A Complete Characterization of Optimal Subgradient Methods for Lipschitz Convex Minimization

Aaron Zoll[*] Benjamin Grimmer[†]

Abstract

We consider the design of optimal fixed-step first-order methods for M-Lipschitz convex optimization given $\|x_0 - x_\star\| \leq D$. Prior works have identified several distinct fixed-step methods, parameterized by a matrix of stepsizes W, with the (information-theoretic) minimax optimal rate $MD/\sqrt{N+1}$ of objective gap convergence. We provide a complete characterization of every optimal fixed-step method. Moreover, we show every optimal fixed-step method can be derived from the constructive approach of [2] and provide a polyhedral representation of the set of optimal methods through proof multipliers. From this characterization, we show that no anytime optimal fixed-step subgradient methods exist.

1 Introduction

We consider the classical problem of Lipschitz (potentially nonsmooth) convex minimization [9]. In arbitrary (potentially large) dimension d, we study problems of the form

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

Given problem parameters $M, D > 0$, a problem instance is a pair (f, x_0) with f convex, M-Lipschitz, and attaining a minimizer $x_\star$ with Euclidean norm $\|x_0 - x_\star\| \leq D$. We denote the set of all instances of any dimension d by $\mathtt{P}_{M,D}$. Herein, we consider iterative methods given access to a first-order oracle mapping, in the ith iteration, x_i to a pair (f_i, g_i) with function value $f_i = f(x_i)$ and subgradient $g_i \in \partial f(x_i)$, lying in the subdifferential $\partial f(x) = \{g : f(y) \geq f(x) + \langle g, y - x\rangle\ \forall y\}$.

Given an iteration budget N, we consider the class of N-step subgradient span algorithms for which $x_n \in x_0 - \operatorname{span}\{g_0, \ldots, g_{n-1}\}$, denoted by $\mathtt{A}_{\text{span}}$. Given an algorithm, we measure its performance by its worst-case final objective gap over all problem instances. Optimal algorithm design then seeks methods with the best worst-case performance.

This "minimax optimal" rate of convergence is known to be exactly $\frac{MD}{\sqrt{N+1}}$. Namely,

$$\min_{a\in\mathtt{A}_{\text{span}}} \max_{(f,x_0)\in\mathtt{P}_{M,D}} f(x_N) - f(x_\star) = \frac{MD}{\sqrt{N+1}}. \tag{1.1}$$

Equality is attained by a simple averaged subgradient method, iterating from $y_0 = x_0$ with $y_n = y_{n-1} - \frac{D}{M\sqrt{N+1}} g_{n-1}$ for $g_{n-1} \in \partial f(y_{n-1})$, reporting final iterate as the average $x_N = \frac{1}{N+1}\sum_{i=0}^N y_i$. This method has $f_N - f_\star \leq \frac{MD}{\sqrt{N+1}}$ by [1, Section 3.1] and no method can do better by the lower bound of [4, Theorem A.1].

[*]Johns Hopkins University, Department of Applied Mathematics and Statistics, azoll1@jhu.edu
[†]Johns Hopkins University, Department of Applied Mathematics and Statistics, grimmer@jhu.edu

This algorithm takes the form of a fixed-step first-order method (FSFOM). These methods, parameterized by a lower triangular matrix W, are defined by the iteration

$$x_n = x_0 - \sum_{i=0}^{n-1} W_{n,i} g_i$$

for $n = 1, \ldots, N$, where $g_i \in \partial f(x_i)$ are subgradients computed at each iteration. For ease of presentation, the rows of W are indexed from 1 to N while the columns are indexed from 0 to $N-1$. Denote the set of all $N \times N$ lower triangular matrices by $\mathbb{L}^N$ and those with positive lower triangular entries by $\mathbb{L}^N_{++}$. For example, the above averaged subgradient method corresponds to the matrix with constant $W_{i,j} = \frac{D}{M\sqrt{N+1}}$ for $0 \le j < i \le N-1$ and $W_{N,j} = \frac{D}{M\sqrt{N+1}} \frac{N-j}{N+1}$.

Surprisingly, many minimax optimal fixed-step methods exist. Zamani and Glineur [14] provided an optimal subgradient method $x_n = x_{n-1} - h_{n-1} g_{n-1}$ with nonconstant stepsizes (but no final averaging). Their method is parameterized by the matrix with $W_{i,j} = \frac{D}{M\sqrt{N+1}} \frac{N-j}{N+1}$. Gösgens and Van Parys [5] identified infinitely many optimal subgradient step methods which interpolate between the averaged method and the nonconstant stepsize method above. Separately, the constructive design approach of Drori and Taylor [2] produced an optimal "momentum-type" method. Their algorithm, derived from their Subspace-Search Elimination Procedure (SSEP), is parameterized by the matrix with $W_{i,j} = \frac{D}{M\sqrt{N+1}} \frac{i-j}{i+1}$.

Note that beyond the restricted family of fixed-step methods, other minimax optimal subgradient methods are known. For example, see the Kelley cutting-plane Like Method of [4]. In fact, this method attains a strong dynamic notion of optimality called subgame perfection [6].

Here we characterize all minimax optimal fixed-step methods, denoted $\mathcal{W}_\star^{M,D,N}$. First, we prove optimality for a considered family of algorithms. These are precisely the methods able to be generated by the constructive approach of [2]. Doing so provides a convex (polyhedral) representation of a family of optimal methods. Complementing this, we show no other optimal methods exist, establishing this family as a complete characterization of all optimal fixed-step methods.

1.1 Related Work and Preliminaries

Performance Estimation. The foundational work [3] introduced Performance Estimation Problems (PEPs) as a principled way to analyze first-order methods over traditional problem classes. Subsequently, [10, 11] established the exactness of this tool via interpolation theory. For some W, a PEP finds the worst final objective gap $f(x_N) - f(x_\star)$ over all possible problem instances. Formally,

$$p_\star(W) = \begin{cases} \max & f(x_N) - f(x_\star) \\ \text{s.t.} & x_n = x_0 - \sum_{i=0}^{n-1} W_{n,i} g_i \qquad \forall n = 1, 2, \ldots, N \\ & (f, x_0) \in \mathtt{P}_{M,D},\ g_i \in \partial f(x_i) \quad \forall i = 0, 1, \ldots, N. \end{cases} \tag{1.2}$$

The key observation to tractably solve PEPs is that the algorithm's trajectory and performance depend only on the first-order information at the iterates $x_0, \ldots, x_N$ and at the minimizer $x_\star$. For ease in referring to such points, we consider the index set $\mathcal{I}_N^\star = \{\star, 0, 1, \ldots, N\}$. The relevant quantities to algorithmic performance are then

$$x_i, \qquad f_i = f(x_i), \quad g_i \in \partial f(x_i) \qquad \forall i \in \mathcal{I}_N^\star \tag{1.3}$$

where we require $g_\star = 0$ at the minimizer $x_\star$. Following the interpolation theory of [10, Theorem 3.5], we know that for observations $\{(x_i, f_i, g_i)\}_{i \in \mathcal{I}_N^\star}$, there exists an M-Lipschitz, convex function, minimized at $x_\star$ for which (1.3) holds if and only if the following values are nonnegative for all $i,j \in \mathcal{I}_N^\star$

$$\mathcal{C}_{i,j} := f_i - f_j - \langle g_j, x_i - x_j \rangle, \qquad \mathcal{M}_i := M^2 - \|g_i\|^2.$$

Combined with the distance bound $\mathcal{D} := D^2 - \|x_0 - x_\star\|^2 \geq 0$, Taylor, Hendrickx, and Glineur [10,11] showed one can define an equivalent optimization problem to (1.2) over finitely many variables.

The standard approach to simplifying such PEP reformulations is to consider a Gram change of variables. Define $P = [x_0 - x_\star \mid g_0 \mid g_1 \mid \ldots \mid g_N]$ and the Gram matrix of all inner products between these vectors as $G = P^\top P$. After substituting the equality definitions of $x_n = x_0 - \sum_{i=0}^{n-1} W_{n,i} g_i$ and $g_\star = 0$ into the remaining inequality constraints, observe that $\mathcal{D}$, $\mathcal{C}_{i,j}$, and $\mathcal{M}_i$ are linear in $F = [f_\star, f_0, \ldots, f_N]^\top$ and G. We denote these functions by $\mathcal{D}(F,G)$, $\mathcal{C}_{i,j}(F,G)$ and $\mathcal{M}_i(F,G)$. Additionally, as a Gram matrix, we require G to be positive semidefinite. Changing to these variables results in a semidefinite program, denoted by

$$p_\star(W) = p_{gram}(W) = \begin{cases} \max\limits_{F,G} & f_N - f_\star \\ \text{s.t.} & \mathcal{D}(F,G) \geq 0 \\ & \mathcal{C}_{i,j}(F,G) \geq 0 \quad \forall i,j \in \mathcal{I}_N^\star \\ & \mathcal{M}_i(F,G) \geq 0 \quad \forall i = 0,\ldots,N \\ & G \succeq 0. \end{cases} \tag{1.4}$$

Although this reformulation is obtained by a relaxation, tightness (the equality above) follows from noting that any positive semidefinite G can be factored into $P^\top P$ and any feasible F, G can be used to construct an interpolating problem instance in dimension at most $d = N+2$.

We denote the dual variables to the constraints as nonnegative $\sigma, \lambda_{i,j}, \mu_i$ and positive semidefinite Z respectively. Such multipliers prove a guarantee $f_N - f_\star \leq r$ if the identity in F and G of

$$\sigma\mathcal{D} + \sum_{i,j \in \mathcal{I}_N^\star} \lambda_{i,j}\mathcal{C}_{i,j} + \sum_{i=0}^{N} \mu_i \mathcal{M}_i + \langle Z, G \rangle = r - (f_N - f_\star) \tag{1.5}$$

holds (substituting $x_n = x_0 - \sum_{i=0}^{n-1} W_{n,i} g_i$ and $g_\star = 0$). Note that $\mathcal{C}_{i,i}$ is identically zero, so we can fix $\lambda_{i,i}$ as zero for each index i throughout without loss.

For any feasible (F,G), the nonnegativity of the left-hand-side of the identity (1.5) then ensures the right-hand-side is nonnegative, proving the rate. The dual problem then minimizes r over all feasible certificates σ, λ, μ, Z. For an optimal algorithm and proof, this rate is $r = \frac{MD}{\sqrt{N+1}}$.

Lemma 1.1. *Given any $M, D > 0$ and any matrix $W \in \mathbb{L}_{++}^N$ parameterizing a FSFOM, there exist nonnegative $\sigma, \lambda_{i,j}, \mu_i$ and positive semidefinite Z such that* (1.5) *holds with $r = p_\star(W)$.*

Proof. We prove this by establishing strong duality with dual attainment via a Slater point (strictly satisfying all non-affine constraints). Fix any orthonormal vectors $e_1, \ldots, e_{N+2}$. Set $f_n = 0$, $g_n = \frac{M}{2} e_{n+1}$, $x_n = x_0 - \sum_{i=0}^{n-1} W_{n,i} g_i$ for $n = 0, \ldots, N$, with $x_0 = 0$, $x_\star = -\frac{D}{2\sqrt{N+2}} \sum_{i=1}^{N+2} e_i$, $f_\star = -\frac{MD}{4\sqrt{N+2}}$, and $g_\star = 0$. Let (F,G) be their Gram representation. By construction, $\|x_0 - x_\star\| < D$ and $\|g_i\| < M$, so $\mathcal{D}(F,G), \mathcal{M}_i(F,G) > 0$. Note that

$$\mathcal{C}_{i,j}(F,G) = f_i - f_j - \langle g_j, x_i - x_j \rangle = \begin{cases} \frac{M^2}{4} W_{i,j} & i > j \\ 0 & \text{else,} \end{cases} \quad \mathcal{C}_{i,\star}(F,G) = \frac{MD}{4\sqrt{N+2}}, \quad \mathcal{C}_{\star,i}(F,G) = 0$$

where we recall $f_i = f_j$, the method fixes $x_n = x_0 - \sum_{i=0}^{n-1} W_{n,i} g_i$, and the gradients are orthogonal. Since $x_0 - x_\star$ lies outside span $\{e_1, \dots, e_{N+1}\}$ and the gradients are orthogonal, $x_0 - x_\star, g_0, \dots, g_N$ are linearly independent and $G \succ 0$. Since the constraints $\mathcal{D}$, $\mathcal{C}_{i,j}$, $\mathcal{M}_i$ are all affine in (F, G), and $G \succ 0$, the feasible (F, G) is a Slater point for (1.4). Since $p_\star(W)$ is finite, [8, Theorem 28.2] applies and strong duality holds with dual attainment. Therefore, there exist nonnegative $\sigma, \lambda_{i,j}, \mu_i$ and positive semidefinite Z such that (1.5) holds with $r = p_\star(W)$. $\square$

Converting From Stepsize Matrices to Proof Multipliers Recently, a complete characterization of all optimal nonexpansive fixed-point methods was given by [13]. A key step was identifying a relation between a matrix of stepsizes (there denoted H) and the associated proof multipliers λ. From this, they characterized every optimal fixed-point method based on associated polynomial equalities and inequalities, called H-invariances and H-certificates.

Here, we follow the same tactic. In particular, we will show that each optimal method has a unique dual optimal PEP proof, all using the same common values of

$$\sigma = \frac{M}{2D\sqrt{N+1}}, \qquad \lambda_{\star,i} = \frac{1}{N+1}, \qquad \mu_i = \frac{D}{2M(N+1)^{3/2}}, \qquad Z = \sigma w w^\top \tag{1.6}$$

with $w = [1, -\frac{D}{M\sqrt{N+1}}, \dots, -\frac{D}{M\sqrt{N+1}}]^\top$ and setting each $\lambda_{i,\star}$ and $\lambda_{i,j}$ with $i > j$ to zero. So optimal methods only differ in their optimal proof structure in the remaining $\lambda_{i,j}$ with $0 \le i < j \le N$. We find a bijection between these upper triangular proof multipliers λ and the lower triangular stepsizes W. Define $W(\lambda)$ recursively by

$$W_{n,j} = \frac{1}{\Lambda_n}\left(\frac{D}{M(N+1)^{3/2}} + \sum_{i=j+1}^{n-1} \lambda_{i,n} W_{i,j}\right), \qquad 0 \le j < n \le N, \tag{1.7}$$

where $\Lambda_n = \frac{1}{N+1} + \sum_{i=0}^{n-1} \lambda_{i,n}$, and define $\lambda(W)$ recursively by

$$\lambda_{j+1,n} = \frac{1}{W_{j+1,j}}\left(\frac{D W_{n,j}}{M(N+1)^{3/2} W_{n,n-1}} - \frac{D}{M(N+1)^{3/2}} - \sum_{i=j+2}^{n-1} \lambda_{i,n} W_{i,j}\right), \qquad j = n-2, \dots, 0,$$

$$\lambda_{0,n} = \frac{D}{M(N+1)^{3/2} W_{n,n-1}} - \frac{1}{N+1} - \sum_{i=1}^{n-1} \lambda_{i,n}. \tag{1.8}$$

These mappings are exactly each other's inverse on the domains given below.

Proposition 1.2. *Given $\lambda_{i,j}$ for $0 \le i < j \le N$ with $\Lambda_n = \frac{1}{N+1} + \sum_{i=0}^{n-1} \lambda_{i,n} > 0$, and define the lower triangular $W = W(\lambda) \in \mathbb{L}^N$ by (1.7). Conversely, for lower triangular W with $W_{n,n-1} > 0$ define $\lambda = \lambda(W)$ by (1.8). Then $W(\cdot)$ and $\lambda(\cdot)$ are inverses, providing bijections between*

$$\{\lambda \mid \Lambda_n > 0,\ n = 1, \dots, N\} \quad \text{and} \quad \left\{W \in \mathbb{L}^N \mid W_{n,n-1} > 0,\ n = 1, \dots, N\right\}.$$

Proof. We first verify the mappings are well-defined on the stated domains. At $j = n-1$ the sum in (1.7) is empty, so $W_{n,n-1} = \frac{D}{M(N+1)^{3/2}\Lambda_n}$, and hence $W_{n,n-1} > 0$ if and only if $\Lambda_n > 0$.

We now show these maps are inverses directly. Suppose $\Lambda_n > 0$ for all n and set $W = W(\lambda)$, which by the mapping in (1.7) has positive subdiagonal. Fixing n, the case $j = n-1$ of (1.7) gives $\Lambda_n = \frac{D}{M(N+1)^{3/2} W_{n,n-1}}$, the quantity appearing throughout (1.8). Inducting on j from $n-2$ to 0, suppose $\lambda(W)_{i,n} = \lambda_{i,n}$ for $i = j+2, \dots, n-1$. Then

$$\lambda(W)_{j+1,n} = \frac{1}{W_{j+1,j}}\left(\frac{D W_{n,j}}{M(N+1)^{3/2} W_{n,n-1}} - \frac{D}{M(N+1)^{3/2}} - \sum_{i=j+2}^{n-1} \lambda_{i,n} W_{i,j}\right) = \lambda_{j+1,n},$$

by $\Lambda_n W_{n,j} - \frac{D}{M(N+1)^{3/2}} = \sum_{i=j+1}^{n-1} \lambda_{i,n} W_{i,j}$. Finally, $\lambda(W)_{0,n} = \Lambda_n - \frac{1}{N+1} - \sum_{i=1}^{n-1} \lambda_{i,n} = \lambda_{0,n}$.

Now suppose W has positive subdiagonal and set $\lambda = \lambda(W)$. Summing the definition of $\lambda_{0,n}$ in (1.8) with $\frac{1}{N+1} + \sum_{i=1}^{n-1} \lambda_{i,n}$ gives $\Lambda_n = \frac{D}{M(N+1)^{3/2} W_{n,n-1}}$. We show $W(\lambda) = W$ by induction on n. For $n = 1$, $W(\lambda)_{1,0} = \frac{D}{M(N+1)^{3/2}\Lambda_1} = W_{1,0}$. Assume $W(\lambda)_{i,j} = W_{i,j}$ for all $i < n$ and consider $j < n$. If $j = n-1$, then $W(\lambda)_{n,n-1} = \frac{D}{M(N+1)^{3/2}\Lambda_n} = W_{n,n-1}$. Otherwise,

$$\begin{aligned} W(\lambda)_{n,j} &= \frac{1}{\Lambda_n}\left(\frac{D}{M(N+1)^{3/2}} + \sum_{i=j+1}^{n-1} \lambda_{i,n} W_{i,j}\right) \\ &= \frac{1}{\Lambda_n}\left(\frac{D}{M(N+1)^{3/2}} + \sum_{i=j+2}^{n-1} \lambda_{i,n} W_{i,j} + \lambda_{j+1,n} W_{j+1,j}\right) = W_{n,j}, \end{aligned}$$

where the last equality considers the definition of $\lambda_{j+1,n}$ in (1.8). $\square$

1.2 Main Theorem

Below we characterize the entire set of minimax optimal methods W as exactly those whose associated upper triangular λ block lies in the polyhedron

$$\mathcal{P}^N = \left\{ \lambda \mid \lambda \geq 0,\ \sum_{i=j+1}^{N} \lambda_{j,i} - \sum_{i=0}^{j-1} \lambda_{i,j} = \frac{1}{N+1},\ j = 0, \dots, N-1 \right\}. \tag{1.9}$$

We prove this characterization in two steps. Section 2 shows that each $\lambda \in \mathcal{P}^N$ (combined with the common values (1.6)) provides a certificate of optimality for $W(\lambda)$. These proofs are exactly of the form generated by the dual in the constructive approach of [2]. Section 3 shows that for any optimal method W, any associated optimal proof multipliers (from Lemma 1.1) must take the common values (1.6) and set the remaining block of λ on $i, j = 0, \dots, N$ to be the upper triangular $\lambda(W)$. As a by-product, their optimal proof is unique (up to vacuous freedom in $\lambda_{i,i}$). This is proven by a novel use of complementary slackness between minimax optimal methods and maximin hard instances.

Theorem 1.3. *For any $M, D > 0$ and $N \in \mathbb{N}$, the set of optimal methods is $\mathcal{W}_\star^{M,D,N} = W(\mathcal{P}^N)$.*

Corollary 1.4. *For $W \in \mathcal{W}_\star^{M,D,N}$, there exist unique dual optimal multipliers σ, λ, μ, Z, given by* (1.6) *with the remaining λ block on $i, j = 0, \dots, N$ as the upper triangular matrix $\lambda(W)$.*

Lastly, we note two connections and insights from relating these results to prior work. Expanding the recursive form (1.8) shows the multipliers $\lambda(W)$ are equal to

$$\begin{aligned} \lambda_{j,n} &= \frac{D}{M(N+1)^{3/2}} \sum_{\ell \geq 1} (-1)^{\ell-1} \sum_{n=k_0>k_1>\cdots>k_\ell=j} \frac{\prod_{r=1}^{\ell} \left(W_{k_{r-1},k_r-1} - W_{k_{r-1},k_r}\right)}{\prod_{r=0}^{\ell} W_{k_r,k_r-1}}, \quad 1 \leq j < n \leq N \\ \lambda_{0,n} &= \frac{D}{M(N+1)^{3/2}} \sum_{\ell \geq 0} (-1)^{\ell} \sum_{n=k_0>k_1>\cdots>k_\ell\geq 1} \frac{\prod_{r=1}^{\ell} \left(W_{k_{r-1},k_r-1} - W_{k_{r-1},k_r}\right)}{\prod_{r=0}^{\ell} W_{k_r,k_r-1}} - \frac{1}{N+1}. \end{aligned} \tag{1.10}$$

The nonnegativity of these values gives the necessary polynomial inequalities in W, mirroring the H-certificates of [13]. The H-invariances of [13] provided further necessary equalities for optimal methods that fix the performance of each optimal method on the hard instance of [7]. Similarly, we find W-invariances for subgradient methods that fix the last row of any optimal method. Namely, from the equalities for $\lambda(W) \in \mathcal{P}^N$, one can derive[1] the equalities that

$$W_{N,j} = \frac{D}{M\sqrt{N+1}} \frac{N-j}{N+1}, \qquad 0 \leq j \leq N-1. \tag{1.11}$$

Following [13], Yoon and Grimmer [12] developed a composition theory, which provided a combinatorial handle on extremal optimal fixed-point methods. Future work may be able to derive similar subgradient method theory by studying $\mathcal{P}^N$. (When mapped to the polyhedron $\mathcal{P}^N$, one can verify the averaged subgradient method and the momentum-type method of [2] correspond to extreme points while the optimal subgradient method without averaging of [14] corresponds to the barycenter of the vertices.)

Zamani and Glineur [14] established that no method iterating $x_n = x_{n-1} - h_{n-1}g_{n-1}$ can guarantee optimal performance at every n. That is, optimal fixed $h_0, \dots, h_{N-1}$ must depend on the horizon N. A similar calculation from Theorem 1.3 lets us widen their conclusion to fixed-step methods: from the W-invariance (1.11), the only possible anytime fixed-step method must iterate

$$x_n = x_0 - \sum_{i=0}^{n-1} \frac{D}{M\sqrt{n+1}} \frac{n-i}{n+1} g_i \tag{1.12}$$

for $n = 1, 2, \dots$. Taking the associated W matrix with $W_{n,i} = \frac{D}{M\sqrt{n+1}}\frac{n-i}{n+1}$ up to $N = 5$, $\lambda(W) \notin \mathcal{P}^5$. This can be verified by noting that for such methods, the induced proof $\lambda(W)$ has

$$\lambda_{0,2} = \frac{3\sqrt{3} - 2\sqrt{2} - \sqrt{N+1}}{(N+1)^{3/2}},$$

which is negative at $N = 5$. This boundary is sharp as the matrix up to $N = 4$ has $\lambda(W) \in \mathcal{P}^4$.

Corollary 1.5. *For $N \geq 5$, no method W has iterates $x_0, \dots, x_N$ each optimal at the nth iteration.*

2 Proof of Sufficiency for Theorem 1.3

Below we prove that each method defined by $W(\lambda)$ for any $\lambda \in \mathcal{P}^N$ is optimal. Note that our proofs take the same form as those generated by the constructive approach of [2]. Up to changes in notation, each considered λ is part of a dual PEP solution for their associated Greedy First-Order Method. From this, each $W = W(\lambda)$ of our considered form could be constructively generated by [2, Corollary 1]. For simplicity, we have given a self-contained proof of their optimality.

Consider any $\lambda \in \mathcal{P}^N$ and let $W = W(\lambda)$, which is well-defined as $\Lambda_n = \frac{1}{N+1} + \sum_{i=0}^{n-1} \lambda_{i,n} > 0$ by the nonnegativity of λ. Fix the values for $\sigma, \lambda_{\star,i}, \mu_i$ as in (1.6) and define $Z = \sigma w w^\top$ with $w = \left[1, -\frac{D}{M\sqrt{N+1}}, \dots, -\frac{D}{M\sqrt{N+1}}\right]^\top$. We claim the core PEP identity holds

$$\sigma\mathcal{D} + \sum_{i,j \in \mathcal{I}_N^\star} \lambda_{i,j}\mathcal{C}_{i,j} + \sum_{i=0}^{N} \mu_i \mathcal{M}_i + \langle Z, G\rangle = \frac{MD}{\sqrt{N+1}} - (f_N - f_\star). \tag{2.1}$$

[1]Notice that (1.9) ensures that $\Lambda_i = \sum_{n=i+1}^{N} \lambda_{i,n}$ for $i < N$ and $\Lambda_N = 1$. Then summing (1.7) over $n = j+1, \dots, N$,

$$\sum_{n=j+1}^{N} \Lambda_n W_{n,j} = (N-j)D/[M(N+1)^{3/2}] + \sum_{n=j+1}^{N}\sum_{i=j+1}^{n-1} \lambda_{i,n} W_{i,j}.$$

Subtracting the common $\sum_{n=j+1}^{N-1} \Lambda_n W_{n,j}$ terms from both sides and noting $\Lambda_N = 1$ gives the claim.

Once proven, the fact that the method with stepsizes W attains the optimal $\frac{MD}{\sqrt{N+1}}$ worst-case rate is immediate by the nonnegativity of the left-hand-side.

Observe that the identity (2.1) is affine in F and G (once one substitutes the equalities $g_\star = 0$ and $x_n = x_0 - \sum_{i=0}^{n-1} W_{n,i} g_i$). Hence, we can verify the identity by verifying that each associated coefficient matches between the two sides. The constant terms agree as their correspondence amounts to

$$\sigma D^2 + \sum_{i=0}^{N} \mu_i M^2 = \frac{MD}{\sqrt{N+1}} \tag{2.2}$$

which is immediate from the values of σ and μ_i in (1.6).

Next, for each j, we verify that the coefficients of each f_j on the left-hand-side reduce to the coefficients on the right-hand-side. Namely, only the constraints $\mathcal{C}_{i,j}$ depend on f terms, so we have

$$\sum_{i,j \in \mathcal{I}_N^\star} \lambda_{i,j}(f_i - f_j) = \sum_{j=0}^{N} \left(\sum_{i>j} \lambda_{j,i} - \sum_{i<j} \lambda_{i,j} - \lambda_{\star,j} \right) f_j + \sum_{j=0}^{N} \lambda_{\star,j} f_\star = -(f_N - f_\star)$$

where the first equality rearranges summations, noting $\lambda_{i,\star}$ and $\lambda_{i,j}$ with $i > j$ are zero, and the second applies the defining equalities of $\lambda \in \mathcal{P}^N$ and our choice of $\lambda_{\star,i} = \frac{1}{N+1}$.

Now, we can complete our proof by verifying the coefficients agree on each term in G. In particular, we show each term vanishes on the left-hand-side as the right-hand-side has no dependence on G. Let $Z_{\star,\star}$ and $Z_{\star,n}$ denote the coefficients on $\|x_0 - x_\star\|^2$ and $\langle x_0 - x_\star, g_n \rangle$ in $\langle Z, G \rangle$. First, note that the coefficient on $\|x_0 - x_\star\|^2$ is zero as

$$\sigma(-\|x_0 - x_\star\|^2) + Z_{\star,\star}\|x_0 - x_\star\|^2 = \left(-\frac{M}{2D\sqrt{N+1}} + \frac{M}{2D\sqrt{N+1}} \right) \|x_0 - x_\star\|^2 = 0.$$

Considering the coefficients on $\|g_n\|^2$, we simplify the left-hand-side as follows

$$\mu_n(-\|g_n\|^2) + Z_{n,n}\|g_n\|^2 = \left(-\frac{D}{2M(N+1)^{3/2}} + \frac{D}{2M(N+1)^{3/2}} \right) \|g_n\|^2 = 0.$$

Lastly, we consider the coefficients on off-diagonal components of G. For each n, consider the terms associated with $\langle g_n, x_0 - x_\star \rangle$ and $\langle g_n, g_j \rangle$ with $j < n$. On these terms, the left-hand-side reduces as

$$\begin{aligned} &\left\langle g_n, (\lambda_{\star,n} + 2Z_{\star,n})(x_0 - x_\star) + \sum_{j=0}^{n-1} \left(\sum_{i=j+1}^{n-1} \lambda_{i,n} W_{i,j} - \Lambda_n W_{n,j} + 2Z_{j,n} \right) g_j \right\rangle \\ &= \left\langle g_n, \sum_{j=0}^{n-1} \left(\sum_{i=j+1}^{n-1} \lambda_{i,n} W_{i,j} - \Lambda_n W_{n,j} + \frac{D}{M(N+1)^{3/2}} \right) g_j \right\rangle = 0 \end{aligned}$$

where the first equality applies $\lambda_{\star,n} + 2Z_{\star,n} = 0$, and the second equality notes these terms vanish as $W = W(\lambda)$ satisfying (1.7). Hence, each coefficient in the identity (2.1) agrees, proving the identity and $W \in \mathcal{W}_\star^{M,D,N}$.

3 Proof of Necessity for Theorem 1.3

Throughout this section, let $M, D > 0$ and $N \in \mathbb{N}$, and fix some $W \in \mathcal{W}_\star^{M,D,N}$. First, we recall a well-known hard problem instance for subgradient methods [4, Theorem A.1]: For any $k \in \mathbb{N}$, let $e_1, \ldots, e_k$ denote an orthonormal basis for $\mathbb{R}^k$. We define a corresponding adversarial problem instance as

$$f(x) = \max\left\{\max_{1\le i\le k}\langle x, Me_i\rangle, -\frac{MD}{\sqrt{k}}\right\}, \quad x_0 = 0, \quad g(x) = \begin{cases} Me_{i_\star} & \text{if } f(x) > -\frac{MD}{\sqrt{k}} \\ 0 & \text{otherwise,} \end{cases} \tag{3.1}$$

where $i_\star = \min\{i : f(x) = \langle x, Me_i\rangle\}$. Notably, f is convex and M-Lipschitz, minimizer $x_\star = -\frac{D}{\sqrt{k}}[1,\ldots,1]^\top$ satisfies $\|x_0 - x_\star\| = D$, and $f(x_\star) = -\frac{MD}{\sqrt{k}}$. For $k = N+1$, no N-step subgradient span method can produce a point with $f(x)$ less than zero. Hence, this instance proves the minimax optimal rate's lower bound. By considering the hard instance with $k = N$, we further find that any optimal W must have strictly positive lower triangular entries.

Lemma 3.1. $W \in \mathbb{L}^N_{++}$.

Proof. Suppose to the contrary that $W_{n,j} \le 0$ for some $j \le n-1$. In particular, let $\hat{n}$ denote the earliest iteration where such a nonpositive value $W_{\hat{n},j}$ occurs. We then derive a contradiction by showing that when applied to the hard problem instance (3.1) with $k = N$, this $\hat{n}$th iteration will cause the method W to be strictly suboptimal.

First, we consider the case when $\hat{n} < N$. Notice that for any iteration $n < \hat{n}$, since $W_{n,j} > 0$, the finite maximum defining f at x_n is inactive on its first n components. Hence, the adversarial oracle will set $g_n = Me_{n+1}$. However, at iteration $\hat{n}$, the oracle will set $g_{\hat{n}} = Me_{j+1}$ for some previous $j < \hat{n}$ with nonpositive $W_{\hat{n},j}$. As a result, $g_{\hat{n}}$ lies in the span of previously seen subgradients. Therefore, at iteration N, the span of $\{g_i\}_{i=0}^{N-1}$ contains at most $N-1 < k$ of the orthonormal directions. Consequently, $\langle e_i, x_N\rangle = 0$ for some i, causing $f(x_N) \ge 0$. In the case where $\hat{n} = N$, some j has $\langle x_N, e_{j+1}\rangle = -MW_{N,j} \ge 0$. Either way, $f(x_N) - f(x_\star) \ge 0 - (-\frac{MD}{\sqrt{N}}) > \frac{MD}{\sqrt{N+1}}$, which contradicts optimality. □

Since W has positive lower triangular entries, Lemma 1.1 applies. Thus we can fix a selection of nonnegative $\sigma, \lambda_{i,j}, \mu_i$ and positive semidefinite Z such that the certificate (1.5) holds. Since $W \in \mathcal{W}^{M,D,N}_\star$, we have that $r = p_\star(W) = \frac{MD}{\sqrt{N+1}}$.

Continuing our examination of the known hard instance with $k = N+1$, we find that structure is forced among these dual values as well via complementary slackness. As a shorthand, let (F^W, G^W) denote the function values and Gram inner products observed when W is applied to the hard instance (3.1). We denote the realized constraints by $\mathcal{D}^W := \mathcal{D}(F^W, G^W)$, $\mathcal{C}^W_{i,j} := \mathcal{C}_{i,j}(F^W, G^W)$, and $\mathcal{M}^W_i := \mathcal{M}_i(F^W, G^W)$.

Lemma 3.2. *The dual optimal multipliers $\lambda_{i,j}$ for the constraints $\mathcal{C}_{i,j} \ge 0$ satisfy $\lambda_{i,j} = 0$ for all $i > j$ and $\lambda_{i,\star} = 0$ for all $i = 0, \ldots, N$.*

Proof. Since $W \in \mathbb{L}^N_{++}$ by Lemma 3.1, the iterate x_n produced by W will always have $f(x_n)$ inactive on its first n components. Thus each iteration n will have $f^W_n = 0$ and $g^W_n = Me_{n+1}$. The following values are then forced, $x^W_0 - x^W_\star = \frac{D}{\sqrt{N+1}}[1,\ldots,1]^\top$ and $x^W_n = -\sum_{i=0}^{n-1} W_{n,i} g^W_i$. Plugging these into $\mathcal{C}_{i,j}$ shows that

$$\mathcal{C}^W_{i,j} = -\left\langle g^W_j, x^W_i\right\rangle = \begin{cases} M^2 W_{i,j} & i > j \\ 0 & \text{else} \end{cases}, \quad \mathcal{C}^W_{i,\star} = -f^W_\star = \frac{MD}{\sqrt{N+1}}, \quad \mathcal{C}^W_{\star,i} = f^W_\star - \left\langle g^W_i, x^W_\star\right\rangle = 0.$$

Since $W \in \mathbb{L}^N_{++}$, the slacks $\mathcal{C}^W_{i,j}$ for $i > j$ and $\mathcal{C}^W_{i,\star}$ are strictly positive. Since the right-hand-side of (1.5) is exactly zero on this tight instance, each nonnegative term on the left-hand-side must be zero. Hence, by complementary slackness, $\lambda_{i,j} = 0$ for $i > j$ and $\lambda_{i,\star} = 0$. □

Lemma 3.3. *The dual-optimal multiplier $Z \succeq 0$ for the constraint $G \succeq 0$ satisfies $Z = \sigma\, ww^\top$, where $\sigma \geq 0$ is the dual multiplier of $\mathcal{D}$ and $w = [1, -\frac{D}{M\sqrt{N+1}}, \dots, -\frac{D}{M\sqrt{N+1}}]^\top$.*

Proof. We continue considering the values $g_n^W = Me_{n+1}$ and $x_0^W - x_\star^W = \frac{D}{\sqrt{N+1}}[1, \dots, 1]^\top$ derived in the previous lemma. Observe that these subgradients are orthogonal and $x_0^W - x_\star^W$ lies in their span. So $\operatorname{rank}(G^W) = N + 1$ and $\ker G^W = \operatorname{span}\{w\}$. By the same complementary slackness reasoning, we must have $\left\langle Z, G^W \right\rangle = 0$. Since $Z, G^W \succeq 0$, it holds that $ZG^W = 0$, making $\operatorname{range}(G^W) \subseteq \ker Z$.

Noting $\operatorname{range}(G^W)$ has codimension one, $\operatorname{rank}(Z) \leq 1$, and in particular, $\operatorname{range}(Z) \subseteq \ker G^W = \operatorname{span}\{w\}$. Therefore, $Z = \beta ww^\top$ for some $\beta \geq 0$. Considering the identity (1.5), the coefficient on $\|x_0 - x_\star\|^2$ is independent of the constraints $\mathcal{C}_{i,j}$ and $\mathcal{M}_i$. Hence, $\sigma\mathcal{D} = \sigma(D^2 - \|x_0 - x_\star\|^2)$ is the sole source of $\|x_0 - x_\star\|^2$ besides $\langle Z, G\rangle$. Thus $-\sigma + Z_{\star,\star} = 0$ and so $Z = \sigma ww^\top$. □

By considering the core PEP identity (1.5), restated below for ease, we will find the multiplier values for W are uniquely determined and, in particular, the corresponding λ block must lie in $\mathcal{P}^N$ and equal $\lambda(W)$, completing our proof. Recall that every optimal proof certificate satisfies

$$\sigma\mathcal{D} + \sum_{i,j \in \mathcal{I}_N^\star} \lambda_{i,j}\mathcal{C}_{i,j} + \sum_{i=0}^N \mu_i \mathcal{M}_i + \langle Z, G\rangle = \frac{MD}{\sqrt{N+1}} - (f_N - f_\star).$$

By Lemma 3.2, $\lambda_{i,j} = 0$ for $i > j$ and $\lambda_{i,\star} = 0$, so the only remaining convexity terms are $\lambda_{i,j}$ for $i < j$ and $\lambda_{\star,i}$.

Substituting $x_n = x_0 - \sum_{i=0}^{n-1} W_{n,i} g_i$, and expanding $\langle g_i, x_i - x_\star\rangle = \langle g_i, x_0 - x_\star\rangle - \sum_{k=0}^{i-1} W_{i,k}\langle g_i, g_k\rangle$ in each $\mathcal{C}_{\star,i}$, the part of $\sigma\mathcal{D} + \sum_{i,j}\lambda_{i,j}\mathcal{C}_{i,j} + \sum_i \mu_i\mathcal{M}_i$ depending on G is

$$-\sigma\|x_0 - x_\star\|^2 - \sum_{i=0}^N \mu_i\|g_i\|^2 + \sum_{i=0}^N \lambda_{\star,i}\langle g_i, x_0 - x_\star\rangle + \sum_{j<n}\left(\sum_{i=j+1}^{n-1} \lambda_{i,n}W_{i,j} - \Lambda_n W_{n,j}\right)\langle g_n, g_j\rangle,$$

where $\Lambda_n = \lambda_{\star,n} + \sum_{i=0}^{n-1}\lambda_{i,n}$. Considering the form of Z from Lemma 3.3, writing $G = P^\top P$ for $P = [x_0 - x_\star \mid g_0 \mid g_1 \mid \dots \mid g_N]$, we note

$$\langle Z, G\rangle = \sigma\left\| x_0 - x_\star - \frac{D}{M\sqrt{N+1}}\sum_{i=0}^N g_i\right\|^2.$$

Adding this term, and noting each Gram coordinate's coefficient vanishes in (1.5), we conclude

$$\lambda_{\star,i} = 2\sigma\frac{D}{M\sqrt{N+1}}, \qquad \mu_i = \sigma\left(\frac{D}{M\sqrt{N+1}}\right)^2, \qquad \Lambda_n W_{n,j} = \sum_{i=j+1}^{n-1}\lambda_{i,n}W_{i,j} + 2\sigma\left(\frac{D}{M\sqrt{N+1}}\right)^2, \tag{3.2}$$

by considering the coefficients on the $\langle g_i, x_0 - x_\star\rangle$, $\|g_i\|^2$, and $\langle g_n, g_j\rangle$ respectively.

Considering the coefficients on the f_i terms, it follows that

$$\sum_{j=0}^N\left(\sum_{i>j}\lambda_{j,i} - \sum_{i<j}\lambda_{i,j} - \lambda_{\star,j}\right) f_j + \sum_{i=0}^N \lambda_{\star,i} f_\star + (f_N - f_\star) = 0.$$

Setting each coefficient to zero requires

$$\sum_{i>j}\lambda_{j,i} - \sum_{i<j}\lambda_{i,j} = \lambda_{\star,j}, \quad \forall j = 0, \dots, N-1, \qquad \sum_{i<N}\lambda_{i,N} = 1 - \lambda_{\star,N}, \qquad \sum_{i=0}^N \lambda_{\star,i} = 1 \tag{3.3}$$

by considering the f_i with $i < N$, f_N and $f_\star$ terms respectively.

Finally, comparing constant terms in the identity, $\frac{MD}{\sqrt{N+1}} = \sigma D^2 + M^2 \sum_{i=0}^N \mu_i$. From the values of $\lambda_{\star,i}$ and μ_i above, we conclude that $\sigma = \frac{M}{2D\sqrt{N+1}}$, $\lambda_{\star,i} = \frac{1}{N+1}$ and $\mu_i = \frac{D}{2M(N+1)^{3/2}}$. With these values of $\lambda_{\star,i}$, the constraints on λ in (3.3), along with nonnegativity from Lemma 1.1, collapse to the constraints defining $\mathcal{P}^N$. Therefore $\lambda \in \mathcal{P}^N$. Lastly, considering the value of σ, from (3.2), $W = W(\lambda)$ is generated from (1.7), so $W \in W(\mathcal{P}^N)$.

To prove Corollary 1.4, observe that we have established the uniqueness of the dual multipliers, fixing $\lambda_{i,j}$ for $0 \leq i < j \leq N$ and setting the rest as in (1.6). Then, since $W = W(\lambda)$ and $W(\cdot)$ is a bijection onto $\Lambda_n > 0$ by Proposition 1.2, $\lambda = \lambda(W)$.

Acknowledgments. This work was supported in part by the Air Force Office of Scientific Research. Benjamin Grimmer was additionally supported as a fellow of the Alfred P. Sloan Foundation.

References

[1] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8:231–357, 2015.

[2] Y. Drori and A. B. Taylor. Efficient first-order methods for convex minimization: a constructive approach. *Mathematical Programming*, 184:183–220, 2020.

[3] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145:451–482, 2014.

[4] Y. Drori and M. Teboulle. An optimal variant of kelley's cutting-plane method. *Mathematical Programming*, 160:321–351, 2016.

[5] M. Gösgens and B. P. G. Van Parys. Subgradient methods for nonsmooth convex functions with adversarial errors. *arXiv preprint: 2510.03072*, 2025.

[6] B. Grimmer and A. Wang. Subgame perfect methods in nonsmooth convex optimization. *arXiv preprint: 2511.13639*, 2025.

[7] J. Park and E. K. Ryu. Exact optimal accelerated complexity for fixed-point iterations. *Proceedings of the 39th International Conference on Machine Learning*, 162:17420–17457, 2022.

[8] R. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

[9] N. Z. Shor. *Minimization methods for non-differentiable functions.* Springer Series in Computational Mathematics, 1985.

[10] A. B. Taylor, J. M. Hendrickx, and F. Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27:1283–1313, 2017.

[11] A. B. Taylor, J. M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161:307–345, 2017.

[12] T. Yoon and B. Grimmer. A theory of composition and duality of extremal optimal fixed-point algorithms. *arXiv preprint: 2605.02231*, 2026.

[13] T. Yoon, E. K. Ryu, and B. Grimmer. H-invariance theory: A complete characterization of minimax optimal fixed-point algorithms. *arXiv preprint: 2511.14915*, 2025.

[14] M. Zamani and F. Glineur. Exact convergence rate of the last iterate in subgradient methods. *SIAM Journal on Optimization*, 35:2182–2201, 2025.