Accuracy Certificates for Convex Optimization at Accelerated Rates via Primal-Dual Averaging

Matthew X. Burns* Jiaming Liang †

April 20, 2026

Abstract

Many works in convex optimization provide rates for achieving a small primal gap. However, this quantity is typically unavailable in practice. In this work, we show that solving a regularized surrogate with algorithms based on simple primal-dual averaging provides non-asymptotic convergence guarantees for a *computable* optimality certificate. We first analyze primal and dual methods based on one average, namely modified dual averaging and generalized conditional gradient, and establish $\tilde{\mathcal{O}}(\varepsilon^{-1})$ certificate complexities. Motivated by asymmetries in the one-average case, we analyze a self-dual, two-average method that preserves symmetry while losing certificate guarantees. To recover certificate convergence, we propose a three-average method that achieves an accelerated $\tilde{\mathcal{O}}(\varepsilon^{-1/2})$ certificate complexity. Furthermore, we prove primal-dual algorithm correspondences for the one, two, and three-average cases. In particular, the primal three-average accelerated method mirrors the well-known gradient extrapolation method in the dual. By interpreting our results through the lens of zero-sum matrix games and Fisher markets, we further connect primal-dual averaging methods to game theory and market dynamics.

Key words. primal-dual averaging, conditional gradient method, accelerated method, gradient extrapolation method, accuracy certificate

1 Introduction

Continuous convex optimization has become ubiquitous in large-scale computing, with applications in machine learning [Lan20, SS12, SNW11, WR22], compressive sensing [CRT06, CR08], statistical inference [JN20], and imaging science [BT09, CP11] (to name a few). In several settings, we require solutions to satisfy a defined optimality bound. Since we cannot run methods indefinitely, it is crucial to have some *certificate* of optimality. In this work, we show that simple methods based on regularization and primal-dual averaging can provide provably convergent, computable optimality certificates. Our focus is the convex smooth composite optimization (CSCO) problem

$$\phi_* = \min_{x\in\mathbb{R}^n} \{\phi(x) := f(x) + h(x)\}, \tag{1}$$

where f is a closed proper convex function that is L-smooth with respect to the primal norm $\|\cdot\|$, and h is merely closed proper convex with bounded domain. For fixed $\varepsilon > 0$, we say that a point $x \in \mathbb{R}^n$ is an ε-solution if it achieves an ε-small primal gap, i.e., $\phi(x) - \phi_* \leq \varepsilon$.

*Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627 (email: mburns13@ur.rochester.edu).

†Goergen Institute for Data Science and Artificial Intelligence and Department of Computer Science, University of Rochester, Rochester, NY 14620 (email: jiaming.liang@rochester.edu). This work was supported by AFOSR grant FA9550-25-1-0182.

Instead of solving (1) directly, our strategy is to solve the regularized problem

$$\phi_*^\alpha = \min_{x\in\mathbb{R}^n}\{\phi^\alpha(x) := f(x) + h(x) + \alpha w(x)\}, \tag{2}$$

where $w : \mathbb{R}^n \to [0, +\infty]$ is non-negative[1], closed, and 1-strongly convex with respect to $\|\cdot\|$ on $\operatorname{dom} h$ and satisfies the following: (1) the generalized linear minimization oracle (GLMO)

$$\min_{x\in\mathbb{R}^n}\left\{\langle v, x\rangle + h(x) + \alpha w(x)\right\},$$

is efficiently computable for all $\alpha > 0$ and $v \in \mathbb{R}^n$, and (2) w is bounded on $\operatorname{dom} h$ with $M := \max_{x\in\operatorname{dom} h} w(x) < \infty$.

By choosing $\alpha = \mathcal{O}(\varepsilon/M)$ and solving (2) to $\mathcal{O}(\varepsilon)$ accuracy, we can obtain an $\mathcal{O}(\varepsilon)$ solution to (1) (see Lemma 2.1). Adding $\mathcal{O}(\varepsilon)$ regularization can have additional computational and theoretical benefits, such as parallel computation in discrete optimal transport [Cut13], nonergodic convergence rates in bilinear saddle-point problems [CWC21], and more stable variable selection in sparse regression [ZH05].

Recent work [GP23] demonstrates that we can obtain stronger primal-dual convergence results by considering the Fenchel-Rockafellar dual to (2),

$$\phi_*^\alpha = \max_{z\in\mathbb{R}^n}\{-(h^\alpha)^*(-z) - f^*(z)\} = -\min_{z\in\mathbb{R}^n}\{\psi^\alpha(z) := (h^\alpha)^*(-z) + f^*(z)\} = -\psi_*^\alpha, \tag{3}$$

where we define the aggregate function $h^\alpha(\cdot) = h(\cdot) + \alpha w(\cdot)$ for convenience and f^* is the convex conjugate of f. By our assumptions on f, h, and w, we have that $(h^\alpha)^*$ is $(1/\alpha)$-smooth and $f^*(\cdot)$ is $(1/L)$-strongly convex relative to the dual norm $\|\cdot\|_*$.

Numerous prior works have revealed deep connections between seemingly disparate optimization algorithms in the CSCO setting by demonstrating that they are "dual" to each other: one algorithm solving the primal (2) generates the same iterate sequences as another algorithm solving the dual (3). Duality correspondences often lead to simplified convergence proofs [LF21, Tib17], optimality certificate guarantees [Bac15], and novel algorithm variants [Tib17, BL26, WAL23]. Existing primal-dual algorithm pairs include generalized conditional gradient (GCG) and mirror descent [Bac15], ADMM and Dykstra's algorithm [Tib17], cyclic coordinate descent and ADMM [Tib17], a stochastic variant of GCG and randomized coordinate descent [LF21], and a primal-dual cutting-plane method and GCG [FS24, Lia25]. Beyond the setting of (1), duality correspondences can provide more efficient algorithms in constrained problems [BL26] and one-level convex reformulations in bilevel optimization [AA11].

Our contributions In this work, we characterize primal-dual pairs in three algorithmic classes targeting the regularized problem (2), methods with one, two, and three averages, along with their complexity and optimality certificate guarantees. The algorithms considered are summarized in Table 1. For one average, we show that a modified dual averaging (MDA) method in the primal is equivalent to GCG in the dual. Furthermore, both algorithms provide computable primal-dual ε-optimality certificates with $\tilde{\mathcal{O}}(\varepsilon^{-1})$ complexity. For two averages, we show that a previously proposed method [ZZ23], which we term aggregated GCG (AggGCG), is self-dual. While the algorithm possesses appealing symmetry, it does not admit the same straightforward primal-dual certificate convergence as the one-average case. Finally, we propose a three-average accelerated (TAA) method. TAA recovers the computable primal-dual certificate of the one-average case with an accelerated $\tilde{\mathcal{O}}(\varepsilon^{-1/2})$ complexity. Furthermore, an interesting finding is that, through duality,

[1]This assumption can be made without loss of generality by translation, since strong-convexity implies that w is bounded from below on $\operatorname{dom} h$.

Averages	Algorithm	Ref	Perspective	Complexity	Cert. Complexity
1	MDA	[Nes09, Lia25]	Primal	$\tilde{\mathcal{O}}(\varepsilon^{-1})$	$\tilde{\mathcal{O}}(\varepsilon^{-1})$
	GCG	[BLM09]	Dual	$\tilde{\mathcal{O}}(\varepsilon^{-1})$	$\tilde{\mathcal{O}}(\varepsilon^{-1})$
2	AggGCG	[ZZ23]	Primal/Dual	$\tilde{\mathcal{O}}(\varepsilon^{-1})$	-
3	TAA	**This Work**	Primal	$\tilde{\mathcal{O}}(\varepsilon^{-1/2})$	$\tilde{\mathcal{O}}(\varepsilon^{-1/2})$
	GEM	[LZ18]	Dual	$\tilde{\mathcal{O}}(\varepsilon^{-1/2})$	$\tilde{\mathcal{O}}(\varepsilon^{-1/2})$

Table 1: High-level summary of primal-dual algorithm pairs considered in this work. We view a "Primal" algorithm as solving problem (2) and a "Dual" algorithm as solving problem (3). "Complexity" refers to the complexity of computing an ε-solution to (1), while "Cert. Complexity" refers to the complexity of computing a verifiable primal-dual certificate of ε-optimality (See Definition 2.4). Acronyms are as follows: "MDA" is "Modified Dual Averaging", "GCG" is "Generalized Conditional Gradient", "AggGCG" is "Aggregated GCG", "TAA" is "Three-Average Acceleration", and "GEM" is the "Gradient Extrapolation Method".

TAA is equivalent to a variant of the well-known gradient extrapolation method (GEM) [LZ18]. As an additional benefit, we show that the GEM also admits a computable optimality certificate with $\tilde{\mathcal{O}}(\varepsilon^{-1/2})$ complexity, a novel result in the CSCO setting. We provide additional insight into our results by game-theoretic interpretations for each pair of primal-dual correspondences and Fisher market interpretations for each of the three averaging methods in the primal space.

Outline Section 2 provides the key definitions and running examples. Section 3 considers the simple case of one-average methods, providing primal-dual equivalence results and $\tilde{\mathcal{O}}(\varepsilon^{-1})$ certificate complexities for MDA and GCG. Moving to two averages, Section 4 formalizes the self-duality of AggGCG along with a $\tilde{\mathcal{O}}(\varepsilon^{-1})$ primal gap complexity bound for (1). Motivated by AggGCG's lack of computable certificate guarantees, Section 5 proposes the TAA method. Along with its certificate complexity, we further state its formal equivalence to the GEM. Conclusions and directions for future work are discussed in Section 7. Technical proofs and self-contained analyses for each method are deferred to the appendices.

2 Preliminaries

Let $\mathbb{R}^n$ be the n-dimensional Euclidean space equipped with the standard inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. The n-dimensional unit simplex is given by Δ^n. For a closed convex function f, we denote the *subdifferential* of f at x by $\partial f(x)$, and the *linearization* of f at $x_0 \in \operatorname{dom} f$ as $\ell_f(\cdot; x_0) := f(x_0) + \langle f'(x_0), \cdot - x_0 \rangle$, where $f'(x_0) \in \partial f(x_0)$. If f is continuously differentiable, then $\partial f(x) = \{\nabla f(x)\}$. For a convex function f, we define the *Bregman divergence* of f as $\mathrm{D}_f(x\|y) := f(x) - \ell_f(x; y)$. We define the *convex conjugate* of a closed and proper function f as $f^*(y) = \max_{x \in \mathbb{R}^n} \{\langle y, x \rangle - f(x)\}$. The conjugate f^* is convex, and for a closed proper convex f we have the identity $f = (f^*)^*$ (see [Bec17, Theorems 4.3 and 4.8]). We define the *domain* of f as $\operatorname{dom} f = \{x \in \mathbb{R}^n : f(x) < \infty\}$. For $\mu \geq 0$, we say that f is *μ-strongly convex* on $Q \subseteq \operatorname{dom} f$ with respect to the norm $\|\cdot\|$ if $\mathrm{D}_f(x\|y) \geq \mu \|x - y\|^2/2$ for all $x, y \in Q$. Additionally, for $L > 0$ we say that f is *L-smooth* with respect to the norm $\|\cdot\|$ on Q if $\mathrm{D}_f(x\|y) \leq L\|x - y\|^2/2$ for all $x, y \in Q$.

With basic definitions established, we motivate the substitution of (2) to solve (1) by connecting a primal-dual solution of the regularized problem (2) to a primal solution of the original objective (1). See Appendix A for the proof.

Lemma 2.1. Assume that $w : \mathbb{R}^n \to \mathbb{R}_+$ is a non-negative function and $M := \max_{x \in \operatorname{dom} h} w(x) < \infty$ is bounded. Given $\varepsilon > 0$ and choosing $\alpha \le \varepsilon/(2M)$, if the primal-dual pair $(x, s) \in \operatorname{dom} h \times \mathbb{R}^n$ satisfies $\phi^\alpha(x) + \psi^\alpha(s) \le \varepsilon/2$, then we have $\phi(x) - \phi_* \le \varepsilon$.

Now we define a function that will be critical in our definition of an optimality certificate.

Definition 2.2 (Aggregated Cutting Plane (ACP) Model)**.** Given an initial point y_0, a set of points $\{x_i\}_{i=0}^{k-1} \in \operatorname{dom} \phi$ and a set of convex combination parameters $\zeta \in [0,1]^k$, we define the associated ACP model as

$$\Gamma_0(x) = h^\alpha(x) + \ell_f(x; y_0), \quad \Gamma_{j+1}(x) = (1 - \zeta_j)\Gamma_j(x) + \zeta_j(h^\alpha(x) + \ell_f(x; x_j)), \tag{4}$$

where $0 \le j \le k-1$. We say that Γ_k is the ACP model induced by $(y_0, \{x_i\}_{i=0}^{k-1}, \zeta)$. If $k = 0$, then we simply say that $\Gamma := \Gamma_0$ is the single-cutting plane (SCP) model induced by y_0.

The ACP model has been used in prior work on proximal bundle methods for convex nonsmooth optimization [LM24, LGM24], however it has several appealing properties for broader convex optimization, as we summarize in the following lemma (whose proof is deferred to Appendix A).

Lemma 2.3. Let $\Gamma_k(\cdot)$ be the ACP model for (2) induced by $(y_0, \{x_i\}_{i=0}^{k-1}, \zeta)$ for $y_0 \in \operatorname{dom} h$, $\{x_i\}_{i=0}^{k-1} \subseteq \operatorname{dom} h$ and $\zeta \in [0,1]^k$. Then, the following hold:

a) for all $x \in \operatorname{dom} h$, $\Gamma_k(x) \le \phi^\alpha(x)$;

b) Γ_k is α-strongly convex.

Furthermore, define $\{s_k\}_{k \ge 0}$ as

$$s_0 = \nabla f(y_0), \quad s_{j+1} = (1 - \zeta_j)s_j + \zeta_j \nabla f(x_j) \tag{5}$$

for $0 \le j \le k-1$. Then, the following bound holds for all $u \in \operatorname{dom} h$

c) $\phi^\alpha(u) + \psi^\alpha(s_k) \le \phi^\alpha(u) - \min_{x \in \mathbb{R}^n} \Gamma_k(x)$.

Motivated by Lemma 2.3(c), we define a computable primal-dual optimality certificate. We define the certificate from the primal perspective, however Lemma A.3(c) shows that we can equivalently define the certificate from the dual perspective.

Definition 2.4 (Primal-Dual Certificate)**.** Let $u \in \operatorname{dom} h$ and suppose $\alpha \le \varepsilon/(2M)$. We say (u, Γ_k) is an ε-certificate for (1) if

$$\phi^\alpha(u) - \min_{x \in \mathbb{R}^n} \Gamma_k(x) \le \frac{\varepsilon}{2},$$

where Γ_k is an ACP model induced by $(y_0, \{x_i\}_{i=0}^{k-1}, \zeta)$.

By Lemmas 2.1 and 2.3(c), an ε-certificate (u, Γ_k) implies that u is an ε-solution to (1). Optimality certificates based on the minimum of aggregated linearizations have been used repeatedly in convex optimization [NOR10, RN23, GL25], known as "accuracy certificates." Our definition of a certificate aligns with these works, unifying recent literature on proximal bundle methods [LM24, LGM24, Lia25] and classical accuracy certificates.

Throughout the work, we provide additional motivation and intuition for the primal-dual correspondences by adopting a game-theoretic lens. Specifically, we consider an entropically smoothed, zero-sum matrix game

$$\min_{x \in \mathbb{R}^n} \max_{z \in \mathbb{R}^n} \left\{ \langle x, Az \rangle + \delta_{\Delta^n}(x) - \delta_{\Delta^n}(z) + \alpha H(x) - L^{-1} H(z) \right\}, \tag{6}$$

where $H(x) = \langle x, \log x \rangle$ is the negative entropy function and we assume that the payoff matrix $A \in \mathbb{R}^{n \times n}$ has unit matrix norm for convenience. Computing the primal ϕ^α and dual d^α functions[2] gives

$$\phi^\alpha(x) = \overbrace{L^{-1} \log \left(\sum_{i=1}^{n} \exp[L A_{:,i}^\top x] \right)}^{f(A^\top x)} + \overbrace{\delta_{\Delta^n}(x)}^{h(x)} + \alpha \overbrace{H(x)}^{w(x)}, \tag{7}$$

$$d^\alpha(z) = -\alpha \log \left(\sum_{j=1}^{n} \exp[-\alpha^{-1} A_{j,:}^\top z] \right) - L^{-1} H(z) - \delta_{\Delta^n}(z), \tag{8}$$

where $A_{:,i}$ and $A_{j,:}$ represent the i^{th} column and j^{th} row of A, respectively. Defining ψ^α as in (3) and applying standard conjugate calculus, we have

$$d^\alpha(z) = -(h^\alpha)^*(-Az) - f^*(z) = -\psi^\alpha(z), \tag{9}$$

where f and h^α are as in (7).

In this setting, we denote $x \in \Delta^n$ and $z \in \Delta^n$ as mixed strategies of the minimizing and maximizing players, respectively, while gradients $\nabla_x f(A^\top x)$ and $\nabla_z (h^\alpha)^*(-Az)$ are predicted opponent responses to the primal and dual players, respectively. Since the primal (7) and dual (8) functions are negative Fenchel dual to each other by (9), the prediction exactly accords with reality. For simplicity, we henceforth use the terms "player prediction" and "opponent response" interchangeably without reiterating their equivalence. Moreover, we can show that the functions (7) and (8) satisfy the assumptions of (2) and (3), so all convergence analysis for the general CSCO problem (1) can be translated to game setting.

3 One-Average Methods

We begin with a simple baseline consisting of one GLMO and one averaging step. When the averaging step is performed over gradients, we obtain the DA [Nes09] variant shown in Algorithm 1.

Algorithm 1 MDA

Initialize: given $y_0 \in \operatorname{dom} h$, $\alpha > 0$, set $\eta = \alpha/(L+\alpha)$, $s_0 = \nabla f(y_0)$, and compute

$$x_0 = \operatorname*{argmin}_{x \in \mathbb{R}^n} \{\langle s_0, x \rangle + h^\alpha(x)\}.$$

for $k \geq 0$ **do**

Compute

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \{\langle s_{k+1}, x \rangle + h^\alpha(x)\},$$

where $s_{k+1} = (1-\eta) s_k + \eta \nabla f(x_k)$.

end for

Algorithm 2 GCG

Initialize: $z_0 \in \operatorname{dom} f^*$, set $\eta = \alpha/(L+\alpha)$.

for $k \geq 0$ **do**

Compute

$$\bar{z}_k = \operatorname*{argmin}_{v \in \mathbb{R}^n} \{\langle -\nabla (h^\alpha)^*(-z_k), v \rangle + f^*(v)\}, \tag{10}$$

$$z_{k+1} = \eta \bar{z}_k + (1-\eta) z_k. \tag{11}$$

end for

Adopting the perspective of the dual problem (3), we can instead employ GCG as shown in Algorithm 2. Here, the "primal" variables z_{k+1} are averaged while a single "gradient" is used to

[2] See [Bec17, Section 4.4.10 and Theorem 4.14]

compute $\bar{z}_k$. Comparing the one-average methods, we obtain the following equivalence result. A similar correspondence was obtained in [Lia25] in the case $w(\cdot) = \|\cdot - x_0\|^2/2$ for some x_0 in the context of proximal bundle methods. The proof is deferred to Appendix B.

Proposition 3.1. Fix $y_0 \in \operatorname{dom} h$ and set $z_0 = \nabla f(y_0)$. Then, on every iteration $k \geq 0$ of Algorithm 1 and Algorithm 2, we have the following correspondence

$$\nabla f(x_k) = \bar{z}_k, \quad x_k = \nabla (h^\alpha)^*(-z_k), \quad s_k = z_k.$$

In the context of the zero-sum game (6), we can better understand the correspondence between Algorithms 1 and 2 as a sequential game dynamic, summarized in the following graph

$$y_0 \longrightarrow (s_0 = z_0) \longrightarrow (x_k \overset{\textbf{(a)}}{=} \nabla(h^\alpha)^*(-z_k)) \to (\nabla f(x_k) \overset{\textbf{(b)}}{=} \bar{z}_k) \to (s_{k+1} \overset{\textbf{(c)}}{=} z_{k+1}).$$

In this game, the average iterates $\{z_k\}$ are mixed strategies, while $\{x_k\}$, $\{\bar{z}_k\}$ are responses to opponent moves.

To start the game, the primal player moves first by selecting strategy y_0, and then predicts the dual player's response s_0 simultaneously with the dual player's actual response z_0. We recall that, by (9), each player's prediction exactly aligns with reality (i.e., the opponent's actual response) due to conjugate symmetry of the problem.

With the game initialized, each round occurs the same way. First, in **(a)** the dual player predicts the primal response to their mixed strategy ($x_k = \nabla(h^\alpha)^*(-z_k)$). Next, in **(b)** the primal player predicts the dual response to their previous move ($\nabla f(x_k) = \bar{z}_k$). Finally, in **(c)** the mixed strategy ($s_{k+1} = z_{k+1}$) is updated using the last dual response and the next round begins. Each player therefore employs simple "regularized best response + averaging" strategies, where the regularized best response is computed via the GLMO and the primary difference is the role of the averaging step. The primal player uses averaging to smooth opponent responses, the dual player uses averaging to compute their own mixed strategy.

With the correspondence established in Proposition 3.1, we state the following certificate complexities for each 1-average method. Full proofs are deferred to Appendix B.

Theorem 3.2. For $y_0 \in \operatorname{dom} h$, for all $k \geq 0$ define $y_{k+1} = (1-\eta)y_k + \eta x_{k+1}$. Given $\varepsilon > 0$, choosing $\alpha = \varepsilon/(2M)$, the pair (y_k, Γ_k) is an ε-certificate for (1) in $k = \tilde{\mathcal{O}}(1 + ML\varepsilon^{-1})$ iterations of Algorithm 1, where Γ_k is the ACP model for (2) induced by $(y_0, \{x_i\}_{i=0}^{k-1}, \{\eta\}^k)$.

Proof Sketch. The majority of the proof of Theorem 3.2 is to show the inequality

$$\Gamma_{k+1}(x_{k+1}) \geq (1-\eta)\Gamma_k(x_k) + \eta\phi^\alpha(x_{k+1}),$$

which utilizes the optimality condition of x_{k+1}, the strong convexity of w, and the smoothness/convexity of $\phi^\alpha(\cdot)$. Recursively expanding and applying the definition of y_k along with the convexity of $\phi^\alpha(\cdot)$ gives a linear convergence rate in the model gap with contraction factor $(1+\alpha/L)^{-1}$. Our choice of α and standard analysis then gives the complexity.

Theorem 3.3. Given $\varepsilon > 0$, choosing $\alpha = \varepsilon/(2M)$, the pair (z_k, Γ_k^*) is an ε-certificate for (1) in $k = \tilde{\mathcal{O}}(1 + ML\varepsilon^{-1})$ iterations of Algorithm 2, where Γ_k^* is the SCP model for (3) induced by z_k.

Proof Sketch. The proof of Theorem 3.3 is somewhat "backwards" compared to the proof of MDA, where the model gap $\phi^\alpha(y_k) - \min_{x\in\mathbb{R}^n} \Gamma_k(x)$ was directly bounded. First, we prove a convergence rate for the primal-dual gap $\phi^\alpha(\tilde{y}_k) + \psi^\alpha(z_k)$ in Proposition B.4 (where $\tilde{y}_k$ is defined in (38)).

We then use the primal-dual convergence to bound the single-cut model gap via a modification of standard results in GCG analysis (see Lemmas B.1 and B.2).

The convergence results for the one-average algorithm pair exhibit a notable asymmetry. While Theorems 3.2 and 3.3 show primal convergence in an average iterate (y_k for the former, $\tilde{y}_k$ for the latter), this sequence never explicitly appears in either algorithm. Instead, gradient evaluations and updates occur using the "regularized best response" sequence $x_k = \nabla(h^\alpha)^*(-z_k)$.

In the next section, we consider a fully symmetric dynamic where gradients are evaluated at the true test sequence $\{y_k\}$.

4 A Two-Average Method

In this section, we study a method which utilizes two averages, one in the primal and one in the dual, shown in Algorithm 3, which has been previously studied in [ZZ23]. Full proofs can be found in Appendix C. Comparing to Algorithm 1, there are two primary differences. First, gradients are evaluated at the smoothed sequence $\{y_k\}$ rather than the best-response sequence $\{x_k\}$. Since $\{y_k\}$ again serves as the primal test point sequence, as Theorem 4.2 will show, Algorithm 3 addresses the asymmetry noted in the prior section. Second, the order of updates is "lagged" in comparison to Algorithm 1. The dual average s_k, which includes $\nabla f(y_{k-1})$ instead of the available $\nabla f(y_k)$, is used to compute x_{k+1}. Two-average methods have also been proposed in the nonsmooth setting [NS15] which do not have this "staggered" scheme. However, the delayed updates of Algorithm 3 enable the use of simple smoothness inequalities in both primal and dual, as shown in Lemma C.1.

Algorithm 3 AggGCG (Primal)

Initialize: $y_0 \in \operatorname{dom} h$, $s_0 \in \mathbb{R}^n$, $\eta = \alpha/(L+\alpha)$

for $k \geq 0$ **do**

1. Compute

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \langle s_k, x \rangle + h^\alpha(x) \right\}.$$

2. Compute

$$y_{k+1} = (1-\eta)y_k + \eta x_{k+1}, \tag{12}$$
$$s_{k+1} = (1-\eta)s_k + \eta \nabla f(y_k). \tag{13}$$

end for

Algorithm 4 AggGCG (Dual)

Initialize: $z_0 \in \operatorname{dom} f^*$, $v_0 \in \mathbb{R}^n$, $\eta = \alpha/(L+\alpha)$

for $k \geq 0$ **do**

1. Compute

$$\bar{z}_{k+1} = \operatorname*{argmin}_{z \in \mathbb{R}^n} \left\{ -\langle v_k, z \rangle + f^*(z) \right\}.$$

2. Compute

$$z_{k+1} = (1-\eta)z_k + \eta \bar{z}_{k+1},$$
$$v_{k+1} = (1-\eta)v_k + \eta \nabla(h^\alpha)^*(-z_k).$$

end for

While it may not seem obvious at first glance, the update ordering makes Algorithm 3 *self-dual*, that is, its dual algorithm utilizes exactly the same updates, as shown in Algorithm 4.

Proposition 4.1. Set $s_0 = z_0$ and $v_0 = y_0$. Then, on every iteration $k \geq 0$ of Algorithm 3 and Algorithm 4, we have the following correspondence

$$y_k = v_k, \quad s_k = z_k, \quad x_{k+1} = \nabla(h^\alpha)^*(-z_k), \quad \nabla f(y_k) = \bar{z}_{k+1}. \tag{14}$$

Since Algorithm 3 is self-dual, we only state convergence results for the primal variant for simplicity of presentation.

Theorem 4.2. Given $\varepsilon > 0$, choosing $\alpha = \varepsilon/(2M)$, then Algorithm 3 computes a pair (y_k, s_k) satisfying $\phi^\alpha(y_k) + \psi^\alpha(s_k) \leq \varepsilon/2$ in $k = \tilde{\mathcal{O}}(1 + ML\varepsilon^{-1})$ iterations. Consequently, the complexity to obtain an ε-solution to (1) is $\tilde{\mathcal{O}}(1 + ML\varepsilon^{-1})$.

Proof Sketch: Our argument is largely similar to [ZZ23]. We begin by using smoothness of f and $(h^\alpha)^*$ and the strong convexity of h^α, f^* to provide a series of inequalities. Exploiting the primal-dual relations between the pairs (x_{k+1}, s_k) and $(y_k, \nabla f(y_k))$ allows us to eliminate terms and, using our choice of α, conclude with the single-step bound

$$\phi^\alpha(y_{k+1}) + \psi^\alpha(s_{k+1}) \leq (1-\eta)(\phi^\alpha(y_k) + \psi^\alpha(s_k)),$$

which provides the complexity result when recursively expanded with our choice of α.

To make the correspondence in Proposition 4.1 concrete, we revisit the zero-sum game (6). Unlike the one-average case, which corresponds to sequential play, the correspondence in (14) corresponds to *simultaneous play*. The game begins with strategies y_0 and z_0 for the primal and dual players, respectively. Each round then proceeds in like manner, illustrated below. Updates in the same column are interpreted to occur simultaneously.

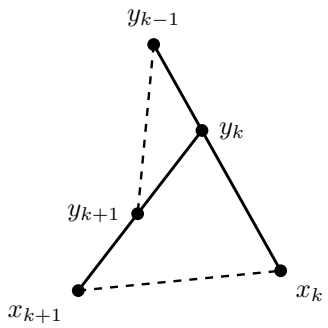
$$(z_0 = s_0,\ y_0 = v_0) \to \begin{matrix} \textbf{(a)} \\ \begin{pmatrix} \bar{z}_{k+1} = \nabla f(y_k) \\ x_{k+1} = \nabla (h^\alpha)^*(-z_k) \end{pmatrix} \end{matrix} \rightarrow \begin{matrix} \textbf{(b)} \\ \begin{pmatrix} z_{k+1} = s_{k+1} \\ y_{k+1} = v_{k+1} \end{pmatrix} \end{matrix}$$

Each round begins in **(a)** with the players simultaneously predicting the opposing strategy and determining their own response based on the previous mixed strategies. The responses/predictions are then used to update the mixed strategies in **(b)**, which are then used in the next round. The simultaneous play dynamics of Algorithm 3 were also examined in [ZZ23], where the authors demonstrated that Algorithm 3 is in fact equivalent to a deterministic variant of the popular Logistic Fictitious Play [FK93] algorithm.

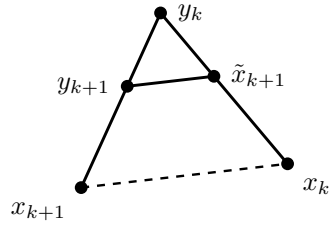
Theorem 4.2 guarantees convergence in the primal-dual gap $\phi^\alpha(y_k) + \psi^\alpha(s_k)$, however it does not provide guarantees for a computable primal-dual certificate. While a more careful analysis may provide a suitable certificate, the difference from the one-average Algorithm 1 (which provides a natural certificate) is notable and admits a simple geometric explanation. In Algorithm 1, evaluating the gradient at x_k naturally allows us to control $\ell_f(x_{k+1}; x_k) + \frac{\alpha(1-\eta)}{2\eta}\|x_k - x_{k+1}\|^2$ by smoothness and a suitable choice of η. However, for Algorithm 3, there is no such trivial relationship between the linearization point y_{k-1}, the test point y_{k+1}, and the exact minimizers x_{k+1} and x_k, as visualized in Figure 1(a). We therefore cannot straightforwardly exploit the smoothness of f, as done in the proof of Theorem 3.2.

5 A Three-Average Method

Motivated by the shortcomings of AggGCG, we propose TAA, and examine its dual counterpart. TAA restores convergence guarantees for a computable certificate while achieving optimal complexity like Nesterov's accelerated gradient (AG) method. Full proofs can be found in Appendix D.



(a) Geometry of the AggGCG update.



(b) Geometry of the TAA update.

Figure 1: Geometric illustration of the AggGCG and TAA updates. In AggGCG, there is no trivially exploitable relationship between the minimizer displacement $x_{k+1}-x_k$ and the linearization point displacement $y_{k+1}-y_{k-1}$. In TAA, the linearization point $\tilde{x}_{k+1}$ restores a simple geometric relationship by interpolating between the last test point y_k and minimizer x_k. As a result, the magnitude of the vectors $y_{k+1}-\tilde{x}_{k+1}$ and $x_{k+1}-x_k$ can be related by a λ-dependent rescaling.

Algorithm 5 TAA

Initialize: given $y_0 \in \operatorname{dom} h$, $L > 0$, $\alpha > 0$, set $s_0 = \nabla f(y_0)$ and compute

$$x_0 = \operatorname*{argmin}_{x\in\mathbb{R}^n} \left\{\langle s_0, x\rangle + h^\alpha(x)\right\}, \quad \lambda = \frac{2\alpha}{\alpha + \sqrt{\alpha^2 + 4L\alpha}}. \tag{15}$$

for $k \geq 0$ **do**

1. Compute

$$\tilde{x}_{k+1} = (1-\lambda)y_k + \lambda x_k, \tag{16}$$
$$s_{k+1} = (1-\lambda)s_k + \lambda \nabla f(\tilde{x}_{k+1}), \tag{17}$$
$$x_{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^n} \left\{\langle s_{k+1}, x\rangle + h^\alpha(x)\right\}, \tag{18}$$
$$y_{k+1} = (1-\lambda)y_k + \lambda x_{k+1}. \tag{19}$$

end for

Instead of linearizing at the previous test point y_{k-1}, TAA linearizes at a convex combination point $\tilde{x}_{k+1}$. As visualized in Figure 1(b), the triangles $(y_k, y_{k+1}, \tilde{x}_{k+1})$ and (y_k, x_{k+1}, x_k) are clearly similar. Furthermore, the rescaling coefficient is dependent only on the convex combination parameter λ. As a result, we obtain a simple geometric relation when analyzing Algorithm 5. Note that TAA involves three averaging steps: two in the primal and one in the dual. Despite differences from existing variants of the AG method, TAA still achieves nearly optimal complexity for CSCO, as shown below.

Theorem 5.1. Given $\varepsilon > 0$, choosing $\alpha = \varepsilon/(2M)$, then the pair (y_k, Γ_k) is an ε-certificate for (1) in $k = \tilde{\mathcal{O}}(1 + \sqrt{ML/\varepsilon})$ iterations of Algorithm 5, where Γ_k is the ACP model for (2) induced by $(y_0, \{\tilde{x}_{i+1}\}_{i=0}^{k-1}, \{\lambda\}^k)$.

Proof Sketch: The proof of Theorem 5.1 follows a similar pattern as Theorem 3.2. However, instead of showing a recursive inequality in the $\{x_k\}$ sequence, we show the inequality

$$\Gamma_{k+1}(x_{k+1}) \geq \phi^\alpha(y_{k+1}) - (1-\lambda)^{k+1}(\phi^\alpha(y_0) - \Gamma_0(x_0)),$$

which directly results in a convergence rate. The complexity bound follows from our choice of λ.

Theorem 5.1 demonstrates that Algorithm 5 achieves near-optimal, accelerated complexity for solving (1). However, the algorithm is substantially different from other acceleration schemes based on proximal mappings. To clarify the source of TAA's acceleration, we consider the GEM [LZ18] applied to the dual problem (3), shown in Algorithm 6. To use the GEM, we need a suitable Bregman regularizer. Accordingly, we assume that the function $\nu^* : \mathbb{R}^n \to (-\infty, \infty]$ is 1-strongly convex with respect to the dual norm $\|\cdot\|_*$ and is L-smooth relative to f^*[3] with $\operatorname{dom} f^* \subseteq \operatorname{dom} \nu^*$. Note that either $\nu^* = Lf^*$ or $\nu^* = \frac{1}{2}\|\cdot\|_*^2$ would satisfy these requirements.

Algorithm 6 GEM

Initialize: given $\alpha > 0$ and $g_0 \in \operatorname{dom} f^*$, set $z_0 = g_0$, $v_{-1} = v_0 = \nabla(h^\alpha)^*(-g_0)$, $\tau_0 = \alpha/L$, $A_0 = 1$, and $a_{-1} = 0$

for $k \geq 0$ **do**

1. Compute

$$\tau_{k+1} = \tau_k + \frac{\alpha a_k}{L}, \quad a_k = \frac{\tau_k + \sqrt{\tau_k^2 + 4\tau_k A_k}}{2}, \tag{20}$$

$$A_{k+1} = A_k + a_k, \quad \hat{v}_k = v_k + \frac{a_{k-1}}{a_k}(v_k - v_{k-1}). \tag{21}$$

2. Compute

$$g_{k+1} = \operatorname*{argmin}_{g \in \mathbb{R}^n} \left\{ a_k[\langle -\hat{v}_k, g\rangle + f^*(g)] + \frac{\tau_k}{\alpha} \operatorname{D}_{\nu^*}(g\|g_k) \right\}, \tag{22}$$

$$z_{k+1} = \frac{A_k}{A_{k+1}} z_k + \frac{a_k}{A_{k+1}} g_{k+1}, \quad v_{k+1} = \nabla(h^\alpha)^*(-z_{k+1}). \tag{23}$$

end for

The GEM is known to have the same optimal complexity as the AG method. The next result presents the complexity of our strongly convex GEM variant for computing an ε-certificate to (1).

Theorem 5.2. Assume that $\max_{g \in \operatorname{dom} f^*} \operatorname{D}_{\nu^*}(g\|g_0) = D < \infty$ is bounded. Given $\varepsilon > 0$ and choosing $\alpha = \varepsilon/(2M)$, then the pair (z_k, Γ_k^*) is an ε-certificate for (1) in $k = \tilde{\mathcal{O}}(1 + \sqrt{ML/\varepsilon})$ iterations of Algorithm 6, where Γ_k^* is the ACP model for (3) induced by $(z_0, \{z_{i+1}\}_{i=0}^{k-1}, \{a_i/A_{i+1}\}_{i=0}^{k-1})$.

Proof Sketch. The first part of the proof of Theorem 5.2 uses the smoothness of $(h^\alpha)^*$, the optimality conditions on g_{k+1}, and Fenchel duality to prove the single-step bound

$$\begin{aligned} A_k \psi^\alpha(z_k) - \frac{\alpha a_k^2}{2\tau_{k+1}} \|v_{k+1} - v_k\|^2 + \frac{\tau_k}{\alpha} \operatorname{D}_{\nu^*}(g\|g_k) - \frac{\tau_k}{\alpha} \operatorname{D}_{\nu^*}(g_{k+1}\|g_k) - \frac{\tau_{k+1}}{\alpha} \operatorname{D}_{\nu^*}(g\|g_{k+1}) \\ \geq A_{k+1}\psi^\alpha(z_{k+1}) + a_k\left[h^\alpha(v_{k+1}) + \langle v_{k+1}, g\rangle - f^*(g)\right] + a_k\langle v_{k+1} - \hat{v}_k, g_{k+1} - g\rangle. \end{aligned}$$

Recursively summing this inequality from 0 to $k-1$ and applying the strong convexity of ν^*, the construction of $\hat{v}_k$, and duality relations yield the certificate convergence bound.

We finally show the most fascinating duality result of the paper: Algorithm 6 is, in fact, the dual of Algorithm 5, clarifying the source of TAA's acceleration.

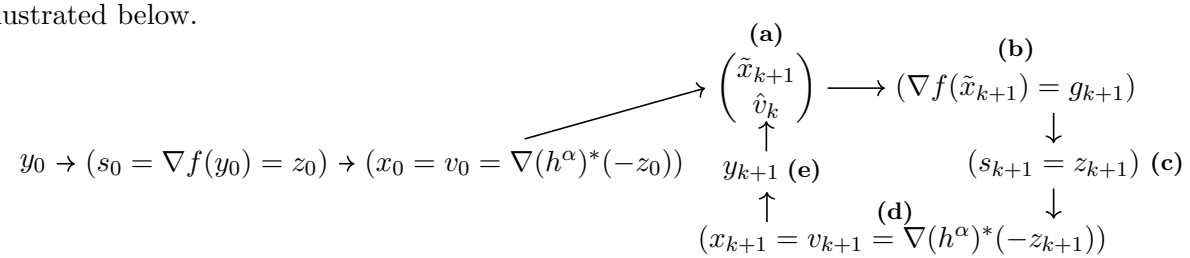
Proposition 5.3. Fix $y_0 \in \operatorname{dom} h$ and suppose that we choose $\nu^* = Lf^*$ in Algorithm 6. Then, setting $s_0 = z_0 = g_0 = \nabla f(y_0)$, $\tilde{x}_0 = y_0$, Algorithm 6 solving (3) is equivalent to Algorithm 5

[3] That is, $Lf^*(\cdot) - \nu^*(\cdot)$ is a convex function. Note that this definition does not assume that ν^* is differentiable, as in [LFN18]. Equivalently, we can say that f^* is L^{-1}-strongly convex relative to ν^*.

solving (2) with the following correspondence [4]

$$\nabla f(\tilde{x}_k) = g_k, \quad s_k = z_k, \quad x_k = v_k.$$

We can once again examine the duality relation in Proposition 5.3 from the game-theoretic perspective of (6). Unlike Algorithm 3, we move back into sequential play with the turn structure illustrated below.



The primal player moves first with the starting strategy y_0. The dual player responds by setting their initial strategy $z_0 = \nabla f(y_0)$ and the primal player responds with $x_0 = \nabla(h^\alpha)^*(-z_0)$. Then, play begins. In **(a)**, the primal player begins by computing the "lookahead" strategy $\tilde{x}_{k+1}$ which acts as a prediction for their next strategy (y_{k+1}) before seeing any dual response. Simultaneously, the dual player forms an optimistic model of the primal player response using the extrapolated point $\hat{v}_k$. In **(b)**, the primal player then predicts the dual player's move as $\nabla f(\tilde{x}_{k+1})$, while the dual player performs an optimistic update to compute the response g_{k+1}. As we show in the proof of Proposition 5.3, the optimistic response g_{k+1} exactly coincides with the lookahead prediction $\nabla f(\tilde{x}_{k+1})$. Next, in **(c)** the dual player updates their mixed strategy z_{k+1}, the primal player computes their response x_{k+1} in **(d)**, and finally the primal player updates their mixed strategy y_{k+1} in **(e)**. The dual player does not directly see the sequence $\{y_k\}$, however as the proof of Proposition 5.3 shows, the optimistic prediction $\hat{v}_k$ and response g_{k+1} implicitly provide information about y_k.

In comparison with the one and two-average methods, which use zero-regret dynamics based on averaging and best-response, the three-average case leverages optimistic learning dynamics: the GEM directly from the $\hat{v}_k$ extrapolation and TAA indirectly through the underlying dual process.

6 Connections to Fisher Market Dynamics

The motivation for the algorithmic progression MDA→AggGCG→TAA was primarily technical: first to show convergence in the linearization sequence, then to recover optimality certificates. In this section, we show that the progression also has an intuitive appeal grounded in Fisher market dynamics. We consider a smoothed variation of the Fisher market equilibrium with linear utility functions proposed by [CJS25][5],

$$\phi^\alpha(\mu) = \overbrace{\sum_{j=1}^{n} \exp(\mu_j) + \delta \sum_{i=1}^{m} B_i \log\left(\sum_{j=1}^{n} \exp\big(\delta^{-1}(\log b_{ij} - \mu_j)\big)\right)}^{f(\mu)} + \overbrace{\delta_{[\underline{\mu},\overline{\mu}]^n}(\mu)}^{h(\mu)} + \overbrace{\frac{\alpha}{2}\|\mu - \mu_{\text{ref}}\|^2}^{\alpha w(\mu)},$$

where μ_j is the logarithmic unit price of item j, b_{ij} is buyer i's valuation for item j, B_i is buyer i's budget, $\underline{\mu} \le \overline{\mu}$ define box constraints, and $\delta > 0$ is an entropic smoothing parameter. The added

[4] Since f^* is not necessarily differentiable, there is some ambiguity in the definition of $\mathrm{D}_{\nu^*}(g\|g_k)$ in Step **2** of Algorithm 6. The choice used in the inductive proof of Proposition 5.3 is to define $\mathrm{D}_{f^*}(g\|g_k)$ using $\tilde{x}_k \in \partial f^*(g_k)$.

[5] The original formulation in [CJS25] did not contain the α perturbation

regularization term admits a natural interpretation as a penalization for multiplicative deviations from a benchmark price. In this example, *primal variables are prices, dual variables represent demand* (incorporating individual allocations from each buyer).

MDA (One-Average). In each round, the market generates provisional prices x_k by a best-fit approximation to smoothed demand s_k. Buyers then respond to this *provisional* price in each round with the proposed demand $\nabla f(x_k)$. However, Theorem 3.2 does not show convergence in the provisional prices, but instead the exponentially smoothed price sequence $\{y_k\}$. The market dynamics (Algorithm 1) are therefore operationally asymmetric, with the market responding to a smoothed demand model s_k without explicitly posting the smoothed price y_k.

AggGCG (Two-Averages). Algorithm 3 remedies the asymmetry in MDA market dynamics, but loses convergence guarantees for optimality certificates. In the MDA results, the ACP model included the latest buyer allocation $\nabla f(x_{k+1})$. As a result, there was an explicit, primal-dual relationship between the latest provisional price x_k and smoothed demand forecast s_k. Algorithm 3 breaks this connection by linearizing at y_{k-1} to generate x_k, which we can interpret as the buyers responding to stale prices y_{k-1} from round $k-1$. Similarly, the market response corresponds to the smoothed demand s_{k-1} from the previous round. As a result, we lose the primal-dual connections between the latest price/demand pair that admit a natural certificate.

TAA (Three-Averages). The technical motivation for a third average was primarily geometric as visualized in Figure 1(b). The move to TAA also admits a more interesting market interpretation as *forecasting*. Each round k of Algorithm 5 begins by predicting the next set of prices $\tilde{x}_{k+1}$. Buyers respond to the forecast price signal with $\nabla f(\tilde{x}_{k+1})$, which is then incorporated into the smoothed demand s_{k+1}. The market computes the best-fit prices x_{k+1} for demand s_{k+1}, which are then added to the smoothed price y_{k+1}. In this dynamic, buyers do not directly respond to the convergent price sequence $\{y_k\}$, but instead to the primal-side price forecasts $\{\tilde{x}_{k+1}\}$. We can then interpret TAA as a cautious forecast model, where both primal and dual-side forecasts are hedged by averaging. Moreover, with the additional forecast sequence $\{\tilde{x}_{k+1}\}$, TAA achieves market equilibrium at faster rates than MDA and AggGCG.

7 Conclusion

In this work, we explore primal-dual correspondences in regularized convex optimization with a focus on certifiable optimality. Motivated by primal-dual correspondences in one and two averages, we propose the TAA method which leverages three averages to obtain near-optimal complexity for solving (1). While TAA is new, we show that its dual counterpart is the well-studied GEM. We show that both TAA and the GEM obtain computable ε-certificates for primal-dual gap convergence with $\tilde{\mathcal{O}}(\varepsilon^{-1/2})$ complexity, a novel result for the GEM to our knowledge. Our certificate guarantees further enhance the utility for the GEM as a subsidiary solver by providing upper bounds on a computable termination condition. Furthermore, we provide in-depth intuition and motivation for our results with concrete examples in Fisher markets and zero-sum matrix games.

There are a number of directions for future work. Our analysis relies on knowledge of the problem smoothness L. However, L may be unknown at runtime, particularly for large-scale problem instances. Future work on universal variants of TAA would substantially improve its scope of application. Similarly, our analysis focused on *smooth* problems, whereas an increasing number of methods target the class of *relatively smooth* problems [LFN18]. Recent work [LTP23] has generalized the typical smoothness/strong-convexity duality relation to the relative case, suggesting that algorithmic correspondences may exist under these generalized relations.

A Technical Results

We begin by providing the proof of Lemma 2.1, which connects the regularized problem (2) to the true target (1).

Proof of Lemma 2.1: By Fenchel-Rockafellar duality, we have

$$\psi^\alpha(s) \geq \min_{z\in\mathbb{R}^n}\{\psi^\alpha(z)\} \overset{(3)}{=} -\phi_*^\alpha,$$

which implies that

$$\phi^\alpha(x) - \phi_*^\alpha \leq \phi^\alpha(x) + \psi^\alpha(s). \tag{24}$$

Let $x_* \in \operatorname*{Argmin}_{x\in\operatorname{dom} h} \phi(x)$. Note that, by the definition of ϕ_*^α and the boundedness of $w(x)$ on $\operatorname{dom} h$,

$$\phi_*^\alpha \leq \phi(x_*) + \alpha w(x_*) \leq \phi_* + \alpha M. \tag{25}$$

This inequality and the non-negativity of w thus yield

$$\phi(x) - \phi_* - \alpha M \overset{(25)}{\leq} \phi(x) + \alpha w(x) - \phi_*^\alpha \overset{(24)}{\leq} \phi^\alpha(x) + \psi^\alpha(s) \leq \frac{\varepsilon}{2},$$

where the second inequality follows from (24). By the choice $\alpha \leq \varepsilon/(2M)$, we obtain

$$\phi(x) - \phi_* \leq \frac{\varepsilon}{2} + \alpha M \leq \varepsilon,$$

which completes the proof. ∎

Next, we recall technical results relating a closed convex function f to its Fenchel dual f^*.

Lemma A.1 ([Bec17, Theorem 4.20])**.** Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be proper, closed, and convex. Then, the following are equivalent for any x, $y \in \mathbb{R}^n$:

a) $\langle x, y\rangle = f(x) + f^*(y)$,

b) $y \in \partial f(x)$,

c) $x \in \partial f^*(y)$.

Lemma A.2 ([Bec17, Corollary 4.21])**.** Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be proper, closed, and convex. Then, for any $x, y \in \mathbb{R}^n$:

$$\partial f(x) = \operatorname*{Argmax}_{z\in\mathbb{R}^n}\{\langle z, x\rangle - f^*(z)\}, \quad \partial f^*(y) = \operatorname*{Argmax}_{u\in\mathbb{R}^n}\{\langle y, u\rangle - f(u)\}.$$

With basic duality results established, we now provide a proof of Lemma 2.3.

Proof of Lemma 2.3: Statements (a) and (b) follow by the definition of Γ_k in (4), the convexity of f, and the α-strong convexity of h^α.

c) Define $v_k = \operatorname{argmin}_{x\in\mathbb{R}^n}\Gamma_k(x)$. Then, by the definitions of Γ_k and s_k in (4) and (5), respectively, we can show

$$0 \in \partial\Gamma_k(v_k) = s_k + \partial h^\alpha(v_k),$$

which implies that $-s_k \in \partial h^\alpha(v_k)$. By Lemma A.1, we obtain

$$h^\alpha(v_k) + \langle s_k, v_k\rangle = -(h^\alpha)^*(-s_k). \tag{26}$$

Similarly, we observe by Lemma A.1 that

$$f(x_i) - \langle \nabla f(x_i), x_i \rangle = -f^*(\nabla f(x_i)), \tag{27}$$

for all $i \in \{0, \dots, k-1\}$. Define the sequence of scalars $\{\chi_i\}_{i=0}^k$ recursively as

$$\chi_0 = f(y_0) - \langle \nabla f(y_0), y_0 \rangle, \quad \chi_{i+1} = (1-\zeta_i)\chi_i + \zeta_i\big(f(x_i) - \langle \nabla f(x_i), x_i \rangle\big).$$

Expanding the inequality, using (27) and the definition of s_k in (5), and applying the convexity of f^* we obtain

$$\chi_k \leq -f^*(s_k). \tag{28}$$

Then, by the definitions of v_k and Γ_k, we have

$$-\min_{x \in \mathbb{R}^n} \Gamma_k(x) = -\Gamma_k(v_k) = -\langle s_k, v_k \rangle - h^\alpha(v_k) - \chi_k \overset{(26)(28)}{\geq} (h^\alpha)^*(-s_k) + f^*(s_k) = \psi^\alpha(s_k).$$

Therefore, statement (c) immediately follows. ∎

For completeness, we state the counterpart of Lemma 2.3 for a dual ACP model. The proof is identical to the primal case, and is therefore omitted.

Lemma A.3. Let $\Gamma_k^*(\cdot)$ be the ACP model for (3) induced by $(z_0, \{g_i\}_{i=0}^{k-1}, \xi)$ for $z_0 \in \operatorname{dom} f^*$, $\{g_i\}_{i=0}^{k-1} \subseteq \operatorname{dom} f^*$ and $\xi \in [0,1]^k$. Then, the following statements hold:

a) for all $g \in \operatorname{dom} f^*$, $\Gamma_k^*(g) \leq \psi^\alpha(g)$;

b) Γ_k^* is $(1/L)$-strongly convex.

Furthermore, define $\{v_k\}_{k \geq 0}$ as

$$v_0 = \nabla (h^\alpha)^*(-z_0), \quad v_{j+1} = (1-\xi_j)v_j + \xi_j \nabla (h^\alpha)^*(-g_j),$$

for $0 \leq j \leq k-1$. Then, the following bound holds for all $g \in \operatorname{dom} f^*$

c) $\phi^\alpha(v_k) + \psi^\alpha(g) \leq \psi^\alpha(g) - \min_{u \in \mathbb{R}^n} \Gamma_k^*(u)$.

Finally, we provide a technical lemma that will be used in the analysis of Algorithm 6. The statement is a slight generalization of [CT06, Lemma 3.2] to the case of relative strong convexity and the case when ν is not necessarily differentiable. Non-differentiability does not affect our results, as our analysis of Algorithm 6 does not require any properties of the Bregman function aside from the 1-strong convexity of ν^*.

Lemma A.4 (Three-Points Inequality)**.** Let $\Phi : \mathbb{R}^n \to (-\infty, \infty]$ and $\omega : \mathbb{R}^n \to (-\infty, \infty]$ be closed, proper, and convex functions with $\operatorname{dom} \Phi \subseteq \operatorname{dom} \omega$ and $\operatorname{dom} \Phi$ has a nonempty relative interior. Assume that Φ is μ-strongly convex relative to ω, that is, $\Phi(\cdot) - \mu\omega(\cdot)$ is a convex function. Fix $x_0 \in \operatorname{dom} \Phi$, $\beta > 0$ and define

$$x^+ = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \Phi(x) + \beta \, \mathrm{D}_\omega(x \| x_0) \right\},$$

where $\mathrm{D}_\omega(x\|x_0)$ is defined using an arbitrary $\omega'(x_0) \in \partial\omega(x_0)$. Then, for any $u \in \operatorname{dom} \Phi$,

$$\Phi(x^+) + \beta \, \mathrm{D}_\omega(x^+ \| x_0) + (\beta + \mu) \, \mathrm{D}_\omega(u \| x^+) \leq \Phi(u) + \beta \, \mathrm{D}_\omega(u \| x_0),$$

where $\mathrm{D}_\omega(\cdot\|x^+)$ is defined using some suitable $\omega'(x^+) \in \partial\omega(x^+)$.

Proof: Define $\varphi(\cdot) = \Phi(\cdot) + \beta \, \mathrm{D}_\omega(\cdot \| x_0)$. By the optimality condition on x^+, we have $0 \in \partial\varphi(x^+)$. Since $\Phi(\cdot) - \mu\omega(\cdot)$ is convex, we have that $\varphi(\cdot) - (\beta+\mu)\omega(\cdot)$ is convex. By [Bec17, Theorem 3.40], we have for all $x \in \operatorname{dom}\varphi$,

$$(\beta+\mu)\partial\omega(x) + \partial[\varphi(\cdot) - (\beta+\mu)\omega(\cdot)](x) = \partial\varphi(x).$$

Then, using that $0 \in \partial\varphi(x^+)$ and rearranging, we have

$$-(\beta+\mu)\omega'(x^+) \in \partial(\varphi(\cdot) - (\beta+\mu)\omega(\cdot))(x^+)$$

for some $\omega'(x^+) \in \partial\omega(x^+)$. It follows from the definition of the subdifferential that

$$\varphi(x^+) - (\beta+\mu)\omega(x^+) - (\beta+\mu)\langle\omega'(x^+), u - x^+\rangle \le \varphi(u) - (\beta+\mu)\omega(u).$$

Substituting the definition of $\varphi(\cdot)$ and noting

$$\omega(u) - \omega(x^+) - \langle\omega'(x^+), u - x^+\rangle = \mathrm{D}_\omega(u\|x^+)$$

gives the result. ∎

B One-Average Analysis

We begin by providing a proof of the algorithm correspondence in Proposition 3.1.
Proof of Proposition 3.1: First, note that if $s_k = z_k$ for $k \ge 0$, then

$$\begin{aligned} x_k &= \operatorname*{argmin}_{x\in\mathbb{R}^n} \{\langle s_k, x\rangle + h^\alpha(x)\} \\ &= \operatorname*{argmax}_{x\in\mathbb{R}^n} \{\langle -s_k, x\rangle - h^\alpha(x)\} = \nabla(h^\alpha)^*(-s_k) = \nabla(h^\alpha)^*(-z_k). \end{aligned} \tag{29}$$

Similarly, by the optimality condition on $\bar{z}_k$, the convexity of f and Lemma A.2, we have

$$\bar{z}_k = \operatorname*{argmax}_{z\in\mathbb{R}^n} \{\langle \nabla(h^\alpha)^*(-z_k), z\rangle - f^*(z)\} = \nabla f(\nabla(h^\alpha)^*(-z_k)) \overset{(29)}{=} \nabla f(x_k). \tag{30}$$

Therefore, it is sufficient to prove that $s_k = z_k$ for all $k \ge 0$, which we will show by induction. The base case is trivial by our choice of $z_0 = \nabla f(y_0) = s_0$. Now, assume $k \ge 0$ and the inductive hypothesis $s_k = z_k$, which implies (29) and (30). Then,

$$z_{k+1} = (1-\eta)z_k + \eta\bar{z}_k \overset{(30)}{=} (1-\eta)s_k + \eta\nabla f(x_k) = s_{k+1},$$

where the second equality follows from the inductive hypothesis and (30). We thus complete the proof. ∎

B.1 Analysis of MDA

In this subsection, we prove the MDA certificate complexity claimed in Theorem 3.2.

To begin, we define Γ_k as the ACP model induced by $(y_0, \{x_i\}_{i=0}^{k-1}, \{\eta\}^k)$. That is, $\Gamma_0(\cdot) = \ell_f(\cdot; y_0) + h^\alpha(\cdot)$ and for all $k \ge 0$,

$$\Gamma_{k+1}(\cdot) = (1-\eta)\Gamma_k(\cdot) + \eta\gamma_k(\cdot), \quad \gamma_k(\cdot) = \ell_f(\cdot; x_k) + h^\alpha(\cdot).$$

Observe that, by construction, we have for all $x \in \operatorname{dom} h$

$$s_k + \partial h^\alpha(x) = \partial \Gamma_k(x),$$

hence $x_k = \operatorname{argmin}_{x \in \mathbb{R}^n} \Gamma_k(x)$ by the optimality conditions on x_k in Algorithm 1. For convenience, we define the quantities for $k \geq 0$

$$m_k = \min_{x \in \mathbb{R}^n} \Gamma_k(x) = \Gamma_k(x_k), \quad t_k = \phi^\alpha(y_k) - m_k,$$

where

$$y_{k+1} = (1-\eta) y_k + \eta x_{k+1}.$$

Observe that y_k does not directly include the point x_0, however the dual average s_k includes gradient information from x_0.

Proof of Theorem 3.2: By the definition of Γ_{k+1} and m_{k+1}, for $k \geq 0$ we obtain

$$\begin{aligned}
m_{k+1} &= (1-\eta)\Gamma_k(x_{k+1}) + \eta \gamma_k(x_{k+1}) \\
&\geq (1-\eta) m_k + \eta \gamma_k(x_{k+1}) + \frac{(1-\eta)\alpha}{2} \|x_k - x_{k+1}\|^2 && \text{(i)} \\
&= (1-\eta) m_k + \eta \left[\gamma_k(x_{k+1}) + \frac{(1-\eta)\alpha}{2\eta} \|x_k - x_{k+1}\|^2 \right] \\
&= (1-\eta) m_k + \eta \left[\gamma_k(x_{k+1}) + \frac{L}{2} \|x_k - x_{k+1}\|^2 \right] && \text{(ii)} \\
&\geq (1-\eta) m_k + \eta \phi^\alpha(x_{k+1}), && \text{(iii)}
\end{aligned}$$

where (i) follows by the 1-strong convexity of w, (ii) by the choice $\eta = \alpha/(\alpha + L)$, and (iii) by the L-smoothness of f. Then, expanding the inequality from 1 to $k+1$, we have

$$\begin{aligned}
m_{k+1} &\geq (1-\eta)^{k+1} m_0 + \eta \sum_{i=1}^{k+1} (1-\eta)^{k+1-i} \phi^\alpha(x_i) \\
&= -(1-\eta)^{k+1} t_0 + (1-\eta)^{k+1} \phi^\alpha(y_0) + \eta \sum_{i=1}^{k+1} (1-\eta)^{k+1-i} \phi^\alpha(x_i) \\
&\geq -(1-\eta)^{k+1} t_0 + \phi^\alpha(y_{k+1}),
\end{aligned}$$

where the second line follows by the definition of t_0, and the third by the convexity and the definition of y_{k+1} and induction. Therefore, we obtain

$$t_{k+1} = \phi^\alpha(y_{k+1}) - m_{k+1} \leq (1-\eta)^{k+1} t_0.$$

Since $1 - \eta = L/(\alpha + L) = (1 + \alpha/L)^{-1}$, we have for all $k \geq 0$

$$\phi^\alpha(y_k) - \min_{x \in \mathbb{R}^n} \Gamma_k(x) \leq \frac{t_0}{\left(1 + \frac{\alpha}{L}\right)^k}.$$

As a result, we obtain a point y_k satisfying $\phi^\alpha(y_k) - \min_{x \in \operatorname{dom} h} \Gamma_k(x) \leq \varepsilon/2$ after

$$k = \mathcal{O}\left(1 + \frac{L}{\alpha} \log\left(\frac{t_0}{\varepsilon}\right)\right)$$

iterations. Choosing $\alpha = \varepsilon/(2M)$ with Definition 2.4 yields the complexity result. ∎

B.2 Analysis of GCG

In this subsection, we prove the complexity bound stated in Theorem 3.3.

We begin by defining the Wolfe gap for (3) as

$$S(z) = \max_{v \in \mathbb{R}^n} \{\langle -\nabla(h^\alpha)^*(-z), z - v\rangle + f^*(z) - f^*(v)\}. \tag{31}$$

The following lemma provides a useful relation between the Wolfe gap and the primal-dual gap.

Lemma B.1. For $z \in \operatorname{dom} f^*$, define $y = \nabla(h^\alpha)^*(-z)$. Then, we have

$$S(z) = \psi^\alpha(z) + \phi^\alpha(y).$$

Proof: Using (31) and the definition of y, we have

$$\begin{aligned} S(z) &\overset{(31)}{=} \max_{v \in \mathbb{R}^n} \{\langle -y, z - v\rangle + f^*(z) - f^*(v)\} \\ &= f^*(z) + \langle y, -z\rangle + \max_{v \in \mathbb{R}^n} \{\langle y, v\rangle - f^*(v)\} \\ &= f^*(z) + h^\alpha(y) + (h^\alpha)^*(-z) + f(y) \\ &\overset{(2),(3)}{=} \psi^\alpha(z) + \phi^\alpha(y), \end{aligned}$$

where the third line follows by Lemma A.1(a) and the fourth line follows by the definitions of ϕ^α and ψ^α in (2) and (3), respectively. ∎

The Wolfe gap therefore acts as a primal-dual certificate of optimality, as has been shown in numerous prior works [Lia25, Jag13]. Accordingly, we can explicitly connect the Wolfe gap (31) to the notion of an SCP model, as the following lemma shows. For notational convenience, we define the linearization $\ell_{(h^\alpha)^*}(\cdot; z)$ as

$$\ell_{(h^\alpha)^*}(\cdot; z) = (h^\alpha)^*(-z) + \langle -\nabla(h^\alpha)^*(-z), \cdot - z\rangle. \tag{32}$$

Lemma B.2. Fix $z \in \operatorname{dom} f^*$. Then, we have

$$S(z) = \psi^\alpha(z) - \min_{v \in \mathbb{R}^n} \Gamma^*(v), \tag{33}$$

where $\Gamma^*(\cdot) = \ell_{(h^\alpha)^*}(\cdot; z) + f^*(\cdot)$ is the SCP model of (3) induced by z.

Proof: By the SCP construction in Definition 2.2 with $k = 0$, Γ^* has the simplified, one-cut form

$$\Gamma^*(\cdot) = \ell_{(h^\alpha)^*}(\cdot; z) + f^*(\cdot). \tag{34}$$

Then,

$$\begin{aligned} \psi^\alpha(z) - \min_{v \in \mathbb{R}^n} \Gamma^*(v) &\overset{(34)}{=} \psi^\alpha(z) - \min_{v \in \mathbb{R}^n} \{(h^\alpha)^*(-z) + \langle -\nabla(h^\alpha)^*(-z), v - z\rangle + f^*(v)\} \\ &= f^*(z) - \min_{v \in \mathbb{R}^n} \{\langle -\nabla(h^\alpha)^*(-z), v - z\rangle + f^*(v)\} \\ &= \max_{v \in \mathbb{R}^n} \{\langle -\nabla(h^\alpha)^*(-z), z - v\rangle + f^*(z) - f^*(v)\} \overset{(31)}{=} S(z), \end{aligned}$$

where the first equality follows by the definition of Γ^* and the final equality by the definition of $S(z)$, which concludes the proof. ∎

The following lemma is adapted from [Bec17, Lemma 13.7] to the case where f^* is L^{-1}-strongly convex with our choice of η.

Lemma B.3. Choosing $\eta = \alpha/(L+\alpha)$, we have for all iterations $k \geq 0$

$$\psi^\alpha(z_{k+1}) \leq \psi^\alpha(z_k) - \eta S(z_k).$$

Proof: It follows from the α^{-1}-smoothness of $(h^\alpha)^*(-\cdot)$ and the definition of z_{k+1} in (11) that

$$\begin{aligned}(h^\alpha)^*(-z_{k+1}) \leq & (h^\alpha)^*(-z_k) + \langle -\nabla(h^\alpha)^*(-z_k), z_{k+1} - z_k\rangle + \frac{1}{2\alpha}\|z_{k+1} - z_k\|_*^2 \\ \overset{(11)}{=} & (h^\alpha)^*(-z_k) + \eta\langle -\nabla(h^\alpha)^*(-z_k), \bar{z}_k - z_k\rangle + \frac{\eta^2}{2\alpha}\|\bar{z}_k - z_k\|_*^2. \end{aligned} \tag{35}$$

Similarly, by the L^{-1}-strong convexity of f^*, we have

$$f^*(z_{k+1}) \leq f^*(z_k) + \eta(f^*(\bar{z}_k) - f^*(z_k)) - \frac{\eta(1-\eta)}{2L}\|z_k - \bar{z}_k\|_*^2. \tag{36}$$

Combining the above bounds, we have

$$\begin{aligned}\psi^\alpha(z_{k+1}) \overset{(3)}{=} & (h^\alpha)^*(-z_{k+1}) + f^*(z_{k+1}) \\ \overset{(35)(36)}{\leq} & (h^\alpha)^*(-z_k) + \eta\langle -\nabla(h^\alpha)^*(-z_k), \bar{z}_k - z_k\rangle + \frac{\eta^2}{2\alpha}\|\bar{z}_k - z_k\|_*^2 \\ & + f^*(z_k) + \eta(f^*(\bar{z}_k) - f^*(z_k)) - \frac{\eta(1-\eta)}{2L}\|z_k - \bar{z}_k\|_*^2 \\ = & \psi^\alpha(z_k) - \eta[\langle -\nabla(h^\alpha)^*(-z_k), z_k - \bar{z}_k\rangle + f^*(z_k) - f^*(\bar{z}_k)] \\ & + \frac{1}{2}\left(\alpha^{-1}\eta^2 - L^{-1}\eta(1-\eta)\right)\|z_k - \bar{z}_k\|_*^2 \\ = & \psi^\alpha(z_k) - \eta[\langle -\nabla(h^\alpha)^*(-z_k), z_k - \bar{z}_k\rangle + f^*(z_k) - f^*(\bar{z}_k)], \end{aligned} \tag{37}$$

where the last equality follows from the choice $\eta = \alpha/(L+\alpha)$ and

$$\alpha^{-1}\eta^2 - L^{-1}\eta(1-\eta) = \frac{\alpha}{(L+\alpha)^2} - \frac{\alpha}{(L+\alpha)^2} = 0.$$

Using (10) and (31), we have

$$S(z_k) = \langle -\nabla(h^\alpha)^*(-z_k), z_k - \bar{z}_k\rangle + f^*(z_k) - f^*(\bar{z}_k),$$

which together with (37) proves the claim of the lemma. ∎

Then, we have the following convergence guarantee.

Proposition B.4. Define $\tilde{y}_0 = \nabla(h^\alpha)^*(-z_0)$, and for all $k \geq 0$,

$$\tilde{y}_{k+1} = (1-\eta)\tilde{y}_k + \eta\nabla(h^\alpha)^*(-z_k). \tag{38}$$

Then, for all $k \geq 0$, the primal-dual pair $(\tilde{y}_k, z_k)$ satisfies

$$\psi^\alpha(z_k) + \phi^\alpha(\tilde{y}_k) \leq \frac{\psi^\alpha(z_0) + \phi^\alpha(\tilde{y}_0)}{\left(1 + \frac{\alpha}{L}\right)^k}.$$

Proof: By Lemma B.1, for $k \geq 0$ we have

$$-\eta S(z_k) = -\eta\psi^\alpha(z_k) - \eta\phi^\alpha(\nabla(h^\alpha)^*(-z_k)) \overset{(38)}{\leq} -\phi^\alpha(\tilde{y}_{k+1}) + (1-\eta)\phi^\alpha(\tilde{y}_k) - \eta\psi^\alpha(z_k), \tag{39}$$

where the inequality follows by the convexity of ϕ^α and the definition of $\tilde{y}_{k+1}$ in (38). Applying Lemma B.3, we obtain

$$\psi^\alpha(z_{k+1}) \leq \psi^\alpha(z_k) - \eta S(z_k) \overset{(39)}{\leq} \psi^\alpha(z_k) - \phi^\alpha(\tilde{y}_{k+1}) + (1-\eta)\phi^\alpha(\tilde{y}_k) - \eta\psi^\alpha(z_k),$$

which implies

$$\psi^\alpha(z_{k+1}) + \phi^\alpha(\tilde{y}_{k+1}) \leq (1-\eta)[\psi^\alpha(z_k) + \phi^\alpha(\tilde{y}_k)].$$

Recursively expanding the inequality from 0 to $k-1$ yields

$$\psi^\alpha(z_k) + \phi^\alpha(\tilde{y}_k) \leq (1-\eta)^k[\psi^\alpha(z_0) + \phi^\alpha(\tilde{y}_0)].$$

Using $(1-\eta) = L/(\alpha+L) = (1+\alpha/L)^{-1}$ gives the claimed convergence bound. ∎

We are now ready to prove Theorem 3.3.

Proof of Theorem 3.3: First, note that by (3) and Proposition B.4, we have for all $k \geq 0$

$$\psi^\alpha(z_k) - \psi_*^\alpha \leq \psi^\alpha(z_k) + \phi^\alpha(\tilde{y}_k) \leq \frac{\psi^\alpha(z_0) + \phi^\alpha(\tilde{y}_0)}{(1+\frac{\alpha}{L})^k}. \tag{40}$$

Then, by Lemma B.3, we have

$$0 \leq \psi^\alpha(z_{k+1}) - \psi_*^\alpha \leq \psi^\alpha(z_k) - \psi_*^\alpha - \eta S(z_k) \overset{(33)}{=} \psi^\alpha(z_k) - \psi_*^\alpha - \eta\left(\psi^\alpha(z_k) - \min_{v\in\mathbb{R}^n} \Gamma_k^*(v)\right).$$

where the equality follows by Lemma B.2. Hence, we obtain

$$\psi^\alpha(z_k) - \min_{v\in\mathbb{R}^n} \Gamma_k^*(v) \leq \frac{\psi^\alpha(z_k) - \psi_*^\alpha}{\eta} \overset{(40)}{\leq} \frac{\psi^\alpha(z_0) + \phi^\alpha(\tilde{y}_0)}{\eta(1+\frac{\alpha}{L})^k}.$$

By standard analysis and our choice of η, the complexity to obtain $\psi^\alpha(z_k) - \min_{v\in\mathbb{R}^n} \Gamma_k^*(v) \leq \varepsilon/2$ is therefore

$$k = \mathcal{O}\left(1 + \frac{L}{\alpha}\log\left(\frac{(L+\alpha)(\psi^\alpha(z_0) + \phi^\alpha(\tilde{y}_0))}{\alpha\varepsilon}\right)\right).$$

The result follows by defining the dual certificate (analogous to Definition 2.4) as the pair (z_k, Γ_k^*) and the choice $\alpha = \varepsilon/(2M)$. ∎

C Two-Average Analysis

We begin by providing a formal proof of the self-duality claimed in Proposition 4.1.

Proof of Proposition 4.1: We prove that $y_k = v_k$ and $s_k = z_k$ for all $k \geq 0$ by induction. The base case $k = 0$ follows by our initialization $s_0 = z_0$ and $y_0 = v_0$. Then, assume that for some $k \geq 0$, the equalities $s_k = z_k$ and $y_k = v_k$ hold. First, observe that by Lemma A.2 and the smoothness of $(h^\alpha)^*$,

$$x_{k+1} = \operatorname*{argmax}_{x\in\mathbb{R}^n} \{-\langle s_k, x\rangle - h(x) - \alpha w(x)\} = \nabla(h^\alpha)^*(-s_k) = \nabla(h^\alpha)^*(-z_k). \tag{41}$$

Then, by the choice of v_{k+1} in Step **2** of Algorithm 4, we have

$$y_{k+1} = \eta x_{k+1} + (1-\eta)y_k \overset{(41)}{=} \eta\nabla(h^\alpha)^*(-z_k) + (1-\eta)v_k = v_{k+1}, \tag{42}$$

where the second equality follows by (41) and the inductive hypothesis. Then, we have

$$\bar{z}_{k+1} = \operatorname{argmin}\{\langle -v_k, z\rangle + f^*(z)\} \overset{(42)}{=} \operatorname{argmax}\{\langle y_k, z\rangle - f^*(z)\} = \nabla f(y_k). \tag{43}$$

Finally, the dual average satisfies

$$z_{k+1} = \eta \bar{z}_{k+1} + (1-\eta) z_k \overset{(43)}{=} \eta \nabla f(y_k) + (1-\eta) z_k = \eta \nabla f(y_k) + (1-\eta) s_k = s_{k+1},$$

completing the inductive step. We therefore conclude that the equivalence $(y_k, s_k) = (v_k, z_k)$ holds for all iterations $k \geq 0$. Then, (41) and (43) imply $(x_{k+1}, \nabla f(y_k)) = (\nabla (h^\alpha)^*(-z_k), \bar{z}_{k+1})$ for all $k \geq 0$, concluding the proof. ∎

With the correspondence proven, the rest of the section is devoted to proving the primal-dual gap convergence rates claimed in Theorem 4.2. Our argument and results are similar to previous work [ZZ23], hence we include the proofs primarily for completeness. For simplicity, we adopt the primal perspective of Algorithm 3.

Lemma C.1. For all iterations $k \geq 0$ of Algorithm 3, the following inequalities hold:

a)

$$f(y_{k+1}) \leq (1-\eta) f(y_k) - \eta f^*(\nabla f(y_k)) + \eta \langle \nabla f(y_k), x_{k+1} \rangle + \frac{L\eta^2}{2} \|x_{k+1} - y_k\|^2;$$

b)

$$(h^\alpha)^*(-s_{k+1}) \leq (1-\eta)(h^\alpha)^*(-s_k) - \eta h^\alpha(x_{k+1}) - \eta \langle x_{k+1}, \nabla f(y_k) \rangle + \frac{\eta^2}{2\alpha} \|\nabla f(y_k) - s_k\|_*^2;$$

c)

$$h^\alpha(y_{k+1}) \leq \eta h^\alpha(x_{k+1}) + (1-\eta) h^\alpha(y_k) - \frac{\alpha\eta(1-\eta)}{2} \|x_{k+1} - y_k\|^2;$$

d)

$$f^*(s_{k+1}) \leq \eta f^*(\nabla f(y_k)) + (1-\eta) f^*(s_k) - \frac{\eta(1-\eta)}{2L} \|\nabla f(y_k) - s_k\|_*^2.$$

Proof: a) Since f is L-smooth, we have

$$\begin{aligned}
f(y_{k+1}) \leq & f(y_k) + \langle \nabla f(y_k), y_{k+1} - y_k \rangle + \frac{L}{2} \|y_{k+1} - y_k\|^2 \\
\overset{(12)}{=} & f(y_k) + \eta \langle \nabla f(y_k), x_{k+1} - y_k \rangle + \frac{L\eta^2}{2} \|x_{k+1} - y_k\|^2 \\
= & f(y_k) - \eta \langle \nabla f(y_k), y_k \rangle + \eta \langle \nabla f(y_k), x_{k+1} \rangle + \frac{L\eta^2}{2} \|x_{k+1} - y_k\|^2 \\
= & (1-\eta) f(y_k) - \eta f^*(\nabla f(y_k)) + \eta \langle \nabla f(y_k), x_{k+1} \rangle + \frac{L\eta^2}{2} \|x_{k+1} - y_k\|^2,
\end{aligned}$$

where the last line follows by Lemma A.1 with $x = y_k$, $y = \nabla f(y_k)$.
b) Since $(h^\alpha)^*(-\cdot)$ is α^{-1}-smooth, we can follow the same steps as in (a) to obtain

$$\begin{aligned}
(h^\alpha)^*(-s_{k+1}) \leq & (h^\alpha)^*(-s_k) + \langle -\nabla (h^\alpha)^*(-s_k), s_{k+1} - s_k \rangle + \frac{1}{2\alpha} \|s_{k+1} - s_k\|_*^2 \\
\overset{(13)}{=} & (h^\alpha)^*(-s_k) + \eta \langle -\nabla (h^\alpha)^*(-s_k), \nabla f(y_k) - s_k \rangle + \frac{\eta^2}{2\alpha} \|\nabla f(y_k) - s_k\|_*^2 \\
= & (h^\alpha)^*(-s_k) + \eta \langle \nabla (h^\alpha)^*(-s_k), s_k \rangle - \eta \langle \nabla (h^\alpha)^*(-s_k), \nabla f(y_k) \rangle + \frac{\eta^2}{2\alpha} \|\nabla f(y_k) - s_k\|_*^2 \\
= & (1-\eta)(h^\alpha)^*(-s_k) - \eta h^\alpha(x_{k+1}) - \eta \langle x_{k+1}, \nabla f(y_k) \rangle + \frac{\eta^2}{2\alpha} \|\nabla f(y_k) - s_k\|_*^2,
\end{aligned}$$

where the last line follows by Lemma A.1 with $x = x_{k+1}$, $y = s_k$.
Statement c) follows directly from the α-strong convexity of h^α with respect to $\|\cdot\|$ and (12), and statement d) follows from the L^{-1}-strong convexity of f^* with respect to $\|\cdot\|_*$ and (13). ∎

Using the smoothness/convexity bounds in Lemma C.1, we are now ready to prove Theorem 4.2.
Proof of Theorem 4.2: Using the choice $\eta = \alpha/(L+\alpha)$ (hence $1-\eta = L/(L+\alpha)$), we obtain

$$L\eta^2 - \eta(1-\eta)\alpha = \frac{L\alpha^2}{(L+\alpha)^2} - \frac{L\alpha^2}{(L+\alpha)^2} = 0, \tag{44}$$

and therefore

$$\frac{\eta^2}{\alpha} - \frac{\eta(1-\eta)}{L} = \frac{1}{L\alpha}\left(L\eta^2 - \eta(1-\eta)\alpha\right) \overset{(44)}{=} 0. \tag{45}$$

Then, from Lemma C.1(a) and (c), we have

$$\begin{aligned}\phi^\alpha(y_{k+1}) =& f(y_{k+1}) + h^\alpha(y_{k+1}) \le (1-\eta)\phi^\alpha(y_k) - \eta f^*(\nabla f(y_k)) + \eta h^\alpha(x_{k+1}) + \eta\langle \nabla f(y_k), x_{k+1}\rangle \\ &+ \frac{1}{2}(L\eta^2 - \alpha(1-\eta)\eta)\|x_{k+1} - y_k\|^2 \\ \overset{(44)}{=}& (1-\eta)\phi^\alpha(y_k) - \eta f^*(\nabla f(y_k)) + \eta h^\alpha(x_{k+1}) + \eta\langle \nabla f(y_k), x_{k+1}\rangle,\end{aligned} \tag{46}$$

and by Lemma C.1(b) and (d), we have

$$\begin{aligned}\psi^\alpha(s_{k+1}) =& f^*(s_{k+1}) + (h^\alpha)^*(-s_{k+1}) \le (1-\eta)\psi^\alpha(s_k) - \eta h^\alpha(x_{k+1}) \\ &+ \eta f^*(\nabla f(y_k)) - \eta\langle x_{k+1}, \nabla f(y_k)\rangle + \frac{1}{2}(\alpha^{-1}\eta^2 - L^{-1}(1-\eta)\eta)\|\nabla f(y_k) - s_k\|_*^2 \\ \overset{(45)}{=}& (1-\eta)\psi^\alpha(s_k) - \eta h^\alpha(x_{k+1}) + \eta f^*(\nabla f(y_k)) - \eta\langle x_{k+1}, \nabla f(y_k)\rangle.\end{aligned} \tag{47}$$

Summing (46) and (47) and canceling terms gives

$$\phi^\alpha(y_{k+1}) + \psi^\alpha(s_{k+1}) \le (1-\eta)[\phi^\alpha(y_k) + \psi^\alpha(s_k)] \le (1-\eta)^{k+1}[\phi^\alpha(y_0) + \psi^\alpha(s_0)].$$

By standard analysis and our choice of η, the complexity to obtain $\phi^\alpha(y_k) + \psi^\alpha(s_k) \le \varepsilon/2$ is therefore

$$k = \mathcal{O}\left(1 + \frac{L}{\alpha}\log\left(\frac{\phi^\alpha(y_0) + \psi^\alpha(s_0)}{\varepsilon}\right)\right),$$

which gives the first complexity bound in view of $\alpha = \varepsilon/(2M)$. The second complexity bound follows directly from Lemma 2.1. ∎

D Three-Average Analysis

As in the previous two sections, we begin by proving the claimed primal-dual correspondence stated in Proposition 5.3.
Proof of Proposition 5.3: First, we observe that Lemma D.2(a) and (20) imply

$$a_k = A_k\left(\frac{\alpha + \sqrt{\alpha^2 + 4\alpha L}}{2L}\right).$$

Then, we obtain

$$\frac{a_k}{A_{k+1}} = \frac{\alpha + \sqrt{\alpha^2 + 4\alpha L}}{2L + \alpha + \sqrt{\alpha^2 + 4\alpha L}} = \frac{2\alpha}{\alpha + \sqrt{\alpha^2 + 4\alpha L}} \overset{(15)}{=} \lambda.$$

It thus follows from Algorithm 5 that for all $k \geq 0$,

$$\tilde{x}_{k+1} = \frac{A_k}{A_{k+1}} y_k + \frac{a_k}{A_{k+1}} x_k, \quad y_{k+1} = \frac{A_k}{A_{k+1}} y_k + \frac{a_k}{A_{k+1}} x_{k+1}; \tag{48}$$
$$s_{k+1} = \frac{A_k}{A_{k+1}} s_k + \frac{a_k}{A_{k+1}} \nabla f(\tilde{x}_{k+1}).$$

We now prove the correspondence $(g_k, z_k, v_k) = (\nabla f(\tilde{x}_k), s_k, x_k)$ holds for all iterations $k \geq 0$ by induction. The base case follows by our choice of initialization with $g_0 = \nabla f(y_0) = \nabla f(\tilde{x}_0)$ and by Lemma A.2 applied to x_0 as defined in (15). Now, assume that $(g_k, z_k, v_k) = (\nabla f(\tilde{x}_k), s_k, x_k)$ for some $k \geq 0$ and define $x_{-1} = x_0$.

Note that the subdifferential $\partial f^*(g_k)$ is not necessarily a singleton, therefore there is some ambiguity in defining $\mathrm{D}_{\nu^*}(g \| g_k)$ as used in Step **2** of Algorithm 6. Since $\tilde{x}_k \in \partial f^*(g_k)$ by the inductive hypothesis and Lemma A.1, we choose $(f^*)'(g_k) = \tilde{x}_k$ for the linearization term, so

$$\mathrm{D}_{\nu^*}(g \| g_k) = L \, \mathrm{D}_{f^*}(g \| g_k) = L \left[f^*(g) - f^*(g_k) - \langle (f^*)'(g_k), g - g_k \rangle \right].$$

Then, by the optimality condition of (22) with $\nu^* = Lf^*$, for some $(f^*)'(g_{k+1}) \in \partial f^*(g_{k+1})$ we have

$$0 = \frac{L}{\alpha} \left(\tau_k + \frac{\alpha}{L} a_k \right) (f^*)'(g_{k+1}) - \frac{L\tau_k}{\alpha} (f^*)'(g_k) - a_k \hat{v}_k.$$

Now, we have $\tau_k + \alpha a_k L^{-1} = \tau_{k+1} = \alpha L^{-1} A_{k+1}$. Then, rearranging, we obtain

$$\begin{aligned} A_{k+1}(f^*)'(g_{k+1}) &= L\alpha^{-1}\tau_{k+1}(f^*)'(g_{k+1}) = L\alpha^{-1}\tau_k(f^*)'(g_k) + a_k\hat{v}_k \\ &\overset{(21)}{=} L\alpha^{-1}\tau_k(f^*)'(g_k) + a_k v_k + a_{k-1}(v_k - v_{k-1}) \\ &= A_k\tilde{x}_k + a_k x_k + a_{k-1}(x_k - x_{k-1}) \\ &\overset{(48)}{=} A_k y_k + a_k x_k \overset{(48)}{=} A_{k+1}\tilde{x}_{k+1}, \end{aligned}$$

where the second line follows from (21), the third line from the inductive hypothesis, and the fourth line from (48). Therefore $\tilde{x}_{k+1} \in \partial f^*(g_{k+1})$, hence by Lemma A.1 we obtain

$$\nabla f(\tilde{x}_{k+1}) = g_{k+1}. \tag{49}$$

With this correspondence, we note that

$$s_{k+1} = \frac{A_k}{A_{k+1}} s_k + \frac{a_k}{A_{k+1}} \nabla f(\tilde{x}_{k+1}) \overset{(49)}{=} \frac{A_k}{A_{k+1}} z_k + \frac{a_k}{A_{k+1}} g_{k+1} = z_{k+1},$$

where the equality follows by the inductive hypothesis and (49). Finally, (18) and Lemma A.2 imply that

$$x_{k+1} = \nabla (h^\alpha)^*(-s_{k+1}) = \nabla (h^\alpha)^*(-z_{k+1}) = v_{k+1}.$$

We therefore have $(g_{k+1}, z_{k+1}, v_{k+1}) = (\nabla f(\tilde{x}_{k+1}), s_{k+1}, x_{k+1})$, completing the proof. ∎

D.1 Analysis of TAA

Throughout this subsection, we define $\Gamma_k(\cdot)$ as the ACP model induced by $(y_0, \{\tilde{x}_{i+1}\}_{i=0}^{k-1}, \{\lambda\}^k)$. Recall that this implies the recursive definition

$$\Gamma_{k+1}(\cdot) = (1-\lambda)\Gamma_k(\cdot) + \lambda \gamma_k(\cdot), \quad \gamma_k(\cdot) = \ell_f(\cdot; \tilde{x}_{k+1}) + h^\alpha(\cdot), \tag{50}$$

with $\Gamma_0(\cdot) = \ell_f(\cdot; y_0) + h^\alpha(\cdot)$. By simple induction and the definition of s_k in (17), we can show that $x_k = \operatorname{argmin}_{x \in \mathbb{R}^n} \Gamma_k(\cdot)$ for all $k \geq 0$. Then, we have the following proposition.

Proposition D.1. For all $k \geq 0$, the following inequality holds

$$\min_{x \in \mathbb{R}^n} \Gamma_k(x) \geq \phi^\alpha(y_k) - (1-\lambda)^k \Delta,$$

where $\Delta = \phi^\alpha(y_0) - \Gamma_0(x_0)$.

Proof: First, note that by our choice of λ and simple algebra, we can show the relationship

$$\frac{L}{\alpha} = \frac{1-\lambda}{\lambda^2}. \tag{51}$$

We now prove the claim by induction on k. The base case trivially holds, since

$$\phi^\alpha(y_0) - (1-\lambda)^0 \Delta = \Gamma_0(x_0) \overset{(15)}{=} \min_{x \in \mathbb{R}^n} \Gamma_0(x).$$

Now suppose that the claim holds for some $k \geq 0$. Then, since $x_{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} \Gamma_{k+1}(x)$ by construction, we have

$$\begin{aligned}
\min_{x \in \mathbb{R}^n} \Gamma_{k+1}(x) = \Gamma_{k+1}(x_{k+1}) &\overset{(50)}{=} (1-\lambda)\Gamma_k(x_{k+1}) + \lambda \gamma_k(x_{k+1}) \\
&\geq (1-\lambda) \min_{x \in \mathbb{R}^n} \Gamma_k(x) + \lambda \gamma_k(x_{k+1}) + \frac{(1-\lambda)\alpha}{2} \|x_k - x_{k+1}\|^2 \\
&\geq (1-\lambda)\phi^\alpha(y_k) - (1-\lambda)^{k+1}\Delta + \lambda \gamma_k(x_{k+1}) + \frac{\alpha(1-\lambda)}{2} \|x_k - x_{k+1}\|^2 \\
&\overset{(50)}{\geq} (1-\lambda)\gamma_k(y_k) - (1-\lambda)^{k+1}\Delta + \lambda \gamma_k(x_{k+1}) + \frac{\alpha(1-\lambda)}{2} \|x_k - x_{k+1}\|^2
\end{aligned}$$

where second line follows by Lemma 2.3(b), the third by the inductive hypothesis and the fourth by the convexity of f. Then, further applying the convexity of γ_k and (19), we obtain

$$\begin{aligned}
\min_{x \in \mathbb{R}^n} \Gamma_{k+1}(x) &\geq \gamma_k(y_{k+1}) - (1-\lambda)^{k+1}\Delta + \frac{\alpha(1-\lambda)}{2} \|x_k - x_{k+1}\|^2 \\
&\overset{(16)(19)}{=} \gamma_k(y_{k+1}) - (1-\lambda)^{k+1}\Delta + \frac{\alpha(1-\lambda)}{2\lambda^2} \|y_{k+1} - \tilde{x}_{k+1}\|^2 \\
&\overset{(51)}{=} \gamma_k(y_{k+1}) - (1-\lambda)^{k+1}\Delta + \frac{L}{2} \|y_{k+1} - \tilde{x}_{k+1}\|^2 \\
&\overset{(50)}{\geq} \phi^\alpha(y_{k+1}) - (1-\lambda)^{k+1}\Delta,
\end{aligned}$$

where the first equality follows by the updates in (16) and (19), the second by (51), and the final line by the L-smoothness of f and the definition of γ_k in (50). We therefore conclude the proof. ∎

Remark: The key step in the previous proof was the identity $\lambda(x_{k+1} - x_k) = y_{k+1} - \tilde{x}_{k+1}$ resulting from the geometric similarity illustrated in Figure 1(b).

Using the upper bound in Proposition D.1, we have (1) a linear convergence rate for the primal-dual certificate (y_k, Γ_k) and (2) a corresponding iteration-complexity for finding an ε-solution to (1).

Proof of Theorem 5.1: Applying Proposition D.1 and rearranging terms, we have

$$\phi^\alpha(y_k) - \min_{x \in \mathbb{R}^n} \Gamma_k(x) \leq (1-\lambda)^k \Delta.$$

It follows from the definition of λ in (15) that $1 - \lambda \leq \left(1 + \frac{\sqrt{\alpha}}{2\sqrt{L}}\right)^{-2}$, which together with the above inequality implies that

$$\phi^\alpha(y_k) - \min_{x \in \mathbb{R}^n} \Gamma_k(x) \leq \frac{\Delta}{\left(1 + \frac{\sqrt{\alpha}}{2\sqrt{L}}\right)^{2k}}.$$

We therefore obtain $\phi^\alpha(y_k) - \min_{x\in\mathbb{R}^n} \Gamma_k(x) \le \varepsilon/2$ in

$$k = \mathcal{O}\left(1 + \sqrt{\frac{L}{\alpha}} \log\left(\frac{\phi^\alpha(y_0) - \Gamma_0(x_0)}{\varepsilon}\right)\right)$$

iterations. Choosing $\alpha = \varepsilon/(2M)$ and applying Definition 2.4 gives the result. ∎

D.2 Analysis of GEM

In this subsection, we analyze the non-asymptotic behavior of GEM for solving (3). We begin with standard technical lemmas for accelerated methods.

Lemma D.2. For all $k \ge 0$, the following hold

a) $L^{-1}A_k = \alpha^{-1}\tau_k$;

b) $a_k^2 = \tau_k A_{k+1} = \tau_{k+1} A_k$;

c) $A_k \ge \left(1 + \frac{\sqrt{\alpha}}{2\sqrt{L}}\right)^{2k}$.

Proof: a) Note that expanding the recursion for τ_k in (20) from 0 to $k-1$, we have

$$\alpha^{-1}\tau_k = \alpha^{-1}\tau_0 + L^{-1}\sum_{i=0}^{k-1} a_i = \alpha^{-1}\tau_0 + L^{-1}(A_k - A_0).$$

Using $\alpha^{-1}\tau_0 = L^{-1} = L^{-1}A_0$ gives,

$$\alpha^{-1}\tau_k = \alpha^{-1}\tau_0 - L^{-1}A_0 + L^{-1}A_k = L^{-1}A_k.$$

b) The expression for a_k in (20) gives

$$a_k^2 = \tau_k a_k + \tau_k A_k \overset{(21)}{=} \tau_k A_{k+1} = (\alpha L^{-1} A_k)(\alpha^{-1} L \tau_{k+1}) = A_k \tau_{k+1},$$

where the third equality follows from part (a).
c) From (20), we have

$$a_k \ge \frac{\tau_k}{2} + \sqrt{\tau_k A_k},$$

which implies that

$$A_{k+1} = a_k + A_k \ge \frac{\tau_k}{2} + \sqrt{\tau_k A_k} + A_k \ge \left(\sqrt{A_k} + \frac{\sqrt{\tau_k}}{2}\right)^2 = A_k\left(1 + \frac{\sqrt{\alpha}}{2\sqrt{L}}\right)^2,$$

where the last equality follows by part (a). Expanding the inequality from 0 to k and using $A_0 = 1$ yields the claim. ∎

With the necessary technical lemmas established, we begin by examining the optimality conditions of the extrapolated update (22).

Lemma D.3. Define $\gamma_k(\cdot) := \langle -\hat{v}_k, \cdot\rangle + f^*(\cdot)$. Then, for all $g \in \operatorname{dom} f^*$, we have

$$A_k\gamma_k(z_k) + a_k\gamma_k(g) + \frac{\tau_k}{\alpha}\mathrm{D}_{\nu^*}(g\|g_k) \ge A_{k+1}\gamma_k(z_{k+1}) + \frac{\tau_k}{\alpha}\mathrm{D}_{\nu^*}(g_{k+1}\|g_k) + \frac{\tau_{k+1}}{\alpha}\mathrm{D}_{\nu^*}(g\|g_{k+1}) \quad (52)$$

Proof: Noting that f^* is $(1/L)$-strongly convex relative to ν^*, and applying Lemma A.4 to the g_{k+1} update in (22), with $\Phi(\cdot) := a_k\gamma_k(\cdot)$, $\omega := \nu^*$, $\beta = \tau_k/\alpha$ and $\mu = a_k/L$, we have for all $g \in \operatorname{dom} f^*$,

$$a_k\gamma_k(g_{k+1}) + \frac{\tau_k}{\alpha}\operatorname{D}_{\nu^*}(g_{k+1}\|g_k) + \alpha^{-1}\left(\frac{\alpha a_k}{L} + \tau_k\right)\operatorname{D}_{\nu^*}(g\|g_{k+1}) \le a_k\gamma_k(g) + \frac{\tau_k}{\alpha}\operatorname{D}_{\nu^*}(g\|g_k).$$

Adding $A_k\gamma_k(z_k)$ to both sides gives

$$\begin{aligned}
&A_k\gamma_k(z_k) + a_k\gamma_k(g) + \frac{\tau_k}{\alpha}\operatorname{D}_{\nu^*}(g\|g_k)\\
\ge& a_k\gamma_k(g_{k+1}) + A_k\gamma_k(z_k) + \frac{\tau_k}{\alpha}\operatorname{D}_{\nu^*}(g_{k+1}\|g_k) + \alpha^{-1}\left(\frac{\alpha a_k}{L} + \tau_k\right)\operatorname{D}_{\nu^*}(g\|g_{k+1})\\
\overset{(23)}{\ge}& A_{k+1}\gamma_k(z_{k+1}) + \frac{\tau_k}{\alpha}\operatorname{D}_{\nu^*}(g_{k+1}\|g_k) + \frac{\tau_{k+1}}{\alpha}\operatorname{D}_{\nu^*}(g\|g_{k+1}),
\end{aligned}$$

where the second inequality follows by the convexity of $\gamma_k(\cdot)$ and the definition of τ_{k+1} in (20). ∎

The following lemma gives an equivalent but more convenient form of (52) in Lemma D.3.

Lemma D.4. For all iterations $k \ge 0$ of Algorithm 6, the following inequality holds

$$\begin{aligned}
&A_k\left[\ell_{(h^\alpha)^*}(z_k; z_{k+1}) + f^*(z_k)\right] + \frac{\tau_k}{\alpha}\operatorname{D}_{\nu^*}(g\|g_k) - \frac{\tau_k}{\alpha}\operatorname{D}_{\nu^*}(g_{k+1}\|g_k) - \frac{\tau_{k+1}}{\alpha}\operatorname{D}_{\nu^*}(g\|g_{k+1})\\
&\quad\ge A_{k+1}\psi^\alpha(z_{k+1}) + a_k\left[h^\alpha(v_{k+1}) + \langle v_{k+1}, g\rangle - f^*(g)\right] + a_k\langle v_{k+1} - \hat{v}_k, g_{k+1} - g\rangle. \qquad (53)
\end{aligned}$$

Proof: Using the definition of γ_k, we observe that for all $g \in \operatorname{dom} f^*$, the following identity holds

$$\begin{aligned}
\gamma_k(g) &= \langle -\hat{v}_k, g\rangle + f^*(g) = \langle v_{k+1} - \hat{v}_k, g\rangle + \langle -v_{k+1}, g - z_{k+1}\rangle - \langle v_{k+1}, z_{k+1}\rangle + f^*(g)\\
&\overset{(23)}{=} \langle v_{k+1} - \hat{v}_k, g\rangle + \ell_{(h^\alpha)^*}(g; z_{k+1}) - (h^\alpha)^*(-z_{k+1}) - \langle v_{k+1}, z_{k+1}\rangle + f^*(g),
\end{aligned}$$

where the final equality holds by $v_{k+1} = \nabla(h^\alpha)^*(-z_{k+1})$ in (23) and the definition of $\ell_{(h^\alpha)^*}(\cdot; z)$ in (32). Expanding each instance of $\gamma_k(\cdot)$ in (52), we obtain

$$\begin{aligned}
&A_k\left[\langle v_{k+1} - \hat{v}_k, z_k\rangle + \ell_{(h^\alpha)^*}(z_k; z_{k+1}) - (h^\alpha)^*(-z_{k+1}) - \langle v_{k+1}, z_{k+1}\rangle + f^*(z_k)\right]\\
&\quad + a_k\left[\langle v_{k+1} - \hat{v}_k, g\rangle + \ell_{(h^\alpha)^*}(g; z_{k+1}) - (h^\alpha)^*(-z_{k+1}) - \langle v_{k+1}, z_{k+1}\rangle + f^*(g)\right] + \frac{\tau_k}{\alpha}\operatorname{D}_{\nu^*}(g\|g_k)\\
\ge& A_{k+1}\left[-\langle v_{k+1}, z_{k+1}\rangle + \langle v_{k+1} - \hat{v}_k, z_{k+1}\rangle + f^*(z_{k+1})\right] + \frac{\tau_k}{\alpha}\operatorname{D}_{\nu^*}(g_{k+1}\|g_k) + \frac{\tau_{k+1}}{\alpha}\operatorname{D}_{\nu^*}(g\|g_{k+1}).
\end{aligned}$$

Rearranging yields

$$\begin{aligned}
&A_k\left[\ell_{(h^\alpha)^*}(z_k; z_{k+1}) + f^*(z_k)\right] + \frac{\tau_k}{\alpha}\operatorname{D}_{\nu^*}(g\|g_k) - \frac{\tau_{k+1}}{\alpha}\operatorname{D}_{\nu^*}(g\|g_{k+1}) - \frac{\tau_k}{\alpha}\operatorname{D}_{\nu^*}(g_{k+1}\|g_k)\\
\ge& A_{k+1}\psi^\alpha(z_{k+1}) + a_k\left[-\ell_{(h^\alpha)^*}(g; z_{k+1}) - f^*(g)\right] + \langle v_{k+1} - \hat{v}_k, A_{k+1}z_{k+1} - A_kz_k - a_kg\rangle\\
=& A_{k+1}\psi^\alpha(z_{k+1}) + a_k\left[\langle v_{k+1}, g\rangle + h^\alpha(v_{k+1}) - f^*(g)\right] + \langle v_{k+1} - \hat{v}_k, A_{k+1}z_{k+1} - A_kz_k - a_kg\rangle,
\end{aligned}$$

where the equality follows from the fact that

$$\begin{aligned}
\ell_{(h^\alpha)^*}(g; z_{k+1}) &\overset{(32)}{=} (h^\alpha)^*(-z_{k+1}) + \langle -\nabla(h^\alpha)^*(-z_{k+1}), g - z_{k+1}\rangle\\
&\overset{(23)}{=} (h^\alpha)^*(-z_{k+1}) - \langle v_{k+1}, -z_{k+1}\rangle - \langle v_{k+1}, g\rangle\\
&\overset{Lemma\ A.1}{=} -h^\alpha(v_{k+1}) - \langle v_{k+1}, g\rangle.
\end{aligned}$$

Finally, we note that

$$A_{k+1}z_{k+1} - A_kz_k \overset{(23)}{=} a_kg_{k+1},$$

which completes the proof. ∎

Next, we state a simple consequence of our choice of a_k, A_k and the smoothness of $(h^\alpha)^*$.

Lemma D.5. For all iterations $k \geq 0$ of Algorithm 6, the following inequality holds

$$A_k \psi^\alpha(z_k) - \frac{\alpha a_k^2}{2\tau_{k+1}} \|v_{k+1} - v_k\|^2 \geq A_k \left[\ell_{(h^\alpha)^*}(z_k; z_{k+1}) + f^*(z_k) \right]. \tag{54}$$

Proof: By the $(1/\alpha)$-smoothness of $(h^\alpha)^*$ (see [Bec17, Theorem 5.8(iii)]), we have

$$\begin{aligned} \ell_{(h^\alpha)^*}(z_k; z_{k+1}) &\leq (h^\alpha)^*(-z_k) - \frac{\alpha}{2} \|\nabla (h^\alpha)^*(-z_k) - \nabla (h^\alpha)^*(-z_{k+1})\|^2 \\ &\overset{(23)}{=} (h^\alpha)^*(-z_k) - \frac{\alpha}{2} \|v_k - v_{k+1}\|^2. \end{aligned} \tag{55}$$

It thus follows from Lemma D.2(b) that

$$A_k \psi^\alpha(z_k) - \frac{\alpha a_k^2}{2\tau_{k+1}} \|v_k - v_{k+1}\|^2 = A_k \left[\psi^\alpha(z_k) - \frac{\alpha}{2} \|v_k - v_{k+1}\|^2 \right] \overset{(55)}{\geq} A_k \left[f^*(z_k) + \ell_{(h^\alpha)^*}(z_k; z_{k+1}) \right].$$

■

The following lemma provides a convenient algebraic identity resulting from the definition of the extrapolation point $\hat{v}_k$.

Lemma D.6. For every $k \geq 0$, define

$$s_k := a_k (v_{k+1} - v_k) \tag{56}$$

with $s_{-1} := 0$. Then, for all $g \in \mathbb{R}^n$ we have

$$\sum_{i=0}^{k-1} a_i \langle v_{i+1} - \hat{v}_i, g_{i+1} - g \rangle = \langle s_{k-1}, g_k - g \rangle - \sum_{i=1}^{k-1} \langle s_{i-1}, g_{i+1} - g_i \rangle. \tag{57}$$

Proof: By the definitions of s_k and $\hat{v}_k$ in (56) and (21), respectively, we have

$$s_k - s_{k-1} = a_k (v_{k+1} - v_k) - a_{k-1}(v_k - v_{k-1}) \overset{(21)}{=} a_k (v_{k+1} - \hat{v}_k).$$

Then, summing from 0 to $k-1$ and using $s_{-1} = 0$, we have

$$\begin{aligned} \sum_{i=0}^{k-1} a_i \langle v_{i+1} - \hat{v}_i, g_{i+1} - g \rangle &= \sum_{i=0}^{k-1} \left(\langle s_i, g_{i+1} - g \rangle - \langle s_{i-1}, g_{i+1} - g \rangle \right) \\ &= -\langle s_{k-1}, g \rangle + \sum_{i=0}^{k-1} \langle s_i - s_{i-1}, g_{i+1} \rangle = \langle s_{k-1}, g_k - g \rangle - \sum_{i=1}^{k-1} \langle s_{i-1}, g_{i+1} - g_i \rangle. \end{aligned}$$

We therefore prove the claim. ■

Combining the previous lemmas, we are now ready to prove Theorem 5.2.

Proof of Theorem 5.2: First, applying Lemmas D.4 and D.5, we obtain for all iterations $i \geq 0$ and all $g \in \operatorname{dom} f^*$

$$\begin{aligned} &A_i \psi^\alpha(z_i) - \frac{\alpha a_i^2}{2\tau_{i+1}} \|v_{i+1} - v_i\|^2 + \frac{\tau_i}{\alpha} \mathrm{D}_{\nu^*}(g \| g_i) - \frac{\tau_i}{\alpha} \mathrm{D}_{\nu^*}(g_{i+1} \| g_i) - \frac{\tau_{i+1}}{\alpha} \mathrm{D}_{\nu^*}(g \| g_{i+1}) \\ &\overset{(53)(54)}{\geq} A_{i+1} \psi^\alpha(z_{i+1}) + a_i \left[h^\alpha(v_{i+1}) + \langle v_{i+1}, g \rangle - f^*(g) \right] + a_i \langle v_{i+1} - \hat{v}_i, g_{i+1} - g \rangle. \end{aligned}$$

Summing from $i = 0$ to $k-1$, rearranging, dropping the non-negative term $\mathrm{D}_{\nu^*}(g_1\|g_0)$, and using $\tau_0 = \alpha/L$ gives

$$\begin{aligned}
A_0\psi^\alpha(z_0) + L^{-1}\,\mathrm{D}_{\nu^*}(g\|g_0) \geq\; & A_k\psi^\alpha(z_k) + \frac{\tau_k}{\alpha}\,\mathrm{D}_{\nu^*}(g\|g_k) + \sum_{i=0}^{k-1} a_i[h^\alpha(v_{i+1}) + \langle v_{i+1}, g\rangle - f^*(g)] \\
& + \frac{\alpha a_{k-1}^2}{2\tau_k}\|v_k - v_{k-1}\|^2 + \sum_{i=0}^{k-1} a_i\langle v_{i+1} - \hat{v}_i, g_{i+1} - g\rangle + \sum_{i=1}^{k-1}\left(\frac{\tau_i}{\alpha}\,\mathrm{D}_{\nu^*}(g_{i+1}\|g_i) + \frac{\alpha a_i^2}{2\tau_i}\|v_i - v_{i-1}\|^2\right) \\
\overset{(56)(57)}{=}\; & A_k\psi^\alpha(z_k) + \frac{\alpha}{2\tau_k}\|s_{k-1}\|^2 + \frac{\tau_k}{\alpha}\,\mathrm{D}_{\nu^*}(g\|g_k) + \sum_{i=0}^{k-1} a_i[h^\alpha(v_{i+1}) + \langle v_{i+1}, g\rangle - f^*(g)] \\
& + \langle s_{k-1}, g_k - g\rangle + \sum_{i=1}^{k-1}\left(\frac{\tau_i}{\alpha}\,\mathrm{D}_{\nu^*}(g_{i+1}\|g_i) - \langle s_{i-1}, g_{i+1} - g_i\rangle + \frac{\alpha}{2\tau_i}\|s_{i-1}\|^2\right),
\end{aligned}$$

where the equality follows by Lemma D.6 and the definition of s_k. Then, using the 1-strong convexity of ν^*, we obtain

$$\begin{aligned}
A_0\psi^\alpha(z_0) + L^{-1}\,\mathrm{D}_{\nu^*}(g\|g_0) \overset{(56)}{\geq}\; & A_k\psi^\alpha(z_k) + \sum_{i=0}^{k-1} a_i[h^\alpha(v_{i+1}) + \langle v_{i+1}, g\rangle - f^*(g)] \\
& + \frac{\tau_k}{2\alpha}\|g - g_k\|^2 + \langle s_{k-1}, g_k - g\rangle + \frac{\alpha}{2\tau_k}\|s_{k-1}\|^2 \\
& + \sum_{i=1}^{k-1}\left(\frac{\tau_i}{2\alpha}\|g_{i+1} - g_i\|^2 - \langle s_{i-1}, g_{i+1} - g_i\rangle + \frac{\alpha}{2\tau_i}\|s_{i-1}\|^2\right) \\
\geq\; & A_k\psi^\alpha(z_k) + \sum_{i=0}^{k-1} a_i[h^\alpha(v_{i+1}) + \langle v_{i+1}, g\rangle - f^*(g)], \qquad (58)
\end{aligned}$$

where the final line follows by the Cauchy-Schwarz inequality. It follows from Lemma A.1, (23), and the definition of $\ell_{(h^\alpha)^*}(\cdot; z)$ in (32) that for any $i \geq 0$,

$$\begin{aligned}
h^\alpha(v_i) + \langle v_i, g\rangle &= h^\alpha(v_i) - \langle v_i, -z_i\rangle + \langle v_i, g - z_i\rangle \\
&\overset{(23)}{=} -(h^\alpha)^*(-z_i) - \langle -\nabla(h^\alpha)^*(-z_i), g - z_i\rangle \\
&\overset{(32)}{=} -\ell_{(h^\alpha)^*}(g; z_i). \qquad (59)
\end{aligned}$$

Applying this identity to the summation on the right-hand side of (58), we have

$$\begin{aligned}
\sum_{i=0}^{k-1} a_i[h^\alpha(v_{i+1}) + \langle v_{i+1}, g\rangle - f^*(g)] \overset{(59)}{=}\; & -A_k\left(\sum_{i=0}^{k-1}\frac{a_i}{A_k}[\ell_{(h^\alpha)^*}(g; z_{i+1}) + f^*(g)]\right. \\
& \left. + \frac{A_0}{A_k}\left(\ell_{(h^\alpha)^*}(g; z_0) + f^*(g)\right)\right) + A_0(\ell_{(h^\alpha)^*}(g; z_0) + f^*(g)) \\
\overset{(59)}{=}\; & -A_k\Gamma_k^*(g) - A_0(h^\alpha(v_0) + \langle v_0, g\rangle - f^*(g)). \qquad (60)
\end{aligned}$$

where the first equality follows from (59) and the second by the definition of the model $\Gamma_k^*(\cdot)$ induced by $(z_0, \{z_{i+1}\}_{i=0}^{k-1}, \{a_i/A_{i+1}\}_{i=0}^{k-1})$ with $A_0 = 1$ and (59) applied in reverse.

Applying (60) to (58) and maximizing both sides over $g \in \operatorname{dom} f^*$ gives

$$A_0(\psi^\alpha(z_0) + \phi^\alpha(v_0)) + L^{-1} \max_{g \in \operatorname{dom} f^*} \mathrm{D}_{\nu^*}(g \| g_0) \geq A_k \left(\psi^\alpha(z_k) - \min_{g \in \mathbb{R}^n} \Gamma_k^*(g) \right),$$

where we use

$$h^\alpha(v_0) + \max_{g \in \operatorname{dom} f^*} \{\langle v_0, g \rangle - f^*(g)\} = \phi^\alpha(v_0).$$

Rearranging, using Lemma D.2(c), and noting that $A_0 = 1$ gives the certificate convergence

$$\psi^\alpha(z_k) - \min_{g \in \mathbb{R}^n} \Gamma_k^*(g) \leq \frac{\phi^\alpha(v_0) + \psi^\alpha(z_0) + DL^{-1}}{\left(1 + \frac{\sqrt{\alpha}}{2\sqrt{L}}\right)^{2k}}.$$

The number of iterations k to obtain $\psi^\alpha(z_k) - \min_g \Gamma_k^*(g) \leq \varepsilon/2$ is therefore

$$k = \mathcal{O}\left(1 + \sqrt{\frac{L}{\alpha}} \log\left(\frac{\phi^\alpha(v_0) + \psi^\alpha(z_0) + DL^{-1}}{\varepsilon}\right)\right)$$

by standard analysis. The result follows by defining the dual certificate (analogous to Definition 2.4) as the pair (z_k, Γ_k^*) and the choice $\alpha = \varepsilon/(2M)$. ∎

References

[AA11] A. Aboussoror and S. Adly. A Fenchel-Lagrange duality approach for a bilevel programming problem with extremal-value function. *Journal of Optimization Theory and Applications*, 149(2):254–268, 2011.

[Bac15] F. Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.

[Bec17] A. Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

[BL26] M. X. Burns and J. Liang. Improved analysis of restarted accelerated gradient and augmented Lagrangian methods via inexact proximal point frameworks. *arXiv preprint arXiv:2602.17878*, 2026.

[BLM09] K. Bredies, D. A. Lorenz, and P. Maass. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computational Optimization and Applications*, 42(2):173–193, 2009.

[BT09] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[CJS25] H. Chen, C. Jiang, and A. M-C. So. Accelerated price adjustment for Fisher markets with exact recovery of competitive equilibrium. In *Conference on Web and Internet Economics (WINE)*, 2025.

[CP11] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[CR08] E. Candès and B. Recht. Exact matrix completion via convex programming. *Foundations of Computational Mathematics*, 9:717–772, 2008.

[CRT06] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[CT06] G. Chen and M. Teboulle. Convergence Analysis of a Proximal-Like Minimization Algorithm Using Bregman Functions. *SIAM Journal on Optimization*, 2006.

[Cut13] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[CWC21] S. Cen, Y. Wei, and Y. Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34:27952–27964, 2021.

[FK93] D. Fudenberg and D. M. Kreps. Learning mixed equilibria. *Games and Economic Behavior*, 5(3):320–367, 1993.

[FS24] D. Fersztand and X. A. Sun. The proximal bundle algorithm under a Frank-Wolfe perspective: an improved complexity analysis. *arXiv preprint arXiv:2411.15926*, 2024.

[GL25] B. Grimmer and D. Li. Some primal-dual theory for subgradient methods for strongly convex optimization. *Mathematical Programming*, 214(1):759–788, 2025.

[GP23] D. H. Gutman and J. F. Peña. Perturbed Fenchel duality and first-order methods. *Mathematical Programming*, 198(1):443–469, 2023.

[Jag13] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013.

[JN20] A. Juditsky and A. Nemirovski. Statistical inference via convex optimization. 2020.

[Lan20] G. Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer, 2020.

[LF21] H. Lu and R. M. Freund. Generalized stochastic Frank–Wolfe algorithm with stochastic "substitute" gradient for structured convex optimization. *Mathematical Programming*, 187(1):317–349, 2021.

[LFN18] H. Lu, R. M. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[LGM24] J. Liang, V. Guigues, and R. D. C. Monteiro. A single cut proximal bundle method for stochastic convex composite optimization. *Mathematical programming*, 208(1):173–208, 2024.

[Lia25] J. Liang. Primal-dual proximal bundle and conditional gradient methods for convex problems. *Mathematical Programming*, 2025.

[LM24] J. Liang and R. D. C. Monteiro. A unified analysis of a class of proximal bundle methods for solving hybrid convex composite optimization problems. *Mathematics of Operations Research*, 49(2):832–855, 2024.

[LTP23] E. Laude, A. Themelis, and P. Patrinos. Dualities for non-Euclidean smoothness and strong convexity under the light of generalized conjugacy. *SIAM Journal on Optimization*, 33(4):2721–2749, 2023.

[LZ18] G. Lan and Y. Zhou. Random gradient extrapolation for distributed and stochastic optimization. *SIAM Journal on Optimization*, 28(4):2753–2782, 2018.

[Nes09] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[NOR10] A. Nemirovski, S. Onn, and U. G. Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1):52–78, 2010.

[NS15] Y. Nesterov and V. Shikhman. Quasi-monotone subgradient methods for nonsmooth convex minimization. *Journal of Optimization Theory and Applications*, 165(3):917–940, 2015.

[RN23] A. Rodomanov and Y. Nesterov. Subgradient ellipsoid method for nonsmooth convex problems. *Mathematical Programming*, 199(1):305–341, 2023.

[SNW11] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*. MIT press, 2011.

[SS12] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[Tib17] R. J. Tibshirani. Dykstra's algorithm, ADMM, and coordinate descent: Connections, insights, and extensions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[WAL23] J-K Wang, J. Abernethy, and K. Y. Levy. No-regret dynamics in the Fenchel game: A unified framework for algorithmic convex optimization. *Mathematical Programming*, pages 1–66, 2023.

[WR22] S. J. Wright and B. Recht. *Optimization for Data Analysis*. Cambridge University Press, 2022.

[ZH05] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

[ZZ23] R. Zhao and Q. Zhu. A generalized Frank–Wolfe method with "dual averaging" for strongly convex composite optimization. *Optimization Letters*, 17(7):1595–1611, 2023.