beta = 10$ is covered by $\beta = 5$.

Theorem 4.1. *For any $C \in \mathcal{C}_{\alpha,\beta,D}$ with $q \in \text{int}_\delta C$ and $\|x_0 - q\| \leq R$, the iteration (4.3) with stepsize $h = \max\{\delta, \frac{1}{\beta}\}$ must halt with $x_i \in \text{int } C$ by iteration $i \leq N_{\max} := \lfloor \frac{(R+h-\delta)^2}{h^2} \rfloor$. Moreover, there exists C, q, x_0 and separating hyperplane oracle such that for any method (4.1) has $x_0, \dots, x_{N_{\max}-1} \notin \text{int } C$.*

In this case, the proof of this minimax optimality theorem is sufficiently simple that one could have reached these conclusions without computer-assistance. Nevertheless, these results demonstrate the new possibilities of our set interpolation theory. The numerical insights generated in the following two sections escalate in complexity beyond what is reasonable to do “by-hand”.

5 Frank-Wolfe as an SDP with Nonconvex Constraints

In this section, we consider performance estimation of the Frank-Wolfe (conditional gradient) method on smooth and strongly convex sets, extending the prior analysis of Taylor, Hendrickx, and Glineur [3] in the case of convex sets. This method applies to constrained optimization problems

$$\min_{x \in C} f(x) \tag{5.1}$$

where f is at least convex and L -smooth and C is at least a convex set with diameter at most D . At various points, we will consider problem classes with additional structures like μ -strong convexity of f and α -strong convexity and β -smoothness of C . We will let $x_\star \in C$ denote a minimizer of (5.1) which exists by compactness.

Given a gradient oracle for f and a linear minimization oracle for C , the Frank-Wolfe method generates iterates $x_1, \dots, x_N \in C$ via

$$z_k \in \operatorname{argmin}_{y \in C} \langle \nabla f(x_k), y \rangle \tag{FW}$$

$$x_{k+1} = (1 - h_k)x_k + h_k z_k$$

for some stepsize schedule $h = (h_0, \dots, h_{N-1})$. Such stepsizes are sometimes called “open-loop”. Typically, stepsizes follow $h_k = \frac{2}{k+2}$ or more generally $h_k = \frac{\ell}{k+\ell}$ for some $\ell > 0$. In all cases, we will assume that $0 \leq h_k \leq 1$ for all k , ensuring x_{k+1} remains a convex combination of x_0 and all z_i .

5.1 Frank-Wolfe PEP Formulation and Computational Approaches

For a fixed algorithm, determined by fixing h above, its worst-case performance after N steps can be formulated as a performance estimation problem. The worst-case minimum objective gap seen is given by

$$p_{\text{FW}}(N, h; \mu, L, \alpha, \beta, D) = \begin{cases} \max_{f, C, x_0} & \min_{k=0, \dots, N} f(x_k) - f(x_\star) \\ \text{s.t.} & x_{i+1} = (1 - h_i)x_i + h_i z_i \\ & z_i \in \operatorname{argmin}_{y \in C} \langle \nabla f(x_i), y \rangle \\ & x_0 \in C \\ & C \in \mathcal{C}_{\alpha, \beta, D} \\ & f \in \mathcal{F}_{\mu, L} \\ & x_\star \in \operatorname{argmin}_{y \in C} f(y) . \end{cases} \quad (5.2)$$

From first-order optimality, $z_i \in \operatorname{argmin}_{y \in C} \langle \nabla f(x_i), y \rangle$ can be replaced by the equivalent condition $-\nabla f(x_i) \in N_C(z_i)$. Then our interpolation theory provides a reformulation as a solvable finite problem. See Appendix C for further details supporting this reformulation. For ease of exposition, in what follows, we consider the specific case of optimizing L -smooth functions over β -smooth sets, but analogous derivations follow for any combination of smoothness and strong convexity of functions and sets. We will separately consider the cases $\nabla f(x_\star) = 0$ and $\nabla f(x_\star) \neq 0$, here focusing on $\nabla f(x_\star) \neq 0$. This separation is further motivated by classic results in the literature [33–36] where the behavior of the Frank-Wolfe algorithm depends significantly on whether $x_\star \in \operatorname{int} C$.

Using both set and function interpolation theory, we can reformulate (5.2) as the following QCQP with additional separable nonconvex equality constraints. Letting $\theta = (h, \mu, L, \alpha, \beta, D)$, with $\alpha = 0$ for ease of presentation,

$$p_{\text{FW}}(N; \theta) = \begin{cases} \max_{\substack{x_k, g_k, f_k \\ w_k, z_i, n_i}} & f_{\min} - f_\star \\ \text{s.t.} & f_{\min} - f_k \leq 0 & \forall k \in \mathcal{K} \\ & f_k - f_l + \langle g_k, x_l - x_k \rangle + \frac{1}{2L} \|g_k - g_l\|^2 \leq 0 & \forall k, l \in \mathcal{K}_\star \\ & x_{i+1} = (1 - h_i)x_i + h_i z_i & \forall i \in \mathcal{I} \\ & \langle -g_i, z_j - \frac{1}{\beta} n_j - (z_i - \frac{1}{\beta} n_i) \rangle \leq 0 & \forall i, j \in \mathcal{I}_\star \\ & \langle -g_i, w_k - (z_i - \frac{1}{\beta} n_i) \rangle \leq 0 & \forall i \in \mathcal{I}_\star, k \in \mathcal{K}_\star \\ & \|x_k - w_k\|^2 \leq \frac{1}{\beta^2} & \forall k \in \mathcal{K}_\star \\ & \|z_i - \frac{1}{\beta} n_i - (z_j - \frac{1}{\beta} n_j)\|^2 \leq (D - \frac{2}{\beta})^2 & \forall i, j \in \mathcal{I}_\star \\ & \|z_i - \frac{1}{\beta} n_i - w_k\|^2 \leq (D - \frac{2}{\beta})^2 & \forall i \in \mathcal{I}_\star, k \in \mathcal{K}_\star \\ & \|w_k - w_l\|^2 \leq (D - \frac{2}{\beta})^2 & \forall k, l \in \mathcal{K}_\star \\ & \langle g_i, n_i \rangle \leq 0 & \forall i \in \mathcal{I}_\star \\ & \|n_i\|^2 = 1 & \forall i \in \mathcal{I}_\star \\ & \langle g_i, n_i \rangle^2 = \|g_i\|^2 & \forall i \in \mathcal{I}_\star . \end{cases} \quad (5.3)$$

Note that the last three constraints above are equivalent to the condition $n_i = \frac{-g_i}{\|g_i\|}$. In the case that $g_\star = 0$, these constraints must be removed as $n_\star$ is undefined, requiring a branch in our formulation based on $g_\star$.

One can repeat the “standard” Gram matrix reformulation approach discussed in Section 4. Assuming $d \geq 4N + 6$, one can derive an equivalent problem with variables G and F as

$$\begin{aligned} F &= [f_0|f_1|\dots|f_N|f_{\min}] \in \mathbb{R}^{1 \times (N+2)} \\ \Lambda &= [x_0|g_\star|g_0|g_1|\dots|g_N|z_0|\dots|z_{N-1}|n_\star|n_0|\dots|n_{N-1}|w_\star|w_0|\dots|w_N] \in \mathbb{R}^{d \times (4N+6)} \\ G &= \Lambda^T \Lambda \in \mathbb{S}_+^{4N+6}. \end{aligned}$$

We omit the restatement of this program after this standard change of variables. This Grammian change-of-variables reformulates all of the constraints of (5.3) as linear constraints in our new variables, with the exception of the equality constraint $\langle g_i, n_i \rangle^2 = \|g_i\|^2$ which is quadratic in G . This constraint is nonconvex, and therefore cannot be expressed through a convex SDP. We present three methods of solving or bounding the PEP (5.3), in order of increasing computational difficulty. In any case, we introduce the redundant constraints $\langle -g_i, n_j \rangle \leq \langle -g_i, n_i \rangle$ for all $i, j \in \mathcal{I}_\star$, which prove useful in our relaxations below.

As a simplest approach, one can relax the constraint $\langle g_i, n_i \rangle^2 = \|g_i\|^2$ to $\langle g_i, n_i \rangle^2 \leq \|g_i\|^2$, which is convex in G , providing a upper bound for (5.2). This is similar to the original performance estimation approach of Drori and Teboulle [1]. While excluded from our numerics, this relaxation enables computations of performance upper bounds up to larger values of N (~ 50).

Alternatively, one can reformulate (5.3) by introducing new variables ψ_i to parameterize the nonconvex constraint. Given $\psi = (\psi_\star, \psi_0, \dots, \psi_{N-1}) \in \mathbb{R}_{\geq 0}^{N+1}$, one can define the parameterized problem $\hat{p}_{\text{FW}}(\psi; \theta)$ by adding the constraints $\|g_i\|^2 = \psi_i$ and $\langle -g_i, n_i \rangle = \sqrt{\psi_i}$, which are linear in G . Finding a local solution to (5.3) then amounts to zeroth-order local maximization of $\hat{p}_{\text{FW}}(\psi; \theta)$ with respect to ψ , yielding a lower bound for (5.3). This approach was computationally feasible to run up to $N \leq 15$.

Finally, complementing the above local computation of lower bounds, one can globally optimize (5.3) via a simple branch-and-bound-type procedure. By Cauchy-Schwarz, $\langle -g_i, n_i \rangle \leq \|g_i\|$. Then one can apply piecewise McCormick relaxations [37, 38] to create a partition of convex relaxations of our equality condition $\langle -g_i, n_i \rangle = \|g_i\|$. One can then globally optimize by iteratively refining this partition for each i . Given this procedure’s computational cost scales exponentially in N , our computations of such global upper bounds were limited to $N \leq 5$. When tractable, these numerical upper bounds (denoted by dots in our figures) always closely aligned with our numerical lower bounds, together certifying tight bounds on the true optimal value.

5.2 Survey of Smooth/Strongly Convex Settings for Acceleration

Bounds on Frank-Wolfe’s worst-case convergence rates are well-studied. For the most general setting of convex bounded C and L -smooth convex f , the best known upper bound on convergence rates in this setting is due to Jaggi [39], establishing that with the “standard” stepsize sequence $h_k = \frac{2}{k+2}$, the objective gap converges at rate⁴ at least $\frac{2LD^2}{N+2}$. In [3, Section 4.5], Taylor et al. previously showed through PEP experiments that this rate is slack. The best known lower bound for Frank-Wolfe over convex sets is $\frac{LD^2}{4N}$, from Lan [40], establishing potential to improve by up to a factor of eight.

Given more structure on C and/or f , one could hope for further gains. Frank-Wolfe’s worst-case convergence rate is known to accelerate up to $O(1/N^2)$ in a few settings: given a strongly convex

⁴Note that while the rate proven by Jaggi describes the objective gap at the terminal iterate, it holds as an equally valid bound for our objective $\min_k f(x_k) - f(x_\star)$.

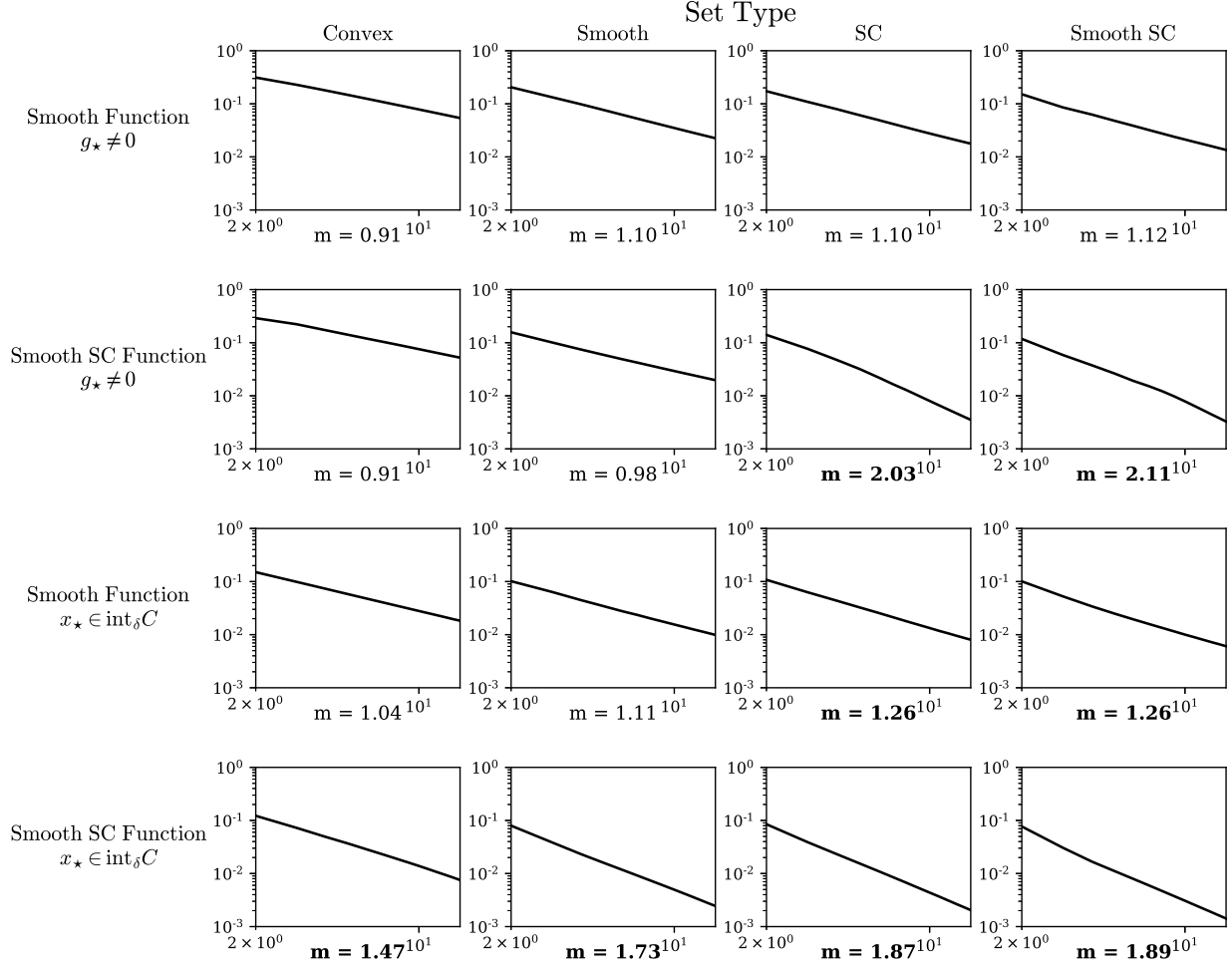


Figure 6: Survey of Frank-Wolfe’s performance with standard stepsizes on 16 different problem classes. Experiments are run with stepsizes $h_k = \frac{2}{k+2}$ and with $L = 1$, $\mu = 0.5$, $\beta = 5$, $\alpha = 1$, $D = 1$ and $\delta = 0.25$ where applicable. We highlight in bold the problem settings in which the estimated rate $O(1/N^m)$ noticeably exceeds the standard $O(1/N)$ rate.

f and C [35, 36] or given a strongly convex f and $x_\star \in \text{int}_\delta C$ [33, 34, 36]. However, it remains undetermined if acceleration can be achieved in any intermediate settings, including any case with smooth constraint sets.

Using our expanded performance estimation toolbox, we explore 16 permutations of problem instances: C being smooth or not; C being strongly convex or not; f being strongly convex or not; the minimizer having $g_\star \neq 0$ or having $g_\star = 0$ with $x_\star$ strictly interior to C . This enables us to assess if other settings achieve $O(1/N^2)$ convergence. Note that line-search or alternate assumptions on f (e.g., uniformly bounded gradients: $\|\nabla f(x)\| \geq \epsilon > 0$ for all $x \in C$) have shown improved (and sometimes linear) convergence [33, 35, 40, 41] but lie outside the scope of this preliminary survey. In Figure 6, we show our results for each of these settings for $N = 2, \dots, 15$ with stepsizes $h_k = \frac{2}{k+2}$, and estimate a rate $O(1/N^m)$ by regression on the values from $N = 8, \dots, 15$.

Our results show no clear evidence of any new problem settings that achieve $O(1/N^2)$ acceleration, with most problem settings having $m \approx 1$. However, in the settings with smooth functions, strongly convex sets, and an interior optimal point, our results seem to noticeably outperform the standard

$O(1/N)$ rate, instead being closer to $O(1/N^{1.2})$. This presents an intriguing area for future analysis. It is worth noting that some known acceleration results (e.g., [36]) require a burn-in phase, so it remains possible that acceleration in some of these interim settings could be missed by our survey.

Remark 7 (Improved Convergence Rates from Set Smoothness). We now briefly consider Frank-Wolfe’s performance over β -smooth sets. To date, no theory shows Frank-Wolfe benefits from smoothness in the constraint set C . Since the lower bound $\frac{LD^2}{4N}$ of [40] uses the simplex as a constraint set (which is nonsmooth), smoothness has the potential to break this bound. To focus on the effect of smoothness, we fix $\alpha = 0$, enabling the necessary and sufficient Theorem 3.3 to apply. Figure 7(a) provides an immediate positive answer to whether smoothness of constraints improves Frank-Wolfe’s performance. Local solves for our exact, nonconvex PEP formulation show up to an order of magnitude improvement in performance by $N = 15$ over the convex setting $\beta = \infty$. For $\beta = 2$ smooth sets, performance strictly outperforms Lan’s lower bound.

Remark 8 (Gaps in Accelerated Theory for Strongly Convex Functions). In [35], Garber and Hazan proved that for strongly convex functions, Frank-Wolfe attains an accelerated convergence rate of $O(1/N^2)$ if the constraint set C is strongly convex ($\alpha > 0$), or if $x_\star \in \text{int}_\delta C$ ($\delta > 0$). Their result relied on stepsize selection by linesearch, but Wirth et al. later proved in [36] that similar $O(1/N^2)$ convergence rates were attained in these settings using the open-loop sequence $h_k = \frac{4}{k+4}$. Specifically, in [36, Theorem E.1], if C is additionally α -strongly convex, then Frank-Wolfe has

$$f(x_N) - f(x_\star) \leq \frac{\frac{128L^2}{\alpha^2\mu} + 8LD^2}{(N+2)^2} . \quad (5.4)$$

Alternatively, in [36, Theorem 3.6], they show that if $x_\star \in \text{int}_\delta C$, then letting $M = \lceil 64LD^2/(\mu\delta^2) \rceil$, for any $N \geq M$, Frank-Wolfe satisfies

$$f(x_N) - f(x_\star) \leq \max \left\{ \frac{(M+3)^2}{(N+2)^2} (f(x_M) - f(x_\star)), \quad \frac{\frac{128L^2D^6}{\mu\delta^4} + 8LD^2}{(N+2)^2} \right\} . \quad (5.5)$$

In Figure 7(b), we compare the PEP result for $h_k = \frac{4}{k+4}$ ($\ell = 4$) with the convergence rate (5.4). In this setting of strongly convex sets with bounded diameter, we can no longer apply our necessary and sufficient interpolation of Theorem 3.3. As a consequence, by applying our interpolation conditions of $\text{Interp}(\alpha, 0, D; 1)$, we only obtain upper bounds on the method’s worst-case performance. These approximate PEP results outperform the guarantee of (5.4) by about two orders of magnitude. Additionally, we include in Figure 7(b) the performance for other values of ℓ (with $h_k = \frac{\ell}{k+\ell}$), seeing $\ell = 2$ performed best among the tested values and, in particular, outperforms $\ell = 4$. We repeat this analysis for the setting of a strongly convex function with an interior optimal point instead of a strongly convex constraint set. Note our necessary and sufficient Theorem 3.3 then applies. Our results in Figure 7(c) indicate that much tighter analysis is possible here as well⁵ with $\ell = 1$ performing best.

6 Gradient Methods for Epismooth Functions

As a final application, we show that our set interpolation theory can provide novel insights even in the context of unconstrained minimization. In particular, consider a gradient method (GM) applied

⁵Since (5.5) only applies after $M = \lceil 64LD^2/(\mu\delta^2) \rceil \geq 256$ iterations, we cannot compare directly with (5.5) due to our computational limits around $N = 15$.

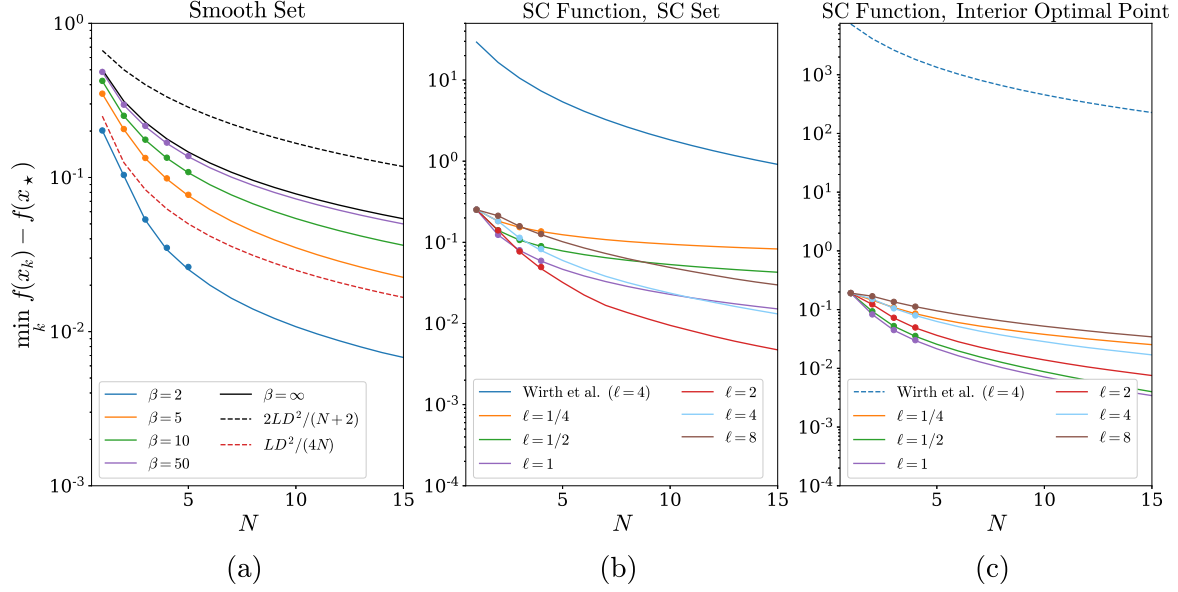


Figure 7: (a) PEP results for Frank-Wolfe over β -smooth sets with $h_k = \frac{2}{k+2}$ and for comparison, the lower and upper bounds of [39] and [40] for convex sets. (b) Results for μ -strongly convex functions over α -strongly convex sets with $h_k = \frac{\ell}{k+\ell}$ and for comparison, the bound (5.4), fixing $\mu = 0.5$ and $\alpha = 1.0$. (c) PEP results for μ -strongly convex functions with $x_* \in \text{int}_\delta C$ with $h_k = \frac{\ell}{k+\ell}$, fixing $\mu = 0.5$ and $\delta = 0.25$. Although the bound (5.5) does not apply until $N \geq 256$, it is included, dotted. All results above are obtained via local optimization but verified globally for small N , as indicated by individual points on the plot.

to minimize some differentiable, convex function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$. Much of the modern theory for gradient methods assumes f is L -smooth. This is a major restriction from the family of all differentiable functions, ruling out, for instance, all polynomials with degree greater than two.

Here, we consider applying first-order methods to a larger class of differentiable functions, namely functions with a smooth epigraph. We say f is **L -epigraphically smooth** (**L -epismooth**) if its epigraph $\text{epi } f = \{(x, t) \mid f(x) \leq t\}$ is an L -smooth set. Theorem 3.7 with $\eta = 1$ establishes that all L -smooth functions are L -epismooth, so this is a strictly broader class of functions. Note epismoothness of f does not guarantee that f has a full domain of $\mathbb{R}^d$ (consider a simple U-shaped function). Consequently, one cannot guarantee the iteration (GM) is well-defined in general, so additional care is needed. To handle this, we consider modified gradient methods of the form

$$x_{i+1} = x_i - \frac{1}{L} \sum_{j=0}^i H_{i,j} \frac{\nabla f(x_j)}{\sqrt{\|\nabla f(x_j)\|^2 + 1}}. \quad (\text{Epi-GM})$$

Normalizing our stepsizes above allows methods to keep $x_{i+1} - x_i$ bounded even as $\nabla f(x_k)$ becomes arbitrarily large (as shown in Figure 8). This property can then be used to ensure the iterates do not leave the domain of f : For example, fixing H to be the identity matrix, the above iteration becomes the gradient descent iteration $x_{i+1} = x_i - \frac{\nabla f(x_i)}{L\sqrt{\|\nabla f(x_i)\|^2 + 1}}$, which repeatedly moves to the minimizer of the ball majorizing f , ensuring progress every iteration.

To illustrate the breadth of this class of functions, the following proposition provides a useful means of checking if a function is **locally** epismooth.

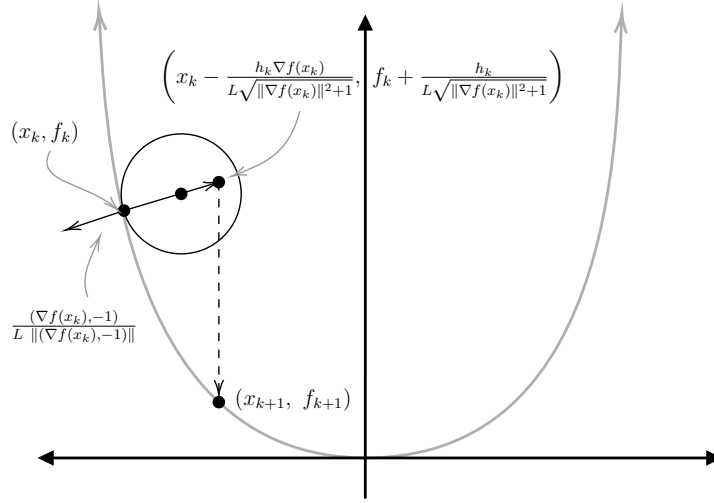


Figure 8: Modified gradient method for epismooth functions (GM).

Proposition 6.1. *For any $f \in \mathcal{C}^2$ and x in the domain of f , let*

$$M(x) = \frac{(I + \nabla f(x) \nabla f(x)^T)^{-1/2} (\nabla^2 f(x)) (I + \nabla f(x) \nabla f(x)^T)^{-1/2}}{\sqrt{\|\nabla f(x)\|^2 + 1}}. \quad (6.1)$$

Then f is locally L -epismooth at x if and only if $L \geq \lambda_{\max}(M)$.

Proof. We define a local ball upper bound of f at x by

$$f(y) \leq \mathbf{b}_x(y; L) := f(x) + \frac{1}{L \sqrt{\|\nabla f(x)\|^2 + 1}} - \sqrt{\frac{1}{L^2} - \|y - x + \frac{\nabla f(x)}{L \sqrt{\|\nabla f(x)\|^2 + 1}}\|^2} \quad (6.2)$$

where we say $\mathbf{b}_x(y; L) = \infty$ outside of its domain. From the definition of smoothness, f is L -epismooth if and only if for all x, y , $f(y) \leq \mathbf{b}_x(y; L)$. Observe that $\mathbf{b}_x(x; L) = f(x)$ and $\nabla \mathbf{b}_x(x; L) = \nabla f(x)$. Therefore, the upper bound $\mathbf{b}_x(y; L)$ holds locally if and only if $\nabla^2 f(x) \preceq \nabla^2 \mathbf{b}_x(x; L)$. We can compute $\nabla^2 \mathbf{b}_x(x; L) = L \sqrt{\|\nabla f(x)\|^2 + 1} (I + \nabla f(x) \nabla f(x)^T)$. Then rearranging and taking the maximum eigenvalue, we obtain our result. $\square$

Using this result, we can identify simple functions that are not smooth in the function sense of uniformly Lipschitz gradient but satisfy local epismoothness on their domain. For example, consider

$$f(x) = \begin{cases} -\log x & \text{if } x > 0 \\ \infty & \text{if } x \leq 0. \end{cases}$$

Using (6.1), we can calculate $\lambda_{\max}(M(x)) = \frac{1}{(1 + \frac{1}{x^2})^{3/2}} \leq \frac{2}{3\sqrt{3}}$ for all $x > 0$. So f is locally $\frac{2}{3\sqrt{3}}$ -epismooth everywhere on its domain. As a second, nontrivial example, one can verify for example $\|Ax - b\|_4^4$ is epismooth.

6.1 Worst-Case Performance of Epismooth Gradient Methods

The worst-case performance of a gradient method defined by a predetermined stepsize matrix H on a smooth function and on an epismooth function are closely related. In Section 2, we introduced

the PEP for smooth functions (2.3), and we now present an analogous PEP for the epismooth case. In both cases, we assume the initial point x_0 satisfies $\|x_0 - x_\star\| \leq R$ for some minimizer $x_\star$ of f . Lastly, to ensure that x_0 is in the domain of f (i.e., $f(x_0) < \infty$), we further require that $R \leq \frac{1}{L}$. Then the worst-case performance of (Epi-GM) for a given matrix H on an L -epismooth function is

$$p_{\text{ES}}(L, R) = \begin{cases} \max_{x_i, f} & f(x_N) - f(x_\star) \\ \text{s.t.} & \text{epi} f \in \mathcal{C}_{0, L, \infty} \\ & x_{i+1} = x_i - \frac{1}{L} \sum_{j=0}^i H_{i,j} \frac{\nabla f(x_j)}{\sqrt{\|\nabla f(x_j)\|^2 + 1}} \quad \forall i \in [0 : N-1] \\ & \|x_0 - x_\star\| \leq R, \quad \nabla f(x_\star) = 0. \end{cases} \quad (6.3)$$

For simplicity, we have suppressed the parameters N and H from our notation for p_{ES} . We again denote the PEP for performance of (GM) given H on an L -smooth function by $p_{\text{S}}(L, R)$. Both of these settings possess useful rescaling properties, proven in Appendix D.1.

Lemma 6.2. *For all $\eta > 0$ and any choice of N, L, R , one has that*

$$\begin{aligned} p_{\text{S}}(L, \eta R) &= \eta^2 p_{\text{S}}(L, R), & p_{\text{ES}}(L, \eta R) &= \eta p_{\text{ES}}(\eta L, R). \\ p_{\text{S}}(\eta L, R) &= \eta p_{\text{S}}(L, R), \end{aligned}$$

Observe that together the rescaling properties for p_{S} imply $p_{\text{S}}(L, \eta R) = \eta p_{\text{S}}(\eta L, R)$, matching that of the epismooth setting. However, notably, the individual rescaling properties for p_{S} do not hold for p_{ES} . These individual rescaling properties of p_{S} establish that it suffices to characterize $p_{\text{S}}(1, 1)$ to fully understand p_{S} . Since this does not hold for p_{ES} , any future characterizations of epismooth minimization methods must depend on the ratio of L and R .

Under an appropriate regularity condition, we find that the convergence of any epismooth minimization method converges to its classic smooth convergence rate as the ratio between L and R grows. That is, if initialized sufficiently close to $x_\star$, perhaps from some initial “burn-in” procedure, guarantees proven for L -smooth minimization can be lifted to L -epismooth minimization.

Formally, this result relies on the following regularity condition: A gradient method defined by H is **eventually-epismooth-stable** if for any L , there exist constants $\bar{R}, C > 0$ such that applying (Epi-GM) to any L -epismooth function with $\|x_0 - x_\star\| \leq R < \bar{R}$ must have $\|x_i - x_\star\| \leq CR$, $\|g_i\| \leq CLR$, and $|f_i| \leq CLR^2$ for all $i = 0, \dots, N$. This condition ensures the following asymptotic relation, proven in Appendix D.3.

Theorem 6.3. *For any eventually-epismooth-stable method defined by H ,*

$$\lim_{\eta \rightarrow 0} \frac{p_{\text{ES}}(L, \eta R)}{\eta^2} = \lim_{\eta \rightarrow 0} \frac{p_{\text{S}}(L, \eta R)}{\eta^2} = p_{\text{S}}(L, R).$$

Hence, after an initial burn-in producing x_0 with $\|x_0 - x_\star\|$ sufficiently small, optimization strategies for smooth functions such as momentum [16, 27] or recent long-step gradient descent techniques [21, 23] may be applicable to epismooth functions. One simple example of an eventually-epismooth-stable algorithm is applying (Epi-GM) with stepsizes $h \in (0, 2)$ (i.e., $H = \text{diag}(h, \dots, h)$). One can directly show this iteration guarantees descent and a nonincreasing distance to $x_\star$, as illustrated in Figure 8, ensuring the needed bounds.

Applying our interpolation theory, we verify this convergence of epismooth and smooth worst-case performances. Following Theorem 3.3, the same Grammian reformulation presented in previous sections provides a reformulation of $p_{\text{ES}}(L, R)$ as an SDP with an additional rank-1 constraint. We

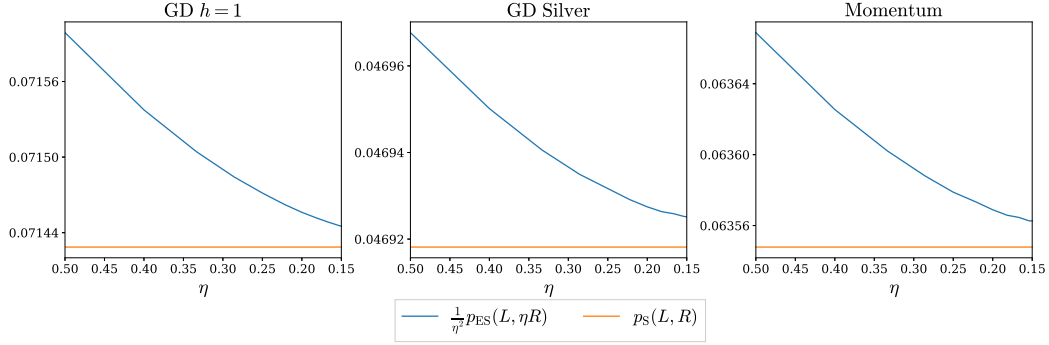


Figure 9: Convergence of $\frac{1}{\eta^2} p_{\text{ES}}(L, \eta R)$ and $p_{\text{S}}(L, R)$ as η decreases, fixing $N = 3$. Data is shown for three different (Epi-GM) algorithms: gradient descent with $h = 1$, gradient descent with Silver stepsizes [21], and Nesterov fast gradient method [27].

show a brief derivation in Appendix D.2. Unfortunately, the rank-1 constraint makes this problem nonconvex. Hence we approach it directly as a QCQP using black-box optimization software [25, 26].

Figure 9 considers the worst-case performance of several classic smooth optimization methods. For the stepsize matrices H corresponding to gradient descent (with $h_i = 1$), gradient descent with Silver stepsizes, and Nesterov’s fast gradient method, the worst-case behavior of (GM) and (Epi-GM) over their respective function classes converge as $\eta \rightarrow 0$, agreeing with Theorem 3.7.

Acknowledgments. This work was supported in part by the Air Force Office of Scientific Research under award number FA9550-23-1-0531. Benjamin Grimmer was additionally supported as a fellow of the Alfred P. Sloan Foundation.

References

- [1] Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: A novel approach. *Mathematical Programming*, 145(1–2):451–482, 2014.
- [2] 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.
- [3] 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.
- [4] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.
- [5] Adrien B. Taylor, Julien M. Hendrickx, and François Glineur. Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. *Journal of Optimization Theory and Applications*, 178:455–476, 2018.
- [6] Donghwan Kim. Accelerated proximal point method for maximally monotone operators. *Mathematical Programming*, 1190:57–87, 2021.
- [7] Moslem Zamani and François Glineur. Exact convergence rate of the subgradient method by using Polyak step size. *arXiv:2407.15195*, 2024.
- [8] Adrien Taylor and Francis Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pages 2934–2992. PMLR, 2021.

- [9] Etienne de Klerk, François Glineur, and Adrien B. Taylor. On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. *Optimization Letters*, 11:1185–1199, 2017.
- [10] Yoel Drori and Adrien B. Taylor. Efficient first-order methods for convex minimization: a constructive approach. *Mathematical Programming*, 184(1–2):183–220, 2020.
- [11] Adrien Taylor and Yoel Drori. An optimal gradient method for smooth strongly convex minimization. *Mathematical Programming*, 199:557–594, 2023.
- [12] Felix Lieder. On the convergence rate of the halpern-iteration. *Optimization Letters*, 15:405–418, 2021.
- [13] Jisun Park and Ernest K Ryu. Exact optimal accelerated complexity for fixed-point iterations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17420–17457. PMLR, 17–23 Jul 2022.
- [14] Baptiste Goujaud, Céline Moucer, François Glineur, Julien Hendrickx, Adrien Taylor, and Aymeric Dieuleveut. PEPit: computer-assisted worst-case analyses of first-order optimization methods in Python. *Mathematical Programming Computation*, 16:1–31, 08 2024.
- [15] Shuvomoy Das Gupta, Bart P.G. Van Parys, and Ernest K. Ryu. Branch-and-bound performance estimation programming: A unified methodology for constructing optimal optimization method. *Mathematical Programming*, 204:567–639, 2024.
- [16] Donghwan Kim and Jeffrey A. Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1–2):81–107, 2016.
- [17] Yoel Drori. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39:1–16, 2017.
- [18] Donghwan Kim and Jeffrey A. Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *Journal of Optimization Theory and Applications*, 188:192–219, 2021.
- [19] Bryan Van Scoy, Randy A. Freeman, and Kevin M. Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54, 2018.
- [20] Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging: Multi-step descent and the silver stepsize schedule. *J. ACM*, 72(2), March 2025.
- [21] Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging: Silver stepsize schedule for smooth convex optimization. *Mathematical Programming*, 213:1105–1118, 2025.
- [22] Benjamin Grimmer, Kevin Shu, and Alex L. Wang. Accelerated gradient descent via long steps. *arXiv:2309.09961*, 2023.
- [23] Benjamin Grimmer, Kevin Shu, and Alex L. Wang. Accelerated objective gap and gradient norm convergence for gradient descent via long steps. *INFORMS Journal on Optimization*, 7:156–169, 2025.
- [24] MOSEK ApS. *MOSEK Optimizer API for Julia manual. Version 10.0.*, 2024.
- [25] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2024.
- [26] Iain Dunning, Joey Huchette, and Miles Lubin. JuMP: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017.
- [27] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.
- [28] Jean-Philippe Vial. Strong convexity of sets and functions. *Journal of Mathematical Economics*, 9(1):187–205, 1982.

- [29] Jean-Philippe Vial. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.
- [30] Vladimir V. Goncharov and Grigorii E. Ivanov. *Strong and Weak Convexity of Closed Sets in a Hilbert Space*, pages 259–297. Springer International Publishing, 2017.
- [31] Ning Liu and Benjamin Grimmer. Gauges and accelerated optimization over smooth and/or strongly convex sets. *arXiv:2303.05037*, 2023.
- [32] Heinrich Jung. Ueber die kleinste kugel, die eine räumliche figur einschliesst. *Journal für die reine und angewandte Mathematik*, 123:241–257, 1901.
- [33] Philip Wolfe. Convergence theory in nonlinear programming. *Integer and Nonlinear Programming*, pages 1–36, 1970.
- [34] J. GuéLat and P. Marcotte. Some comments on Wolfe’s ‘away step’. *Mathematical Programming*, 35:110–119, 1986.
- [35] 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.
- [36] Elias Wirth, Thomas Kerdreux, and Sebastian Pokutta. Acceleration of Frank-Wolfe algorithms with open-loop step-sizes. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 77–100. PMLR, 2023.
- [37] Garth P. McCormick. Computability of global solutions to factorable nonconvex programs: Part I - convex underestimating problems. *Mathematical Programming*, 10:147–175, 1976.
- [38] Maria Lorena Bergamini, Pio Aguirre, and Ignacio Grossmann. Logic-based outer approximation for globally optimal synthesis of process networks. *Computers & Chemical Engineering*, 29(9):1914–1933, 2005.
- [39] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 427–435. PMLR, 2013.
- [40] Guanhui Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv:1309.5550*, 2014.
- [41] Elias Wirth, Javier Pena, and Sebastian Pokutta. Accelerated affine-invariant convergence rates of the Frank-Wolfe algorithm with open-loop step-sizes. *Mathematical Programming*, 2025.

A Deferred Interpolation Proofs

A.1 Proof of Theorem 3.7

We first introduce notation and a key lemma. Given $\{(x_i, g_i, f_i)\}_{i \in \mathcal{I}}$, define

$$\begin{aligned} \hat{Q}_{\mu, L}^{i, j}(x, g, f) &= f_j - f_i - \langle g_i, x_j - x_i \rangle \\ &\quad - \frac{1}{2(1 - \mu/L)} \left(\frac{1}{L} \|g_i - g_j\|^2 + \mu \|x_i - x_j\|^2 - \frac{2\mu}{L} \langle g_i - g_j, x_i - x_j \rangle \right). \end{aligned}$$

Recall from Theorem 2.2 that $\hat{Q}_{\mu, L}^{i, j}(x, g, f) \geq 0$ for all $i, j \in \mathcal{I}$ if and only if $\{(x_i, g_i, f_i)\}_{i \in \mathcal{I}}$ is $\mathcal{F}_{\mu, L}$ -interpolable. Next, let $z_i = (x_i, f_i)$ and $v_i = (g_i, -1)$, and let $n_i = \frac{v_i}{\|v_i\|}$. We further define

$$Q_{\mu, L}^{i, j}(x, g, f) = \frac{\mu}{2(1 - \mu/L)} \left(\left(\frac{1}{\mu} - \frac{1}{L} \right)^2 - \left\| z_j - \frac{1}{L} n_j - z_i + \frac{1}{\mu} n_i \right\|^2 \right)$$

$$= \frac{L - \mu}{2L\mu} - \frac{\mu}{2(1 - \mu/L)} \left(\left\| x_j - \frac{g_j}{L\|v_j\|} - x_i + \frac{g_i}{\mu\|v_i\|} \right\|^2 + \left(f_j + \frac{1}{L\|v_j\|} - f_i - \frac{1}{\mu\|v_i\|} \right)^2 \right).$$

From Theorem 3.2, we know that $Q_{\mu,L}^{i,j}(x, g, f) \geq 0$ for all $i, j \in \mathcal{I}$ if and only if $(\{(x_i, f_i), (g_i, -1)\}_{i \in \mathcal{I}}, \emptyset)$ is $\mathcal{C}_{\mu,L,\infty}$ -interpolable (This is a direct application of (Interp1)). We can relate these conditions through the following lemma, which is proven by a direct Taylor series expansion.

Lemma A.1. *For any $\eta \geq 0$,*

$$Q_{\eta^2\mu, \eta^2L}^{i,j}\left(\frac{x}{\eta}, \eta g, f\right) = \hat{Q}_{\mu,L}^{i,j}(x, g, f) + \eta^2 c(x, g, f, \mu, L, \eta)$$

where $\eta^2 c(x, g, f, \mu, L, \eta) \rightarrow 0$ as $\eta \rightarrow 0$.

Next we consider smoothness of the epigraph of f , $\text{epi}f$. This idea is explored further in Section 6 but is necessary here for our proof. Note for any differentiable convex function f , $(\nabla f(x), -1) \in N_{\text{epi}f}(x, f(x))$. Then, by definition, $\text{epi}f$ being an L -smooth set is equivalent to a local ball upper bound $\mathbf{b}_x(y; L)$ holding for all x, y (See (6.2) for explicit expression). This ball upper bound $\mathbf{b}_x(y; L)$ is strictly larger than the analogous quadratic bound for smooth functions. That is, for all x, y , we have

$$\mathbf{b}_x(y; L) \geq \mathbf{q}_x(y; L) := f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|x - y\|^2.$$

Now, we are ready to prove Theorem 3.7.

($\Rightarrow$) Suppose that $\{(x_i, g_i, f_i)\}_{i \in \mathcal{I}}$ is $\mathcal{F}_{0,L}$ -interpolable. So there exists an L -smooth convex function f such that $f(x_i) = f_i$ and $\nabla f(x_i) = g_i$. Observe that for any $\eta > 0$, $\hat{f}$ defined by $\hat{f}(x) = f(\eta x)$ is an $\eta^2 L$ -smooth convex function with $\hat{f}(\frac{x_i}{\eta}) = f(x_i) = f_i$ and $\nabla \hat{f}(\frac{x_i}{\eta}) = \eta \nabla f(\eta \frac{x_i}{\eta}) = \eta g_i$. Then for all x, y , $\hat{f}(y) \leq \mathbf{q}_x(y; \eta^2 L) \leq \mathbf{b}_x(y; \eta^2 L)$. As a result, $\text{epi} \hat{f}$ is $\eta^2 L$ -smooth. Finally, since for any x , $(\nabla \hat{f}(x), -1) \in N_{\text{epi} \hat{f}}(x, \hat{f}(x))$, this implies that $\text{epi} \hat{f}$ interpolates $(\{(\frac{x_i}{\eta}, f_i), (\eta g_i, -1)\}_{i \in \mathcal{I}}, \emptyset)$.

($\Leftarrow$) Suppose that $(\{(\frac{x_i}{\eta}, f_i), (\eta g_i, -1)\}_{i \in \mathcal{I}}, \emptyset)$ is $\mathcal{C}_{0, \eta^2 L, \infty}$ -interpolable for all $\eta > 0$. By our interpolability theorems, we know that $Q_{0, \eta^2 L}^{i,j}(\frac{x}{\eta}, \eta g, f) \geq 0$ for all i, j . Applying Lemma A.1 yields, for any $\eta > 0$,

$$\hat{Q}_{0,L}^{i,j}(x, g, f) = Q_{0, \eta^2 L}^{i,j}\left(\frac{x}{\eta}, \eta g, f\right) + \eta^2 c(x, g, f, \mu, L, \eta) \geq \eta^2 c(x, g, f, \mu, L, \eta).$$

Considering this as η tends to 0 implies $\hat{Q}_{0,L}^{i,j}(x, g, f) \geq 0$ from which Theorem 2.2 ensures $\{(x_i, g_i, f_i)\}_{i \in \mathcal{I}}$ is $\mathcal{F}_{0,L}$ -interpolable.

B Deferred Calculations on Separating Hyperplane Algorithms

B.1 An SDP Reformulation of the Stopping Decision Problem

We follow the approach in many previous works [2, 3, 10] by converting our finite-dimensional problem (4.2) into an SDP. We follow similar notation to that of [10] and introduce it below. First, define

$$\Lambda = [x_0 | z_0 | z_1 | \dots | z_N | n_0 | n_1 | \dots | n_N | w] \in \mathbb{R}^{d \times (2N+4)}$$

$$G = \Lambda^T \Lambda \in \mathbb{S}_+^{2N+4}$$

Next, using the standard unit basis vectors e_i , define special selection vectors for extracting particular elements of our matrix G :

$$\begin{aligned} \mathbf{x}_0 &= e_1 \in \mathbb{R}^{2N+4} \\ \mathbf{z}_i &= e_{i+2} \in \mathbb{R}^{2N+4} & i \in [0:N] \\ \mathbf{n}_i &= e_{i+(N+3)} \in \mathbb{R}^{2N+4} & i \in [0:N] \\ \mathbf{w} &= e_{2N+4} \in \mathbb{R}^{2N+4} \\ \mathbf{x}_{i+1} &= \mathbf{x}_i - \sum_{j=0}^i H_{i,j} \mathbf{n}_j \in \mathbb{R}^{2N+4} & i \in [0:N-1] . \end{aligned}$$

This construction yields the useful identities $\Lambda \mathbf{x}_i = x_i$, $\Lambda \mathbf{z}_i = z_i$, $\Lambda \mathbf{n}_i = n_i$, and $\Lambda \mathbf{w} = w$ (lastly, recall our assumption that $q = 0$). Denote the symmetric outer product by $x \odot y = \frac{1}{2}xy^T + \frac{1}{2}yx^T$. Then we can directly express the standard dot product of our vectors as $\text{Tr}(G(\mathbf{x}_i \odot \mathbf{z}_j)) = \langle x_i, z_j \rangle$ for any $i, j \in [0:N]$ (and this holds analogously for other vector combinations). Lastly, we invoke the assumption that the problem dimension d satisfies $d \geq 2N+4$ to guarantee that any identified G can be factorized into some $\Lambda \in \mathbb{R}^{d \times (2N+4)}$, which is common practice throughout PEP literature [1] (See Remark 1).

We can then write our feasibility condition (4.2) as

$$\exists G \in \mathbb{S}_+^{2N+4} \text{ s.t. } \left\{ \begin{array}{l} G \succeq 0 \\ \text{Tr} \left(G \left((\mathbf{z}_i - \frac{1}{\alpha} \mathbf{n}_i - \mathbf{z}_j + \frac{1}{\beta} \mathbf{n}_j) \odot (\mathbf{z}_i - \frac{1}{\alpha} \mathbf{n}_i - \mathbf{z}_j + \frac{1}{\beta} \mathbf{n}_j) \right) \right) \leq \frac{1}{\gamma^2} \\ \text{Tr} \left(G \left((\mathbf{z}_i - \frac{1}{\alpha} \mathbf{n}_i - \mathbf{w}) \odot (\mathbf{z}_i - \frac{1}{\alpha} \mathbf{n}_i - \mathbf{w}) \right) \right) \leq (\frac{1}{\gamma} - s)^2 \\ \text{Tr} (G(\mathbf{w} \odot \mathbf{w})) \leq (\frac{1}{\beta} - \delta + s)^2 \\ \text{Tr} (G(\mathbf{n}_i \odot (\mathbf{z}_i - \mathbf{x}_i))) \leq 0 \\ \text{Tr} (G(\mathbf{n}_i \odot \mathbf{n}_i)) = 1 \\ \text{Tr} (G(\mathbf{x}_0 \odot \mathbf{x}_0)) \leq R^2 \end{array} \right. \quad (\text{B.1})$$

with $\gamma = (\frac{1}{\alpha} - \frac{1}{\beta})^{-1}$ and $s = \max\{0, \delta - \frac{1}{\beta}\}$.

B.2 Proof of Theorem 4.1

We first claim that there exists $\hat{q}$ such that $B(\hat{q}, h) \subseteq C$ and $\|\hat{q} - q\| \leq h - \delta$. If $\frac{1}{\beta} \leq \delta$, then $h = \delta$ and we simply let $\hat{q} = q$ and the claim is trivial. Otherwise, if $\frac{1}{\beta} > \delta$, let $z \in \text{argmin}_{y \in \text{bdry } C} \|y - q\|$. The optimality of z ensures $z - q \in N_C(z)$ and hence, by smoothness of C , $B(z - \frac{1}{\beta} \frac{z-q}{\|z-q\|}, \frac{1}{\beta}) \subseteq C$. If $\|z - q\| > \frac{1}{\beta}$, then $B(q, \frac{1}{\beta}) \subseteq C$, so we again let $\hat{q} = q$, with $h = \frac{1}{\beta}$. Lastly, suppose $\|z - q\| \leq \frac{1}{\beta}$. Then setting $\hat{q} = z - \frac{1}{\beta} \frac{z-q}{\|z-q\|}$ yields

$$\|q - \hat{q}\| = \|z - \frac{1}{\beta} \frac{z-q}{\|z-q\|} - q\| = \frac{1}{\beta} - \|z - q\| \leq \frac{1}{\beta} - \delta = h - \delta$$

where the inequality follows from $z \in \text{bdry } C$, so $\|z - q\| \geq \delta$. Since $B(\hat{q}, \frac{1}{\beta}) \subseteq C$, this proves our first claim.

Proof of Claimed Stopping Time. Suppose that by iteration k , the method has not yet terminated. Since $x_k \notin \text{int } C$, $x_k \notin \text{int } B(\hat{q}, h)$. By definition of n_k as a separating hyperplane, we know

$\langle n_k, y - x_k \rangle \leq 0$ for all $y \in B(\hat{q}, h)$. Considering $y = \hat{q} + hn_k$, this yields $\langle n_k, x_k - \hat{q} \rangle \geq h$. Using this fact, every iteration k satisfies

$$\|x_{k+1} - \hat{q}\|^2 = \|x_k - \hat{q}\|^2 + h^2 - 2h\langle n_k, x_k - \hat{q} \rangle \leq \|x_k - \hat{q}\|^2 - h^2.$$

Inductively applying this ensures that if all $x_0, \dots, x_N \notin \text{int } C$, then

$$\|x_N - \hat{q}\|^2 \leq \|x_0 - \hat{q}\|^2 - Nh^2 \leq (R + h - \delta)^2 - Nh^2 \quad (\text{B.2})$$

where we use the fact that if $\delta > \frac{1}{\beta}$, then $R + h - \delta = R \geq \|x_0 - q\| = \|x_0 - \hat{q}\|$, and if $\delta \leq \frac{1}{\beta}$, then $\|x_0 - \hat{q}\| \leq \|x_0 - q\| + \|q - \hat{q}\| \leq R + (\frac{1}{\beta} - \delta) = R + h - \delta$. Noting $x_N \notin \text{int } C$ implies that $x_N \notin \text{int } B(\hat{q}, h)$, i.e. $\|x_N - \hat{q}\| \geq h$, it follows from $(R + h - \delta)^2 - Nh^2 \geq \|x_N - \hat{q}\|^2 \geq h^2$ that $N \leq \frac{(R+h-\delta)^2}{h^2} - 1$. Hence after $N = \lfloor \frac{(R+h-\delta)^2}{h^2} \rfloor$ iterations, some iterate x_k must have lied in the interior of C , halting the algorithm.

Proof of Matching Lower Bound. Finally, we establish an exactly matching lower bound on any method of the form (4.1). For simplicity, we will assume that $N = \frac{(R+h-\delta)^2}{h^2}$ is an integer, however, one can adjust our construction to handle non-integer cases. Below we construct a hard problem instance such that for any separating hyperplane method, $x_k \notin \text{int } C$ for all $k \leq N - 1$.

Let $x_0 = 0$, $q = (R/\sqrt{N}, \dots, R/\sqrt{N}) \in \mathbb{R}^N$, and $\hat{q} = (h, \dots, h)$. Further define $C = B(\hat{q}, h) \subseteq \mathbb{R}^N$ and observe that $q \in \text{int}_\delta C$ and $\|x_0 - q\| = (R/\sqrt{N})\sqrt{N} = R$. For each k , we can select $n_k = -e_{k+1}$, the negative unit basis vector in $\mathbb{R}^N$. Note this is a valid separating hyperplane due to x_k 's support being its first k entries as $\langle n_k, x_k \rangle = 0 \geq \langle n_k, y \rangle$ for all $y \in C$. Moreover, this choice of n_k ensures the support of x_{k+1} will only increase by one. Noting all $k \leq N - 1$ have final entry equal to zero, $\|x_k - \hat{q}\| \geq h$ and so $x_k \notin \text{int } C$. Therefore no separating hyperplane method can identify an interior point faster than the above simple constant stepsize method.

C Deferred Calculations on Frank-Wolfe Methods

To fully justify our PEP formulation, we need to derive two simple propositions. We will once again consider separately the cases where $g_\star = 0$ and $g_\star \neq 0$ (See Section 5.1). In the results below, let $\mathcal{I} = [0 : N - 1]$ and $\mathcal{K} = [0 : N]$, along with $\mathcal{I}_\star = \mathcal{I} \cup \{\star\}$ and $\mathcal{K}_\star = \mathcal{K} \cup \{\star\}$.

Proposition C.1. *Consider any points $z_i, x_\star$, and $x_{i+1} = (1 - h_i)x_i + h_i z_i$ and vectors g_i and define*

$$S_{g_\star \neq 0} = (\{(z_i, -g_i)\}_{i \in \mathcal{I}} \cup \{(x_\star, -g_\star)\}, \{(x_0, 0)\})$$

where $g_\star \neq 0$. Then there exist $f \in \mathcal{F}_{\mu, L}$ and $C \in \mathcal{C}_{\alpha, \beta, D}$ satisfying $z_i \in \arg\min_{y \in C} \langle \nabla f(x_i), y \rangle$ for all $i \in \mathcal{I}$, $x_\star \in \arg\min_{y \in C} f(y)$, and $x_0 \in C$ if and only if $S_{g_\star \neq 0}$ is $\mathcal{C}_{\alpha, \beta, D}$ -interpolable and $\{(x_k, g_k, f_k)\}_{k \in \mathcal{K}_\star}$ is $\mathcal{F}_{\mu, L}$ -interpolable, where $f(x_k) = f_k$ and $\nabla f(x_k) = g_k$ for all $k \in \mathcal{K}_\star$.

Proof. ($\Rightarrow$) Suppose that there exist $f \in \mathcal{F}_{\mu, L}$ and $C \in \mathcal{C}_{\alpha, \beta, D}$ such that $x_\star \in \arg\min_{y \in C} f(y)$ with $\nabla f(x_\star) \neq 0$ and applying Frank-Wolfe with $x_0 \in C$ yields $z_i \in \arg\min_{y \in C} \langle \nabla f(x_i), y \rangle$ for all $i \in \mathcal{I}$. By construction, setting $g_k = \nabla f(x_k)$ and $f(x_k) = f_k$ for all $k \in \mathcal{K}_\star$ ensures $\{(x_k, g_k, f_k)\}_{k \in \mathcal{K}_\star}$ is $\mathcal{F}_{\mu, L}$ -interpolable. Since C is convex, the optimality condition defining $z_i \in \arg\min_{y \in C} \langle g_i, y \rangle$ implies that $z_i \in C$ and $-g_i \in N_C(z_i)$. Similarly, the optimality condition for $x_\star$ ensures that $-g_\star \in N_C(x_\star)$. Since $x_0 \in C$, we have that $S_{g_\star \neq 0}$ is $\mathcal{C}_{\alpha, \beta, D}$ -interpolable.

($\Leftarrow$) Suppose that $S_{g_\star \neq 0}$ is $\mathcal{C}_{\alpha, \beta, D}$ -interpolable and $\{(x_k, g_k, f_k)\}_{k \in \mathcal{K}_\star}$ is $\mathcal{F}_{\mu, L}$ -interpolable. That is, there exists a μ -strongly convex, L -smooth function f such that $f(x_k) = f_k$ and $\nabla f(x_k) = g_k$ for all $k \in \mathcal{K}_\star$ and an α -strongly convex, β -smooth set C such that $x_0 \in C$ and $z_i \in C$ and

$-g_i \in N_C(z_i)$ for all $i \in \mathcal{I}_*$. Since C is convex, we know that $-\nabla f(x_i) \in N_C(z_i)$ implies that $z_i \in \operatorname{argmin}_{y \in C} \langle \nabla f(x_i), y \rangle$. Similarly, $-\nabla f(x_*) \in N_C(x_*)$ implies that $x_* \in \operatorname{argmin}_{y \in C} f(y)$. Hence the desired f and C exist. $\square$

Proposition C.2. Consider any points z_i, x_* , and $x_{i+1} = (1 - h_i)x_i + h_i z_i$ and vectors g_i and define

$$S_{g_*=0}^\delta = (\{(z_i, -g_i)\}_{i \in \mathcal{I}}, \{(x_0, 0), (x_*, \delta)\}) .$$

Then there exist $f \in \mathcal{F}_{\mu, L}$ and $C \in \mathcal{C}_{\alpha, \beta, D}$ satisfying $z_i \in \operatorname{argmin}_{y \in C} \langle \nabla f(x_i), y \rangle$ for all $i \in \mathcal{I}$ with $x_0 \in C$ and $x_* \in \operatorname{argmin}_{y \in C} f(y)$ with $x_* \in \operatorname{int}_\delta C$ if and only if $\nabla f(x_*) = 0$, $S_{g_*=0}^\delta$ is $\mathcal{C}_{\alpha, \beta, D}$ -interpolable, and $\{(x_k, g_k, f_k)\}_{k \in \mathcal{K}_*}$ is $\mathcal{F}_{\mu, L}$ -interpolable, where $f(x_k) = f_k$ and $\nabla f(x_k) = g_k$ for all $k \in \mathcal{K}_*$.

Proof. We apply the same approach as above, but now use the fact that $x_* = \operatorname{argmin}_{y \in C} f(y)$ and $x_* \in \operatorname{int}_\delta C$ imply $\nabla f(x_*) = 0$, and $\nabla f(x_*) = 0$ implies $x_* = \operatorname{argmin}_{y \in C} f(y)$. $\square$

D Deferred Calculations on Epismooth Gradient Methods

D.1 Proof of Lemma 6.2

Denote $b_i = \frac{\nabla f(x_i)}{\sqrt{\|\nabla f(x_i)\|^2 + 1}} \in \mathbb{R}^d$ and $t_i = \frac{-1}{\sqrt{\|\nabla f(x_i)\|^2 + 1}} \in \mathbb{R}$. Then this follows from a simple change of variables, $\tilde{x}_i = \frac{x_i}{\eta}$ and $\tilde{f}_i = \frac{f_i}{\eta}$, as

$$\begin{aligned} p_{\text{ES}}(L, \eta R) &= \left\{ \begin{array}{ll} \max_{x_i, b_i, f_i} & f_N \\ \text{s.t.} & x_{i+1} = x_i - \frac{1}{L} \sum_{j=0}^i H_{i,j} b_j \quad \forall i \in \mathcal{I} \\ & \|x_0 - x_*\|^2 \leq \eta^2 R^2 \\ & b_* = 0, \quad f_* = 0, \quad x_* = 0 \\ & t_k = -\sqrt{1 - \|b_k\|^2} \quad \forall k \in \mathcal{K}_* \\ & \langle b_k, x_l - \frac{1}{L} b_l - x_k + \frac{1}{L} b_k \rangle \\ & \quad + \langle t_k, f_l - \frac{1}{L} t_l - f_k + \frac{1}{L} t_k \rangle \leq 0 \quad \forall k, l \in \mathcal{K}_* \end{array} \right. \\ &= \left\{ \begin{array}{ll} \max_{\tilde{x}_i, b_i, \tilde{f}_i} & \eta \tilde{f}_N \\ \text{s.t.} & \tilde{x}_{i+1} = \tilde{x}_i - \frac{1}{\eta L} \sum_{j=0}^i H_{i,j} b_j \quad \forall i \in \mathcal{I} \\ & \|\tilde{x}_0 - \tilde{x}_*\|^2 \leq R^2 \\ & b_* = 0, \quad \tilde{f}_* = 0, \quad \tilde{x}_* = 0 \\ & t_k = -\sqrt{1 - \|b_k\|^2} \quad \forall k \in \mathcal{K}_* \\ & \langle b_k, \tilde{x}_l - \frac{1}{\eta L} b_l - \tilde{x}_k + \frac{1}{\eta L} b_k \rangle \\ & \quad + \langle t_k, \tilde{f}_l - \frac{1}{\eta L} t_l - \tilde{f}_k + \frac{1}{\eta L} t_k \rangle \leq 0 \quad \forall k, l \in \mathcal{K}_* \end{array} \right. \\ &= \eta p_{\text{ES}}(\eta L, R) . \end{aligned}$$

Similarly, for p_{S} , the first result follows using the change of variables $\tilde{x}_i = \frac{x_i}{\eta}$, $\tilde{g}_i = \frac{g_i}{\eta}$ and $\tilde{f}_i = \frac{f_i}{\eta^2}$, and the second from $\tilde{g}_i = \frac{g_i}{\eta}$ and $\tilde{f}_i = \frac{f_i}{\eta}$.

D.2 Derivation of SDP with Rank Constraints for Epismooth Gradient Methods

Assume without loss of generality that $x_\star = 0$ and $f_\star = 0$. Using b_i and t_i as defined in Appendix D.1, we define

$$\Lambda = [x_0 | b_0 | \dots | b_N] \in \mathbb{R}^{d \times (N+2)}$$

$$v = (f_0, \dots, f_N, t_\star, t_0, \dots, t_N) \in \mathbb{R}^{1 \times (2N+3)}$$

with $G = \Lambda^T \Lambda \in \mathbb{S}_+^{N+2}$ and $F = v^T v \in \mathbb{S}_+^{2N+3}$. We define $\mathbf{x}_i$ and $\mathbf{b}_i$ as selection vectors relative to G , similar to Appendix B, and we define $\mathbf{f}_i$ and $\mathbf{t}_i$ as selection vectors relative to F . We can then encode $p_{\text{ES}}(L, R)$ as an SDP with an additional rank-1 constraint:

$$(p_{\text{ES}}(L, R))^2 = \left\{ \begin{array}{ll} \max_{F, G, v} & \text{Tr}(F(\mathbf{f}_N \odot \mathbf{f}_N)) \\ \text{s.t.} & G \succeq 0, \quad F \succeq 0 \\ & \text{Tr}(G((\mathbf{x}_0 - \mathbf{x}_\star) \odot (\mathbf{x}_0 - \mathbf{x}_\star))) \leq R^2 \\ & \text{Tr}(G(\mathbf{b}_k \odot \mathbf{b}_k) + F(\mathbf{t}_k \odot \mathbf{t}_k)) = 1 \quad \forall k \in \mathcal{K}_\star \\ & \text{Tr}\left(G\left(\mathbf{b}_k \odot \left(\mathbf{x}_1 - \frac{1}{L}\mathbf{b}_1 - \mathbf{x}_k + \frac{1}{L}\mathbf{b}_k\right)\right) \right. \\ & \quad \left. + F\left(\mathbf{t}_k \odot \left(\mathbf{f}_1 - \frac{1}{L}\mathbf{t}_1 - \mathbf{f}_k + \frac{1}{L}\mathbf{t}_k\right)\right)\right) \leq 0 \quad \forall k, l \in \mathcal{K}_\star \\ & \text{Tr}(F(\mathbf{t}_\star \odot \mathbf{t}_\star)) = 1 \\ & \text{Tr}(F(\mathbf{t}_\star \odot \mathbf{f}_0)) \leq 0 \\ & \text{Tr}(F(\mathbf{t}_k \odot \mathbf{f}_0)) \leq 0 \quad \forall k \in \mathcal{K} \\ & F = v^T v. \end{array} \right. \quad (\text{D.1})$$

In order to enforce $t_\star = -1$, we use the two constraints $t_\star^2 = 1$ and $t_\star f_0 \leq 0$. The first ensures that $t_\star = \pm 1$. By convexity of f , we know that $f_0 \geq f_\star = 0$. If $f_0 = 0$, then $b_0 = 0$, and consequently $f_N = 0$, so this is irrelevant for the worst-case instance. So we know effectively that $f_0 > 0$ and conclude that $t_\star = -1$. Similarly, to enforce that $t_k \leq 0$, we require $t_k f_0 \leq 0$.

D.3 Proof of Theorem 6.3

From Lemma 6.2, we know that $\frac{1}{\eta^2} p_{\text{S}}(L, \eta R) = p_{\text{S}}(L, R)$ for all η . So the second equality is immediate. It remains to show $\lim_{\eta \rightarrow 0} \frac{p_{\text{ES}}(L, \eta R)}{\eta^2} = p_{\text{S}}(L, R)$.

Let $\rho_\eta := (x^{(\eta)}, g^{(\eta)}, f^{(\eta)})$ denote the sequence of maximizers of $p_{\text{ES}}(\eta^2 L, \frac{R}{\eta})$ as $\eta \rightarrow 0$. We consider the rescaled sequence $s_\eta = (\eta x^{(\eta)}, \frac{g^{(\eta)}}{\eta}, f^{(\eta)})$. By our method being eventually-epismooth-stable and the fact that ρ_η is a solution to $p_{\text{ES}}(\eta^2 L, \frac{R}{\eta})$, we must have $\|x_i^{(\eta)}\| \leq C \frac{R}{\eta}$, $\|g_i^{(\eta)}\| \leq C \eta^2 L \frac{R}{\eta}$, and $|f_i^{(\eta)}| \leq C \eta^2 L \frac{R^2}{\eta^2}$ for all i . Consequently, we have $\|\eta x_i^{(\eta)}\| \leq CR$, $\|\frac{g_i^{(\eta)}}{\eta}\| \leq CLR$, and $|f_i^{(\eta)}| \leq CLR^2$ for all i . This shows that s_η belongs to a compact set for all η . We consider $\limsup_{\eta \rightarrow 0} p_{\text{ES}}(\eta L, \frac{R}{\eta})$ and define η_k as the subsequence attaining the lim sup. By compactness, consider a further subsequence $\eta_{k'}$ such that $s_{\eta_{k'}}$ converges to some limit point s^* .

Next, we claim that s^* is a feasible point for $p_{\text{S}}(L, R)$. We know that each $\rho_\eta = (x^{(\eta)}, g^{(\eta)}, f^{(\eta)})$ satisfies the constraints

$$\left\{ \begin{array}{l} x_{i+1}^{(\eta)} = x_i^{(\eta)} - \frac{1}{\eta^2 L} \sum_{j=0}^i H_{i,j} \frac{g_j^{(\eta)}}{\sqrt{\|g_j^{(\eta)}\|^2 + 1}} \\ \|x_0^{(\eta)} - x_\star^{(\eta)}\|^2 \leq \frac{R^2}{\eta^2} \\ Q_{0, \eta^2 L}^{i,j}(x^{(\eta)}, g^{(\eta)}, f^{(\eta)}) \geq 0. \end{array} \right.$$

Then rescaling with $\tilde{x}^{(\eta)} = \eta x^{(\eta)}$, $\tilde{g}^{(\eta)} = \frac{g^{(\eta)}}{\eta}$, and $\tilde{f}^{(\eta)} = f^{(\eta)}$ yields

$$\left\{ \begin{array}{l} \tilde{x}_{i+1}^{(\eta)} = \tilde{x}_i^{(\eta)} - \frac{1}{L} \sum_{j=0}^i H_{i,j} \frac{\tilde{g}_j^{(\eta)}}{\sqrt{\eta^2 \|\tilde{g}_j^{(\eta)}\|^2 + 1}} \\ \|\tilde{x}_0^{(\eta)} - \tilde{x}_\star^{(\eta)}\|^2 \leq R^2 \\ Q_{0,\eta^2 L}^{i,j}(\frac{\tilde{x}^{(\eta)}}{\eta}, \eta \tilde{g}^{(\eta)}, \tilde{f}^{(\eta)}) \geq 0 . \end{array} \right.$$

From Lemma A.1, taking the limit as $\eta \rightarrow 0$, we have that $\lim_{\eta \rightarrow 0} Q_{\eta^2 \mu, \eta^2 L}^{i,j}(\frac{x}{\eta}, \eta g, f) \geq 0$ if and only if $\hat{Q}_{\mu,L}^{i,j}(x, g, f) \geq 0$. Applying this and the continuity of our constraints, $s^* = (\tilde{x}^*, \tilde{g}^*, \tilde{f}^*)$ must satisfy

$$\left\{ \begin{array}{l} \tilde{x}_{i+1}^* = \tilde{x}_i^* - \frac{1}{L} \sum_{j=0}^i H_{i,j} \tilde{g}_j^* \\ \|\tilde{x}_0^* - \tilde{x}_\star^*\|^2 \leq R^2 \\ \hat{Q}_{0,L}^{i,j}(\tilde{x}^*, \tilde{g}^*, \tilde{f}^*) \geq 0 . \end{array} \right.$$

Therefore s^* is a feasible solution to $p_S(L, R)$ and so $p_S(L, R) \geq \tilde{f}_N^* - \tilde{f}_\star^*$.

We know from the proof of Lemma A.1 that for all η , $p_{ES}(\eta^2 L, \frac{R}{\eta}) \geq p_S(\eta^2 L, \frac{R}{\eta})$. Combining this with our rescaling result in Lemma 6.2, we have

$$\liminf_{\eta \rightarrow 0} p_{ES}(\eta^2 L, \frac{R}{\eta}) \geq \liminf_{\eta \rightarrow 0} p_S(\eta^2 L, \frac{R}{\eta}) = p_S(L, R) .$$

Combining our results above, we can squeeze the limit as follows

$$\liminf_{\eta \rightarrow 0} p_{ES}(\eta^2 L, \frac{R}{\eta}) \geq p_S(L, R) \geq \tilde{f}_N^* - \tilde{f}_\star^* = \lim_{k \rightarrow \infty} f_N^{(\eta'_k)} - f_\star^{(\eta'_k)} = \limsup_{\eta \rightarrow 0} p_{ES}(\eta^2 L, \frac{R}{\eta}) .$$

Therefore, $\lim_{\eta \rightarrow 0} p_{ES}(\eta^2 L, \frac{R}{\eta}) = p_S(L, R)$. Applying Lemma 6.2 twice gives the final claim.