Implicit Primal-Dual Guarantees in Unconstrained First-Order Minimization

Benjamin Grimmer* Alex L. Wang†

Abstract

This work considers the design of first-order convex optimization algorithms and convergence proofs. In particular, we consider nonsmooth Lipschitz and smooth problems accessed through a subgradient or gradient oracle, respectively. For the general class of fixed-step first-order methods, prior work on Performance Estimation Problems (PEPs) has shown that structured, tight convergence proofs typically exist. Under mild conditions, we further show that any first-order method guaranteeing a bound on the primal objective gap $f(x_N) - f(x_\star)$ assuming only a bound on $\|x_0 - x_\star\|$ actually has a stronger guarantee on an explicit, computable primal-dual gap at the same rate. These implicit optimal dual certificates, which take the form of affine lower bounds, also provide insight into the role of auxiliary sequences in momentum methods.

1 Introduction

We consider the setting of unconstrained convex optimization via first-order methods. Namely, we consider primal problems of the form

$$f_\star = \min_{x \in \mathbb{R}^d} f(x) \tag{1}$$

for a given convex function $f\colon \mathbb{R}^d \to \mathbb{R}$, attained at some minimizer, denoted $x_\star$. A problem instance consists of a function f, minimizer $x_\star$, and initialization x_0 satisfying $\|x_0 - x_\star\| \leq D$ for some $D > 0$. The two problem classes we cover are M-Lipschitz nonsmooth convex optimization and L-smooth convex optimization.

We will consider iterative methods for solving (1) given access to either a subgradient oracle or a gradient oracle, depending on the structure assumed on f. Assume that the iterative method produces iterates $x_0, \ldots, x_N$ and let $f_i = f(x_i)$ denote the function values and $g_i \in \partial f(x_i)$ denote the subgradients (or gradients) of f at x_i.

Our primary goal is to understand the nature of convergence guarantees that iterative methods can provide. A substantial body of work in the theoretical optimization literature has developed first-order methods in these settings guaranteeing

$$f_N - f_\star \leq r. \tag{2}$$

Here r represents a convergence rate, typically decreasing to zero as N grows. In this work, we show that objective guarantees of the form (2) leave substantial structure on the table and can be strengthened "for free", i.e., without loss in the numerical rate. We say our strengthened bounds are *implicit* since they are implied without additional effort from existing work proving (2).

Specifically, consider an *arbitrary* first-order method and suppose that (2) holds. Our strengthened bounds relate the primal iterate x_N with a "minimizer estimate" $z_{N+1} \in \mathbb{R}^d$ and a dual certificate[1] (formally introduced in Section 3). The dual certificate, α, is an affine function constructed only from *observed first-order information* with the property that $\alpha(x_\star) \leq f(x_\star)$. The following two strengthened forms of the guarantee (2) hold: for some $\sigma \geq 0$,

$$\underbrace{f_N - f_\star}_{\text{Primal Objective Gap}} \leq \underbrace{f_N - \alpha(x_\star)}_{\text{Primal-Dual Gap}} + \sigma \underbrace{\|z_{N+1} - x_\star\|^2}_{\text{Minimizer Estimate}} \leq r$$

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

[1]This is a dual certificate for the given problem instance, *not* a general PEP certificate.

and

$$\underbrace{f_N - f_\star}_{\text{Primal Objective Gap}} \quad \leq \quad \underbrace{f_N - \min_{\|x-x_0\|\leq D} \alpha(x)}_{\text{Computable Primal-Dual Gap}} \quad \leq r.$$

The first strengthened bound includes an additional minimizer estimate $z_{N+1} \in \mathbb{R}^d$. As a result, on any instance where (2) is tight and $\sigma > 0$, one must have $z_{N+1} = x_\star$. Conversely, if $\sigma > 0$ and $x_\star \neq z_{N+1}$, the algorithm must strictly outperform the bound (2) by an amount growing quadratically in their distance. The second strengthened bound shows that the primal objective gap, $f_N - f_\star$, can be replaced by a computable primal-dual gap without loss in the numerical rate.

The constructions for α, z_{N+1}, σ can be made explicit when the algorithm in question is a *fixed-step first-order method* (FSFOM) under mild nondegeneracy conditions. FSFOMs are parameterized by predetermined lower triangular weight matrices $W \in \mathbb{R}^{N\times N}$ and generate their iterates according to the rule:

$$x_n = x_0 - \sum_{i=0}^{n-1} W_{n-1,i} g_i. \tag{3}$$

The FSFOM setup models many classical methods for smooth and nonsmooth convex optimization. It includes simple one-step algorithms like gradient descent and the subgradient method as well as momentum-type algorithms like Nesterov's Fast Gradient Method [1], the Optimized Gradient Method (OGM) [2] and the subgradient SSEP method [3]. In the last decade, FSFOMs have received particular focus as the Performance Estimation Problem (PEP) framework ensures that they typically have structured convergence proof certificates, providing a convergence guarantee r as in (2). Such certificates can be numerically computed by semidefinite programming.

The explicit constructions of α, z_{N+1}, σ allow us to derive equivalent primal-dual understandings of many classic methods. These follow directly from examining existing PEP proofs, strengthening their classic guarantee. As one particularly surprising insight, the minimizer estimate z_{N+1} is exactly the auxiliary point generated by momentum methods. For OGM as an example, our theory identifies that the method can be equivalently understood as tracking a primal solution x_n and dual affine function α_n, updated by alternating averaging with descent

$$\begin{aligned} x_n &= \frac{\tau_{n-1}}{\tau_n}\left(x_{n-1} - \frac{1}{L} g_{n-1} - \frac{\tau_n - \tau_{n-1}}{L}\nabla\alpha_n\right) + \frac{\tau_n - \tau_{n-1}}{\tau_n} x_0 \\ \alpha_{n+1} &= \frac{\tau_{n-1}}{\tau_n}\alpha_n + \frac{\tau_n - \tau_{n-1}}{\tau_n}\left(f(x_n) + \langle g_n, \cdot - x_n\rangle + \frac{1}{2L}\|g_n\|^2\right), \end{aligned} \tag{4}$$

where L is the smoothness constant of f and τ_n is a parameter sequence defined in (41). In this form, OGM's guarantee can be strengthened, ensuring

$$\underbrace{f_N - f_\star}_{\text{Primal Objective Gap}} \leq \underbrace{f(x_N) - \min_{\|x-x_0\|\leq D} \alpha_{N+1}(x)}_{\text{Computable Primal-Dual Gap}} \leq \frac{LD^2}{2\tau_N} \sim \frac{LD^2}{N^2},$$

which is exactly the minimax optimal rate for primal objective gap convergence.

Outline. Sections 2 and 3 introduce the two key tools our development utilizes, the structured convergence analysis framework of performance estimation and an elementary primal-dual formulation of unconstrained minimization problems (1), respectively. Section 4 shows that improved guarantees of the above forms are necessary for any general first-order method in smooth or nonsmooth optimization. Sections 5 and 6 then specialize to the setting of fixed-step methods, where we can characterize α, z_{N+1}, σ more explicitly, for Lipschitz convex minimization and smooth convex minimization, respectively.

1.1 Related Work

This work is in line with the recent movement in the first-order optimization literature to use duality in algorithm design and analysis. A structured family of methods including the classic subgradient method and fast gradient method was analyzed in a unified fashion via a perturbed Fenchel duality approach

by Gutman and Peña [4]. A wider approach to algorithmic design from duality was proposed by Diakonikolas and Orecchia [5]. There, they provided a general framework for analyzing continuous-time dynamics for which an associated duality gap converges. From this, associated iterative methods have primal-dual convergence guarantees given by controlling or canceling discretization error terms. This approximate duality gap technique applies beyond the scope of problems considered here, providing algorithms and analyses, for example, covering composite/constrained problems, convex-concave saddle point problems, and variational inequalities.

Particular focus has been placed on the development of a dual understanding of momentum in smooth optimization. Nesterov [1]'s fast gradient method possesses an estimate sequence analysis that aggregates lower bounding affine models (with an initial quadratic term). Here our lower bounds are similarly built by aggregating affine models, e.g., (4). An accelerated algorithm based on optimally averaging quadratic lower bounds (for smooth strongly convex problems) was proposed by Drusvyatskiy, Fazel, and Roy [6]. Allen-Zhu and Orecchia [7] analyzed Nesterov's acceleration through a linear coupling approach: making primal progress through gradient descent steps and dual progress through mirror descent steps.

More recent works have taken primal-dual and game-theoretic approaches to understand momentum. Lan and Zhou [8] reformulate their main problem into a primal-dual saddle point problem and then describe momentum as extrapolated steps on the primal side. From this, they proposed new accelerated schemes using gradient extrapolations on the dual side. Wang, Abernethy, and Levy [9] directly reformulate convex optimization as a Fenchel game where no-regret strategies correspond to classical methods. Guarantees for these methods can then be extracted from associated regret bounds. Most recently, Burns and Liang [10] designed a new accelerated method able to provide computable certificates of accuracy. There, they interpret the resulting primal-dual averaging schemes through zero-sum games and Fisher markets.

These prior works either provided primal-dual approaches to design new algorithms or new understandings of classical methods. In contrast, we provide a general mechanism from which dual phenomena must arise for first-order methods. Stronger primal-dual convergence guarantees necessarily exist for arbitrary algorithms with objective gap guarantees. For any fixed-step method with an appropriate PEP proof, our theory further gives these necessary primal-dual strengthened bounds in closed form.

2 A Review of Performance Estimation Problems

The performance estimation framework, pioneered by Drori and Teboulle [11] and Taylor, Hendrickx, and Glineur [12, 13], provides a means to construct a worst-case instance $(f, x_0, x_\star)$ for a fixed-step method (defined by stepsize weights W) against a given structured problem class. The key observation is that when such a method is run, the only relevant aspects of the problem instance are the function values and (sub)gradients at the iterates x_i. We denote these by

$$x_i, \qquad f_i = f(x_i), \qquad g_i \in \partial f(x_i).$$

Further, denote the first-order information that the problem instance provides at the given minimizer by $x_\star$, $f_\star$, and $g_\star = 0$. Let $\mathcal{I} = \{\star, 0, \dots, N\}$.

Using this first-order information, a few useful nonnegative quantities can be defined. The distance bound implies

$$\mathcal{D} := D^2 - \|x_0 - x_\star\|^2 \geq 0. \tag{5}$$

Convexity of the function f, and in particular the subgradient inequality, guarantees that the following quantity is nonnegative for every pair of indexed points

$$\mathcal{C}_{i,j} := f_i - f_j - \langle g_j, x_i - x_j \rangle \geq 0 \qquad \forall i, j \in \mathcal{I}. \tag{6}$$

In the case that f is M-Lipschitz, one can additionally guarantee that

$$\mathcal{M}_i = M^2 - \|g_i\|^2 \geq 0 \qquad \forall i \in \mathcal{I}. \tag{7}$$

For L-smooth convex functions with $L > 0$ (defined as ∇f being L-Lipschitz), the convexity inequality can be strengthened to the following cocoercivity-type inequality

$$\mathcal{Q}_{i,j} := f_i - f_j - \langle g_j, x_i - x_j \rangle - \frac{1}{2L}\|g_i - g_j\|^2 \geq 0 \qquad \forall i, j \in \mathcal{I}. \tag{8}$$

The fundamental insight behind the PEP approach to analyzing FSFOMs is the fact that the nonnegativity of these quantities suffices for any analysis based only on observed values. The following interpolation theorems formalize this.

Theorem 2.1 (Theorem 3.5 of [12]). *The first-order observations $(x_i, f_i, g_i)_{i\in\mathcal{I}}$ can be interpolated by an M-Lipschitz convex function if and only if $\mathcal{C}_{i,j} \geq 0$ for all $i, j \in \mathcal{I}$ and $\mathcal{M}_i \geq 0$ for all $i \in \mathcal{I}$.*

Theorem 2.2 (Theorem 4 of [13]). *The first-order observations $(x_i, f_i, g_i)_{i\in\mathcal{I}}$ can be interpolated by an L-smooth convex function if and only if $\mathcal{Q}_{i,j} \geq 0$ for all $i, j \in \mathcal{I}$.*

Using these interpolation theorems, Taylor, Hendrickx, and Glineur showed that one can equivalently cast the problem of computing a worst-case instance as computing worst-case first-order oracle values satisfying these conditions. For example, for a FSFOM with weights W, the worst-case final objective gap over M-Lipschitz convex problems is equal to

$$\begin{aligned}
\texttt{PEP}_M(W) := &\begin{cases} \max_{f,(x_i,g_i)_{i\in\mathcal{I}}} & f(x_N) - f(x_\star) \\ \text{s.t.} & x_n = x_0 - \sum_{i=0}^{n-1} W_{n-1,i} g_i \quad \forall n = 1,\dots,N \\ & g_i \in \partial f(x_i) \qquad \forall i \in \mathcal{I} \\ & f \text{ is convex, } M\text{-Lipschitz, and minimized at } x_\star \\ & \|x_0 - x_\star\| \leq D \end{cases} \\
= &\begin{cases} \max_{(x_i,f_i,g_i)_{i\in\mathcal{I}}} & f_N - f_\star \\ \text{s.t.} & x_n = x_0 - \sum_{i=0}^{n-1} W_{n-1,i} g_i \quad \forall n = 1,\dots,N \\ & g_\star = 0 \\ & \mathcal{C}_{i,j} \geq 0 \qquad \forall i, j \in \mathcal{I} \\ & \mathcal{M}_i \geq 0 \qquad \forall i \in \mathcal{I} \\ & \mathcal{D} \geq 0. \end{cases}
\end{aligned} \tag{9}$$

This problem can be written as a semidefinite program. To see this, consider substituting the defining equalities for $g_\star$ and x_n, $n = 1,\dots,N$, in the remaining constraints. The resulting problem is linear in the function values $f = (f_\star, f_0, \dots, f_N)$ and linear in the quadratic form

$$G = P^\top P, \qquad P = \begin{bmatrix} x_0 - x_\star, g_0, \dots, g_N \end{bmatrix} \tag{10}$$

which must be positive semidefinite by definition. Provided $d \geq N + 2$, any positive semidefinite G can be factored to recover the underlying P. Hence, for d large enough, $\texttt{PEP}_M(W)$ is a semidefinite program. For $d < N + 2$, the resulting semidefinite program still provides a valid upper bound.

The dual of this semidefinite program then optimizes over upper bounds r on the worst-case value of $f_N - f_\star$. This dual selects nonnegative multipliers $\lambda_{i,j}$, γ_i, and σ for the inequality constraints and a positive semidefinite multiplier Z for the $G \succeq 0$ constraint. These multipliers are dual feasible if they satisfy the polynomial identity

$$r - (f_N - f_\star) = \sum_{i,j\in\mathcal{I}} \lambda_{i,j}\mathcal{C}_{i,j} + \sum_{i\in\mathcal{I}} \gamma_i \mathcal{M}_i + \sigma\mathcal{D} + \langle Z, G\rangle,$$

fixing $x_n = x_0 - \sum_{i=0}^{n-1} W_{n-1,i} g_i$ and $g_\star = 0$. Here in this identity, we treat the values f_i and the entries of G as free variables. In these terms, the identity is affine. Note this identity combined with the nonnegativity of the right-hand-side proves the guarantee r. If strong duality (with dual attainment) holds, then $r = \texttt{PEP}_M(W)$ and an optimal solution $(\lambda_{i,j}, \gamma_i, \sigma, Z)$ provides a structured proof of this convergence rate.

Analogous primal and dual semidefinite programs can be constructed for L-smooth convex functions: simply replace the constraints $\mathcal{C}_{i,j} \geq 0$ and $\mathcal{M}_i \geq 0$ with the constraints $\mathcal{Q}_{i,j} \geq 0$.

3 The Primal-Dual View of Unconstrained Minimization

Here, we develop convergence theory in terms of stronger primal-dual guarantees than the primal-only objective guarantees natural in the PEP framework. Our development follows the primal-dual

minimax perspective taken by [14] in the analysis of certain subgradient methods. At first glance, an unconstrained convex minimization problem of the form (1) does not possess a natural primal-dual perspective. However, for full domain convex functions (which must be closed convex and proper), one can describe f as the supremum over its affine minorants [15, Theorem 12.1]. Letting A_f denote the set of all affine minorants

$$A_f = \{\alpha \colon \mathbb{R}^d \to \mathbb{R} \mid \alpha(x) = \langle a, x\rangle + b,\ \alpha \leq f\},$$

one has that $f(x) = \sup_{\alpha \in A_f} \alpha(x)$. Hence, our basic minimization problem can be written as

$$f_\star = \min_x f(x) = \min_x \sup_{\alpha \in A_f} \alpha(x).$$

Given any $\alpha \in A_f$, a lower bound on $f_\star$ is immediately provided by $\alpha(x_\star)$. A lower bound independent of $x_\star$ can be produced by considering the dual maximin problem of

$$f_\star = \min_x \sup_{\alpha \in A_f} \alpha(x) \geq \sup_{\alpha \in A_f} \min_x \alpha(x) = \sup_{\alpha \in A_f} \begin{cases} \alpha(x_0) & \text{if } \nabla\alpha = 0 \\ -\infty & \text{otherwise} \end{cases} =: d_\star$$

where the second equality observes that the minimum of any nonconstant affine function is $-\infty$. Note that since α is affine, we omit the evaluation point from the gradient $\nabla\alpha$ in our notation.

From the above calculation, the dual problem seeks the largest constant function (i.e., $\nabla\alpha = 0$) underneath f. The dual maximum is attained by $\alpha_\star(x) = \langle 0, x\rangle + f_\star$. So strong duality holds, with $f_\star = d_\star$. The above dual is too simple to provide a meaningful source of lower bounding certificates, however, as the bound provided by a dual certificate α is $-\infty$ for all but constant functions.

This primal-dual lens can be enriched in two ways. First, when a bound D on the primal distance to optimal is known, one can add the primal constraint $\|x - x_0\| \leq D$. Second, we may relax the dual constraint $\alpha \in A_f$ to only requiring that $\alpha(x_\star) \leq f(x_\star)$ (i.e., α lower bounds f at its minimizer). The former will be useful in addressing M-Lipschitz settings, while the latter will find use in L-smooth settings. The following two subsections develop these models.

3.1 Stronger Certificates Given a Primal Distance Bound D

Given some $D > 0$ such that the problems of interest have a minimizer $x_\star$ with bounded distance $\|x_0 - x_\star\| \leq D$, we can strengthen our dual. Without loss, we can add the primal minimization constraint that x lies in the ball $\|x - x_0\| \leq D$. Hence, we can write a primal-dual problem pair for a given problem instance with bounded distance as

$$f_\star = \min_{\|x - x_0\| \leq D} \sup_{\alpha \in A_f} \alpha(x) \geq \sup_{\alpha \in A_f} \min_{\|x - x_0\| \leq D} \alpha(x) =: d_\star. \tag{11}$$

Since α is affine, we can solve the inner minimization over x in closed form. It is attained at $x = x_0 - D\nabla\alpha/\|\nabla\alpha\|$ for any given nonconstant α, making the dual problem $\sup_{\alpha \in A_f} \alpha(x_0) - D\|\nabla\alpha\|$. Again, strong duality holds here, attained by the constant function $\alpha_\star(x) = \langle 0, x\rangle + f_\star$.

From (11), we have a structured primal-dual pair of outputs that algorithms can provide to certify solution quality. Given a dual solution $\alpha \in A_f$, one can lower bound $f_\star$ by the following two bounds

$$f_\star \geq \alpha(x_\star) \geq \min_{\|x - x_0\| \leq D} \alpha(x) \tag{12}$$

where the first is tighter but the second is independent of $x_\star$. Hence, given a primal solution x and a dual solution α, the primal objective gap at x is bounded by the primal-dual gaps of

$$f(x) - f(x_\star) \leq \underbrace{f(x) - \alpha(x_\star)}_{\text{Primal-Dual Gap}} \leq \underbrace{f(x) - \min_{\|x - x_0\| \leq D} \alpha(x)}_{\text{Computable Primal-Dual Gap}}. \tag{13}$$

Provided a bound D is known, this final quantity is computable and independent of $x_\star$.

Constructing elements of A_f is straightforward for first-order methods. Each first-order oracle query directly provides an affine minorant by the definition of a $g_i \in \partial f(x_i)$. Namely, one must have

$$f(x) \geq f(x_i) + \langle g_i, x - x_i\rangle \qquad \forall x \in \mathbb{R}^d.$$

So at every iteration, an affine minorant $\alpha(x) = f(x_i) + \langle g_i, x - x_i\rangle$ is observed. Moreover, noting that the set A_f is convex, any convex combination of subgradient lower bounds can provide a candidate dual solution α.

3.2 Certificates via "Minorants at the Minimizer" Given Smoothness

Smoothness enables an even stronger class of dual models to be constructed for our unconstrained minimization. For L-smooth convex functions, each gradient provides a stronger lower bound at the minimizer than the simple subgradient inequality. Namely, the cocoercive-type inequality (i.e., $\mathcal{Q}_{\star,i} \geq 0$ in (8)) ensures that every gradient $g_i = \nabla f(x_i)$ has

$$f(x_\star) \geq f(x_i) + \langle g_i, x_\star - x_i \rangle + \frac{1}{2L}\|g_i\|^2.$$

Note this inequality only holds at $x_\star$, not all x. So at every iteration, gradient methods obtain an affine "minorant at the minimizer" given by $\alpha(x) = f(x_i) + \frac{1}{2L}\|g_i\|^2 + \langle g_i, x - x_i \rangle$.

(This extra term in the gradient norm squared was recently observed by [16] to be key to strengthening momentum methods to attain the exactly optimal convergence rate of OGM.)

For our primal-dual reasoning above, all that is needed for α is that it is no greater than f at $x_\star$ for (12) to hold. Hence, for L-smooth convex functions, one can consider any affine "minorant at the minimizer", denoted

$$A_f^\star = \{\alpha \colon \mathbb{R}^d \to \mathbb{R} \mid \alpha(x) = \langle a, x \rangle + b,\ \alpha(x_\star) \leq f(x_\star)\}.$$

We refer to such $\alpha \in A_f^\star$ as an affine "minorant at the minimizer", given that it is only guaranteed to lie below f at $x_\star$.

Note then that $\sup_{\alpha \in A_f^\star} \alpha(x)$ is $+\infty$ for all inputs other than $x_\star$, where it takes value $f_\star$. Although this supremum is typically different from f, minimizing it still returns $f_\star$. So we have strong duality between the primal-dual pair $\min_{\|x - x_0\| \leq D} \sup_{\alpha \in A_f^\star} \alpha(x) = \sup_{\alpha \in A_f^\star} \min_{\|x - x_0\| \leq D} \alpha(x)$.

Again, observing that $A_f^\star$ is convex, one can construct dual solutions from any first-order method by aggregating the affine "minorants at the minimizer" generated by observed gradients at runtime $\alpha(x) = f(x_i) + \frac{1}{2L}\|g_i\|^2 + \langle g_i, x - x_i \rangle$.

4 Existence of Dual Certificates and Minimizer Estimates

In this section, we consider any first-order method (not necessarily fixed-step, deterministic, or otherwise structured). Suppose the method has been run on some problem instance, observing first-order information $(x_i, f_i, g_i)_{i=0}^N$ and reporting a primal objective guarantee against all instances consistent with the given first-order observations of

$$f_N - f_\star \leq r. \tag{14}$$

Note that just as the method is left general, how one derives this convergence guarantee is left general. It need not come from a structured PEP proof of the form discussed in Section 2. Any procedure resulting in a uniform guarantee against every problem instance consistent with the fixed, observed values $(x_i, f_i, g_i)_{i=0}^N$ is fine.

Here we will show that, in both the nonsmooth and smooth settings, one can always deduce stronger primal-dual guarantees from this observed information. The key object in both settings will be to consider all possible placements of $x_\star$ consistent with the first-order information. Optimizing over $x_\star$'s placement will yield dual insights.

4.1 Dual Certificates in Nonsmooth Optimization

First we consider convex (potentially nonsmooth) minimization given a distance bound $D > 0$. Here we do not make any Lipschitz continuity assumptions, although these will be required in Section 5 when we consider fixed-step methods. Suppose an algorithm observed first-order information $(x_i, f_i, g_i)_{i=0}^N$. We know that whatever values $(x_\star, f_\star)$ take, they must satisfy $\mathcal{C}_{i,\star} \geq 0$ and $\mathcal{C}_{\star,i} \geq 0$ for each $i = 0, \ldots, N$. From this, the worst possible final objective gap is given by

$$r_\star = \begin{cases} \max_{x_\star, f_\star} & f_N - f_\star \\ \text{s.t.} & D^2 - \|x_0 - x_\star\|^2 \geq 0 \\ & f_\star - f_i - \langle g_i, x_\star - x_i \rangle \geq 0 \qquad \forall i = 0, \ldots, N \\ & f_i - f_\star \geq 0 \qquad \forall i = 0, \ldots, N. \end{cases} \tag{15}$$

Note this is not a relaxation, but exact, as given maximizers $x_\star, f_\star$, one can set $g_\star = 0$ and apply the Interpolation Theorem 2.1 to construct an instance realizing this worst-case gap. So $r_\star \leq r$.

Let $\ell_i(x)$ denote $f_i + \langle g_i, x - x_i\rangle$, the affine minorant of any interpolating f provided by the subgradients g_i. Note $\mathcal{C}_{\star,i} \geq 0$ ensures that $f_\star \geq \ell_i(x_\star)$ for each $i = 0, \dots, N$. Taking the maximum over indices i, we have that f must lie above the cutting-plane model constructed from each of these affine minorants. In particular, for any placement of $x_\star$, we have

$$f(x_\star) \geq \max_{i=0,\dots,N} \ell_i(x_\star).$$

Below we find that the maximum final objective gap (i.e., smallest possible final value of $f(x_\star)$), consistent with the observed first-order information, is given by minimizing over all placements of $x_\star$ in this cutting plane model. By resolving the associated minimax problem, we arrive at guarantees of our strengthened primal-dual forms.

Theorem 4.1. *Consider any observed first-order information $(x_i, f_i, g_i)_{i=0}^N$ consistent with some convex function possessing a minimizer with $\|x_0 - x_\star\| \leq D$. Then there exist nonnegative weights a_i, summing to one, such that $\alpha(x) = \sum_{i=0}^N a_i(f_i + \langle g_i, x - x_i\rangle) \in A_f$ has*

$$\begin{aligned}
\underbrace{f_N - f_\star}_{\textit{Primal Objective Gap}} &\leq \underbrace{f_N - f_\star}_{\textit{Primal Objective Gap}} + \underbrace{\sigma\|z_{N+1} - x_\star\|^2}_{\textit{Minimizer Estimate}} && (16)\\
&\leq \underbrace{f_N - \alpha(x_\star)}_{\textit{Primal-Dual Gap}} + \underbrace{\sigma\|z_{N+1} - x_\star\|^2}_{\textit{Minimizer Estimate}} && (17)\\
&\leq \underbrace{f_N - \min_{\|x - x_0\| \leq D} \alpha(x)}_{\textit{Computable Primal-Dual Gap}} && (18)\\
&\leq r_\star && (19)
\end{aligned}$$

where if $\nabla\alpha \neq 0$, $\sigma = \frac{\|\nabla\alpha\|}{2D}$, $z_{N+1} = x_0 - \frac{1}{2\sigma}\nabla\alpha$, otherwise $\sigma = 0$ and z_{N+1} can be set arbitrarily.

Proof. First, notice that since $(x_i, f_i, g_i)_{i=0}^N$ is consistent with some function $\hat{f}$ minimized at $\hat{x}_\star$, the problem (15) is feasible—consider setting $f_\star = \hat{f}(\hat{x}_\star)$ and $x_\star = \hat{x}_\star$. Moreover, the final constraints in (15), corresponding to $\mathcal{C}_{i,\star} \geq 0$, can be omitted without loss. This follows since any global maximizer with these omitted has $f_\star \leq \hat{f}(\hat{x}_\star)$, so $f_i - f_\star = (f_i - \hat{f}(\hat{x}_\star)) + (\hat{f}(\hat{x}_\star) - f_\star) \geq 0$ and so is feasible for the full problem.

Thus we can simplify the definition of $r_\star$ as follows

$$\begin{aligned}
r_\star &= \max_{\|x_\star - x_0\| \leq D} \min_{i=0,\dots,N} f_N - \ell_i(x_\star)\\
&= \max_{\|x_\star - x_0\| \leq D} \min_{a \in \Delta_{N+1}} f_N - \sum_{i=0}^N a_i \ell_i(x_\star)\\
&= \min_{a \in \Delta_{N+1}} \max_{\|x_\star - x_0\| \leq D} f_N - \sum_{i=0}^N a_i \ell_i(x_\star)\\
&= \min_{a \in \Delta_{N+1}} f_N - \alpha_a(x_0) + D\|\nabla\alpha_a\|
\end{aligned}$$

where the first equality sets $f_\star$ to its minimum allowable value, the second replaces the finite minimum with minimization over the simplex $\Delta_{N+1} = \{a \in \mathbb{R}^{N+1} \mid \sum_{i=0}^N a_i = 1, a \geq 0\}$, the third line notes that this final problem is a compact, convex-concave minimax problem, so we can exchange the order of minimization and maximization, and the final line computes the minimum of this affine function $\alpha_a = \sum_{i=0}^N a_i \ell_i$ over a ball.

Note that since the simplex is compact, some minimizing $a \in \Delta_{N+1}$ must exist. If this optimal a has $\nabla\alpha_a = 0$, setting $\sigma = 0$, the claimed sequence of inequalities all hold immediately. Now suppose

$\nabla\alpha_a \neq 0$. Then the claimed inequalities hold with $\sigma = \frac{\|\nabla\alpha_a\|}{2D}$ and $z_{N+1} = x_0 - \frac{1}{2\sigma}\nabla\alpha_a$ as

$$\begin{aligned}
r_\star &= f_N - \alpha_a(x_0) + \|\nabla\alpha_a\| D \\
&= f_N - \alpha_a(x_0) + \sigma D^2 + \frac{1}{4\sigma}\|\nabla\alpha_a\|^2 \\
&= f_N - \alpha_a(x_\star) + \sigma(D^2 - \|x_0 - x_\star\|^2) + \sigma\|z_{N+1} - x_\star\|^2 \\
&\geq f_N - \alpha_a(x_\star) + \sigma\|z_{N+1} - x_\star\|^2 \\
&\geq f_N - f_\star + \sigma\|z_{N+1} - x_\star\|^2.
\end{aligned}$$

□

4.2 Dual Certificates in Smooth Convex Optimization

We now repeat the above development for the case where the observed first-order information comes from an L-smooth convex function.

As before, we consider all possible placements of $(x_\star, f_\star)$ consistent with the available first-order information. Recalling the necessary $\mathcal{Q}$ inequalities from (8) (which are sufficient for interpolation), any placement of the minimizer and minimum value $(x_\star, f_\star)$ must satisfy, for each $i = 0, \dots, N$,

$$\begin{aligned}
\mathcal{Q}_{\star,i} &= f_\star - f_i - \langle g_i, x_\star - x_i\rangle - \frac{1}{2L}\|g_i\|^2 \geq 0, \\
\mathcal{Q}_{i,\star} &= f_i - f_\star - \frac{1}{2L}\|g_i\|^2 \geq 0,
\end{aligned}$$

together with the distance bound $\mathcal{D} = D^2 - \|x_0 - x_\star\|^2 \geq 0$. So the largest final objective gap compatible with the observed first-order information is

$$r_\star = \begin{cases}
\max_{x_\star, f_\star} & f_N - f_\star \\
\text{s.t.} & f_\star - f_i - \langle g_i, x_\star - x_i\rangle - \frac{1}{2L}\|g_i\|^2 \geq 0 \qquad \forall i = 0, \dots, N, \\
& f_i - f_\star - \frac{1}{2L}\|g_i\|^2 \geq 0 \qquad \forall i = 0, \dots, N, \\
& D^2 - \|x_0 - x_\star\|^2 \geq 0.
\end{cases} \tag{20}$$

The direct interpretation of (20) is again that $f_N - f_\star \leq r_\star$, so any reported primal guarantee $f_N - f_\star \leq r$ must have $r \geq r_\star$.

Comparing (20) with (15), we see that the only difference is replacing f_i with $f_i + \frac{1}{2L}\|g_i\|^2$ in the first inequalities and replacing f_i with $f_i - \frac{1}{2L}\|g_i\|^2$ in the second inequalities. Again, considering a function $\hat{f}$ consistent with the given information, one can conclude this problem is feasible and that the second collection of inequalities corresponding to $\mathcal{Q}_{i,\star} \geq 0$ can be omitted.

Thus, by taking the same minimax view as before, stronger guarantees can again be extracted. We omit its proof, which follows verbatim after making the above substitutions and considering affine minorants at the minimizer $m_i(x) = f_i + \frac{1}{2L}\|g_i\|^2 + \langle g_i, x - x_i\rangle$ instead of ℓ_i.

Theorem 4.2. *Consider any observed first-order information $(x_i, f_i, g_i)_{i=0}^N$ consistent with some L-smooth convex function possessing a minimizer with $\|x_0 - x_\star\| \leq D$. Then there exist nonnegative weights a_i, summing to one, such that $\alpha(x) = \sum_{i=0}^N a_i(f_i + \frac{1}{2L}\|g_i\|^2 + \langle g_i, x - x_i\rangle) \in A_f^\star$ has*

$$\begin{aligned}
\underbrace{f_N - f_\star}_{\textit{Primal Objective Gap}} &\leq \underbrace{f_N - f_\star}_{\textit{Primal Objective Gap}} + \underbrace{\sigma\|z_{N+1} - x_\star\|^2}_{\textit{Minimizer Estimate}} && (21) \\
&\leq \underbrace{f_N - \alpha(x_\star)}_{\textit{Primal-Dual Gap}} + \underbrace{\sigma\|z_{N+1} - x_\star\|^2}_{\textit{Minimizer Estimate}} && (22) \\
&\leq \underbrace{f_N - \min_{\|x - x_0\| \leq D} \alpha(x)}_{\textit{Computable Primal-Dual Gap}} && (23) \\
&\leq r_\star && (24)
\end{aligned}$$

where if $\nabla\alpha \neq 0$, $\sigma = \frac{\|\nabla\alpha\|}{2D}$, $z_{N+1} = x_0 - \frac{1}{2\sigma}\nabla\alpha$, otherwise $\sigma = 0$ and z_{N+1} can be set arbitrarily.

So in the smooth convex setting, the same dynamic placement argument again yields a minimizer estimate and a dual certificate directly from the observed first-order information. The main distinction from the nonsmooth case is that smoothness allows an extra term in the minorant at the minimizer.

We summarize the results of this section: Consider an arbitrary procedure for producing iterates $x_0, x_1, \ldots, x_N$. After the fact, once one has observed the first-order information from running the algorithm, one can deduce a minimizer estimate z_{N+1} and a dual model α by solving an instance-dependent minimax problem, dual to (15) or (20). These candidate minimizers and dual certificates directly establish stronger performance guarantees than (14) at no cost, i.e., without harming worst-case guarantees. In the remainder of this paper, we show that for nondegenerate FSFOMs, an *instance-independent* construction of α and z_{N+1} can be determined *a priori* from the algorithm's PEP.

5 Dual Constructions for Fixed-Step Subgradient Methods

This section develops general conditions under which subgradient-type methods with a primal objective gap guarantee have the same guarantee on a primal-dual gap, often being strictly stronger. The section then concludes with three example subgradient methods from the literature, now with strengthened primal-dual theory.

5.1 Primal-Dual Guarantees for Lipschitz Convex Minimization

Consider any fixed-step subgradient method, defined by weights W. Recall its worst-case final objective gap is given by $\texttt{PEP}_M(W)$, defined in (9). Further, suppose $d \geq N+2$ and that strong duality holds for its associated semidefinite programming problem. Establishing the strict feasibility of these PEPs suffices to establish strong duality and dual attainment. The lower triangular entries of W being nonzero is sufficient to this end [17, Lemma 1.1]. For example, Theorem 6 of [13] establishes the analogous result for smooth convex settings.

The dual problem then establishes bounds $\texttt{PEP}_M(W) \leq r$ by providing a corresponding dual feasible solution: nonnegative multipliers $\lambda_{i,j}, \gamma_i, \sigma$ (for $\mathcal{C}_{i,j}, \mathcal{M}_i, \mathcal{D}$, respectively) and positive semidefinite $Z \in \mathbb{S}^{N+2}$. Given strong duality and dual attainment, there must exist multipliers satisfying the following identity:

$$\texttt{PEP}_M(W) - (f_N - f_\star) = \sum_{i,j \in \mathcal{I}} \lambda_{i,j} \mathcal{C}_{i,j} + \sum_{i \in \mathcal{I}} \gamma_i \mathcal{M}_i + \sigma \mathcal{D} + \langle Z, G \rangle, \tag{25}$$

where $x_n = x_0 - \sum_{i=0}^{n-1} W_{n-1,i} g_i$ and $g_\star = 0$, and the remaining data (f_i and each inner product in G) are treated as free variables. These multipliers establish that the right-hand side is nonnegative for any first-order information (x_i, f_i, g_i) encountered. Hence, the left-hand side is nonnegative, i.e., they prove a tight convergence rate of $f_N - f_\star \leq \texttt{PEP}_M(W)$.

The tightness of this convergence rate is easily observed. Let x_i, f_i, g_i be first-order data attaining the optimal value of the primal PEP (9). By the Interpolation Theorem 2.1, there exists an M-Lipschitz convex problem matching these first-order values. Then, running the subgradient method defined by W on this particular interpolated instance will observe first-order values

$$x_i^W = x_i, \qquad f_i^W = f_i, \qquad g_i^W = g_i. \tag{26}$$

So the final objective gap realized by W will be $f_N^W - f_\star^W = \texttt{PEP}_M(W)$.

We next apply complementary slackness to simplify a given primal-dual PEP pair (a tight instance and a structured proof). If the realized primal problem data has $\mathcal{C}_{i,j}^W > 0$, then the corresponding dual multiplier $\lambda_{i,j}$ must be zero. In particular, suppose that no iterate is prematurely optimal *on all worst-case functions*. Equivalently, suppose that for each $i = 0, \ldots, N$, there exists some worst-case function f such that $f_i^W > f_\star^W$. Then, since $g_\star = 0$, we have that $\mathcal{C}_{i,\star} > 0$ and so $\lambda_{i,\star} = 0$. Also note that since $\mathcal{C}_{i,i} = 0$ identically, one can set $\lambda_{i,i} = 0$ without loss. So viewing λ as a block matrix

$$\lambda = \begin{bmatrix} \lambda_{\star,\star} & \lambda_{\star,\cdot} \\ \lambda_{\cdot,\star} & \hat{\lambda} \end{bmatrix}, \tag{27}$$

the first column of λ can be taken to be zero. Hence, it suffices to consider only the remainder of the first row $\lambda_{\star,\cdot}$ and the main block $\hat{\lambda}$. We abbreviate this structure with a definition[2].

Definition 5.1. *A subgradient method defined by fixed-step weights W is* nondegenerate *if the semidefinite programming formulation of its PEP has (i) strong duality and primal/dual attainment for the associated PEP and (ii) for each $i = 0, \ldots, N$, there exists an optimal PEP solution with $f_i^W > f_\star^W$.*

Similar to the above complementary slackness deductions, any optimal proof must have $\gamma_\star = 0$ since $\mathcal{M}_\star = M^2 - \|0\|^2 > 0$. The following lemma uses these observations to provide a sufficient form of optimal proof certificates, independent of quantities related to the minimizer $x_\star$.

Lemma 5.2. *Consider any $N \geq 1$, $M, D > 0$, and any nondegenerate subgradient method defined by fixed-step weights W. Then there exists a nonnegative vector $a \in \mathbb{R}^{N+1}$, $\sigma \geq 0$, and a positive semidefinite matrix $\hat{Z} \in \mathbb{S}^{N+1}$ such that $\sum_{i=0}^N a_i = 1$, the following matrix is positive semidefinite*

$$\begin{bmatrix} \sigma & -\frac{1}{2}a^\top \\ -\frac{1}{2}a & \hat{Z} \end{bmatrix}, \tag{28}$$

and

$$\mathtt{PEP}_M(W) - \left(f_N - \sum_{i=0}^N a_i (f_i + \langle g_i, x_0 - x_i \rangle) \right) \geq \sigma D^2 + \langle \hat{Z}, \hat{G} \rangle, \tag{29}$$

where $\hat{G}$ denotes the principal submatrix of G corresponding to subgradient-subgradient inner products.

Proof. Let $\lambda, \gamma, \sigma, Z$ be optimal dual PEP certificates satisfying (25). By the complementary slackness discussion above, we may assume $\lambda_{i,\star} = 0$, $\lambda_{\star,\star} = 0$ and $\gamma_\star = 0$. Decompose λ and Z into blocks as:

$$\lambda = \begin{bmatrix} 0 & a^\top \\ 0 & \hat{\lambda} \end{bmatrix} \qquad \text{and} \qquad Z = \begin{bmatrix} c & b^\top \\ b & \hat{Z} \end{bmatrix}.$$

Then, the inner product $\langle Z, G \rangle$ can be expanded in this block form to be

$$\langle Z, G \rangle = c\|x_0 - x_\star\|^2 + 2\sum_{i=0}^N b_i \langle g_i, x_0 - x_\star \rangle + \langle \hat{Z}, \hat{G} \rangle.$$

Likewise, for each $i = 0, \ldots, N$, we can expand

$$\mathcal{C}_{\star,i} = f_\star - f_i - \langle g_i, x_\star - x_i \rangle = f_\star - f_i - \langle g_i, x_0 - x_i \rangle + \langle g_i, x_0 - x_\star \rangle.$$

Substituting these expressions into (25) and recalling we have removed the first column of λ gives

$$\begin{aligned} \mathtt{PEP}_M(W) - (f_N - f_\star) = &\sum_{i=0}^N a_i (f_\star - f_i - \langle g_i, x_0 - x_i \rangle + \langle g_i, x_0 - x_\star \rangle) \\ &+ \sum_{i,j=0}^N \hat{\lambda}_{i,j} \mathcal{C}_{i,j} + \sum_{i=0}^N \gamma_i \mathcal{M}_i + \sigma(D^2 - \|x_0 - x_\star\|^2) \\ &+ c\|x_0 - x_\star\|^2 + 2\sum_{i=0}^N b_i \langle g_i, x_0 - x_\star \rangle + \langle \hat{Z}, \hat{G} \rangle. \end{aligned}$$

We now compare coefficients of the free variables appearing on both sides. First, the coefficient of $f_\star$ on the left-hand side is 1, while on the right-hand side it is $\sum_{i=0}^N a_i$. Hence,

$$\sum_{i=0}^N a_i = 1.$$

[2] Although all classical subgradient methods studied, to our knowledge, are nondegenerate, simple degenerate methods exist. For instance when $M = D = 1$ and $N = 2$, $W = \begin{bmatrix} 1 & 0 \\ 2 & 3 \end{bmatrix}$ is degenerate as solving the associated PEP SDP, one has $\lambda_{1,\star} = 1/4 > 0$.

Second, since the left-hand side is independent of $\langle g_i, x_0 - x_\star \rangle$, the coefficient of each such term on the right-hand side must vanish, ensuring that

$$a_i + 2b_i = 0 \qquad \forall i = 0, \dots, N.$$

So we conclude that $b = -a/2$. Third, noting that the left-hand side is independent of $\|x_0 - x_\star\|^2$, its coefficient on the right-hand side must vanish. Therefore,

$$-\sigma + c = 0.$$

So we conclude that $c = \sigma$.

Substituting these three identities back into the equality above, and canceling the $f_\star$ terms using $\sum_{i=0}^N a_i = 1$, yields

$$\mathtt{PEP}_M(W) - \left(f_N - \sum_{i=0}^N a_i (f_i + \langle g_i, x_0 - x_i \rangle) \right) = \sum_{i,j=0}^N \hat{\lambda}_{i,j} \mathcal{C}_{i,j} + \sum_{i=0}^N \gamma_i \mathcal{M}_i + \sigma D^2 + \langle \hat{Z}, \hat{G} \rangle.$$

Dropping the first two nonnegative summations yields (29). Finally, since $Z \succeq 0$, we have positive semidefiniteness of $\begin{bmatrix} \sigma & -\frac{1}{2}a^\top \\ -\frac{1}{2}a & \hat{Z} \end{bmatrix}$. □

Theorem 5.1. *Consider any $N \geq 1$, $M, D > 0$, and any nondegenerate subgradient method defined by fixed-step weights W. Then there exists a nonnegative vector $a \in \mathbb{R}^{N+1}$ summing to one and a constant $\sigma > 0$, such that the method applied to any M-Lipschitz convex problem with $\|x_0 - x_\star\| \leq D$ has*

$$\underbrace{f_N - f_\star}_{\textit{Primal Objective Gap}} \leq \underbrace{f_N - \alpha(x_\star)}_{\textit{Primal-Dual Gap}} + \underbrace{\sigma \|z_{N+1} - x_\star\|^2}_{\textit{Minimizer Estimate}} \leq \mathtt{PEP}_M(W) \tag{30}$$

and

$$\underbrace{f_N - f_\star}_{\textit{Primal Objective Gap}} \leq \underbrace{f_N - \min_{\|x - x_0\| \leq D} \alpha(x)}_{\textit{Computable Primal-Dual Gap}} \leq \mathtt{PEP}_M(W) \tag{31}$$

where $\alpha(x) = \sum_{i=0}^N a_i (f(x_i) + \langle g_i, x - x_i \rangle) \in A_f \subset A_f^\star$, and $z_{N+1} = x_0 - \frac{1}{2\sigma} \nabla \alpha$.

Proof. Let $a, \sigma, \hat{Z}$ be as provided by Lemma 5.2, satisfying (28) and (29). Since each function $x \mapsto f_i + \langle g_i, x - x_i \rangle$ is an affine minorant of f, and since $a_i \geq 0$ with $\sum_{i=0}^N a_i = 1$, the function α is also an affine minorant of f. In particular, $\alpha \in A_f \subset A_f^\star$. Further, note the values

$$\alpha(x_0) = \sum_{i=0}^N a_i (f_i + \langle g_i, x_0 - x_i \rangle), \qquad \nabla \alpha = \sum_{i=0}^N a_i g_i.$$

By Lemma 5.2, we have that

$$\mathtt{PEP}_M(W) - (f_N - \alpha(x_0)) \geq \sigma D^2 + \langle \hat{Z}, \hat{G} \rangle.$$

Next, since $\sum_{i=0}^N a_i = 1$, the vector a is nonzero. Therefore, since $\begin{bmatrix} \sigma & -\frac{1}{2}a^\top \\ -\frac{1}{2}a & \hat{Z} \end{bmatrix}$ is positive semidefinite, σ must be strictly positive. By the Schur complement, we then have $\hat{Z} \succeq \frac{1}{4\sigma} a a^\top$. Consequently,

$$\langle \hat{Z}, \hat{G} \rangle \geq \frac{1}{4\sigma} a^\top \hat{G} a = \frac{1}{4\sigma} \left\| \sum_{i=0}^N a_i g_i \right\|^2 = \frac{1}{4\sigma} \|\nabla \alpha\|^2.$$

From the last two displays, we conclude that

$$\mathtt{PEP}_M(W) - (f_N - \alpha(x_0)) \geq \sigma D^2 + \frac{1}{4\sigma} \|\nabla \alpha\|^2. \tag{32}$$

We will bound (32) in two different ways.

First, we can combine (32) with the following two inequalities:

$$f_\star \geq \alpha(x_\star) = \alpha(x_0) + \langle x_\star - x_0, \nabla\alpha\rangle$$
$$\|z_{N+1} - x_\star\|^2 \leq D^2 + \frac{1}{4\sigma^2}\|\nabla\alpha\|^2 + \frac{1}{\sigma}\langle \nabla\alpha, x_\star - x_0\rangle$$

to deduce that

$$\begin{aligned} f_N - f_\star &\leq f_N - \alpha(x_\star) + \sigma\|z_{N+1} - x_\star\|^2 \\ &\leq f_N - \alpha(x_0) + \sigma D^2 + \frac{1}{4\sigma}\|\nabla\alpha\|^2 \\ &\leq \texttt{PEP}_M(W). \end{aligned}$$

Alternatively, by the AM-GM inequality,

$$\texttt{PEP}_M(W) - (f_N - \alpha(x_0)) \geq \sigma D^2 + \frac{1}{4\sigma}\|\nabla\alpha\|^2 \geq D\|\nabla\alpha\|.$$

Rearranging this inequality gives the second claim. □

The first bound in Theorem 5.1 shows that the primal-dual gap $f(x_N) - \alpha(x_\star)$ converges not only as fast as the worst-case primal gap, but outperforms it by $\sigma\|z_{N+1} - x_\star\|^2$. Hence, in the tight worst-case where the primal gap (and hence primal-dual gap) equals $\texttt{PEP}_M(W)$, one must have that the minimizer exactly occurs at $z_{N+1} = x_0 - \frac{1}{2\sigma}\nabla\alpha$. So the dual solution α not only provides a dual lower bound but also indicates where the minimizer must be in tight, worst-case instances. As the true minimizer $x_\star$ differs from the worst-case placement z_{N+1}, the method's guarantee improves at a quadratic rate.

Finally, we note two subtleties in Theorem 5.1. First, while the tightness of the PEP reformulation required $d \geq N+2$, the above upper bounding guarantee holds for problems of any dimension. Second, the necessary dual certificate combines subgradient lower bounds from all of the iterates $x_0, \ldots, x_N$. As a result, numerically constructing this certificate requires one to compute a subgradient at the terminal iterate x_N (if $a_N > 0$), which may otherwise not be needed.

5.2 Example Subgradient Methods and their Implicit Dual Certificates

Below, we consider three subgradient methods from the literature, each known to admit a PEP certificate proving an exactly optimal worst-case rate of $MD/\sqrt{N+1}$. Since each of these methods is exactly optimal, its worst-case is attained by the maximin hard instance presented in [18, Appendix A]. Their known proof and this hard instance establish strong duality and attainment. In particular, on this instance, no method can reach the minimum objective value in $i < N+1$ iterations. Hence, $f_i^W > f_\star^W$. So each optimal method below is nondegenerate.

For each considered method, the affine function certifying this optimal objective gap rate is immediately available by examining the existing convergence proofs of these methods in the literature (in particular, the value of $\lambda_{\star,i}$ which sets a_i). For each method, the considered proof has $\lambda_{\star,i}$ constant, leading the dual model in each case to be the uniform averaging of the subgradient affine models seen. Indeed, this is necessary for all minimax optimal subgradient methods and optimal proofs, as shown by the complete characterization of [17].

The Averaged Subgradient Method. The averaged subgradient method iterates

$$x_n = x_{n-1} - \frac{D}{M\sqrt{N+1}}g_{n-1} \qquad \forall n = 1, \ldots, N-1$$
$$x_N = \frac{1}{N+1}\left[\sum_{i=0}^{N-1} x_i + \left(x_{N-1} - \frac{D}{M\sqrt{N+1}}g_{N-1}\right)\right].$$

This amounts to the subgradient method with a fixed stepsize $\frac{D}{M\sqrt{N+1}}$ for N iterations, where the terminal iterate is set as the average of these iterations. The presentation above combines the final step and averaging step to fit the standard fixed-step model (3).

A known optimal proof for this method, establishing that it has $f(x_N) - f(x_\star) \leq MD/\sqrt{N+1}$, has constant coefficients $\lambda_{\star,i}$ equal to $1/(N+1)$ and $\sigma = M/(2D\sqrt{N+1})$. Hence $a_i = 1/(N+1)$ and so the averaged subgradient method's standard primal objective gap convergence can be strengthened to the guarantees that

$$f(x_N) - \alpha(x_\star) + \sigma \left\| z_{N+1} - x_\star \right\|^2 \leq \frac{MD}{\sqrt{N+1}} \tag{33}$$

$$f(x_N) - \alpha(x_0) + D\|\nabla \alpha\| \leq \frac{MD}{\sqrt{N+1}} \tag{34}$$

where the dual certificate is given by the affine minorant $\alpha(x) = \frac{1}{N+1}\sum_{i=0}^{N}\left(f(x_i) + \langle g_i, x - x_i\rangle\right)$ and the estimate of the minimizer is given by $z_{N+1} = x_0 - \frac{D}{M\sqrt{N+1}}\sum_{i=0}^{N} g_i$.

The Final Step Subgradient Method. Recently, Zamani and Glineur [19] proposed a subgradient method with nonconstant stepsizes whose terminal iterate (without averaging) possesses the same optimal convergence guarantee as seen above. Formally, they consider the iteration for $n = 1, \ldots, N$

$$x_n = x_{n-1} - \frac{D(N+1-n)}{M(N+1)^{3/2}} g_{n-1}.$$

Their proof also uses uniform multipliers $\lambda_{\star,i} = 1/(N+1)$ and $\sigma = M/(2D\sqrt{N+1})$. So their method implicitly constructs the same affine minorant $\alpha(x) = \frac{1}{N+1}\sum_{i=0}^{N}\left(f(x_i) + \langle g_i, x - x_i\rangle\right)$ and constructs the same minimizer estimate $z_{N+1} = x_0 - \frac{D}{M\sqrt{N+1}}\sum_{i=0}^{N} g_i$. These strengthen Theorem 5.1 of [19] to ensure that their final iterate has the same stronger primal-dual gap bounds (33) and (34).

The SSEP Method. Another optimal fixed-step method of a momentum type was designed by Drori and Taylor [3]. They considered the following subgradient SSEP method for $n = 1, \ldots, N$

$$\begin{aligned} y_n &= \frac{n}{n+1} x_{n-1} + \frac{1}{n+1} x_0, \\ d_n &= \frac{1}{n+1}\sum_{i=0}^{n-1} g_i, \\ x_n &= y_n - \frac{D}{M\sqrt{N+1}}\, d_n \end{aligned}$$

which can be recursively expanded to verify that it takes the fixed-step form (3). Corollary 3 of [3] establishes that this has an optimal $MD/\sqrt{N+1}$ convergence rate. Again, their proof uses multipliers $\lambda_{\star,i} = 1/(N+1)$ and $\sigma = M/(2D\sqrt{N+1})$. So the implicit dual constructions take the same form. Hence, their Corollary 3 can be strengthened to ensure (33) and (34).

Our implicit dual theory provides an alternative understanding of the SSEP momentum method. Consider the affine minorant corresponding to the partial dual certificate built prior to iteration n

$$\alpha_n(x) = \frac{1}{n}\sum_{i=0}^{n-1}(f(x_i) + \langle g_i, x - x_i\rangle).$$

Noting that this dual solution has $d_n = \frac{n}{n+1}\nabla\alpha_n$, the SSEP method can be written in primal-dual terms, alternating between averaging and descent on the dual minorant. In these terms, the SSEP method is initialized with $\alpha_1(x) = f(x_0) + \langle g_0, x - x_0\rangle$ and iterates for $n = 1, \ldots, N$

$$\begin{aligned} x_n &= \frac{n}{n+1}\left(x_{n-1} - \frac{D}{M\sqrt{N+1}}\nabla\alpha_n\right) + \frac{1}{n+1}x_0 \\ \alpha_{n+1} &= \frac{n}{n+1}\alpha_n + \frac{1}{n+1}(f(x_n) + \langle g_n, \cdot - x_n\rangle). \end{aligned}$$

6 Dual Constructions for Fixed-Step Gradient Methods

The same progression used for subgradient methods also applies to gradient methods. Consider any fixed-step gradient method (3) with weights W for L-smooth convex minimization. Then its worst-case

performance (invoking Theorem 2.2 as an interpolation result) is given by

$$\begin{aligned}\texttt{PEP}_L(W) := &\begin{cases}\max_{f,x_i} & f(x_N)-f(x_\star)\\ \text{s.t.} & x_n = x_0 - \sum_{i=0}^{n-1} W_{n-1,i} g_i \quad \forall n = 1,\dots,N\\ & g_i = \nabla f(x_i) \qquad \forall i \in \mathcal{I}\\ & f \text{ is convex, } L\text{-smooth, and minimized at } x_\star\\ & \|x_0 - x_\star\| \le D\end{cases}\\ = &\begin{cases}\max_{(x_i,f_i,g_i)_{i\in\mathcal{I}}} & f_N - f_\star\\ \text{s.t.} & x_n = x_0 - \sum_{i=0}^{n-1} W_{n-1,i} g_i \quad \forall n = 1,\dots,N\\ & g_\star = 0\\ & \mathcal{Q}_{i,j} \ge 0 \qquad \forall i,j \in \mathcal{I}\\ & \mathcal{D} \ge 0.\end{cases}\end{aligned} \tag{35}$$

As before, this problem has a well-known reformulation as a semidefinite program. This follows by substituting each x_n and $g_\star$ for their equality definitions. This makes the above problem linear in the function values $f = (f_\star, f_0, \dots, f_N)$ and linear in the quadratic form $G = P^\top P$, with $P = \begin{bmatrix} x_0 - x_\star, g_0, \dots, g_N \end{bmatrix}$ which must be positive semidefinite. Provided $d \ge N+2$, the resulting semidefinite program is equivalent.

The dual semidefinite program, with nonnegative multipliers $\lambda_{i,j}$ for the $\mathcal{Q}_{i,j} \ge 0$ constraints, takes the same general form. A bound $f_N - f_\star \le r$ is proven by nonnegative multipliers $\lambda_{i,j}, \sigma$ and a positive semidefinite Z such that the following identity holds

$$r - (f_N - f_\star) = \sum_{i,j\in\mathcal{I}} \lambda_{i,j}\mathcal{Q}_{i,j} + \sigma\mathcal{D} + \langle Z, G\rangle,$$

fixing $x_n = x_0 - \sum_{i=0}^{n-1} W_{n-1,i} g_i$ and $g_\star = 0$, and treating f_i and each inner product from G as free variables. The dual optimization problem finds the best proof of this form (i.e., minimizing r). Given that strong duality holds, one can find multipliers such that

$$\texttt{PEP}_L(W) - (f_N - f_\star) = \sum_{i,j\in\mathcal{I}} \lambda_{i,j}\mathcal{Q}_{i,j} + \sigma\mathcal{D} + \langle Z, G\rangle. \tag{36}$$

6.1 Primal-Dual Guarantees for Smooth Convex Minimization

We restrict our attention to *nondegenerate* gradient methods[3], defined as those whose semidefinite programming PEP formulation has (i) strong duality and primal/dual attainment for the associated PEP and (ii) for each $i = 0,\dots,N$, a tight instance exists with realized values $f_i^W - \frac{1}{2L}\|g_i^W\|^2 > f_\star^W$.

Note that if $f_i^W - \frac{1}{2L}\|g_i^W\|^2 = f_\star^W$, then the classic descent lemma (i.e., $\mathcal{Q}_{i,\star} \ge 0$, noting $g_\star = 0$) ensures that the algorithm is one gradient descent step (with stepsize $1/L$) away from reaching the exact minimum value.

We can apply the same reasoning used to prove Lemma 5.2 to conclude the following lemma. As before, the constructed weights a_i below are exactly the optimal proof's $\lambda_{\star,i}$ multipliers. In the case of smooth optimization, we further deduce the exact value of σ in terms of D and $\texttt{PEP}_L(W)$.

Lemma 6.1. *Consider any $N \ge 1$, $L, D > 0$, and any nondegenerate gradient method defined by fixed-step weights W. Then there exists a nonnegative vector $a \in \mathbb{R}^{N+1}$, nonnegative multipliers $\hat{\lambda}_{i,j}$ for $i,j = 0,\dots,N$, and a positive semidefinite matrix $\hat{Z} \in \mathbb{S}^{N+1}$ such that $\sum_{i=0}^N a_i = 1$, the following matrix is positive semidefinite*

$$\begin{bmatrix} \sigma & -\frac{1}{2}a^\top \\ -\frac{1}{2}a & \hat{Z}\end{bmatrix}, \tag{37}$$

and

$$\texttt{PEP}_L(W) - \left(f_N - \sum_{i=0}^N a_i\left(f_i + \langle g_i, x_0 - x_i\rangle + \frac{1}{2L}\|g_i\|^2\right)\right) = \sum_{i,j=0}^N \hat{\lambda}_{i,j}\mathcal{Q}_{i,j} + \sigma D^2 + \langle \hat{Z}, \hat{G}\rangle, \tag{38}$$

[3]Again, simple degenerate methods exist. For instance when $L = D = 1$ and $N = 2$, $W = \begin{bmatrix}1 & 0\\ 2 & 3\end{bmatrix}$ is degenerate as solving the associated PEP SDP, one has $\lambda_{1,\star} = 7/2 > 0$.

where $\sigma = \frac{\texttt{PEP}_L(W)}{D^2} > 0$ *and* $\hat{G}$ *denotes the principal submatrix of* G *corresponding to gradient-gradient inner products.*

Proof. The proof of this result follows almost identically to the previous lemma, omitting terms related to Lipschitzness $\gamma_i \mathcal{M}_i$ and replacing each $\mathcal{C}_{i,j}$ by $\mathcal{Q}_{i,j}$. Since the $\mathcal{M}_i$ terms are omitted in (36), one can compare constant terms on the left- and right-hand sides. The only scalar on the right-hand side is σD^2. From this, we deduce that any dual feasible solution must have $\sigma = \frac{\texttt{PEP}_L(W)}{D^2}$. Since we replaced $\mathcal{C}_{i,j}$ by $\mathcal{Q}_{i,j}$, observe that whereas

$$\mathcal{C}_{\star,i} = f_\star - f_i - \langle g_i, x_\star - x_i\rangle = f_\star - f_i - \langle g_i, x_0 - x_i\rangle + \langle g_i, x_0 - x_\star\rangle,$$

we now have an additional $\frac{1}{2L}\|g_i\|^2$ term as

$$\mathcal{Q}_{\star,i} = f_\star - f_i - \langle g_i, x_\star - x_i\rangle - \frac{1}{2L}\|g_i\|^2 = f_\star - f_i - \langle g_i, x_0 - x_i\rangle + \langle g_i, x_0 - x_\star\rangle - \frac{1}{2L}\|g_i\|^2.$$

Carrying this term forward, the remainder of the proof goes through without modification. □

This lemma directly supports a similar main theorem for smooth convex optimization. The proof of this result follows the same sequence of algebraic steps. The individual affine terms in the previous lemma

$$\left(f_i + \langle g_i, x - x_i\rangle + \frac{1}{2L}\|g_i\|^2\right)$$

are no longer guaranteed to be minorants due to the added gradient norm squared term. So they may not lie in A_f. However, cocoercivity (8) (i.e., $\mathcal{Q}_{\star,i} \geq 0$) ensures that they are "minorants at the minimizer", lying in $A_f^\star$. As discussed in Section 3, this difference has no effect on the validity of the resulting primal-dual gap. Noting that $A_f^\star$ is also closed under convex combinations, the same proof as the previous theorem gives the following extension.

Theorem 6.1. *Consider any* $N \geq 1$, $L, D > 0$, *and any nondegenerate gradient method defined by fixed-step weights* W. *Then there exists a nonnegative vector* $a \in \mathbb{R}^{N+1}$, *summing to one, such that the method applied to any* L*-smooth convex problem with* $\|x_0 - x_\star\| \leq D$ *has*

$$\underbrace{f_N - f_\star}_{\textit{Primal Objective Gap}} \leq \underbrace{f_N - \alpha(x_\star)}_{\textit{Primal-Dual Gap}} + \underbrace{\frac{\texttt{PEP}_L(W)}{D^2}\|z_{N+1} - x_\star\|^2}_{\textit{Minimizer Estimate}} \leq \texttt{PEP}_L(W) \tag{39}$$

and

$$\underbrace{f_N - f_\star}_{\textit{Primal Objective Gap}} \leq \underbrace{f_N - (\alpha(x_0) - D\|\nabla\alpha\|)}_{\textit{Computable Primal-Dual Gap}} \leq \texttt{PEP}_L(W) \tag{40}$$

where $\alpha(x) = \sum_{i=0}^N a_i \left(f(x_i) + \langle g_i, x - x_i\rangle + \frac{1}{2L}\|g_i\|^2\right)$, *which lies in* $A_f^\star$, *and* $z_{N+1} = x_0 - \frac{D^2}{2\texttt{PEP}_L(W)}\nabla\alpha$.

Again, note that computing the affine dual certificate for a given algorithm requires computing a gradient at the terminal iterate (provided $a_N > 0$). So in order to report the final objective value $f(x_N)$ and a dual lower bound $(\alpha(x_0) - D\|\nabla\alpha\|)$, one additional round of first-order queries is needed.

6.2 Example Gradient Methods and their Implicit Dual Certificates

Here, we consider three methods, gradient descent, the Optimized Gradient Method, and Nesterov's Fast Gradient Method. The first two have well-known PEP-style convergence proofs and attain their worst-case convergence rate on a Huber-type function. Together, these establish strong duality and attainment. On each method's tight Huber function instance, the iterates are all more than one gradient descent step away from the minimizer. As a result, the first two methods are nondegenerate.

The exactly optimal PEP certificate for Nesterov's Fast Gradient Method, to our knowledge, has not been analytically determined. A feasible, nearly optimal PEP certificate was identified by [2] with each $\lambda_{i,\star} = 0$. So our theory can still be applied to extract primal-dual guarantees from this known, not-quite-tight certificate.

Gradient Descent. First, consider gradient descent, which iterates for $n = 1, \dots, N$

$$x_n = x_{n-1} - \frac{1}{L} g_{n-1}.$$

Drori and Teboulle [11] gave a tight convergence guarantee for gradient descent, ensuring $f_N - f_\star \leq \texttt{PEP}_L(W) = LD^2/(4N+2)$. Their dual proof sets multipliers on the star-row inequalities as

$$\lambda_{\star,0} = \frac{1}{2N}, \qquad \lambda_{\star,i} = \frac{2N+1}{(2N-i)(2N+1-i)} \quad (i = 1, \dots, N-1), \qquad \lambda_{\star,N} = \frac{1}{N+1}.$$

Hence, these are exactly the coefficients a_i in the implicitly constructed dual certificate, and they satisfy $\sum_{i=0}^{N} a_i = 1$. Therefore, gradient descent's known tight primal objective gap guarantee immediately strengthens to our pair of primal-dual guarantees with the dual certificate as the affine function

$$\alpha(x) = \sum_{i=0}^{N} a_i \left(f(x_i) + \langle g_i, x - x_i \rangle + \frac{1}{2L} \|g_i\|^2 \right),$$

and coefficients $a_i = \lambda_{\star,i}$. Note that these are nonnegative and sum to one as needed.

The Optimized Gradient Method (OGM). Second, we consider OGM as defined by [2], which is known to be exactly minimax optimal for L-smooth convex minimization (provided $d \geq N+2$). This method first computes $g_0 = \nabla f(x_0)$, sets $\tau_0 = 2$ and $z_1 = x_0 - \frac{2}{L} g_0$, and then iterates for $n = 1, \dots, N$

$$\tau_n = \begin{cases} \tau_{n-1} + 1 + \sqrt{1 + 2\tau_{n-1}} & \text{if } n < N \\ \tau_{n-1} + \frac{1 + \sqrt{1 + 4\tau_{n-1}}}{2} & \text{if } n = N, \end{cases} \tag{41}$$

$$x_n = \frac{\tau_{n-1}}{\tau_n} \left(x_{n-1} - \frac{1}{L} g_{n-1} \right) + \frac{\tau_n - \tau_{n-1}}{\tau_n} z_n, \tag{42}$$

$$z_{n+1} = z_n - \frac{\tau_n - \tau_{n-1}}{L} g_n. \tag{43}$$

Above, we use the τ_n scalar sequence of [20] to define OGM. This is mathematically equivalent to the recursive definition and inductive analyses using θ_n^2 sequences in the literature [21, 22]. One known convergence proof of OGM's optimal rate sets $\lambda_{\star,0} = \frac{2}{\tau_N}$ and each $\lambda_{\star,n} = \frac{\tau_n - \tau_{n-1}}{\tau_N}$ for $n = 1, \dots, N$, proving $\texttt{PEP}_L(W) = \frac{LD^2}{2\tau_N}$. So the corresponding implicit dual certificate constructed by OGM is the affine function

$$\alpha(x) = \frac{2}{\tau_N} \left(f(x_0) + \langle g_0, x - x_0 \rangle + \frac{1}{2L} \|g_0\|^2 \right) + \sum_{i=1}^{N} \frac{\tau_i - \tau_{i-1}}{\tau_N} \left(f(x_i) + \langle g_i, x - x_i \rangle + \frac{1}{2L} \|g_i\|^2 \right).$$

The slope of this affine function is then exactly related to $(z_{N+1} - x_0)$ by

$$\nabla \alpha = \frac{2}{\tau_N} g_0 + \sum_{i=1}^{N} \frac{\tau_i - \tau_{i-1}}{\tau_N} g_i = -\frac{L}{\tau_N} (z_{N+1} - x_0).$$

Hence, the constructed z_{N+1} from OGM exactly agrees with our theorem's z_{N+1}. So the auxiliary sequence precisely provides the location that the minimizer $x_\star$ must be in tight instances where the worst-case primal gap (and hence primal-dual gap) occurs. Thus, at each step, OGM estimates the worst-case minimizer's location and averages toward it while taking a gradient descent step.

Further, one can view the auxiliary/momentum sequence z_{n+1} as tracking the slope of the aggregate dual solution implicitly constructed by the algorithm's optimal proof. As we observed with the SSEP method, OGM can then be stated in primal-dual terms as alternating averaging with descent of the objective and descent on the dual certificate. With the τ_n sequence defined as above, OGM initializes $\alpha_1(x) = f(x_0) + \langle g_0, x - x_0 \rangle + \frac{1}{2L} \|g_0\|^2$ (which lies in $A_f^\star$) and iterates

$$x_n = \frac{\tau_{n-1}}{\tau_n} \left(x_{n-1} - \frac{1}{L} g_{n-1} \right) + \frac{\tau_n - \tau_{n-1}}{\tau_n} \left(x_0 - \frac{\tau_{n-1}}{L} \nabla \alpha_n \right)$$

$$\alpha_{n+1} = \frac{\tau_{n-1}}{\tau_n} \alpha_n + \frac{\tau_n - \tau_{n-1}}{\tau_n} \left(f(x_n) + \langle g_n, \cdot - x_n \rangle + \frac{1}{2L} \|g_n\|^2 \right),$$

ultimately reporting x_N and α_{N+1}. Taking either of these perspectives, our Theorem 6.1 provides OGM with two strengthened statements of its optimal convergence rate: (i) OGM has primal-dual gap $f(x_N) - \alpha_{N+1}(x_\star)$ bounded by the same optimal rate and strictly improves when the minimizer differs from its auxiliary z_{N+1}, (ii) OGM has a computable primal-dual gap converging at the same optimal rate as it alternates between building x_n and building α_{n+1}.

Nesterov's Fast Gradient Method. Finally, we consider the fast gradient method, iterating

$$\begin{aligned} t_n &= \frac{1+\sqrt{1+4t_{n-1}^2}}{2}, \\ x_n &= \frac{t_n - 1}{t_n}\left(x_{n-1} - \frac{1}{L}g_{n-1}\right) + \frac{1}{t_n}z_n, \\ z_{n+1} &= z_n - \frac{t_n}{L}g_n \end{aligned}$$

with $t_0 = 1$ and $z_1 = x_0 - t_0 g_0/L$. A PEP-style convergence guarantee for this method was given by [2] by considering $\sigma = \frac{L}{2t_N^2}$, a sparse λ selection having only nonzero values for $\lambda_{i,i+1} = \frac{t_i^2}{t_N^2}$ and $\lambda_{\star,i} = \frac{t_i}{t_N^2}$, and finally,

$$Z = \frac{1}{2t_N^2}\left[\begin{pmatrix} \sqrt{L} \\ -\frac{t_0}{\sqrt{L}} \\ \vdots \\ -\frac{t_N}{\sqrt{L}} \end{pmatrix}\begin{pmatrix} \sqrt{L} & -\frac{t_0}{\sqrt{L}} & \cdots & -\frac{t_N}{\sqrt{L}} \end{pmatrix} + \frac{1}{L}\operatorname{diag}(0, t_0^2, \ldots, t_{N-1}^2, 0)\right].$$

(Note this proof matches Nesterov's classic rate, but is not exactly tight in the PEP-sense.) This certificate proves a guarantee of

$$f(x_N) - f(x_\star) \leq \frac{L\|x_0 - x_\star\|^2}{2t_N^2} \leq \frac{2L\|x_0 - x_\star\|^2}{(N+2)^2}.$$

As we did with OGM, applying our reformulation of guarantees, Nesterov's Fast Gradient Method can be seen as implicitly building the dual model

$$\alpha_n(x) = \sum_{i=0}^{n-1} \frac{t_i}{t_{n-1}^2}\left(f(x_i) + \langle g_i, x - x_i\rangle + \frac{1}{2L}\|g_i\|^2\right)$$

at each iteration. The slope of this affine function is given by $\nabla\alpha_n = \sum_{i=0}^{n-1}\frac{t_i}{t_{n-1}^2}g_i = -\frac{L}{t_{n-1}^2}(z_n - x_0)$. Then, one can write the update in primal-dual form. In these terms, Nesterov's method initializes $\alpha_1(x) = f(x_0) + \langle g_0, x - x_0\rangle + \frac{1}{2L}\|g_0\|^2$, and then iterates

$$\begin{aligned} x_n &= \frac{t_n - 1}{t_n}\left(x_{n-1} - \frac{1}{L}g_{n-1}\right) + \frac{1}{t_n}\left(x_0 - \frac{t_{n-1}^2}{L}\nabla\alpha_n\right), \\ \alpha_{n+1} &= \frac{t_n - 1}{t_n}\alpha_n + \frac{1}{t_n}\left(f(x_n) + \langle g_n, \cdot - x_n\rangle + \frac{1}{2L}\|g_n\|^2\right). \end{aligned}$$

While this runs, the same classic $O(1/N^2)$ primal guarantee applies more strongly, ensuring that the primal-dual gap with estimate error $(f(x_N) - \alpha_{N+1}(x_\star) + \frac{L}{2t_N^2}\|z_{N+1} - x_\star\|^2)$ and the computable primal-dual gap $(f(x_N) - (\alpha_{N+1}(x_0) - D\|\nabla\alpha_{N+1}\|))$ are both at most $\frac{LD^2}{2t_N^2}$.

7 Conclusion

Dual interpretations of primal methods have occurred throughout the optimization literature. In nonsmooth optimization, Nesterov [23] connected traditional primal subgradient methods with a dual perspective, enabling new averaging methods. More recently, Grimmer and Li [14] explicitly connected dual averaging to the kind of primal-dual minimax perspectives considered here. In smooth optimization, several previously discussed works [5, 7, 8, 6, 4, 9, 10] presented dual analyses and explanations

of momentum. Here, we provide a general mechanism from which dual phenomena arise. Equivalent primal-dual convergence theory must exist as a necessary consequence of structured PEP proofs and nondegeneracy.

For Lipschitz convex minimization and smooth convex minimization, we have shown that attempts to design nondegenerate algorithms and proofs for reducing the objective gap inherently prove stronger guarantees on primal-dual gaps. Future efforts to design optimized algorithms may benefit from pursuing primal-dual guarantees directly. Building on our understanding of auxiliary sequences in momentum methods as an implicit dual proof construction may also lead to new methods.

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

References

[1] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

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

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

[4] David H. Gutman and Javier F. Peña. Perturbed Fenchel duality and first-order methods. *Mathematical Programming*, 198(1):443–469, 2022.

[5] Jelena Diakonikolas and Lorenzo Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 29(1):660–689, 2019.

[6] Dmitriy Drusvyatskiy, Maryam Fazel, and Scott Roy. An optimal first order method based on optimal quadratic averaging. *SIAM Journal on Optimization*, 28(1):251–271, 2018.

[7] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 3:1–3:22, Dagstuhl, Germany, 2017. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[8] Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical Programming*, 171:167 – 215, 2018.

[9] Jun-Kun Wang, Jacob Abernethy, and Kfir Y. Levy. No-regret dynamics in the Fenchel game: a unified framework for algorithmic convex optimization. *Mathematical Programming*, 205:203–268, 2024.

[10] Matthew X. Burns and Jiaming Liang. Accuracy certificates for convex optimization at accelerated rates via primal-dual averaging. *arXiv:2604.18321*, 2026.

[11] Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: A novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

[12] 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.

[13] Adrien B Taylor, Julien M Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1–2):307–345, 2017.

[14] Benjamin Grimmer and Danlin Li. Some primal-dual theory for subgradient methods for strongly convex optimization. *Mathematical Programming*, 214(1–2):759–788, 2025.

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

[16] Mihai I. Florea and Yurii E. Nesterov. An optimal lower bound for smooth convex functions. *Found. Comput. Math.*, 25(6):1939–1973, 2025.

[17] Aaron Zoll and Benjamin Grimmer. A complete characterization of optimal subgradient methods for Lipschitz convex minimization. *arXiv:2607.19240*, 2026.

[18] Yoel Drori and Marc Teboulle. An optimal variant of Kelley's cutting-plane method. *Mathematical Programming*, 160:321–351, 2016.

[19] Moslem Zamani and François Glineur. Exact convergence rate of the last iterate in subgradient methods. *SIAM Journal on Optimization*, 35(3):2182–2201, 2025.

[20] Benjamin Grimmer, Kevin Shu, and Alex L. Wang. Beyond minimax optimality: A subgame perfect gradient method. *Mathematical Programming*, 2026.

[21] Adrien B Taylor and Francis Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2934–2992, 2019.

[22] Chanwoo Park, Jisun Park, and Ernest K Ryu. Factor-$\sqrt{2}$ acceleration of accelerated gradient methods. *Applied Mathematics & Optimization*, 88(3):77, 2023.

[23] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.