Subgame Perfect Methods in Nonsmooth Convex Optimization

Benjamin Grimmer* Alex L. Wang†

Abstract

This paper considers nonsmooth convex optimization with either a subgradient or proximal operator oracle. In both settings, we identify algorithms that achieve the recently introduced game-theoretic optimality notion for algorithms known as subgame perfection. Subgame perfect algorithms meet a more stringent requirement than just minimax optimality. Not only must they provide optimal uniform guarantees on the entire problem class, but also on any subclass determined by information revealed during the execution of the algorithm. In the setting of nonsmooth convex optimization with a subgradient oracle, we show that the Kelley cutting plane-Like Method due to Drori and Teboulle [1] is subgame perfect. For nonsmooth convex optimization with a proximal operator oracle, we develop a new algorithm, the Subgame Perfect Proximal Point Algorithm, and establish that it is subgame perfect. Both of these methods solve a history-aware second-order cone program within each iteration, independent of the ambient problem dimension, to plan their next steps. This yields performance guarantees that are never worse than the minimax optimal guarantees and often substantially better.

1 Introduction

Consider a (potentially nonsmooth) convex minimization problem

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

with unknown minimizer $x_\star$. We are particularly interested in the high-dimensional regime where d may be arbitrarily large. To design iterative methods for such problems, one typically assumes an oracle access model for f. The two most fundamental oracle models for nonsmooth optimization are the subgradient oracle

$$x \mapsto (f(x), g) \quad \text{where} \quad g \in \partial f(x) \coloneqq \{g \in \mathbb{R}^d \mid f(z) \geq f(x) + \langle g, z - x\rangle \quad \forall z \in \mathbb{R}^d\}$$

and $g \in \partial f(x)$ may be chosen adversarially; and the proximal operator oracle,

$$x \mapsto (f(y), y) \quad \text{where} \quad y = \text{prox}_{L,f}(x) \coloneqq \operatorname*{arg\,min}_{y\in\mathbb{R}^d} f(y) + \frac{L}{2}\|y - x\|^2$$

for a given proximal parameter $L > 0$. Throughout, we will consider algorithms given a fixed budget of N evaluations of one of these two oracles, for producing approximate minimizers.

These two models of optimization have seen substantial study. The study of subgradient methods (those utilizing a subgradient oracle), was pioneered early on by [2, 3, 4]. Similarly, proximal methods (those utilizing a proximal operator oracle), have a rich history of study dating back to [5, 6, 7, 8].

A classical goal in algorithm design is minimax optimal performance on a given problem class. That is, one often wants an algorithm with the best worst-case performance. To be concrete, consider designing a subgradient method, with a budget of N subgradient oracle queries, for minimizing an M-Lipschitz continuous convex function f given an initialization x_0 with $\|x_0 - x_\star\| \leq R$ for some minimizer $x_\star$ of f. Denote the set of all such instances (f, x_0) by $\mathcal{F}_{M,R}$. Denote by $\mathcal{A}$ the set of all (deterministic) subgradient-span methods with iterates satisfying

$$x_n \in x_0 + \text{span}\{g_0, \ldots, g_{n-1}\}, \qquad g_i \in \partial f(x_i) \tag{1}$$

*Johns Hopkins University, Department of Applied Mathematics and Statistics, grimmer@jhu.edu
†Purdue University, Daniels School of Business, wang5984@purdue.edu

for $n = 1, \ldots, N$. That is, each $A \in \mathcal{A}$ is a deterministic map from a history (possibly empty) of query-response pairs $\{(x_0, f_0, g_0), \ldots, (x_{n-1}, f_{n-1}, g_{n-1})\}$ to the next query point $x_n \in x_0 + \text{span}\{g_0, \ldots, g_{n-1}\}$. Here, f_i is shorthand for $f(x_i)$. When $A \in \mathcal{A}$ and $(f, x_0) \in \mathcal{F}_{M,R}$ are clear from context, we will let $x_0, x_1, \ldots, x_N$ denote the iterates produced by A on the instance (f, x_0). The worst-case final objective gap for A is given by

$$\max_{(f,x_0)\in\mathcal{F}_{M,R}} f(x_N) - f(x_\star).$$

An algorithm is "minimax optimal" if it attains the optimal worst-case performance guarantee:

$$\min_{A\in\mathcal{A}} \max_{(f,x_0)\in\mathcal{F}_{M,R}} f(x_N) - f(x_\star). \tag{2}$$

Methods that are within a universal constant factor of being minimax optimal for a wide range of problem/algorithm settings were designed in the 1980s by the seminal works [9, 10]. Since then, exactly minimax optimal algorithms have been designed for many settings [11, 12, 13, 14, 15], supported by matching lower bounding instances [16, 17], in large part spurred by the Performance Estimation Program (PEP) methodology [18, 19, 20].

Alas, minimax optimal algorithms often do not offer the best performance on "typical" or "real-world" problems. In essence, minimax optimality is a uniform requirement and does not preclude an algorithm from performing "suboptimally" when the problem instance $(f, x_0) \in \mathcal{F}_{M,R}$ is not chosen adversarially.

Recently, a theoretically grounded approach to strengthening minimax optimality was proposed by the present authors in [21]. Instead of only requiring optimal guarantees on the *entire* problem class, we require that the algorithm, *at every iteration* n, offers the optimal guarantee against the *subclass* of remaining problem instances agreeing with the oracle responses so far. That is, let $\mathcal{H} = \{(x_i, f_i, g_i)\}_{i=0}^{n-1}$ denote the history of query-response pairs seen up to iteration n. Denote by $\mathcal{F}_{M,R}^{\mathcal{H}} \subseteq \mathcal{F}_{M,R}$ the set of remaining problem instances (f, x_0) agreeing with the given history, i.e., $f_i = f(x_i)$ and $g_i \in \partial f(x_i)$. Similarly, denote by $\mathcal{A}^{\mathcal{H}}$ the set of all subgradient methods providing a continuation of the first n iterates $x_0, \ldots, x_{n-1}$ for the remaining $N - n + 1$ iterations.[1] Then, an algorithm is "subgame perfect" if at every iteration n and for any observed history $\mathcal{H}$ up to iteration n, it attains the best possible worst-case final objective gap in the remaining iterations

$$\min_{A\in\mathcal{A}^{\mathcal{H}}} \max_{(f,x_0)\in\mathcal{F}_{M,R}^{\mathcal{H}}} f(x_N) - f(x_\star). \tag{3}$$

This is a strengthening of minimax optimality, since at iteration $n = 0$, (3) reduces to (2). The terminology "subgame perfect" stems from a game-theoretic perspective on (2) and (3): A minimax optimal method and a worst-case adversarial selection of the problem instance (f, x_0) together provide a saddle point, or Nash Equilibrium, of (2). In other words, a minimax optimal method corresponds to an optimal algorithm for playing (2) *if* (f, x_0) is chosen adversarially. On the other hand, a subgame perfect method and an associated subgame perfect adversary form a Subgame Perfect Nash Equilibrium for the associated sequential optimization game. Intuitively, a subgame perfect method corresponds to an algorithm that optimally capitalizes on imperfect play by the adversary, while maintaining minimax optimality against fully adversarial play. This offers a principled beyond-worst-case guarantee for convex optimization. We refer readers to [21] for a more in-depth discussion of this game-theoretic perspective. Additionally, numerical experiments are given in [21, 22] showing strong practical gains from such refinements.

This subgame perfect criteria was first proposed in the context of smooth convex minimization [21]. In this work, we show that subgame perfect methods can also be designed for the setting of nonsmooth convex minimization. First, we show that the Kelley cutting plane-Like Method (KLM) of Drori and Teboulle [1] is in fact subgame perfect for Lipschitz convex minimization with a subgradient oracle. Second, we design a new subgame perfect algorithm for convex optimization with a proximal oracle, which we call the Subgame Perfect Proximal Point Algorithm (SPPPA).

Contributions and organization. In Section 2, we begin by describing a minimax optimal subgradient method, KLM, of Drori and Teboulle [1] and state its dynamic guarantees. We then show that this method is subgame perfect by constructing a matching dynamic lower bound. In Section 3, we

[1] Formally, $\mathcal{A}^{\mathcal{H}}$ is the subset of $A \in \mathcal{A}$ mapping $\{(x_i, f_i, g_i)\}_{i=0}^{j-1} \mapsto x_j$ for all $j = 1, \ldots, n-1$.

turn to designing a new subgame perfect proximal point-type method. Our method, SPPPA, generalizes the known minimax-optimal proximal point method, improving it with a dynamic reoptimization of its inductive hypothesis at every iteration. In Section 4, we prove that SPPPA is subgame perfect by examining associated dual certificates generated by these reoptimizations at each step to construct matching dynamic lower bounds. We expect this process to generalize to yet more settings in the future, discussed briefly in Section 5.

2 Subgame Perfect Subgradient Method

In this section, we consider unconstrained minimization of a convex M-Lipschitz function

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

That is, $f:\mathbb{R}^d\to\mathbb{R}$ is convex and M-Lipschitz. Note that this ensures $\|g\|\le M$ for every $g\in\partial f(x)$ and $x\in\mathbb{R}^d$. We assume an optimal point $x_\star\in\arg\min f$ exists and that the given initialization x_0 satisfies $\|x_0-x_\star\|\le R$. Throughout, all norms are the Euclidean norm associated with the given inner product $\langle\cdot,\cdot\rangle$.

Given the iterates $x_0,\dots,x_N$, it is convenient to reason about the associated first-order (subgradient) data $\{(x_i,f_i,g_i)\}_{i\in\mathcal{I}}$ with $f_i=f(x_i)$ and $g_i\in\partial f(x_i)$ where $\mathcal{I}=\{0,\dots,N\}$. At times, it will be useful to include the minimizer $x_\star$ in our set of observations; to do so, we set $g_\star=0$, $f_\star=f(x_\star)$ and $\mathcal{I}_\star=\{0,\dots,N,\star\}$. Convexity of f ensures that

$$Q_{i,j}\coloneqq f_i-f_j-\langle g_j,x_i-x_j\rangle\ge 0\qquad\forall i,j\in\mathcal{I}_\star,$$

and the M-Lipschitz continuity of f ensures that

$$S_i\coloneqq M^2-\|g_i\|^2\ge 0\qquad\forall i\in\mathcal{I}_\star.$$

That is, the nonnegativity of the above quantities is a necessary condition on the first-order data to have come from some M-Lipschitz convex function. A special case of the interpolation theorem of [20, Theorem 3.5] states that this condition is also sufficient:

Lemma 2.1 ([20, Theorem 3.5])**.** *Let $\{(x_i,f_i,g_i)\}_{i\in\mathcal{I}_\star}\subseteq\mathbb{R}^d\times\mathbb{R}\times\mathbb{R}^d$. The quantities $Q_{i,j}$ and S_i are nonnegative for all $i,j\in\mathcal{I}_\star$ if and only if there exists an M-Lipschitz convex function f satisfying*

$$f_i=f(x_i)\quad\text{and}\quad g_i\in\partial f(x_i)\qquad\forall i\in\mathcal{I}_\star.$$

This lemma can also be proved directly by setting $f(y)=\max_{i\in\mathcal{I}_\star}\{f_i+\langle g_i,y-x_i\rangle\}$.

2.1 The Kelley-Like Method of Drori and Teboulle [1]

The Kelley cutting plane Method [23] is an early example of a "bundle method": an algorithm maintaining a bundle of first-order information $\{(x_i,f_i,g_i)\}_{i=0}^{n-1}$ at each iteration n, used to inform the selection of the next iterate x_n. Such methods have a long history of practical success, being studied theoretically by [3, 4, 24, 25, 26, 27, 28, 29, 30, 31] among many other works.

Drori and Teboulle [1] designed a variant of the Kelley cutting plane method seeking to optimize the iterate selection against the hardest possible function consistent with the observed subgradients. Their resulting method was developed by reformulating and relaxing the search for the hard instances through a series of mathematical programs, eventually providing a tractable planning problem to be solved at every iteration. From this, they propose their Kelley cutting plane-Like Method (KLM), which solves this planning problem at every iteration to dynamically respond to observed subgradients. In particular, consider an observed history of first-order evaluations prior to iteration n, $\mathcal{H}=\{(x_i,f_i,g_i)\}_{i=0}^{n-1}$ where $f_i=f(x_i)$ and $g_i\in\partial f(x_i)$. KLM sets $f_{n-1/2}=\min_{i=0,\dots,n-1}f_i$ as the minimum function value observed so far, then solves the second-order cone program

$$\Theta_n=\begin{cases}\max\limits_{y\in\mathbb{R}^d,\zeta,t\in\mathbb{R}} & f_{n-1/2}-t\\ \text{s.t.} & t\ge f_i+\langle g_i,y-x_i\rangle\qquad\forall i=0,\dots,n-1\\ & f_{n-1/2}-M\zeta\le t\\ & \|y-x_0\|^2+(N-n+1)\zeta^2\le R^2.\end{cases}\tag{4}$$

Observe that (i) $\Theta_n \geq 0$ since $y = x_\star, \zeta = 0, t = f_{n-1/2}$ is a feasible solution and (ii) there exists an optimal solution with $y \in x_0 + \text{span}\{g_0, \ldots, g_{n-1}\}$. Letting (y, ζ, t) denote some such optimal solution to this problem in the nth iteration, KLM will iterate by setting $x_n = y$. Since $y \in x_0 + \text{span}\{g_0, \ldots, g_{n-1}\}$, this update falls within the (sub)gradient span model (1).

The KLM is formalized in Algorithm 1. As input, this method requires bounds on the problem Lipschitz constant M, the distance to a minimizer R, and the total iteration budget N. In [1], it is noted that one can freely, at some iterations, take a regular subgradient method step instead of the optimized step resulting from the planning subproblem. We omit this from our presentation as the resulting algorithm is no longer subgame perfect.

Algorithm 1 The Kelley cutting plane-Like Method (KLM)

Given convex f, initial iterate x_0, Lipschitz constant M, distance bound R, iteration count N

- For $n = 0$, define $\Theta_0 = MR/\sqrt{N+1}$
- For $n = 1, 2, \ldots, N$
  - Query oracle to get $f_{n-1} = f(x_{n-1})$ and $g_{n-1} \in \partial f(x_{n-1})$.
  - Let $f_{n-1/2} = \min_{i \in 0,\ldots,n-1} f_i$ and solve (4), generating Θ_n and optimal (y, ζ, t).
  - Set $x_n = y$.

The design philosophy of Drori and Teboulle is very similar in nature to the motivation behind the subgame perfect framework of [21], making KLM a plausible candidate for a subgame perfect subgradient method. In [1], they proved the following convergence upper bound for KLM.

Theorem 2.1. *For any $(f, x_0) \in \mathcal{F}_{M,R}$ and $N \geq 0$, the output x_N of KLM is guaranteed to satisfy*

$$f(x_N) - f(x_\star) \leq \Theta_N \leq \cdots \leq \Theta_1 \leq \Theta_0 = \frac{MR}{\sqrt{N+1}}.$$

The derivation of this dynamic sequence of upper bounds is nontrivial, requiring substantial intermediate developments. We refer readers to [1] for its proof.

In addition to this upper bound, Drori and Teboulle provide a lower bounding instance (assuming $d \geq N+1$) showing that no subgradient method can achieve a worst-case convergence guarantee strictly better than $MR/\sqrt{N+1}$. As a result, KLM is minimax optimal for Lipschitz convex minimization in the sense of (2). However, this single hard instance is insufficient to prove that KLM is subgame perfect. Instead, one needs to construct a dynamic lower bound consistent with any observed history up to iteration n matching the upper bound Θ_n at iteration n. The remainder of this section provides such a construction leading to the following strengthened guarantee for KLM:

Theorem 2.2. *Let $N > 0$ and $d \geq N+1$ and consider some iteration $0 \leq n \leq N$. Suppose $\mathcal{H} = \{(x_i, f_i, g_i)\}_{i=0}^{n-1}$ is the set of observed first-order history. Then,*

$$\min_{A \in \mathcal{A}^{\mathcal{H}}} \max_{(f,x_0) \in \mathcal{F}_{M,R}^{\mathcal{H}}} f(A(f, x_0)) - f(x_\star) = \Theta_n.$$

That is, KLM is subgame perfect.

The hardness result of [1] is equivalent to the special case of this theorem with $n = 0$.

2.2 Proof of Theorem 2.2

This section contains a proof of Theorem 2.2. Since the case $n = 0$ is covered by [1], we will assume $1 \leq n \leq N$ and $d \geq N+1$. Fix a set of observed first-order history $\{(x_i, f_i, g_i)\}_{i=0}^{n-1}$.

Our goal is to construct an instance $(f, x_0) \in \mathcal{F}_{M,R}^{\mathcal{H}}$ so that for all $A \in \mathcal{A}^{\mathcal{H}}$:

$$f(A(f, x_0)) - f(x_\star) \geq \Theta_n. \tag{5}$$

Since $d \geq N+1$, we can choose orthonormal vectors $e_n, \ldots, e_N$ perpendicular to $\operatorname{span}\{g_0, \ldots, g_{n-1}\}$. By definition, Θ_n is the optimal value of (4). Let (y, ζ, t) denote an optimizer of (4). Define

$$x_i \coloneqq y - \frac{f_{n-1/2} - t}{M} \sum_{j=n}^{i-1} e_j, \qquad f_i \coloneqq f_{n-1/2}, \qquad g_i \coloneqq M e_i \qquad \text{for all } i = n, \ldots, N, \tag{6}$$

treating the sum above as 0 when it is empty and

$$x_\star \coloneqq y - \frac{f_{n-1/2} - t}{M} \sum_{j=n}^{N} e_j, \qquad f_\star \coloneqq t, \qquad g_\star \coloneqq 0.$$

Combining the given first-order data, $\{(x_i, f_i, g_i)\}_{i=0,\ldots,n-1}$, with the constructed first-order data, $\{(x_i, f_i, g_i)\}_{i=n,\ldots,N,\star}$, we may construct our candidate hard function as the associated max-of-affine function

$$f_{\mathcal{H}}(x) \coloneqq \max_{i \in \mathcal{I}_\star} \{f_i + \langle g_i, x - x_i \rangle\}. \tag{7}$$

One may verify that $x_n, \ldots, x_N$ are the iterates produced by KLM on this instance $(f_{\mathcal{H}}, x_0)$. This justifies the notation $x_n, \ldots, x_N, x_\star$ and we refer to these quantities as the "future iterates." We caution that the n through Nth iterates produced by some other algorithm $A \in \mathcal{A}^{\mathcal{H}}$ may not coincide with the above construction of $x_n, \ldots, x_N$. Nonetheless, it will still hold that the *instance* $(f_{\mathcal{H}}, x_0) \in \mathcal{F}^{\mathcal{H}}_{M,R}$ is hard for all $A \in \mathcal{A}^{\mathcal{H}}$ in the sense of (5).

Below we verify that $(f_{\mathcal{H}}, x_0) \in \mathcal{F}^{\mathcal{H}}_{M,R}$ and prove a needed "zero-chain" property. These facts can then be combined to give a short direct proof of Theorem 2.2.

Proposition 2.2. *Suppose $d \geq N+1$ and $1 \leq n \leq N$. Let $(f_{\mathcal{H}}, x_0)$ denote the instance constructed above from $\mathcal{H} = \{(x_i, f_i, g_i)\}_{i=0}^{n-1}$. It holds that $f_{\mathcal{H}}$ is convex and M-Lipschitz,*

$$f_{\mathcal{H}}(x_i) = f_i \qquad \partial f_{\mathcal{H}}(x_i) \ni g_i \qquad \forall i \in \mathcal{I}_\star,$$

and $\|x_0 - x_\star\| \leq R$. In particular, $(f_{\mathcal{H}}, x_0) \in \mathcal{F}^{\mathcal{H}}_{M,R}$.

Proof. As $f_{\mathcal{H}}$ is a finite maximum of M-Lipschitz affine functions, it follows that it is convex and M-Lipschitz.

We next check that $f_{\mathcal{H}}(x_i) = f_i$ and $\partial f_{\mathcal{H}}(x_i) \ni g_i$ for all $i \in \mathcal{I}_\star$. Inspecting (7), we see it suffices to check that the affine function $f_i + \langle g_i, y - x_i \rangle$ is active at x_i, i.e., that

$$Q_{i,j} \coloneqq f_i - f_j - \langle g_j, x_i - x_j \rangle \geq 0 \qquad \forall i, j \in \mathcal{I}_\star.$$

We break this verification into cases: First, suppose $i \in \{0, \ldots, n-1\}$. Then,

$$\begin{aligned} j \in \{0, \ldots, n-1\} \;&:\; f_i \geq f_j + \langle g_j, x_i - x_j \rangle, \\ j \in \{n, \ldots, N\} \;&:\; f_i \geq f_{n-1/2} = f_j + \langle g_j, x_i - x_j \rangle, \text{ and} \\ j = \star \;&:\; f_i \geq f_{n-1/2} \geq f_\star = f_\star + \langle g_\star, x_i - x_\star \rangle. \end{aligned}$$

The first case comes from the fact that the observed first-order history must come from some convex function; the second case uses the definition of $f_{n-1/2}$ and the orthogonality of $g_j = M e_j$ to the span $\operatorname{span}\{g_0, \ldots, g_{n-1}, \ldots, g_{j-1}\}$ which contains both $x_i - x_0$ and $x_j - x_0$ for the equality; and the third case uses that $\Theta_n \geq 0$ so that $f_\star \leq f_{n-1/2}$ and that $g_\star = 0$.

Next, suppose $i \in \{n, \ldots, N\}$. Then,

$$\begin{aligned} j \in \{0, \ldots, n-1\} \;&:\; f_i = f_{n-1/2} = t + \Theta_n \geq f_j + \langle g_j, y - x_j \rangle + \Theta_n \geq f_j + \langle g_j, x_i - x_j \rangle, \\ j \in \{n, \ldots, N\} \;&:\; f_i = f_{n-1/2} = f_j \geq f_j + \langle g_j, x_i - x_j \rangle, \text{ and} \\ j = \star \;&:\; f_i = f_{n-1/2} = f_\star + \Theta_n \geq f_\star = f_\star + \langle g_\star, x_i - x_\star \rangle. \end{aligned}$$

The first case uses the first constraint in (4) and that $\Theta_n \geq 0$; the second case uses that $\langle g_j, x_i - x_j \rangle$ is zero when $j \geq i$ and negative when $j < i$; and the third case uses that $g_\star = 0$.

Finally, suppose $i = \star$. Then,

$$j \in \{0, \dots, n-1\} \;:\; f_\star = t \geq f_j + \langle g_j, y - x_j \rangle = f_j + \langle g_j, x_\star - x_j \rangle, \text{ and}$$
$$j \in \{n, \dots, N\} \;:\; f_\star = f_{n-1/2} - \Theta_n = f_j + \left\langle Me_j, -\frac{f_{n-1/2} - t}{M} e_j \right\rangle = f_j + \langle g_j, x_\star - x_j \rangle.$$

The first case uses the first constraint in (4) and the orthogonality of g_j to $x_\star - y$.

We deduce that

$$f_{\mathcal{H}}(x_i) = f_i \qquad g_i \in \partial f_{\mathcal{H}}(x_i) \qquad \text{for all } i \in \mathcal{I}_\star.$$

Since $g_\star = 0 \in \partial f_{\mathcal{H}}(x_\star)$, $x_\star$ must be a minimizer of $f_{\mathcal{H}}$.

It remains to verify the needed distance from initialization to a minimizer bound:

$$\|x_0 - x_\star\|^2 = \|y - x_0\|^2 + (N - n + 1)\left(\frac{f_{n-1/2} - t}{M}\right)^2 \leq \|y - x_0\|^2 + (N - n + 1)\zeta^2 \leq R^2,$$

where the equality follows from orthogonality of each e_i in our definition of $x_\star$ and the two inequalities use the constraints respectively $f_{n-1/2} - M\zeta \leq t$ and $\|y - x_0\|^2 + (N - n + 1)\zeta^2 \leq R^2$. □

Proposition 2.3. *Fix $j \in \{n, \dots, N-1\}$ and let $x \in x_0 + \text{span}\{g_0, \dots, g_{j-1}\}$. Then*

$$\{g_0, \dots, g_j\} \cap \partial f_{\mathcal{H}}(x) \neq \emptyset.$$

Proof. Consider an $x \in x_0 + \text{span}\{g_0, \dots, g_{j-1}\}$. Let $i \in \mathcal{I}_\star$ denote an active component of $f_{\mathcal{H}}(x)$ in the definition (7). If $i < j$, then $g_i \in \partial f_{\mathcal{H}}(x)$ and we are done. Otherwise, suppose $i \geq j$. Note by orthogonality, that for every $i \geq j$,

$$f_i + \langle g_i, x - x_i \rangle = f_i = f_{n-1/2}.$$

As a result, $g_j \in \partial f_{\mathcal{H}}(x)$ and we are done. □

We are now ready to prove Theorem 2.2.

Proof of Theorem 2.2. By Proposition 2.2, the instance $(f_{\mathcal{H}}, x_0) \in \mathcal{F}^{\mathcal{H}}_{M,R}$. Now consider an arbitrary deterministic subgradient-span method $A \in \mathcal{A}^{\mathcal{H}}$ responsible for producing iterates $x^A_n, x^A_{n+1}, \dots, x^A_N$. By the subgradient-span condition, $x^A_n \in x_0 + \text{span}\{g_0, \dots, g_{n-1}\}$. Inductively, suppose that for $n \leq j < N$, $x^A_j \in x_0 + \text{span}\{g_0, \dots, g_{j-1}\}$. By Proposition 2.3, an adversarial oracle could provide some subgradient in $\{g_0, \dots, g_j\}$, so that by the subgradient-span condition, $x^A_{j+1} \in x_0 + \text{span}\{g_0, \dots, g_j\}$. We deduce that $x^A_N \in x_0 + \text{span}\{g_0, \dots, g_{N-1}\}$. As a result,

$$f_{\mathcal{H}}(x^A_N) \geq f_N + \langle g_N, x^A_N - x_N \rangle = f_N = f_\star + (f_{n-1/2} - t) = f_\star + \Theta_n,$$

where the inequality restricts to the Nth affine term defining $f_{\mathcal{H}}$ and the first equality uses orthogonality of $g_N = Me_N$ to $\text{span}\{g_0, \dots, g_{N-1}\}$ which contains $x^A_N - x_N$. □

3 The Subgame Perfect Proximal Point Algorithm (SPPPA)

We next design a subgame perfect algorithm in the setting of nonsmooth convex optimization with a proximal operator oracle.

Let $L_0, L_1, \dots, L_N$ denote a fixed sequence of parameters. Our family of algorithms in this setting consists of any algorithm of the form: Given an initialization x_0, iterate for $n = 1, \dots, N$: query the proximal operator oracle at x_{n-1} to receive

$$\begin{cases} y_{n-1} = \text{prox}_{L_{n-1}, f}(x_{n-1}) \\ g_{n-1} = L_{n-1}(x_{n-1} - y_{n-1}) \in \partial f(y_{n-1}) \\ f_{n-1} = f(y_{n-1}) \end{cases} ; \tag{8}$$

then pick the next query point

$$x_n \in x_0 + \text{span}\{g_0, \dots, g_{n-1}\}$$

as a function of the first-order history $\{(x_i, f_i, g_i)\}_{i=0}^{n-1}$. The fact that $g_{n-1} \in \partial f(y_{n-1})$ is an immediate consequence of the optimality condition governing the proximal step's computation. We caution the reader that in the proximal operator setting, we measure objective values at y_i (in contrast to the subgradient oracle setting, where we evaluated objective values at x_i).

Define $y_\star = x_\star \in \arg\min_x f(x)$, $f_\star = f(y_\star) = \min_x f(x)$, and $g_\star = 0 \in \partial f(y_\star)$. We see that $g_i \in \partial f(y_i)$ for all $i \in \mathcal{I}_\star$ so that

$$Q_{i,j} \coloneqq f_i - f_j - \langle g_j, y_i - y_j \rangle \geq 0 \tag{9}$$

for all $i, j \in \mathcal{I}_\star$.

In our design of proximal point methods, we will not assume either Lipschitz continuity of f or a bound on $\|x_0 - y_\star\|$. Instead, the target class of problem instances is all (f, x_0) with f closed, convex, proper, and attaining a minimum at some $y_\star$. In this setting, a natural goal is to seek to reduce the *normalized* suboptimality

$$\frac{f(y_N) - f(y_\star)}{\frac{1}{2}\|x_0 - y_\star\|^2}.$$

Indeed, as $\|x_0 - y_\star\|$ is allowed to be arbitrarily large, it is impossible to provide any bound on the *unnormalized* suboptimality for any proximal point method.

A series of classic works [32, 33, 34, 35] have developed minimax optimal methods for this task (although their exact optimality was only recently proven in [36]). Below, in Section 3.1, we discuss a minimax optimal algorithm, called the Optimized Proximal Point Algorithm (OPPA), and present an inductive proof of its convergence guarantee. In Section 3.2, we will introduce the Subgame Perfect Proximal Point Algorithm (SPPPA), a modification of OPPA that dynamically optimizes the inductive statement within the proof of OPPA using observed first-order information. The remainder of Section 3 will prove dynamic upper bounds on the convergence rate of SPPPA. Dynamic lower bounds will be presented in Section 4, thereby showing that SPPPA is subgame perfect.

Note that the setting of proximal operators has substantial similarities to the setting of smooth convex optimization. In particular, if one fixes all $L_n = L$ to be constant, then proximal operator steps are exactly gradient steps on the objective function's Moreau envelope. Correspondingly, our developed method and analysis here closely mirror those of the subgame perfect gradient method [21], recovering the core operations of SPGM in the special case of a constant proximal parameter sequence.

3.1 The Existing Optimized Proximal Point Algorithm (OPPA)

We now describe a minimax optimal method, which we refer to as the "Optimized Proximal Point Algorithm" (OPPA). This method was first proposed in the case of constant proximal parameters $L_n = L$ by Güler [33]. Generalizations allowing for any L_n as well as inexactness in the evaluation of proximal operators were given by [34] and [35], which reduce to OPPA when proximal operator evaluations are exact.

OPPA maintains four sequences of iterates $x_n \in \mathbb{R}^d$, $y_n \in \mathbb{R}^d$, $z_n \in \mathbb{R}^d$ and $\tau_n \in \mathbb{R}$ via the following induction: x_0 is given; initialize $\tau_0 = 2/L_0$ and $z_1 = x_0 - \tau_0 g_0$; using the proximal operator oracle, inductively define for $n = 1, 2, \ldots$

$$\begin{cases} y_{n-1} &= \operatorname{prox}_{L_{n-1}, f}(x_{n-1}) \\ g_{n-1} &= L_{n-1}(x_{n-1} - y_{n-1}) \in \partial f(y_{n-1}) \\ \tau_n &= \tau_{n-1} + \frac{1}{L_n}(1 + \sqrt{1 + 2L_n\tau_{n-1}}) \\ x_n &= \frac{\tau_{n-1}}{\tau_n} y_{n-1} + \frac{\tau_n - \tau_{n-1}}{\tau_n} z_n \\ z_{n+1} &= z_n - (\tau_n - \tau_{n-1}) g_n \end{cases}. \tag{10}$$

It will be conceptually useful to view z_{n+1} as being defined in the nth iteration, although it cannot explicitly be computed until the $(n+1)$th iteration, after g_n is observed. OPPA's design and proof strategy maintains a specific inductive hypothesis on its iterate sequences. Our design of SPPPA and its proof will use these same ideas.

For $n \geq 0$, define the expression

$$H_n \coloneqq \tau_n(f_\star - f_n) - \frac{1}{2}\|z_{n+1} - y_\star\|^2 + \frac{1}{2}\|x_0 - y_\star\|^2. \tag{11}$$

Lemma 3.1. *For any closed convex proper f with minimizer $x_\star = y_\star \in \mathbb{R}^d$ and any $x_0 \in \mathbb{R}^d$, $H_0 \geq 0$.*

Proof. By definition, $x_0 = y_0 + g_0/L_0$, $z_1 = x_0 - 2g_0/L_0 = y_0 - g_0/L_0$, and $\tau_0 = \frac{2}{L_0}$. Thus,

$$\begin{aligned} H_0 &= \frac{2}{L_0}(f_\star - f_0) - \frac{1}{2}\|y_0 - x_\star - g_0/L_0\|^2 + \frac{1}{2}\|y_0 - x_\star + g_0/L_0\|^2 \\ &= \frac{2}{L_0}(f_\star - f_0) + \frac{2}{L_0}\langle g_0, y_0 - x_\star\rangle = \frac{2}{L_0}Q_{\star,0} \geq 0, \end{aligned}$$

where the last inequality follows from (9). □

Lemma 3.2. *Suppose f is closed convex proper with minimizer $x_\star = y_\star \in \mathbb{R}^d$, $m \in [0, n-1]$, $\tau' \in \mathbb{R}$, and $z' \in \mathbb{R}^d$ satisfy*

$$H' \coloneqq \tau'(f_\star - f_m) - \frac{1}{2}\|z' - y_\star\|^2 + \frac{1}{2}\|x_0 - y_\star\|^2 \geq 0.$$

Define

$$\tau_n = \tau' + \frac{1}{L_n}(1 + \sqrt{1 + 2L_n\tau'}), \quad x_n = \frac{\tau'}{\tau_n}y_m + \frac{\tau_n - \tau'}{\tau_n}z', \quad \text{and} \quad z_{n+1} = z' - (\tau_n - \tau')g_n.$$

Then, $H_n \geq 0$.

Proof. It suffices to check that $H_n = H' + (\tau_n - \tau')Q_{\star,n} + \tau' Q_{m,n}$, which is nonnegative as it is the sum of three nonnegative terms. □

Applying the above lemma with $m = n-1$, $\tau' = \tau_{n-1}$ and $z' = z_n$, we deduce that $H_n \geq 0$ for all $n \geq 0$. This leads us to OPPA's convergence guarantee:

$$\tau_N(f_N - f_\star) \leq \tau_N(f_N - f_\star) + \frac{1}{2}\|z_{N+1} - x_\star\|^2 \leq \frac{1}{2}\|x_0 - x_\star\|^2. \tag{12}$$

To make this convergence guarantee concrete, if $L_n = L$ is constant, then $\tau_N = N^2/2L + o(N^2)$. Thus, the bound becomes $f(y_N) - f(x_\star) \leq L\|x_0 - x_\star\|^2/N^2$ up to lower order terms. It has long been known that this $O(1/N^2)$ is the minimax optimal order of convergence [37, 38]. Many works since [39, 40, 41, 42, 43, 44, 45] have built lower bounding theory for first-order methods. Recently, [36] extended the zero-chain framework of [45] to provide the first exactly matching lower bound for OPPA, proving its minimax optimality.

3.2 Design of SPPPA

SPPPA is presented in Algorithm 2. Similar to OPPA, SPPPA will maintain iterate sequences x_n, y_n, z_n (and a scalar sequence τ_n) so that the expression H_n (defined in (11)) is nonnegative for all $n \geq 0$. We will adopt OPPA's initialization so that $H_0 \geq 0$ (see Lemma 3.1). The key conceptual difference between OPPA and SPPPA is that SPPPA will, in each iteration n, apply Lemma 3.2 to a choice of $m \in [0, n-1]$, τ', and z' determined by the observed first-order responses. This contrasts with OPPA, where m, τ', z' are always set to $n-1$, τ_{n-1}, and z_n respectively.

3.2.1 The Planning Subproblem

We now develop a tractable subproblem, the *planning subproblem*, for optimizing m, τ', z'.

Throughout this subsection, we will assume SPPPA is at iteration n and has observed/constructed $\{(y_i, f_i, g_i, z_{i+1}, \tau_i)\}_{i=0}^{n-1}$ with $f_i = f(y_i)$, $g_i \in \partial f(y_i)$. Furthermore, we may assume by induction that $H_i \geq 0$ for all $0 \leq i \leq n-1$.

Fix an arbitrary $m \in \arg\min_{i \in [0,n-1]} f_i$ and define the following matrices and vectors

$$\begin{gathered} Z = \begin{bmatrix} z_1 - x_0 & \cdots & z_n - x_0 \end{bmatrix} \in \mathbb{R}^{d \times n}, \quad G = \begin{bmatrix} g_0 & \cdots & g_{n-1} \end{bmatrix} \in \mathbb{R}^{d \times n}, \\ \tau = (\tau_0, \ldots, \tau_{n-1})^\top, \quad f = (f_0, \ldots, f_{n-1})^\top, \end{gathered} \tag{13}$$

and as additional helpful quantities

$$q = (q_0, \ldots, q_{n-1})^\top, \quad a = (a_0, \ldots, a_{n-1})^\top, \quad b = (b_0, \ldots, b_{n-1})^\top \tag{14}$$

where $q_i := f_i - \langle g_i, y_i - x_0\rangle$, $a_i := \frac{1}{2}\|z_{i+1} - x_0\|^2 + \tau_i(f_i - f_m)$, and $b_i := q_i - f_m$. Our goal is to find τ' and z' so that $H' \geq 0$. We will attempt to certify the nonnegativity of H' by writing it as

$$H' = \sum_{i=0}^{n-1} \mu_i H_i + \sum_{i=0}^{n-1} \lambda_{\star,i} Q_{\star,i} + \varepsilon \tag{15}$$

for some nonnegative $\mu, \lambda_\star \in \mathbb{R}^n$ and $\varepsilon \in \mathbb{R}$.

Lemma 3.3. *For any vectors $\mu, \lambda_\star \in \mathbb{R}^n$, the identity (15) holds if*

$$\tau' = \langle \tau, \mu\rangle + \langle \mathbf{1}, \lambda_\star\rangle, \qquad z' = x_0 + Z\mu - G\lambda_\star,$$
$$\varepsilon = \langle \mu, a\rangle + \langle \lambda_\star, b\rangle - \tfrac{1}{2}\|Z\mu - G\lambda_\star\|^2.$$

Hence, if $\mu, \lambda_\star \geq 0$ and $\varepsilon \geq 0$, then $H' \geq 0$.

Proof. Verifying this identity simply corresponds to expanding the right-hand side and collecting like terms. For completeness, we present this verification below:

Plugging in the definition $H_i = \tau_i(f_\star - f_i) + \frac{1}{2}\|x_0 - y_\star\|^2 - \frac{1}{2}\|z_{i+1} - y_\star\|^2$ and noting $\sum_i \mu_i(x_0 - z_{i+1}) = -Z\mu$, the first term in our claimed decomposition equals

$$\begin{aligned}\sum_i \mu_i H_i &= \langle \tau, \mu\rangle f_\star - \sum_i \mu_i \tau_i f_i - \tfrac{1}{2}\sum_i \mu_i \|z_{i+1} - x_0\|^2 + \big\langle y_\star - x_0,\ Z\mu\big\rangle \\ &= \langle \tau, \mu\rangle(f_\star - f_m) - \langle \mu, a\rangle + \big\langle y_\star - x_0,\ Z\mu\big\rangle.\end{aligned}$$

Using $q_i = f_i - \langle g_i, y_i - x_0\rangle$ and so $Q_{\star,i} = f_\star - f_i - \langle g_i, y_\star - y_i\rangle = f_\star - q_i - \langle g_i, y_\star - x_0\rangle$, the second term in our claimed decomposition equals

$$\begin{aligned}\sum_i \lambda_{\star,i} Q_{\star,i} &= \langle \mathbf{1}, \lambda_\star\rangle f_\star - \langle \lambda_\star, q\rangle - \big\langle y_\star - x_0,\ G\lambda_\star\big\rangle \\ &= \langle \mathbf{1}, \lambda_\star\rangle(f_\star - f_m) - \langle \lambda_\star, b\rangle - \big\langle y_\star - x_0,\ G\lambda_\star\big\rangle.\end{aligned}$$

Summing these two expressions with $\varepsilon = \langle \mu, a\rangle + \langle \lambda_\star, b\rangle - \frac{1}{2}\|Z\mu - G\lambda_\star\|^2$ gives

$$\begin{aligned}&\sum_{i=0}^{n-1} \mu_i H_i + \sum_{i=0}^{n-1} \lambda_{\star,i} Q_{\star,i} + \varepsilon \\ &= (\langle \tau, \mu\rangle + \langle \mathbf{1}, \lambda_\star\rangle)(f_\star - f_m) + \big\langle y_\star - x_0,\ Z\mu - G\lambda_\star\big\rangle - \tfrac{1}{2}\|Z\mu - G\lambda_\star\|^2 \\ &= (\langle \tau, \mu\rangle + \langle \mathbf{1}, \lambda_\star\rangle)(f_\star - f_m) + \tfrac{1}{2}\|x_0 - y_\star\|^2 - \tfrac{1}{2}\|x_0 + Z\mu - G\lambda_\star - y_\star\|^2.\end{aligned}$$

Plugging in the chosen values of τ', z' from the lemma statement, this is exactly H'. The final conclusion that H' must be nonnegative whenever $\mu, \lambda_\star \geq 0$ and $\varepsilon \geq 0$ follows from the fact that this decomposition shows H' is then equal to a sum of nonnegative quantities. □

The objective function in the planning subproblem is to maximize the value of τ'. Hence, the planning subproblem can be written as

$$\tau' = \sup_{\mu, \lambda_\star \in \mathbb{R}^n_{\geq 0}} \left\{ \langle \tau, \mu\rangle + \langle \mathbf{1}, \lambda_\star\rangle :\ \tfrac{1}{2}\|Z\mu - G\lambda_\star\|^2 \leq \langle \mu, a\rangle + \langle \lambda_\star, b\rangle \right\}. \tag{16}$$

This is a simple convex optimization problem, independent of the ambient dimension d, optimizing a linear function over a feasible region given by nonnegativity and a single rotated second-order cone constraint. Although the optimizer of (16) lacks a closed-form in general, linear optimization over a single convex quadratic constraint (and nonnegativity) is quite standard and can be done using industrial solvers.

Note that $\mu = (0, \ldots, 0, 1), \lambda_\star = (0, \ldots, 0)$ is always a feasible solution in (16), as the constraint becomes

$$\frac{1}{2}\|z_n - x_0\|^2 \leq \frac{1}{2}\|z_n - x_0\|^2 + \tau_{n-1}(f_{n-1} - f_m),$$

which holds by the assumption that $m \in \arg\min_{i \in [0,n-1]} f_i$. We deduce that $\tau' \geq \tau_{n-1}$.

On the other hand, if (16) has unbounded optimal value, then we claim $y_m \in \arg\min_x f(x)$. Indeed, for any feasible $\mu, \lambda_\star$ in (16), it holds that $H' \geq 0$. Thus, by rearranging, we have that

$$f(y_m) - f_\star \leq \frac{1}{2(\langle \tau, \mu\rangle + \langle \mathbf{1}, \lambda_\star\rangle)}\|x_0 - x_\star\|^2.$$

3.2.2 Subgame Perfect Proximal Point Algorithm and its Guarantees

We now formally state SPPPA:

Algorithm 2 The Subgame Perfect Proximal Point Algorithm (SPPPA)

Given closed convex proper function f, initial iterate x_0, proximal parameter sequence $L_0, L_1, \ldots$

- Define $\tau_0 = \frac{2}{L_0}$ and $z_1 = x_0 - \tau_0 g_0$
- For $n = 1, 2, \ldots$.
  - Query $y_{n-1} = \text{prox}_{L_{n-1},f}(x_{n-1})$ and $f_{n-1} = f(y_{n-1})$. Set $g_{n-1} = L_{n-1}(x_{n-1} - y_{n-1})$.
  - Let $m \in \arg\min_{i \in \{0,\ldots,n-1\}} \{f(y_i)\}$.
  - If (16) is unbounded, terminate and output y_m. Else, let $(\mu, \lambda_\star)$ be an optimal solution to (16) and set $\tau' = \langle \tau, \mu \rangle + \langle \mathbf{1}, \lambda_\star \rangle$ and $z' = x_0 + Z\mu - G\lambda_\star$.
  - Define

$$\tau_n = \tau' + \frac{1}{L_n}(1 + \sqrt{1 + 2L_n\tau'})$$
$$x_n = \frac{\tau'}{\tau_n} y_m + \frac{\tau_n - \tau'}{\tau_n} z'$$
$$z_{n+1} = z' - (\tau_n - \tau')g_n.$$

By Lemmas 3.1, 3.2 and 3.3, this method maintains $H_n \geq 0$ at every iteration and hence provides a guarantee in every iteration

$$\frac{f(y_n) - f(y_\star)}{\frac{1}{2}\|x_0 - x_\star\|^2} \leq \frac{1}{\tau_n}.$$

In order to argue that SPPPA is subgame perfect, we will additionally need dynamic guarantees on $f(y_N) - f(y_\star)$ that can be made at earlier iterations $0 \leq n \leq N$. We will state these guarantees in terms of a doubly-indexed expression $\tau_{i,j}$. First, define $\tau_{0,0}, \ldots, \tau_{0,N}$ to be the sequence generated by the OPPA recurrence: $\tau_{0,0} = 2/L_0$ and for all $i \in [1, N]$,

$$\tau_{0,i} = \tau_{0,i-1} + \frac{1}{L_i}(1 + \sqrt{1 + 2L_i\tau_{0,i-1}}).$$

Next, for every fixed $n \in [1, N]$, let $\tilde{\tau}$ denote the optimal value of (16) in iteration n, and define

$$\begin{cases} \tau_{n,n-1} = \tau' \\ \tau_{n,i} = \tau_{n,i-1} + \frac{1}{L_i}(1 + \sqrt{1 + 2L_i\tau_{n,i-1}}). \end{cases} \tag{17}$$

Note that the τ_n sequence maintained in SPPPA coincides with the sequence $\tau_{n,n}$.

Theorem 3.1. *For any closed convex proper f and $x_0 \in \mathbb{R}^d$ and fixed sequence of proximal parameters $L_0, \ldots, L_{N-1}$, SPPPA guarantees*

$$\frac{f(y_N) - f_\star}{\frac{1}{2}\|x_0 - x_\star\|^2} \leq \Psi_N \leq \cdots \leq \Psi_0,$$

where $\Psi_n = 1/\tau_{n,N}$ is the guarantee of SPPPA based on the first-order responses seen up to iteration n and Ψ_0 is the minimax optimal guarantee ensured by OPPA.

Proof. Recall that at iteration n, the optimal value of τ' in (16) is at least τ_{n-1}. This is equivalent to saying that $\tau_{n,n-1} \geq \tau_{n-1,n-1}$ for all $n = 1, \ldots, N$.

Now, observe that for any $L > 0$, the expression $\tau + \frac{1}{L}(1 + \sqrt{1 + 2L\tau})$ is an increasing function of τ. We deduce that for any $n \in [1, N]$,

$$\begin{aligned}\tau_{n,n} &= \tau_{n,n-1} + \frac{1}{L_n}(1 + \sqrt{1 + 2L_n\tau_{n,n-1}}) \\ &\geq \tau_{n-1,n-1} + \frac{1}{L_n}(1 + \sqrt{1 + 2L_n\tau_{n-1,n-1}}) \\ &= \tau_{n-1,n}.\end{aligned}$$

We can chain this argument to get:

$$\begin{aligned}\tau_{n,n} &\geq \tau_{n-1,n-1} + \frac{1}{L_n}(1 + \sqrt{1 + 2L_n\tau_{n-1,n-1}}) \\ &\geq \tau_{n-2,n-1} + \frac{1}{L_n}(1 + \sqrt{1 + 2L_n\tau_{n-2,n-1}}) \\ &= \tau_{n-2,n}.\end{aligned}$$

Repeating this argument shows that $\tau_{0,N} \leq \tau_{1,N} \leq \cdots \leq \tau_{N,N}$, thereby completing the argument. □

As a pragmatic note, any feasible solution to the problem (16) suffices to ensure that the induction $H_n \geq 0$. OPPA corresponds to one specific (often suboptimal) feasible solution. As a practical consequence of this freedom to select suboptimal $(\mu, \lambda_\star)$, one may easily design a limited-memory variant of SPPPA where only memory of the last k data are stored $\{(y_i, f_i, g_i, z_{i+1}, \tau_i)\}_{i=n-k}^{n-1}$. In such a setting, the convex optimization problem (16) is then of fixed dimension $2k$, incurring only a constant per-iteration cost to the algorithm. In the setting of smooth convex optimization, a limited memory subgame perfect method has already been preliminarily, experimentally explored [21]. An adaptive, parameter-free subgame perfect method was developed in [22], showing even stronger practical performance on a wider numerical sample.

4 Subgame Perfection of SPPPA

This section constructs, for any given history of first-order information revealed before the nth iteration, a worst–case convex function for which no method of the form (8) can outperform the SPPPA guarantee. Our argument leverages the zero-chain construction ideas of [45] that were extended to proximal settings by [36]. Our process for dynamically constructing lower bounds parallels that of the smooth convex setting [21].

4.1 Dynamic Construction of Candidate Hard Problem Instance

Fix an $n \in [1, N]$ in this section. Let $\mathcal{H} = \{(x_i, f_i, g_i, y_i, \tau_i, z_{i+1})\}_{i=0}^{n-1}$ be the iterates and responses produced/observed by SPPPA with proximal parameters $(L_i)_{i\geq 0}$, where $y_i = \operatorname{prox}_{L_i,f}(x_i)$ and $g_i = L_i(x_i - y_i) \in \partial f(y_i)$, and $f_i = f(y_i)$. Note that by construction, for all $i, j \in \{0, \ldots, n-1\}$,

$$Q_{i,j} = f(y_i) - f(y_j) - \langle g_j, y_i - y_j\rangle \geq 0,$$

and

$$H_i = \tau_i(f(y_\star) - f(y_i)) + \tfrac{1}{2}\|x_0 - y_\star\|^2 - \tfrac{1}{2}\|z_{i+1} - y_\star\|^2 \geq 0.$$

We will take our hard function f to be a max-of-affine function

$$f_{\mathcal{H}}(x) = \max_{i\in\mathcal{I}_\star}\{f_i + \langle g_i, x - y_i\rangle\}, \tag{18}$$

where the tuples (f_i, g_i, y_i) come from the given history of first-order responses for $i = 0, \ldots, n-1$, and need to be constructed for $i = n, \ldots, N, \star$.

4.1.1 Dual of the Planning Subproblem

As the guarantee of SPPPA depends on the optimal value of the planning subproblem, it is natural to use a dual optimal solution to construct lower bounds. We continue using the notation defined in (13) and (14). Let $(\mu, \lambda_\star)$ denote the maximizers of (16) computed by SPPPA in its nth iteration, generating values $\tau' = \langle \tau, \mu \rangle + \langle \mathbf{1}, \lambda_\star \rangle$ and $z' = x_0 + Z\mu - G\lambda_\star$. We derive the dual to (16) below.

Lemma 4.1. *The dual of (16) is*

$$\inf_{\xi>0,\ w\in\mathbb{R}^d} \left\{ \frac{1}{2\xi}\|w\|^2 \ :\ \tau + \xi a - Z^\top w \le 0, \quad \mathbf{1} + \xi b + G^\top w \le 0 \right\}. \tag{19}$$

If the supremum in (16) is finite, it is attained and strong duality holds. For any primal/dual optimizers $(\mu, \lambda_\star)$ and (ξ, w),

$$w = \xi(z' - x_0).$$

Proof. Consider the Lagrangian with multiplier $\xi \ge 0$:

$$\mathcal{L}(\mu, \lambda_\star; \xi) = \langle \tau, \mu \rangle + \langle \mathbf{1}, \lambda_\star \rangle + \xi\big(\langle \mu, a \rangle + \langle \lambda_\star, b \rangle - \tfrac{1}{2}\|Z\mu - G\lambda_\star\|^2\big), \quad \mu, \lambda_\star \ge 0.$$

Using the Fenchel identity $-\frac{\xi}{2}\|t\|^2 = \inf_{w\in\mathbb{R}^d}\{\frac{1}{2\xi}\|w\|^2 - \langle w, t\rangle\}$ (valid for $\xi \ge 0$), we get

$$\mathcal{L}(\mu, \lambda_\star; \xi) = \inf_{w\in\mathbb{R}^d} \Big\{ \tfrac{1}{2\xi}\|w\|^2 + \langle \tau + \xi a - Z^\top w, \mu \rangle + \langle \mathbf{1} + \xi b + G^\top w, \lambda_\star \rangle \Big\}.$$

Maximizing over $\mu, \lambda_\star \ge 0$ yields finiteness if and only if

$$\tau + \xi a - Z^\top w \le 0, \qquad \mathbf{1} + \xi b + G^\top w \le 0,$$

in which case the supremum over $(\mu, \lambda_\star)$ equals 0. Then a short algebraic simplification gives the dual

$$\inf_{\xi>0,\ w\in\mathbb{R}^d} \left\{ \frac{1}{2\xi}\|w\|^2 \ :\ \tau + \xi a - Z^\top w \le 0, \quad \mathbf{1} + \xi b + G^\top w \le 0 \right\}.$$

If the supremum in (16) is finite, the recession directions that would make it $+\infty$ are excluded, which forces $\xi > 0$; standard conic/Fenchel duality (and closedness of the feasible set) then gives strong duality and attainment on both sides.

By the Fenchel step, at any primal/dual optimizers $(\mu, \lambda_\star)$ and (ξ, w) we have

$$w = \xi(Z\mu - G\lambda_\star) = \xi(z' - x_0). \qquad \square$$

It will be instructive to view the dual variable $w = \xi(z' - x_0)$ as parameterized by z' and $\xi = \frac{1}{f_m - f_\star}$ as parameterized by the variable $f_\star$. This quantity will coincide with the optimal value of the hard function that we will construct in this next section, justifying the notation $f_\star$. The constraints in the dual problem can be rewritten as constraints on $f_\star$ and z': for every $i \in \{0, \dots, n-1\}$,

$$f_\star \ \ge\ f_i + \langle g_i,\ z' - y_i \rangle, \tag{20}$$

$$\tau_i(f_\star - f_i) + \tfrac{1}{2}\|x_0 - z'\|^2 - \tfrac{1}{2}\|z_{i+1} - z'\|^2 \ \ge\ 0. \tag{21}$$

4.1.2 Construction of Our Dynamic Hard Instance

It suffices to consider the case where (16) is bounded as otherwise SPPPA outputs an exact minimizer of f. Let $(\mu, \lambda_\star)$ denote optimizers and let (w, ξ) be optimizers of the primal and dual problems. By Lemma 4.1, $w = \xi(z' - x_0)$. Let $e_n, \dots, e_N$ be unit vectors orthogonal to $\operatorname{span}\{g_0, \dots, g_{n-1}\}$ and mutually orthogonal. It is possible to pick these vectors under the assumption $d \ge N + 1$.

Then we define *future* iterates and first-order observations for $i \in \{n, \dots, N\}$ as follows (first at $i = n$, then inductively for $i > n$)

$$\begin{aligned}
\tau_n &= \tau' + \frac{1}{L_n}\left(1 + \sqrt{1 + 2L_n\tau'}\right) & \tau_i &= \tau_{i-1} + \frac{1}{L_i}\left(1 + \sqrt{1 + 2L_i\tau_{i-1}}\right) \\
x_n &= \frac{\tau'}{\tau_n} y_m + \frac{\tau_n - \tau'}{\tau_n} z' & x_i &= \frac{\tau_{i-1}}{\tau_i} y_{i-1} + \frac{\tau_i - \tau_{i-1}}{\tau_i} z_i \\
g_n &= \sqrt{\frac{f_m - f_\star}{\tau_n - \tau'}}\, e_n & g_i &= \sqrt{\frac{f_{i-1} - f_\star}{\tau_i - \tau_{i-1}}}\, e_i \\
f_n &= f_m - \frac{1}{L_n}\|g_n\|^2 & f_i &= f_{i-1} - \frac{1}{L_i}\|g_i\|^2 \\
y_n &= x_n - \frac{1}{L_n} g_n & y_i &= x_i - \frac{1}{L_i} g_i \\
z_{n+1} &= z' - (\tau_n - \tau') g_n & z_{i+1} &= z_i - (\tau_i - \tau_{i-1}) g_i.
\end{aligned}$$

For $i = \star$, set $f_\star = f_m - \frac{1}{\xi}$, $g_\star = 0$ and $y_\star = z_{N+1}$.

Note that with these definitions, $\tau_{n,i} = \tau_i$ for all $i = n, \dots, N$ so that $\Psi_n = \frac{1}{\tau_{n,N}} = \frac{1}{\tau_N}$.

4.2 Properties of the Candidate Hard Instance

The following three lemmas record useful algebraic properties of our construction. Proofs of each of these results are given in the appendix for completeness.

Lemma 4.2. *For $i \in \{n, \dots, N\}$, the following hold for each index $j < i$*

$$\langle g_j, y_i - y_j \rangle = \begin{cases} \langle g_j, z' - y_j \rangle + \frac{\tau'}{\tau_i} \langle g_j, y_m - z' \rangle & \text{if } j < n \\ -\frac{\tau_i - \tau_j}{\tau_i} \left(f_j - f_\star\right) & \text{if } j \geq n. \end{cases} \tag{22}$$

Lemma 4.3. *For $i \in \{n, \dots, N\}$, $\tau_i\big(f_i - f_\star\big)$ is nondecreasing and $\tau_n\big(f_n - f_\star\big) \geq \tau'(f_m - f_\star)$.*

Lemma 4.4. *It holds that $Q_{m,n} = Q_{n,n+1} = \cdots = Q_{N-1,N} = 0$. For $i \in \{n, \dots, N\}$,*

$$Q_{\star,i} = 0, \qquad H_i = \tau_i(f_\star - f_i) + \tfrac{1}{2}\|x_0 - y_\star\|^2 - \tfrac{1}{2}\|z_{i+1} - y_\star\|^2 = 0.$$

The following pair of propositions establish the key properties enabling $f_{\mathcal{H}}$ to serve as a dynamic hard lower bounding instance for proving subgame perfection. First we show that indeed $f_{\mathcal{H}}$ interpolates the past observed data and the proposed future values in our definition. Second we show that this construction has a zero-chain property, preventing any proximal point-type method (8) from discovering more than one new direction e_i per iteration.

Proposition 4.5. *The function $f_{\mathcal{H}}$ is proper, closed, and convex, satisfying for every $i \in \mathcal{I}$, $f_{\mathcal{H}}(y_i) = f_i$ and $g_i \in \partial f_{\mathcal{H}}(y_i)$.*

Proof. Note that $f_{\mathcal{H}}$ must be closed, convex, and proper since it is defined as a finite maximum of affine functions. As occurred in our previous lower bound construction for KLM in Proposition 2.2, it suffices to verify nonnegativity of each $Q_{i,j}$ to verify $f_{\mathcal{H}}(y_i) = f_i$ and $g_i \in \partial f_{\mathcal{H}}(y_i)$. Again, we do this case-wise:

First consider $i \in \{0, \dots, n-1\}$. If $j \in \{0, \dots, n-1\}$ as well, then $Q_{i,j} \geq 0$ follows since the past observations were generated from some convex function. If $j \in \{n, \dots, N\}$, then

$$\begin{aligned}
Q_{i,j} &= f_i - (f_j + \langle g_j, y_i - y_j \rangle) \\
&= f_i - (f_\star - \langle g_j, y_\star - y_j \rangle + \langle g_j, y_i - y_j \rangle) \\
&= f_i - f_\star + \langle g_j, y_\star - y_i \rangle \\
&= f_i - f_\star - (\tau_j - \tau_{j-1})\|g_j\|^2 \\
&= f_i - f_{j-1} \geq 0
\end{aligned} \tag{23}$$

where the second equality uses that $Q_{\star,j} = 0$ by Lemma 4.4. If $j = \star$, then

$$Q_{i,\star} = f_i - (f_\star + \langle g_\star, y_i - y_\star\rangle) = f_i - f_\star = f_i - f_m + \frac{1}{\xi} \geq 0.$$

Next consider $i \in \{n, \ldots, N\}$. If $j \in \{0, \ldots, n-1\}$, then

$$\begin{aligned}
Q_{i,j} &= f_i - f_j - \langle g_j, y_i - y_j\rangle \\
&= f_i - f_j - \left(\langle g_j, z' - y_j\rangle + \frac{\tau'}{\tau_i}\langle g_j, y_m - z'\rangle\right) \\
&= f_i - \left(1 - \frac{\tau'}{\tau_i}\right)(f_j + \langle g_j, z' - y_j\rangle) - \frac{\tau'}{\tau_i}(f_j + \langle g_j, y_m - y_j\rangle) \\
&\geq f_i - \left(1 - \frac{\tau'}{\tau_i}\right) f_\star - \frac{\tau'}{\tau_i} f_m \\
&= \frac{\tau_i(f_i - f_\star) - \tau'(f_m - f_\star)}{\tau_i} \geq 0,
\end{aligned}$$

where the second equality uses Lemma 4.2, the first inequality uses (20) and $Q_{m,j} \geq 0$ and the second inequality uses Lemma 4.3. If $j \in \{n, \ldots, i-1\}$, then

$$\begin{aligned}
Q_{i,j} &= f_i - f_j - \langle g_j, y_i - y_j\rangle \\
&= f_i - f_j + \frac{\tau_i - \tau_j}{\tau_i}(f_j - f_\star) \\
&= \frac{1}{\tau_i}\left(\tau_i(f_i - f_\star) - \tau_j(f_j - f_\star)\right) \geq 0,
\end{aligned}$$

where the second equality uses Lemma 4.2 and the final inequality follows from Lemma 4.3. If $j \in \{i+1, \ldots, N\}$, then this follows by the same reasoning presented in the chain of equalities (23). If $j = \star$, then

$$Q_{i,\star} = f_i - (f_\star + \langle g_\star, y_i - y_\star\rangle) = f_i - f_\star \geq 0.$$

Finally, consider $i = \star$. Then for $j \in \{0, \ldots, n-1\}$, having $Q_{\star,j} \geq 0$ is precisely guaranteed by the first dual constraint (20). For $j \in \{n, \ldots, N\}$, $Q_{\star,j} = 0$ is guaranteed by Lemma 4.4. □

Proposition 4.6. *Suppose $0 \leq n \leq N$ and the history $\mathcal{H} = \{(x_i, f_i, g_i, y_i, \tau_i, z_{i+1})\}_{i=0}^{n-1}$ is given. Let $f_{\mathcal{H}}$ denote the function constructed in (18) and Section 4.1. Let $j \in \{n, \ldots, N-1\}$. If $x \in x_0 + \operatorname{span}\{g_0, \ldots, g_{j-1}\}$, then*

$$\operatorname{prox}_{L_j, f_{\mathcal{H}}}(x) \in x_0 + \operatorname{span}\{g_0, \ldots, g_j\}.$$

Proof. Fix $j \in \{n, \ldots, N-1\}$ and define

$$\tilde{f}(x) = \max_{i \leq j}\{f_i + \langle g_i, x - y_i\rangle\} \leq f_{\mathcal{H}}(x).$$

Fix an $x \in x_0 + \operatorname{span}\{g_0, \ldots, g_{j-1}\}$. Let $y = \operatorname{prox}_{L_j, \tilde{f}}(x)$. We know that $g = L_j(x - y) \in \partial\tilde{f}(x)$ is a convex combination of $\{g_0, \ldots, g_j\}$. Then, as $x \in x_0 + \operatorname{span}\{g_0, \ldots, g_{j-1}\}$, we deduce that $y = x - g/L_j \in x_0 + \operatorname{span}\{g_0, \ldots, g_j\}$. Recall that we defined g_j to be orthogonal to g_i for all $i < j$. Thus, it follows that $\langle g_j, g\rangle \leq \|g_j\|^2$. Hence for any $i > j$, we observe that

$$\begin{aligned}
\tilde{f}(y) &\geq f_j + \langle g_j, y - y_j\rangle \\
&= f_j + \left\langle g_j, \left(x - \frac{1}{L_j} g\right) - \left(x_j - \frac{1}{L_j} g_j\right)\right\rangle \\
&= f_j + \frac{1}{L_j}\langle g_j, g_j - g\rangle \\
&\geq f_j \\
&\geq f_i + \langle g_i, y_j - y_i\rangle \\
&= f_i + \langle g_i, y - y_i\rangle,
\end{aligned}$$

where the first line lower bounds $\tilde{f}$ by its final linear term, the second and third apply definitions and simplify, the fourth uses $\langle g_j, g\rangle \le \|g_j\|^2$, the fifth uses $Q_{j,i} \ge 0$ and the last uses orthogonality of g_i to $y_j - y$. Similarly, considering $i = \star$, we have that $\tilde{f}(y) \ge f_j \ge f_\star + \langle g_\star, y - y_\star\rangle$. We deduce that $f_{\mathcal{H}}(y) = \tilde{f}(y)$.

Hence, for all $y' \in \mathbb{R}^d$, we have that

$$f_{\mathcal{H}}(y) + \frac{L_j}{2}\|y - x\|^2 = \tilde{f}(y) + \frac{L_j}{2}\|y - x\|^2 \le \tilde{f}(y') + \frac{L_j}{2}\|y' - x\|^2 \le f_{\mathcal{H}}(y') + \frac{L_j}{2}\|y' - x\|^2,$$

where the first inequality uses the definition of y and the second inequality uses the fact that $\tilde{f} \le f_{\mathcal{H}}$ pointwise. We deduce that $\operatorname{prox}_{L_j, f_{\mathcal{H}}}(x) = y \in x_0 + \operatorname{span}\{g_0, \dots, g_j\}$. □

4.3 Proof of Subgame Perfection

Theorem 4.1. *Assume $d \ge N + 1$. Suppose $0 \le n \le N$ and the history $\mathcal{H}$ is given. Let $f_{\mathcal{H}}$ denote the function constructed in (18) and Section 4.1. Then any method A of the form (8) generating $x_0, \dots, x_{n-1}$ when applied to $f_{\mathcal{H}}$ has terminal iterate y_N^A satisfying*

$$\frac{f_{\mathcal{H}}(y_N^A) - f_\star}{\frac{1}{2}\|x_0 - y_\star\|^2} \ge \Psi_n.$$

Consequently, combined with Theorem 3.1, SPPPA is subgame perfect.

Proof. Let x_j^A, y_j^A, g_j^A denote the sequence of iterates and subgradients generated by some method of the form (8) consistent with the given history $\mathcal{H}$ prior to iteration n. Note, for $j \ge n$, that x_j, y_j, g_j may be distinct from x_j^A, y_j^A, g_j^A. By Proposition 4.6, if $x_j^A \in x_0 + \operatorname{span}\{g_0, \dots, g_{j-1}\}$ then $y_j^A = \operatorname{prox}_{L_j, f_{\mathcal{H}}}(x_j^A) \in x_0 + \operatorname{span}\{g_0, \dots, g_j\}$ and hence $g_j^A \in \operatorname{span}\{g_0, \dots, g_j\}$. Since any method of the form (8) sets $x_{j+1}^A \in x_0 + \operatorname{span}\{g_0^A, \dots, g_j^A\}$, induction over $j = n, \dots, N$ yields

$$y_N^A \in x_0 + \operatorname{span}\{g_0^A, \dots, g_N^A\} \subseteq x_0 + \operatorname{span}\{g_0, \dots, g_N\}.$$

Further, $g_N^A = L_N(x_N^A - y_N^A) \in \partial f_{\mathcal{H}}(y_N^A)$ must lie in the convex hull of $\{g_0, \dots, g_N, g_\star\}$. Noting g_N is orthogonal to all other g_i, we have $\langle g_N, g_N^A\rangle \le \|g_N\|^2$. Consequently,

$$\begin{aligned}
f_{\mathcal{H}}(y_N^A) &\ge f_N + \langle g_N, y_N^A - y_N\rangle \\
&= f_N + \left\langle g_N, \left(x_N^A - \frac{1}{L_N} g_N^A\right) - \left(x_N - \frac{1}{L_N} g_N\right)\right\rangle \\
&= f_N + \frac{1}{L_N}\langle g_N, g_N - g_N^A\rangle \\
&\ge f_N = f_\star + \tfrac{1}{2\tau_N}\|x_0 - y_\star\|^2
\end{aligned}$$

where the first considers $i = N$ in the definition of $f_{\mathcal{H}}$, the second and third apply definitions and simplify, and the fourth uses that $\langle g_N, g_N^A\rangle \le \|g_N\|^2$, and the final equality uses that $H_N = 0$ due to Lemma 4.4. Since $\Psi_n = 1/\tau_{n,N} = \frac{1}{\tau_N}$, the claimed lower bound holds. □

5 Conclusion

We established subgame perfect algorithms for the settings of subgradient methods and proximal point algorithms. In particular, for subgradient methods, the Kelley-cutting plane Like Method of [1] is not only minimax optimal (as previously known), but also subgame perfect. For proximal point methods, we introduced a new extension of OPPA, which we call SPPPA, that dynamically optimizes its induction at every step and is subgame perfect. By constructing dynamic lower bounding instances as a function of observed first-order responses, we established that these methods are guaranteed to not only attain the minimax optimal convergence guarantee over the family of every convex problem instance, but over every subclass of problem instances restricted to agree with the first-order information seen up to any iteration n.

This game-theoretic notion was first applied to gradient methods in smooth convex minimization [21] and extended to adaptive backtracking settings in [22]. Many more settings exist where minimax optimal methods have been developed that represent fruitful opportunities to develop subgame perfect methods enabling optimal adaptation. Optimal methods for general fixed point and monotone operator settings have been considered by [46, 47], representing one opportunity. A minimax optimal proximal gradient method (OptISTA) for convex, composite optimization was recently developed by [36]. A subgame perfect extension of this also represents an opportunity. Additionally, it is known that there are multiple minimax optimal subgradient methods (for example, the classic subgradient method and the subspace search elimination method of [48]). It would be of interest to determine if there exist other subgame perfect subgradient methods, distinct from KLM.

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. An optimal variant of Kelley's cutting-plane method. *Mathematical Programming*, 160:321–351, 2016.

[2] Naum Zuselevich Shor. *Minimization Methods for Non-Differentiable Functions*, page 23. Springer Berlin Heidelberg, Berlin, Heidelberg, 1985.

[3] Claude Lemarechal. *An Extension of Davidon Methods to Nondifferentiable Problems*, pages 95–109. Springer Berlin Heidelberg, Berlin, Heidelberg, 1975.

[4] Philip Wolfe. *A Method of Conjugate Subgradients for Minimizing Nondifferentiable Functions*, pages 145–173. Springer Berlin Heidelberg, Berlin, Heidelberg, 1975.

[5] Bernard Martinet. Régularisation d'inéquations variationnelles par approximations successives. rev. française informat. *Recherche Opérationnelle*, 4:154–158, 1970.

[6] Bernard Martinet. Détermination approchée d'un point fixe d'une application pseudo-contractante. *CR Acad. Sci. Paris*, 274(2):163–165, 1972.

[7] R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[8] Haim Brézis and Pierre Louis Lions. Produits infinis de résolvantes. *Israel Journal of Mathematics*, 29:329–345, 1978.

[9] 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.

[10] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

[11] Donghwan Kim and Jeffrey A. Fessler. Optimized first-order methods for smooth convex minimization. *Math. Program.*, 159(1–2):81–107, 2016.

[12] 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.

[13] Saman Cyrus, Bin Hu, Bryan Van Scoy, and Laurent Lessard. A robust accelerated optimization algorithm for strongly convex functions. In *2018 Annual American Control Conference (ACC)*, pages 1376–1381, 2018.

[14] Chanwoo Park, Jisun Park, and Ernest K. Ryu. Factor-$\sqrt{2}$ acceleration of accelerated gradient methods. *Applied Mathematics & Optimization*, 88:1–38, 2021.

[15] Adrien Taylor and Yoel Drori. An optimal gradient method for smooth strongly convex minimization. *Math. Program.*, 199(1–2):557–594, jun 2022.

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

[17] Yoel Drori and Adrien B. Taylor. On the oracle complexity of smooth strongly convex minimization. *J. Complex.*, 68:101590, 2021.

[18] Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Math. Program.*, 145:451–482, 2012.

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

[20] Adrien B Taylor, Julien M 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.

[21] Benjamin Grimmer, Kevin Shu, and Alex L. Wang. Beyond minimax optimality: A subgame perfect gradient method. *arxiv:2412.06731*, 2025.

[22] Alan Luner and Benjamin Grimmer. A practical adaptive subgame perfect gradient method. *arxiv:2510.21617*, 2025.

[23] J. E. Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8(4):703–712, 1960.

[24] Krzysztof C. Kiwiel. An Aggregate Subgradient Method for Nonsmooth Convex Minimization. *Math. Program.*, 27(3):320–341, October 1983.

[25] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Acceleration of the Cutting-Plane Algorithm: Primal Forms of Bundle Methods*, pages 275–330. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993.

[26] Andrzej Ruszczynski. *Nonlinear Optimization*. Princeton University Press, Princeton, NJ, USA, 2006.

[27] Warren Hare and Claudia Sagastizábal. A Redistributed Proximal Bundle Method for Nonconvex Optimization. *SIAM J. Optim.*, 20(5):2442–2473, 2010.

[28] Guanghui Lan. Bundle-Level Type Methods Uniformly Optimal For Smooth And Nonsmooth Convex Optimization. *Mathematical Programming*, 149(1):1–45, Feb 2015.

[29] Yu Du and Andrzej Ruszczyński. Rate of Convergence of the Bundle Method. *J. Optim. Theory Appl.*, 173(3):908–922, June 2017.

[30] Mateo Díaz and Benjamin Grimmer. Optimal convergence rates for the proximal bundle method. *SIAM Journal on Optimization*, 33(2):424–454, 2023.

[31] Jiaming Liang and Renato D. C. Monteiro. A proximal bundle variant with optimal iteration-complexity for a large range of prox stepsizes. *SIAM Journal on Optimization*, 31(4):2955–2986, 2021.

[32] Osman Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, 1991.

[33] Osman Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

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

[35] Mathieu Barré, Adrien B. Taylor, and Francis Bach. Principled analyses and design of first-order methods with inexact proximal operators. *Mathematical Programming*, 201:185–230, 2023.

[36] Uijeong Jang, Shuvomoy Das Gupta, and Ernest K. Ryu. Computer-assisted design of accelerated composite optimization methods: Optista. *Mathematical Programming*, 2025.

[37] Arkadi S Nemirovsky. On optimality of Krylov's information when solving linear operator equations. *Journal of Complexity*, 7(2):121–130, 1991.

[38] Arkadi Nemirovsky. Information-based complexity of linear operator equations. *Journal of Complexity*, 8(2):153–175, 1992.

[39] Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2nd edition, 2018.

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

[41] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184:71–120, 2020.

[42] Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. *International Conference on Machine Learning*, 2020.

[43] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, 185(1-2):315–355, 2021.

[44] Radu-Alexandru Dragomir, Adrien B Taylor, Alexandre d'Aspremont, and Jérôme Bolte. Optimal complexity and certification of bregman first-order methods. *Mathematical Programming*, 194:41–83, 2022.

[45] Yoel Drori and Adrien B Taylor. On the oracle complexity of smooth strongly convex minimization. *Journal of Complexity*, 68:101590, 2022.

[46] Felix Lieder. On the convergence rate of the halpern-iteration. *Optimization letters*, 15(2):405–418, 2021.

[47] Donghwan Kim. Accelerated proximal point method for maximally monotone operators. *Mathematical Programming*, 190(1–2):57–87, 2021.

[48] Yoel Drori and Adrien B. Taylor. Efficient first-order methods for convex minimization: a constructive approach. *Mathematical Programming*, 184(1):183–220, 2020.

6 Deferred Proofs of Lemmas in SPPPA's Lower Bound

6.1 Proof of Lemma 4.2

We handle the $i = n$ and $i > n$ cases separately. First, suppose $i = n$. Since $i > j$, necessarily $j < n$, so only the first case of the lemma statement is relevant. We write

$$\langle g_j, y_n - y_j\rangle = \langle g_j, y_n - z'\rangle + \langle g_j, z' - y_j\rangle.$$

By construction, g_n is orthogonal to $\mathrm{span}\{g_0, \dots, g_{n-1}\}$, and hence to g_j for all $j < n$. Since $y_n = x_n - \frac{1}{L_n} g_n$, we have

$$\langle g_j, y_n - z'\rangle = \langle g_j, x_n - z'\rangle.$$

From the definition of x_n as $\frac{\tau'}{\tau_n} y_m + \frac{\tau_n - \tau'}{\tau_n} z'$, we then have that

$$\langle g_j, x_n - z'\rangle = \frac{\tau'}{\tau_n} \langle g_j, y_m - z'\rangle.$$

Combining the displays gives the claim in this first case of $i = n$.

For the remainder of the proof, assume $i > n$. We expand

$$\langle g_j, y_i - y_j\rangle = \langle g_j, y_i - z_i\rangle + \langle g_j, z_i - y_j\rangle. \tag{24}$$

By definition, $z_i = z' - \sum_{\ell=n}^{i-1} (\tau_\ell - \tau_{\ell-1}) g_\ell$. Thus the second term in this expansion is

$$\langle g_j, z_i - y_j\rangle = \begin{cases} \langle g_j, z' - y_j\rangle & \text{if } j < n \\ f_\star - f_j & \text{if } j \geq n \end{cases} \tag{25}$$

where the $j \geq n$ case uses that $\|g_j\|^2 = \frac{f_{j-1} - f_\star}{\tau_j - \tau_{j-1}}$ and $f_j = f_{j-1} - \frac{1}{L_j}\|g_j\|^2$.

For the first term in this expansion, we use the inductive definitions of x_i, y_i, z_{i+1} and τ_i to write

$$\begin{aligned} \langle g_j, y_i - z_i\rangle &= \langle g_j, x_i - z_i\rangle \\ &= \frac{\tau_{i-1}}{\tau_i} \langle g_j, y_{i-1} - z_i\rangle \\ &= \begin{cases} \frac{\tau_n}{\tau_{n+1}} \langle g_j, y_n - z' + (\tau_n - \tau') g_{i-1}\rangle & \text{if } i = n+1 \\ \frac{\tau_{i-1}}{\tau_i} \langle g_j, y_{i-1} - z_{i-1} + (\tau_{i-1} - \tau_{i-2}) g_{i-1}\rangle & \text{if } i \geq n+2 \end{cases} \end{aligned}$$

Observing that if $j < i - 1$, we have $\langle g_j, g_{i-1}\rangle = 0$, we can unroll this recurrence, giving

$$\langle g_j, y_i - z_i\rangle = \begin{cases} \frac{\tau'}{\tau_i} \langle g_j, y_m - z'\rangle & \text{if } j < n, \\ \frac{\tau_n}{\tau_i} \langle g_n, y_n - z' + (\tau_n - \tau') g_n\rangle & \text{if } j = n, \\ \frac{\tau_j}{\tau_i} \langle g_j, y_j - z_j + (\tau_j - \tau_{j-1}) g_j\rangle & \text{if } j > n. \end{cases}$$

In the latter two cases, we use that $x_n, z' \in x_0 + \mathrm{span}\{g_0, \dots, g_{n-1}\}$ and $x_j, z_j \in x_0 + \mathrm{span}\{g_0, \dots, g_{j-1}\}$ so that $\langle g_n, x_n - x_0\rangle = \langle g_n, z' - x_0\rangle = 0$ and $\langle g_j, x_j - x_0\rangle = \langle g_j, z_j - x_0\rangle = 0$, together with $y_n = x_n - \frac{1}{L_n} g_n$ and $y_j = x_j - \frac{1}{L_j} g_j$, and the identities

$$\|g_n\|^2 = \frac{f_{n-1} - f_\star}{\tau_n - \tau'}, \quad f_n = f_{n-1} - \frac{1}{L_n}\|g_n\|^2, \qquad \|g_j\|^2 = \frac{f_{j-1} - f_\star}{\tau_j - \tau_{j-1}}, \quad f_j = f_{j-1} - \frac{1}{L_j}\|g_j\|^2,$$

to conclude that

$$\langle g_n, y_n - z' + (\tau_n - \tau')g_n\rangle = f_n - f_\star, \qquad \langle g_j, y_j - z_j + (\tau_j - \tau_{j-1})g_j\rangle = f_j - f_\star.$$

Applying these simplifications, we find

$$\langle g_j, y_i - z_i\rangle = \begin{cases} \frac{\tau'}{\tau_i}\langle g_j, y_m - z'\rangle & \text{if } j < n, \\ \frac{\tau_j}{\tau_i}(f_j - f_\star) & \text{if } j \geq n. \end{cases} \tag{26}$$

Combining (25) and (26) yields our claimed formula when $i > n$, completing our proof.

6.2 Proof of Lemma 4.3

Define $A_i = \tau_i(f_i - f_\star)$ as our main quantity of interest in this lemma. For convenience and as a minor abuse of notation, we introduce an index $i = n-1$ with

$$\tau_{n-1} \coloneqq \tau', \qquad f_{n-1} \coloneqq f_m, \qquad A_{n-1} \coloneqq \tau'(f_m - f_\star).$$

With this convention, our construction of future iterate values can be written uniformly as follows: for every $i \in \{n, \dots, N\}$,

$$\delta_i = \tau_i - \tau_{i-1} = \frac{1}{L_i}\left(1 + \sqrt{1 + 2L_i\tau_{i-1}}\right),$$
$$\|g_i\|^2 = \frac{f_{i-1} - f_\star}{\delta_i}, \qquad f_i = f_{i-1} - \frac{1}{L_i}\|g_i\|^2.$$

For any $i \geq n$, we have $f_i - f_\star = f_{i-1} - \frac{1}{L_i}\|g_i\|^2 - f_\star = (f_{i-1} - f_\star)\left(1 - \frac{1}{L_i\delta_i}\right)$, and so

$$\begin{aligned} A_i - A_{i-1} &= \tau_i(f_i - f_\star) - \tau_{i-1}(f_{i-1} - f_\star) \\ &= (f_{i-1} - f_\star)\left(\tau_i\left(1 - \frac{1}{L_i\delta_i}\right) - \tau_{i-1}\right). \end{aligned}$$

By construction, $\delta_i > 0$ and $\|g_i\|^2 \geq 0$, so $f_{i-1} - f_\star = \|g_i\|^2\delta_i \geq 0$. Hence it suffices to show

$$\tau_i\left(1 - \frac{1}{L_i\delta_i}\right) - \tau_{i-1} \;\geq\; 0 \qquad \text{for all } i \in \{n, \dots, N\}. \tag{27}$$

Using $\delta_i = \tau_i - \tau_{i-1}$, we rewrite

$$\tau_i\left(1 - \frac{1}{L_i\delta_i}\right) - \tau_{i-1} = (\tau_i - \tau_{i-1}) - \frac{\tau_i}{L_i\delta_i} = \delta_i - \frac{\tau_i}{L_i\delta_i} = \frac{L_i\delta_i^2 - \tau_i}{L_i\delta_i}.$$

Since $\delta_i, L_i > 0$, it remains to verify that $L_i\delta_i^2 \geq \tau_i$. By definition, δ_i is a root of the quadratic equation

$$\frac{L_i\delta_i^2}{2} - \delta_i - \tau_{i-1} = 0.$$

Recognizing, $\tau_{i-1} + \delta_i = \tau_i$, we deduce that $L_i\delta_i^2 \geq \frac{L_i\delta_i^2}{2} = \tau_i$, completing the proof.

6.3 Proof of Lemma 4.4

First, $Q_{i-1,i} = 0$ for all $i > n$:

$$f_i + \langle g_i, y_{i-1} - y_i\rangle = f_i + \left\langle g_i, y_{i-1} - \left(\frac{\tau_{i-1}}{\tau_i}y_{i-1} + \frac{\tau_i - \tau_{i-1}}{\tau_i}z_i - \frac{1}{L_i}g_i\right)\right\rangle = f_i + \frac{1}{L_i}\|g_i\|^2 = f_{i-1}$$

where the second equality uses that g_i is orthogonal to $\operatorname{span}\{g_0, \dots, g_{i-1}\}$ which contains both $y_{i-1} - x_0$ and $z_i - x_0$. Second, $Q_{\star,i} = 0$ for all $i > n$, as

$$f_i + \langle g_i, y_\star - y_i\rangle = f_{i-1} + \langle g_i, z_{N+1} - y_{i-1}\rangle = f_{i-1} - (\tau_i - \tau_{i-1})\|g_i\|^2 = f_\star$$

where the first equality uses that $Q_{i-1,i} = 0$ and substitutes the definition $y_\star = z_{N+1}$, the second uses the definition of z_{N+1} and orthogonality of each e_i, and the third uses the fact that $\|g_i\|^2 = \frac{f_{i-1}-f_\star}{\tau_i-\tau_{i-1}}$. For both arguments above, the case of $i = n$ is identical, replacing τ_{i-1}, z_i, y_{i-1}, and f_{i-1} with τ', z', y_m, and f_m respectively.

From strong duality (see Lemma 4.1), we know that the primal/dual optimizers $(\mu, \lambda_\star)$ and (ξ, w) with the associated $z' = x_0 + Z\mu - G\lambda_\star$ and $\frac{1}{\xi} = f_m - f_\star$ have

$$\tau' = \frac{\xi}{2}\|z' - x_0\|^2, \qquad \tau'(f_\star - f_m) = -\frac{1}{2}\|z' - x_0\|^2. \tag{28}$$

Evaluating H' at $y_\star = z_{N+1}$, using (28) and the Pythagorean theorem, yields

$$H' = \tau'(f_\star - f_m) + \tfrac{1}{2}\|x_0 - y_\star\|^2 - \tfrac{1}{2}\|z' - y_\star\|^2 = \tau'(f_\star - f_m) + \tfrac{1}{2}\|x_0 - z'\|^2 = 0.$$

Finally, we verify that $H_i = 0$ for all $i \geq n$. We do so inductively:

$$\begin{aligned} H_n &= H' + (\tau_n - \tau')Q_{\star,n} + \tau' Q_{m,n} = 0, \\ H_i &= H_{i-1} + (\tau_i - \tau_{i-1})Q_{\star,i} + \tau_{i-1}Q_{i-1,i} = 0 \qquad \text{for } i > n \end{aligned}$$

where the equality on each line uses Lemma 3.2 and the second equality on each line recognizes expressions that we have (inductively) shown to be zero.