A Universally Optimal Primal-Dual Method for Minimizing Heterogeneous Compositions

Aaron Zoll* Benjamin Grimmer†

Abstract

This paper proposes a universal algorithm for convex minimization problems of the composite form $g_0(x) + h(g_1(x), \dots, g_m(x)) + u(x)$. We allow each g_j to independently range from being nonsmooth Lipschitz to smooth, from convex to strongly convex, described by notions of Hölder continuous gradients and uniform convexity. Note that, although the objective is built from a heterogeneous combination of such structured components, it does not necessarily possess smoothness, Lipschitzness, or any favorable structure overall other than convexity. Regardless, we provide a universal optimal method in terms of oracle access to (sub)gradients of each g_j. The key insight enabling our optimal universal analysis and a core technical contribution is the construction of two new constants, the Approximate Dualized Aggregate smoothness and strong convexity, which combine the benefits of each heterogeneous structure into single quantities amenable to analysis. As a key application, fixing h as the nonpositive indicator function, this model readily captures functionally constrained minimization $g_0(x) + u(x)$ subject to $g_j(x) \leq 0$. In particular, our algorithm and analysis are directly inspired by the smooth constrained minimization method of Zhang and Lan and consequently recover and generalize their accelerated guarantees.

1 Introduction

This paper considers the design of scalable first-order methods for the following quite general class of convex optimization problems. Given closed, convex functions $g_j,\ u\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ for $j = 0, \dots, m$, a closed, convex, component-wise nondecreasing function $h\colon \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, and a closed, convex constraint set $\mathcal{X} \subseteq \mathbb{R}^n$, we consider the convex composite optimization problem

$$p_\star = \min_{x \in \mathcal{X}} F(x) := g_0(x) + h(g_1(x), \dots, g_m(x)) + u(x)\ . \tag{1.1}$$

We are particularly interested in *heterogeneous* settings where the components g_j forming the overall objective F vary in their individual smoothness (or lack thereof) and convexity. The convex composite model (1.1) is well-studied and captures a range of standard optimization models:

- *Minimization of Finite Summations.* As perhaps the most basic composite setup, minimization of finite sums $h(z) = \sum z_j$, where each $z_j = g_j(x)$ is one component of the objective, is widespread in machine learning and data science applications. The optimization of objective functions built from heterogeneous sums of smooth and nonsmooth components was recently considered by the fine-grained theory of [10] and the bundle method theory of [26]. Universal, optimal guarantees for the minimization of any sum of heterogeneously smooth components via Nesterov's universal fast gradient method [30] were given by [15].

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

- *Functionally Constrained Optimization.* Considering the composing function as the indicator function $h(z) = \iota_{z \leq 0}(z)$ for $z_j = g_j(x)$, this model recovers the standard notion of functionally constrained optimization. This setting has been studied significantly, with the recent smooth constrained optimization work of [42] being a particular motivation for this work. Constrained optimization handles a large class of problems with applications to machine learning, statistics, and signal processing [3, 20, 22, 32].

- *Minimization of Finite Maximums.* Our model also captures minimizing finite maximums: $h(z) = \max_j z_j$ of several component functions $z_j = g_j(x)$ [32]. For example, such objectives arise as a fundamental model in game theory, in robust optimization seeking good performance across many objectives [6], and in the radial optimization framework of [16, 17].

- *Smoothed Finite Maximum and Constrained Optimization* Finally, we provide two convex composite examples that address the previous two models in a smoothed setting. First, for applications minimizing the maximum of several functions $g_j(x)$, one can instead minimize an η-smoothing of the max function [5]: for some $\eta > 0$, consider $h_\eta(z) = \eta \log(\sum_{j=1}^m \exp(z_j/\eta))$. As η tends to zero, this converges to $\max_j z_j$ but becomes less smooth. Second, consider $h_\eta(z) = \sum_{j=1}^m \max\{z_j/\eta, 0\}^2$, providing a smooth penalty for any constraint function violating nonpositivity.

Here we address convex composite problems (1.1), assuming u, h, and $\mathcal{X}$ are reasonably simple (i.e., have a computable proximal/projection operator). Note that this captures all four of the above application settings. We allow g_j to vary significantly in structure (i.e., ranging from nonsmooth Lipschitz to having Lipschitz gradients and from simple convexity to strong convexity). Section 4 and Section 5 present the considered heterogeneous models of Hölder smoothness and uniform convexity formally, but we give the following definitions here. We say that f is (L, p)-Hölder smooth for $p \in [0, 1]$ if its gradient is Hölder continuous:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|^p \quad \forall x, y \in \mathrm{dom}(f) \ . \tag{1.2}$$

As an immediate consequence of the fundamental theorem of calculus,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{p+1}\|y - x\|^{p+1} \quad \forall x, y \in \mathrm{dom}(f) \ . \tag{1.3}$$

Conversely, we say that f is (μ, q)-uniformly convex for $q \geq 1$ if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{q+1}\|y - x\|^{q+1} \quad \forall x, y \in \mathrm{dom}(f) \ . \tag{1.4}$$

We note that allowing each g_j to satisfy these conditions with their own (L_j, p_j) and (μ_j, q_j) does not guarantee F possesses any of these favorable structures besides being simply convex. Despite this lack of centralized structure, this work presents a simple first-order method attaining optimal convergence guarantees, combining and leveraging whatever structure is present in each component.

Algorithms that can be applied optimally across a range of structurally different problem settings are known as *universal* methods. Universality is a key property for developing practical algorithms capable of being widely deployed in blackbox fashion. For the case of minimizing a single function f ranging in its Hölder smoothness, optimal universal methods were first pioneered by Lan [14, 21] and Nesterov [13, 30]. Further work on universal methods allowing for convex hybrid composite models [18, 26], heterogeneous summations [15, 38], varied growth structures [19, 31], constrained optimization [8, 20, 42], and stochastic optimization [2, 35] has followed since. To varying degrees,

the above works developed algorithms that are "mostly" parameter-free, potentially relying on a target accuracy ε, an upper bound on the diameter of $\mathcal{X}$, or similar universal problem constants. Without additional parameters, stopping criteria indicating when a target accuracy is reached are often unavailable. Hence, although the above methods apply universally, they vary in how parameter-free they are.

As a concrete example of a universal method, the Universal Fast Gradient Method (UFGM) [30] can optimally minimize $F = g_0 + u$, with the structure of g_0 ranging from smooth to nonsmooth. This setup is modeled by supposing g_0 is convex with (L, p)-Hölder continuous gradient, corresponding to Lipschitz gradients when $p = 1$ and Lipschitz functions value when $p = 0$. The UFGM, given target accuracy $\varepsilon > 0$ as an input, produces a point with at most ε objective gap in either of these settings and in every intermediate one, using at most

$$K_{SM}(\varepsilon, L, p, \|x^0 - x^\star\|) = \mathrm{O}\left(\left(\frac{L}{\varepsilon}\right)^{\frac{2}{1+3p}} \|x^0 - x^\star\|^{\frac{2+2p}{1+3p}}\right) \tag{1.5}$$

(sub)gradient oracle evaluations for g_0. The matching lower bounds cited in [28, page 26] establish that this rate is optimal for every $p \in [0, 1]$. Given additional structure, like (μ, q)-uniform convexity of g_0, a universal restarting scheme like [33] can be applied to achieve the optimal, faster rates of

$$K_{UC}(\varepsilon, L, p, \mu, q) = \begin{cases} \mathrm{O}\left(\left(\frac{L^{1+q}}{\mu^{1+p}\varepsilon^{q-p}}\right)^{\frac{2}{(1+3p)(1+q)}}\right) & \text{if } q > p \\ \mathrm{O}\left(\left(\frac{L^{1+q}}{\mu^{1+p}}\right)^{\frac{2}{(1+3p)(1+q)}} \log\left(\frac{F(x^0)-F(x^\star)}{\varepsilon}\right)\right) & \text{if } q = p \end{cases} \tag{1.6}$$

(sub)gradient oracle evaluations with respect to g_0, up to logarithmic factors.

This work aims to develop a universal method for the composite setting (1.1), allowing heterogeneity in the Hölder smoothness and uniform convexity of each g_j, capturing and generalizing the settings of the above universal methods. Our proposed Universal Fast Composite Method (UFCM) is formally defined in Algorithm 1. When restarting is included, we denote it by R-UFCM, defined in Algorithm 2. Our method is not parameter-free, depending on the following three main parameters: a target accuracy $\varepsilon > 0$, an Approximate Dualized Aggregate smoothness $L^{\mathtt{ADA}}_{\varepsilon,r}$ capturing the combined effect of any upper bounding curvature present among the composition, and finally, an Approximate Dualized Aggregate convexity $\mu^{\mathtt{ADA}}_{\varepsilon}$ capturing the combined effect of any lower bounding curvature. The invention of these unifying constants, abstracting and simplifying any complex dependence on individual components' Hölder smoothness and uniform convexity constants and exponents, is key to our algorithm's success. We formally define the latter parameters in (4.2) and (5.1).

We note that one may tradeoff knowledge of $L^{\mathtt{ADA}}_{\varepsilon,r}$ for knowledge of bounds on the initial distances to optimal, D_x and D_λ, without affecting big-O oracle complexities. Further discussion is presented in Remark 4.6. The design of entirely parameter-free methods, avoiding knowledge of these aggregate constants and distance bounds, is left as an important future direction. The parameter-free techniques of [25, 30, 44] may be useful to this extent.

Measuring the convergence of a method requires a suitable notion of solution quality. Often, iterative methods seek to produce a solution x^t with a bounded objective gap:

$$F(x^t) - p_\star \leq \varepsilon \ . \tag{1.7}$$

However, for general composite problems (1.1), since F is allowed to take infinite value arbitrarily near a minimizer (an important attribute for modeling constrained optimization as discussed above), our iterative schemes for minimizing F do not directly provide a solution x^t with bounded suboptimality.

Instead, we identify (ε, r)-optimal x^t defined as there existing $\hat{g} \in \mathbb{R}^m$ and a subgradient $\hat{\lambda} \in \partial h(\hat{g})$ satisfying

$$\begin{cases} g_0(x^t) + h(\hat{g}) + \langle \hat{\lambda}, g(x^t) - \hat{g} \rangle + u(x^t) - p_\star & \leq \varepsilon \ , \\ r\|g(x^t) - \hat{g}\| & \leq \varepsilon \ . \end{cases} \tag{1.8}$$

Here, $\hat{g}$ informally serves as a perturbed projection of $g(x^t)$ onto the domain of ∂h and r is a hyperparameter that one can fix proportional to $\sqrt{\varepsilon}$ to obtain simply an "ε-optimal" solution where $\|g(x^t) - \hat{g}\|^2 \lesssim \varepsilon$ (Lemma 3.2 introduces r and discusses its meaning as a radius for dual multipliers).

This condition states x^t nearly attains the optimal objective value when the outer composition function h is linearized via a subgradient $\hat{\lambda}$ taken at a nearby $\hat{g}$. For example, in the context of constrained minimization where h is an indicator function for the nonpositive orthant, $\hat{\lambda}$ is precisely a nonnegative vector of Lagrange multipliers, making the above conditions correspond to the approximate attainment of the KKT conditions. In this case, $\hat{g}$ is a perturbed projection of $g(x^t)$ onto the nonpositive orthant. By construction, $\hat{\lambda}$ and $\hat{g}$ are orthogonal, and the conditions for an (ε, r)-optimal solution correspond to

$$\begin{cases} g_0(x^t) + \left\langle \hat{\lambda}, g(x^t) \right\rangle + u(x^t) - p_\star & \leq \varepsilon \ , \\ r\|g(x^t) - \hat{g}\| & \leq \varepsilon \ , \\ \hat{g} \leq 0, \ \hat{\lambda} \geq 0 \ . \end{cases}$$

The first condition states that x^t approximately minimizes the Lagrangian at $\hat{\lambda}$. Approximate primal feasibility follows from the second and third conditions establishing $\operatorname{dist}(g(x^t), \mathbb{R}^m_-) \leq \varepsilon/r$. Dual feasibility follows from the nonnegativity of $\hat{\lambda}$. Finally, approximate complementary slackness follows from the orthogonality of $\hat{\lambda}$ and $\hat{g}$ as $|\langle \hat{\lambda}, g(x^t) \rangle| = |\langle \hat{\lambda}, g(x^t) - \hat{g} \rangle| \leq \|\hat{\lambda}\| \varepsilon / r$.

As a second example, when h is a linear function (e.g. when directly minimizing a sum of component functions), one has $h(g(x^t)) = h(\hat{g}) + \langle \hat{\lambda}, g(x^t) - \hat{g} \rangle$ and so (1.8) reduces to the classic bounded suboptimality measure (1.7). In this case, $\hat{g} = g(x^t)$. Further discussion on the roles and values of $\hat{g}$ and $\hat{\lambda}$ is in Section 3.1

Note in our developed algorithms, both $\hat{g}$ and $\hat{\lambda}$ are not explicitly constructed, and are generally inaccessible computationally. Hence, although our analysis guarantees such values exist certifying approximate optimality, we cannot certify at runtime when the iterate becomes (ε, r)-optimal. This limitation cannot be improved: When h is linear, our optimality condition (1.8) reduces to suboptimality $F(x^t) - p_\star \leq \epsilon$, which cannot be certified without knowledge of $p_\star$ or additional global structure.

1.1 Our Contributions

This work develops a universal primal-dual method for heterogeneous compositions (1.1) with optimal first-order complexity with respect to components g_j. Our proposed UFCM and its restarted variant R-UFCM leverage the sliding technique of [23] and the "Q-analysis" technique of [42], originally designed for smooth constrained optimization. As a key contribution to this end, we propose new notions of the Approximate Dualized Aggregate smoothness constant $L^{\texttt{ADA}}_{\varepsilon,r}$ and the Approximate Dualized Aggregate convexity constant $\mu^{\texttt{ADA}}_{\varepsilon}$. These two constants provide a new unifying technical tool for the analysis of heterogeneous optimization that may be of independent interest. From these, we prove the oracle complexities outlined in Table 1. For example, in the simple setting of minimizing $g_0(x) + u(x)$, these rates recover the optimal suboptimality convergence rates of (1.5) and (1.6).

	First-Order Oracle Calls to Components g	***Proximal Oracle Calls to*** h ***and*** u	
		$L_h > D_\lambda^2/\varepsilon$	$L_h \leq D_\lambda^2/\varepsilon$
$\mu_\varepsilon^{\texttt{ADA}} < \varepsilon/D_x^2$	$\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{\varepsilon}}$	$\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{\varepsilon}} + \frac{M D_x D_\lambda}{\varepsilon}$	$\sqrt{\frac{(L_{\varepsilon,r}^{\texttt{ADA}} + M^2 L_h) D_x^2}{\varepsilon}}$
$\mu_\varepsilon^{\texttt{ADA}} \geq \varepsilon/D_x^2$	$\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}}}{\mu_\varepsilon^{\texttt{ADA}}}} \log\left(\frac{1}{\varepsilon}\right)$	$\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}}}{\mu_\varepsilon^{\texttt{ADA}}}} \log\left(\frac{1}{\varepsilon}\right) + \frac{M D_\lambda}{\sqrt{\mu_\varepsilon^{\texttt{ADA}} \varepsilon}}$	$\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}} + M^2 L_h}{\mu_\varepsilon^{\texttt{ADA}}}} \log\left(\frac{1}{\varepsilon}\right)$

Table 1: Oracle complexities in terms of universal parameters $L_{\varepsilon,r}^{\texttt{ADA}}$ and $\mu_\varepsilon^{\texttt{ADA}}$, proven in Theorem 5.5, up to constants and additive logarithmic terms in ε. Here, D_x and D_λ denote bounds on the initial primal and dual distances to optimality, and M denotes a local Lipschitz constant.

For ease of exposition, we develop our convergence theory incrementally through three main theorems:
–Theorem 3.4 establishes a $O(1/\sqrt{\varepsilon})$ rate towards ε-optimality when each g_j is smooth and convex. Hence, smooth composite optimization is nearly as easy as unconstrained smooth optimization.
–Theorem 4.3 generalizes this analysis to establish optimal rates when each g_j is convex with varying Hölder continuous gradient (1.2), recovering (1.5) as a special case.
–Theorem 5.5 finally leverages standard restarting techniques to establish optimal rates when the components additionally possess varying uniform convexity (1.4), recovering (1.6) as a special case.

Remark 1.1. *Note that these rates are only optimal for the first-order complexity with respect to the components g_j, not necessarily the proximal oracle complexity. The work of [37] shows the latter can be improved to $\mathrm{O}\left(1/\varepsilon\right)$ when g_0 is nonsmooth and Lipschitz whereas our method requires $\mathrm{O}\left(1/\varepsilon^2\right)$ proximal evaluations in such nonsmooth settings.*

1.2 Example of our Universal Constants $L_{\varepsilon,r}^{\texttt{ADA}}$ and $\mu_\varepsilon^{\texttt{ADA}}$ and an Application of Convergence Rates.

Our ability to provide universal guarantees across heterogeneous problem settings is primarily enabled by the design of our Approximate Dualized Aggregate smoothness $L_{\varepsilon,r}^{\texttt{ADA}}$ and strong convexity $\mu_\varepsilon^{\texttt{ADA}}$ constants. Although we defer formal definitions of these to (4.2) and (5.1), here we briefly discuss their essential properties and consequences. These constants are "approximate" in that they depend on the target accuracy ϵ, "dualized" in that they depend on associated optimal dual multipliers $\lambda_j^\star$, and "aggregate" in that they combine these dependencies and every problem parameter (L_j, p_j), (μ_j, q_j), etc. into a single constant. From these constants, we find that the traditional smooth and smooth, strongly convex rates $\mathrm{O}\left(\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{\varepsilon}}\right)$ and $\mathrm{O}\left(\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}}}{\mu_\varepsilon^{\texttt{ADA}}}} \log\left(\frac{1}{\varepsilon}\right)\right)$ hold for generic heterogeneous composite settings.

These unifying constants are graceful in their dependence on dual multipliers $\lambda_j^\star$: the dependence on the jth component's (L_j, p_j) and (μ_j, q_j) vanishes as $\lambda_j^\star$ tends to zero. In constrained optimization, $\lambda_j^\star = 0$ corresponds to the constraint being inactive at the optimal solution. Hence, inactive constraints play a vanishing role in our convergence rates (as one would hope). As a more concrete

example, consider minimizing a finite maximum $h(g_1(x), g_2(x))$ with $h = \max\{z_1, z_2\}$ of an L-smooth function g_1 and an M-Lipschitz nonsmooth function g_2. Here, the optimal dual multiplier $\lambda^\star \in [0, 1]$ describes the activity of each component at the minimizer, $\lambda^\star = 0$ if the problem reduces to minimizing the smooth component, $\lambda^\star = 1$ if the problem reduces to minimizing the Lipschitz component, and $\lambda^\star \in (0, 1)$ if both are active. Corollary 4.5 shows that our gradient complexity guarantees for such problems simply decompose into the sum of each component's complexity separately, weighted by its dual multiplier plus r,

$$\mathrm{O}\left(\sqrt{\frac{(1-\lambda^\star+r)LD_x^2}{\varepsilon}} + \frac{((\lambda^\star+r)M)^2D_x^2}{\varepsilon^2}\right) .$$

Selecting $r = O(\epsilon^{3/4})$, this bound transitions from the optimal accelerated smooth rate $O(1/\sqrt{\epsilon})$to the optimal nonsmooth rate $O(1/\epsilon^2)$ gracefully as $\lambda^\star \in [0, 1]$ varies.

This recovery of the optimal rates (1.5) when there is a single active component establishes near optimality of our guarantees with respect to first-order oracle evaluations of the component functions g_j. Any improvement in our dependencies in $\mathrm{O}\left(\sqrt{\frac{L_{\varepsilon,r}^{\mathtt{ADA}}D_x^2}{\varepsilon}}\right)$ and $\mathrm{O}\left(\sqrt{\frac{L_{\varepsilon,r}^{\mathtt{ADA}}}{\mu_\varepsilon^{\mathtt{ADA}}}}\log\left(\frac{1}{\varepsilon}\right)\right)$ beyond a log term would violate the lower bounds stated by [28].

Outline. Section 2 introduces preliminaries as well as the sliding technique and "Q-analysis" discussed in [42] for solving constrained optimization. Section 3 extends this method to smooth composite optimization, proving optimal guarantees. Section 4 generalizes to functions with Hölder continuous gradient. Finally, Section 5 generalizes to allow heterogeneous levels of uniform convexity.

2 Preliminaries

We define our notation to align with [42]'s prior work in constrained optimization. First, without loss of generality we set $g_0(x) = 0$ as one can consider instead minimizing $0 + \hat{h}(g_0(x), \dots g_m(x)) + u(x)$ with $\hat{h}(z_0, z_1, \dots z_m) = z_0 + h(z_1, \dots z_m)$. Hence, it suffices to consider problems of the form

$$\min_{x\in\mathcal{X}} F(x) := h(g_1(x), \dots, g_m(x)) + u(x) , \tag{2.1}$$

defined by a closed, convex set $\mathcal{X} \subseteq \mathbb{R}^n$ and the following closed, convex functions: regularizing function $u\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$, composing function $h\colon \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, and component functions $g_j\colon \mathcal{X} \to \mathbb{R}$. Below, we describe the additional structures assumed on each function.

Assumed Structure of Objective Components g, h, u. We assume each g_j is locally Lipschitz with some form of bounds on its curvature. We take each g_j to be L_j-smooth (i.e., ∇g_j is L_j-Lipschitz) in Section 3 to set up the algorithmic framework and convergence results. In Sections 4 and 5, we allow the components to have varying levels of smoothness and varying levels of convexity, as defined in (1.2) and (1.4). Whatever structure is present in these components only arises in our convergence theory through the unifying parameters $L_{\varepsilon,r}^{\mathtt{ADA}}$ and $\mu_\varepsilon^{\mathtt{ADA}}$, which aggregate any structures available, enabling our universal method and analysis. We assume $\mathcal{X}$, u, and h are sufficiently simple that their proximal operators can be evaluated, defined for any parameter $\tau > 0$ as

$$\mathrm{prox}_{u,\tau}(x) := \operatorname*{argmin}_{y\in\mathcal{X}} u(y) + \frac{\tau}{2}\|y - x\|^2, \tag{2.2}$$

$$\mathrm{prox}_{h,\tau}(x) := \operatorname*{argmin}_{y\in\mathbb{R}^m} h(y) + \frac{\tau}{2}\|y - x\|^2 \tag{2.3}$$

respectively.

The algorithms designed herein are primal-dual, leveraging the convex (Fenchel) conjugates [11] of h and each g_j. For any convex function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, we denote its conjugate as

$$f^*(s) = \sup_{x \in \mathbb{R}^n} \{\langle s, x\rangle - f(x)\} \ . \tag{2.4}$$

Note Moreau's decomposition [3, Theorem 6.45] shows $\mathrm{prox}_{h^*,\tau}(x) = x - \mathrm{prox}_{h,1/\tau}(\tau x)/\tau$ and so the assumed oracle access to prox_h via (2.3) further provides access to prox_{h^*}.

Finally, we assume h is component-wise nondecreasing, which suffices to ensure the overall objective F is convex. The following pair of standard lemmas formalize the resulting structures.

Lemma 2.1. *[34, Theorem 5.1] If $h\colon \mathbb{R}^m \to \mathbb{R}\cup\{+\infty\}$ is convex and component-wise nondecreasing and $g\colon \mathbb{R}^n \to \mathbb{R}^m$ is component-wise convex, then $c(x) := h(g(x))\colon \mathbb{R}^n \to \mathbb{R}$ is convex.*

Lemma 2.2. *If $h\colon \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is convex and component-wise nondecreasing, then* $\mathrm{dom}(h^*) \subseteq \mathbb{R}^m_+$.

Proof. Since h is component-wise nondecreasing and convex, at any $x \in \mathrm{dom}(h)$,

$$h'(x; -e_j) \le 0, \ \forall j \iff \sup_{s \in \partial h(x)} \langle s, -e_j\rangle \le 0, \ \forall j \iff \forall s \in \partial h(x), \ s \ge 0 \iff \partial h(x) \subseteq \mathbb{R}^m_+ \ .$$

Then, for any $s \in \mathrm{ri}(\mathrm{dom}(h^*))$, there exists $x \in \partial h^*(s)$, and thus $s \in \partial h(x) \subseteq \mathbb{R}^m_+$. □

Lagrangian Reformulations. We can now define a Lagrangian function essential to our algorithm and its analysis. Recalling $f = f^{**}$ for any closed, convex, and proper function [34, Corollary 12.2.1], one has that

$$h(g(x)) = \sup_{\lambda \in \Lambda} \langle \lambda, g(x)\rangle - h^*(\lambda), \quad \text{where } \Lambda := \mathrm{dom}(h^*) \ .$$

The *Standard Lagrangian* reformulation follows as

$$\inf_{x \in \mathcal{X}} h(g(x)) + u(x) = \inf_{x \in \mathcal{X}} \sup_{\lambda \in \Lambda} \{\langle \lambda, g(x)\rangle - h^*(\lambda) + u(x)\} \ . \tag{2.5}$$

Furthermore, since each $\lambda \in \mathbb{R}^m_+$ (see Lemma 2.2), one can dualize each component function g_j, obtaining the equivalent *Extended Lagrangian* reformulation, which our analysis will utilize

$$\inf_{x \in \mathcal{X}} h(g(x)) + u(x) = \inf_{x \in \mathcal{X}} \sup_{(\lambda,\nu) \in \Lambda \times V} \{\mathcal{L}(x; \lambda, \nu) := \langle \lambda, \nu x - g^*(\nu)\rangle - h^*(\lambda) + u(x)\} \ , \tag{2.6}$$

where $V := \mathrm{dom}(g^*)$. Note that $\mathcal{L}(x; \lambda, \nu)$ is convex in x and block-wise concave in λ and ν.

Define $\mathcal{Z} := \mathcal{X} \times \Lambda \times V$ for primal variables $x \in \mathcal{X}$, dual variables $\lambda \in \Lambda = \mathrm{dom}(h^*) \subseteq \mathbb{R}^m_+$, and conjugate variables $\nu \in V = \mathrm{dom}(g^*) \subseteq \mathbb{R}^{m \times n}$. Let $Z^\star$ denote the set of saddle points of (2.6), which we assume throughout is nonempty. In the case where h is linear, this assumption is equivalent to the existence of a minimizer. In the case where $h = \iota_{\le 0}$, the functionally constrained setting, this assumption is equivalent to strong duality holding with primal and dual attainment. Note any such $z^\star \in Z^\star$ must have $0 \in \partial_\lambda \mathcal{L}(x^\star; \lambda^\star, \nu^\star)$ and consequently $\lambda^\star \in \partial h(g(x^\star))$.

As a common generalization of the Euclidean distance, for any convex reference function $g\colon \mathcal{X} \to \mathbb{R}$, we define the associated Bregman divergence as

$$U_g(x; \hat{x}) := g(x) - g(\hat{x}) - \langle g'(\hat{x}), x - \hat{x}\rangle \tag{2.7}$$

for some $g'(\hat{x}) \in \partial g(\hat{x})$. If g is vector-valued, we extend the definition above by $g = (g_1, ..., g_m)$ and $U_g = (U_{g_1}, ..., U_{g_m})$. That is, U_g is vector-valued with each component being the Bregman divergence of the corresponding component of g.

2.1 Key Techniques from Prior Works

Our results rely on four technical tools developed over the last decade that we bring together to handle various facets of the general problem (1.1). We introduce these formally below. In short, Lan's sliding technique [23] allows us to decompose the complexity concerning proximal steps on u and h from that of gradient calls to g; the Q function analysis of [42] provides the primal-dual framework from which UFCM is built; the technique to universally analyze Hölder smooth functions from [30] allows our results to generalize beyond smooth optimization; restarted methods allow us to generalize our results further to benefit from any uniform convexity present among its components.

Sliding Gradient Methods. The sliding technique introduced by Lan [23, 24] iteratively and approximately solves subproblems associated with accelerated proximal gradient methods. This approach was first developed to handle objectives $g_0 + u$, with g_0 smooth and u nonsmooth but with readily available subgradients. The sliding gradient method allows the number of first-order oracle calls to g_0 and u to be decomposed, often significantly reducing the number of calls needed to g_0. In the context of our considered method, a central step of our method requires a proximal step on a certain minimax optimization subproblem involving u and h^*. Sliding performs this step approximately, decomposing computations related to ∇g, $\text{prox}_{u,\tau}$, and $\text{prox}_{h^*,\tau}$.

Q Function Framework for Constrained Optimization. The novel work of [42] considered the problem of minimizing $g_0 + u$ subject to inequality constraints $g_j(x) \leq 0$, corresponding in our model to minimizing $F = g_0(x) + \iota_{z\leq 0}(g_1, \ldots, g_m) + u(x)$. The key step therein is designing algorithms generating iterates z^t driving an associated "gap function" providing a measure of optimality on the extended Lagrangian reformulation to zero[1]:

$$Q(z^t, z) := \mathcal{L}(x^t; \lambda, \nu, \pi) - \mathcal{L}(x; \lambda^t, \nu^t, \pi^t) \ .$$

Their proposed accelerated method for smooth constrained optimization works by optimizing the Q function separately with each block of variables. Their updates concerning π and ν amount to computing gradients of g_0 and g_j. Their updates for x and λ correspond to solving a quadratic program, which a sliding technique is able to decompose.

Our theory recovers these results of Zhang and Lan, improving their guarantees in settings with strongly convex constraints, enabling it to apply universally to compositions (not just constrained optimization) and to problems with Hölder smooth and/or uniformly convex components. Section 3.1 formally develops our generalization of their Q function framework. After developing our convergence theory, Section 5.6 provides a detailed comparison of results.

Universal Methods for Minimization with Hölder Continuous Gradient.

Nesterov's universal fast gradient method [30] provided a generalization of Nesterov's classic fast gradient method [29] capable of minimizing any (L, p)-Hölder smooth objective. The key technical insight originating from [9] that enables this method is a lemma establishing an approximate smoothness result for any such function, meaning the standard quadratic upper bound derived for functions with Lipschitz gradient holds for functions with Hölder gradient up to an additive constant. A variant of this lemma showcasing a standard cocoercivity inequality also generalizes at the cost of an additive constant, derived by [25].

[1]Here π is dual multiplier corresponding to the function g_0, always equal to $\pi^t := \nabla g_0(x^t)$. We omit this variable from the formulation considered throughout this work as without loss of generality, we set $g_0 = 0$.

Lemma 2.3 (Lemma 1, Nesterov [30]). *For any tolerance $\delta > 0$ and (L, p)-Hölder smooth function $f\colon \mathcal{X} \to \mathbb{R}$ with $L_\delta \geq \left[\frac{1-p}{1+p}\frac{1}{\delta}\right]^{\frac{1-p}{1+p}} L^{\frac{2}{1+p}}$,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_\delta}{2}\|y - x\|^2 + \frac{\delta}{2},\ \forall x, y \in \mathrm{dom}(f)\ . \tag{2.8}$$

Lemma 2.4 (Lemma 1, Li and Lan [25]). *In the same setting as Lemma 2.3,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L_\delta}\|\nabla f(x) - \nabla f(y)\|^2 - \frac{\delta}{2},\ \forall x, y \in \mathrm{dom}(f)\ . \tag{2.9}$$

These lemmas facilitate our generalization in Section 4 from smooth components to heterogeneously Hölder smooth components. Our Approximate Dualized Aggregate smoothness constant $L^{\mathtt{ADA}}_{\varepsilon,r}$ is a further generalization of the approximate smoothness constants L_δ seen above. Namely, $L^{\mathtt{ADA}}_{\varepsilon,r}$ further aggregates the Hölder smoothness of each component g_j, weighted approximately by the corresponding optimal dual multiplier $\lambda^\star_j$.

Restarting Gradient Methods. Algorithmic restarting, dating back to at least [28], can be shown to accelerate the convergence rate of first-order methods. The more recent works [33, 36, 39] established improved convergence guarantees given strong or uniform convexity or any general Hölderian growth. The analysis of such schemes tends to rely on ensuring a reduction, often a contraction, in the distance to optimal occurs at each restart. Such schemes have found particular success in primal-dual algorithm design for linear and quadratic programming [1, 27].

In our analysis, two distances to optimal are traced based on the distance from primal iterates x^t to $x^\star$ and the distance from the dual iterates λ^t to $\lambda^\star$. Given any uniform convexity among the components g_j, our Approximate Dualized Aggregate convexity $\mu^{\mathtt{ADA}}_\varepsilon$ describes the improvement in convergence gained from restarting the primal iterate sequence. Given any smoothness L_h in the composing function h, improved convergence follows from restarting the dual iterate sequence. The relative sizes of $\mu^{\mathtt{ADA}}_\varepsilon$ and L_h determine our various rates previously claimed in Table 1.

3 Minimization of Compositions with Smooth Components

For ease of exposition, in this section, we first develop our main algorithm UFCM, assuming each component function g_j is L_j-smooth and convex. The following two sections provide extensions to benefit from any Hölder smoothness and uniform convexity present in each g_j and any smoothness present in h. Section 3.1 formalizes the Q analysis framework for our composite optimization context, and Section 3.2 introduces our unifying Approximate Dualized Aggregate smoothness parameter $L^{\mathtt{ADA}}_{\varepsilon,r}$. Then, Section 3.3 presents our first convergence guarantee, only requiring the Approximate Dualized Aggregate smoothness $L^{\mathtt{ADA}}_{\varepsilon,r}$ (or any upper bound) as input. Finally, Section 3.4 provides the key steps in our analysis, deferring any reasoning directly generalizing the constrained optimization analysis of [42] to the appendix.

3.1 Q Function Framework for Composite Optimization

We can now introduce our generalization of the Q analysis framework of [42] that drives this paper. Based on the extended Lagrangian (2.6), we define an analogous gap function.

Definition 3.1. *Given functions g, h, u defining an instance of (2.1), the gap function is defined as*

$$Q(z, \hat{z}) := \mathcal{L}(x; \hat{\lambda}, \hat{\nu}) - \mathcal{L}(\hat{x}; \lambda, \nu)\ . \tag{3.1}$$

Fixing $h(\cdot)$ as the indicator for the nonpositive orthant recovers their definition.

For the sake of our analysis, we fix an arbitrary saddle point $z^\star := (x^\star; \lambda^\star, \nu^\star)$ with $\nu^\star := \nabla g(x^\star)$. Note $\mathcal{L}(x^\star; \lambda, \nu) \leq \mathcal{L}(x^\star; \lambda^\star, \nu^\star) \leq \mathcal{L}(x; \lambda^\star, \nu^\star)$. Hence, $Q(z, z^\star) \geq 0$ for all $z \in \mathcal{Z}$, making convergence of $Q(z^t, z^\star)$ a potential measure of solution quality. Our analysis considers a slight modification, allowing perturbations of $\lambda^\star$ and $\nu^\star$, giving a condition that implies z^t is an (ε, r)-optimal solution (1.8) for our original composite problem. To this end, we restrict to considering λ within a fixed distance r of $\lambda^\star$ and in the dual domain $\Lambda = \mathrm{dom}(h^*)$, denoted

$$\Lambda_r := B(\lambda^\star, r) \cap \Lambda \ , \tag{3.2}$$

where $B(\lambda^\star, r)$ is the closed ball of radius r centered at $\lambda^\star$.

Given a candidate primal solution x^t, for analysis sake only, we define the following perturbed component function value

$$\hat{g} \in \operatorname*{argmin}_{w \in \mathrm{dom}(h)} h(w) + \langle -\lambda^\star, w \rangle + r\|w - g(x^t)\|, \tag{3.3}$$

which exists as the objective has compact level sets. In particular, $h(w) - \langle \lambda^\star, w \rangle \geq -h^*(\lambda^\star)$, by the Fenchel-Young inequality. It then holds that for any $z \in \mathbb{R}$,

$$\{w \in \mathbb{R}^m : h(w) + \langle -\lambda^\star, w \rangle + r\|w - g(x^t)\| \leq z\} \subseteq \{w \in \mathbb{R}^m : -h^*(\lambda^\star) + r\|w - g(x^t)\| \leq z\}$$

where the larger set is bounded. From this, for analysis sake only, we define the following associated perturbed dual variables as

$$\hat{\lambda} := \begin{cases} \lambda^\star + r\frac{g(x^t) - \hat{g}}{\|g(x^t) - \hat{g}\|} & \hat{g} \neq g(x^t) \\ \lambda^\star + r\zeta & \text{otherwise.} \end{cases},$$

where $\zeta \in B(0, 1)$ is an appropriate perturbation such that $\lambda^\star + r\zeta \in \partial h(\hat{g})$, which is guaranteed by first-order optimality conditions. Note that $\hat{\lambda} \in \partial h(\hat{g})$, implying $\hat{\lambda} \in \Lambda$ (since $\hat{g} \in \partial h^*(\hat{\lambda})$), so $\hat{\lambda} \in \Lambda_r$. The following lemma relates (ε, r)-optimality to the evaluation of Q at z^t with respect to $(x^\star, \hat{\lambda}, \nabla g(x^t))$.

Lemma 3.2. *For any $z^t = (x^t; \lambda^t, \nu^t) \in \mathcal{Z}$ and $\varepsilon > 0$, if $Q(z^t, (x^\star, \hat{\lambda}, \nabla g(x^t))) \leq \varepsilon$, then x^t is (ε, r)-optimal* (1.8).

Proof. Let $\hat{\nu} = \nabla g(x^t)$. Since $\hat{\lambda} \in \partial h(\hat{g}) \cap \Lambda_r$, the first condition for the (ε, r)-optimality of x^t holds as

$$\begin{aligned} \left[h(\hat{g}) + \left\langle \hat{\lambda}, g(x^t) - \hat{g} \right\rangle + u(x^t)\right] - F(x^\star) &\leq \left[h(\hat{g}) + \left\langle \hat{\lambda}, g(x^t) - \hat{g} \right\rangle + u(x^t)\right] - \mathcal{L}(x^\star; \lambda^t, \nu^t) \\ &= \left[\left\langle \hat{\lambda}, g(x^t) \right\rangle - h^*(\hat{\lambda}) + u(x^t)\right] - \mathcal{L}(x^\star; \lambda^t, \nu^t) \\ &= \left[\left\langle \hat{\lambda}, \hat{\nu} x^t - g^*(\hat{\nu}) \right\rangle - h^*(\hat{\lambda}) + u(x^t)\right] - \mathcal{L}(x^\star; \lambda^t, \nu^t) \\ &= Q(z^t, (x^\star; \hat{\lambda}, \hat{\nu})) \leq \varepsilon \ , \end{aligned}$$

where the first inequality simply bounds $F(x^\star)$ below by $\mathcal{L}(x^\star; \lambda^t, \nu^t)$ and the following two equalities apply the Fenchel-Young inequality, holding with equality since $\hat{\lambda} \in \partial h(\hat{g})$ and $\hat{\nu} = \nabla g(x^t)$.

To show the second condition for (ε, r)-optimality holds, if $\hat{g} = g(x^t)$ then this result is trivial. Otherwise, we note that since $(x^\star, \lambda^\star)$ is a saddle point to (2.5),

$$0 \leq \left\langle \lambda^\star, g(x^t) \right\rangle - h^*(\lambda^\star) + u(x^t) - \left[\left\langle \lambda^t, g(x^\star) \right\rangle - h^*(\lambda^t) + u(x^\star)\right] \ . \tag{3.4}$$

Consequently,

$$\begin{aligned}
r\|g(x^t)-\hat{g}\| &\leq \left\langle \frac{r(g(x^t)-\hat{g})}{\|g(x^t)-\hat{g}\|}, g(x^t)-\hat{g} \right\rangle + \left\langle \lambda^\star, g(x^t) \right\rangle - h^*(\lambda^\star) + u(x^t) \\
&\qquad - \left[\left\langle \lambda^t, g(x^\star) \right\rangle - h^*(\lambda^t) + u(x^\star) \right] \\
&= \left\langle \hat{\lambda}, g(x^t) \right\rangle - \left\langle \frac{r(g(x^t)-\hat{g})}{\|g(x^t)-\hat{g}\|}, \hat{g} \right\rangle - \langle \lambda^\star, \hat{g} \rangle + \langle \lambda^\star, \hat{g} \rangle - h^*(\lambda^\star) + u(x^t) \\
&\qquad - \left[\left\langle \lambda^t, g(x^\star) \right\rangle - h^*(\lambda^t) + u(x^\star) \right] \\
&\leq \left\langle \hat{\lambda}, g(x^t) \right\rangle - \left\langle \hat{\lambda}, \hat{g} \right\rangle + h(\hat{g}) + u(x^t) - \left[\left\langle \lambda^t, g(x^\star) \right\rangle - h^*(\lambda^t) + u(x^\star) \right] \\
&\leq Q(z^t, (x^\star, \hat{\lambda}, \hat{\nu})) \leq \varepsilon
\end{aligned}$$

where the first inequality follows from (3.4), and the second and third apply Fenchel-Young. □

3.2 An Approximate Dualized Aggregate Smoothness Constant

If one knew the optimal dual multipliers $\lambda^\star$, the convex composite optimization problem (2.1) could be rewritten as the simpler minimization problem of

$$\min_{x \in \mathcal{X}} \sum_{j=1}^{m} \lambda_j^\star g_j(x) + u(x) \ , \tag{3.5}$$

which can be addressed by accelerated (regularized) smooth optimization methods like FISTA [4]. In this simplified problem, $\sum \lambda_j^\star g_j(x)$ is $\sum \lambda_j^\star L_j$-smooth, aggregating the individual smoothness constants weighted by the optimal dual multiplier. Without knowing $\lambda^\star$, we aim to approximate this aggregate dualized constant. Our theory instead depends on the slightly larger constant given by considering all λ in the neighborhood of $\lambda^\star$ given by Λ_r. Given each g_j is L_j-smooth, we denote this "Approximate Dualized Aggregate" smoothness constant by

$$L_{\varepsilon,r}^{\texttt{ADA}} := \sum_{j=1}^{m} (\lambda_j^\star + r) L_j \ . \tag{3.6}$$

As r tends to zero, $L_{\varepsilon,r}^{\texttt{ADA}}$ converges to the idealized value $\sum \lambda_j^\star L_j$. Note this only depends on the target accuracy $r > 0$, not $\epsilon > 0$. We include this dependence in our notation as the appropriate generalization to Hölder smooth settings given in equation (4.2) will depend on both. Further generality will be introduced when the components possess uniform convexity in equation (5.2). The special case of constrained optimization minimizing $g_0(x) + u(x)$ subject to $g_j(x) \leq 0$ provides a particularly nice application to understand $L_{\varepsilon,r}^{\texttt{ADA}}$. There, $h(z_0, \ldots, z_m) = z_0 + \iota_{z \leq 0}(z_1, \ldots, z_m)$, so $\lambda_0^\star = 1$ while $\lambda_1^\star, \ldots, \lambda_m^\star$ are the optimal dual multipliers for each constraint. With $L_{\varepsilon,r}^{\texttt{ADA}} = L_0 + \sum_{j=1}^{m} (\lambda_j^\star + r) L_j$, the smoothness of the objective always plays a role while only the smoothness of active constraints at the minimizer can nontrivially affect the convergence rate (that is, complementary slackness ensures that $\lambda_j^\star = 0$ for each inactive constraint).

3.3 The Universal Fast Composite Gradient Method

UFCM works primarily by splitting and optimizing $Q(z^t, z)$ on its primal, dual, and conjugate variables separately. This process formalized below is directly analogous to the algorithm design of the

ACGD-S method for smooth constrained optimization of [42], extended to allow a general proximal step on h^* and the usage of our new $L_{\varepsilon,r}^{\texttt{ADA}}$ constant. With these established, parameter choices only need slight modifications. Hence, UFCM generalizes ACGD-S to heterogeneous and composite settings. We define these three components such that $Q(z^t, z) = Q_\nu(z^t, z) + Q_x(z^t, z) + Q_\lambda(z^t, z)$ as

$$Q_\nu(z^t, z) = \mathcal{L}(x^t; \lambda, \nu) - \mathcal{L}(x^t; \lambda, \nu^t) = \left\langle \lambda, \nu x^t - g^*(\nu) \right\rangle \boxed{- \left\langle \lambda, \nu^t x^t - g^*(\nu^t) \right\rangle},$$

$$Q_x(z^t, z) = \mathcal{L}(x^t; \lambda^t, \nu^t) - \mathcal{L}(x; \lambda^t, \nu^t) = \boxed{\left\langle \sum_{j=1}^m \lambda_j^t \nu_j^t, x^t \right\rangle + u(x^t)} - \left\langle \sum_{j=1}^m \lambda_j^t \nu_j^t, x \right\rangle - u(x),$$

$$Q_\lambda(z^t, z) = \mathcal{L}(x^t; \lambda, \nu^t) - \mathcal{L}(x^t; \lambda^t, \nu^t) = \left\langle \lambda, \nu^t x^t - g^*(\nu^t) \right\rangle - h^*(\lambda) \boxed{- \left[\left\langle \lambda^t, \nu^t x^t - g^*(\nu^t) \right\rangle - h^*(\lambda^t) \right]}.$$

Each boxed term above corresponds to the component depending on the next iterate ν^t, x^t, λ^t. We aim to minimize each subproblem with respect to z^t; thus, we minimize each boxed value. Informally, UFCM proceeds by first computing a momentum step in x, denoted by $\tilde{x}^t = x^{t-1} + \theta_t(x^{t-1} - x^{t-2})$, and then computing (potentially many) proximal operator-type steps in each of ν, x, λ corresponding to

$$\nu_j^t \leftarrow \operatorname*{argmax}_{\nu_j \in V_j} \left\langle \nu_j, \tilde{x}^t \right\rangle - g_j^*(\nu_j) - \tau_t U_{g_j^*}(\nu_j; \nu_j^{t-1}),$$
$$(x^t, \lambda^t) \leftarrow \operatorname*{argmin}_{x \in \mathcal{X}} \max_{\lambda \in \Lambda} \left\langle \lambda, \nu^t x - g^*(\nu^t) \right\rangle + u(x) - h^*(\lambda) + \tfrac{\eta_t}{2} \|x - x^{t-1}\|^2$$

In the above, θ_t parametrizes the momentum step and the nonnegative parameters τ_t and η_t are stepsizes for the proximal steps. Recall $U_{g_j^*}$ is the Bregman divergence generated by g_j^*. Note that solving ν^t utilizes a Bregman divergence instead of the standard Euclidean distance, as it can be shown recursively that this is identical to a gradient evaluation of g_j at a particular averaged point [41, Lemma 2].

Solving the second subproblem is not as simple. We utilize the sliding technique to take alternating proximal steps with respect to x and λ, without addition to the gradient oracle complexity. We further employ two more nonnegative parameters, $\beta^{(t)}$ and $\gamma^{(t)}$, respectively handling the proximal steps on u and h^* for the inner loop iterates.

Formally, UFCM defined in Algorithm 1 proceeds by iteratively applying a momentum update and the above update to ν in the outer loop. The inner loop using the sliding technique to apply several proximal steps to compute the above update to (x, λ) without requiring any additional first-order evaluations of g. As computational notes, Line 11 saves previous iterates $y_0^{(t+1)}$, $\lambda_0^{(t+1)}$, and $\lambda_{-1}^{(t+1)}$ for use in the next inner loop. The subtle change from ν^t to ν^{t-1} in the two cases defined in Line 7 is common for sequential dual type algorithms using the sliding technique [40, 42, 43].

3.4 Guarantees for Composite Optimization with Smooth Components

We begin this section by introducing the two oracle complexities we bound with respect to finding an (ε, r)-optimal solution. We denote the gradient complexity of UFCM by $N_{\varepsilon,r}$ if for any $T \geq N_{\varepsilon,r}$, $\bar{x}^T$ is guaranteed to be an (ε, r)-optimal point. Likewise, we denote the proximal complexity of UFCM by $P_{\varepsilon,r}$ if at most $\lceil P_{\varepsilon,r} \rceil$ proximal evaluations of u and h are guaranteed to be performed in the first $\lceil N_{\varepsilon,r} \rceil$ outer loop iterations of UFCM.

Algorithm 1 Universal Fast Composite Method (UFCM)

Input $z^0 \in \mathcal{X} \times \Lambda$, outer loop iteration count T, and smoothness constant $L_{\varepsilon,r}^{\texttt{ADA}}$

Initialize $x^{-1} = \underline{x}^0 = y_0^{(1)} = x^0 \in \mathcal{X}$, $\lambda_{-1}^{(1)} = \lambda_0^{(1)} = \lambda^0 \in \Lambda$, and parameters $\{\theta_t\}, \{\eta_t\}, \{\tau_t\}$, $\{\omega_t\}$ as a function of $L_{\varepsilon,r}^{\texttt{ADA}}$

Set $\nu^0 = \nabla g(x^0)$.
for $t = 1,\ 2,\ 3,\ ...,\ T$ **do**
  Set $\underline{x}^t \leftarrow (\tau_t \underline{x}^{t-1} + \tilde{x}^t)/(1+\tau_t)$ where $\tilde{x}^t = x^{t-1} + \theta_t(x^{t-1} - x^{t-2})$
  Set $\nu^t \leftarrow \nabla g(\underline{x}^t)$
  Calculate inner loop iteration limit S_t, parameters $\beta^{(t)}$, $\gamma^{(t)}$, and $\rho^{(t)}$
  for $s = 1, 2, ..., S_t$ **do**
    Set $\tilde{h}^{(t),s} = \begin{cases} (\nu^t)^\top \lambda_0^{(t)} + \rho^{(t)}(\nu^{t-1})^\top(\lambda_0^{(t)} - \lambda_{-1}^{(t)}) & \text{if } s = 1, \\ (\nu^t)^\top \lambda_{s-1}^{(t)} + (\nu^t)^\top(\lambda_{s-1}^{(t)} - \lambda_{s-2}^{(t)}) & \text{otherwise} \end{cases}$
    Solve $y_s^{(t)} \leftarrow \underset{y \in \mathcal{X}}{\operatorname{argmin}} \left\langle \tilde{h}^{(t),s}, y \right\rangle + u(y) + \frac{\eta_t}{2}\|y - x^{t-1}\|^2 + \frac{\beta^{(t)}}{2}\|y - y_{s-1}^{(t)}\|^2$
    Solve $\lambda_s^{(t)} \leftarrow \underset{\lambda \in \Lambda}{\operatorname{argmax}} \left\langle \lambda, \nu^t(y_s^{(t)} - \underline{x}^t) + g(\underline{x}^t) \right\rangle - h^*(\lambda) - \frac{\gamma^{(t)}}{2}\|\lambda - \lambda_{s-1}^{(t)}\|^2$
  end for
  Set $\lambda_0^{(t+1)} = \lambda_{S_t}^{(t)}$, $\lambda_{-1}^{(t+1)} = \lambda_{S_t-1}^{(t)}$, $y_0^{(t+1)} = y_{S_t}^{(t)}$
  Set $x^t = \sum_{s=1}^{S_t} y_s^{(t)}/S_t$ and $\tilde{\lambda}^t = \sum_{s=1}^{S_t} \lambda_s^{(t)}/S_t$
end for
return $(\bar{x}^T, \bar{\lambda}^T) := \sum_{t=1}^T \omega_t \left(x^t, \tilde{\lambda}^t\right) / \left(\sum_{t=1}^T \omega_t\right)$

To ensure that UFCM converges to an (ϵ, r)-optimal solution, we place several requirements on the selection of its parameters. For each outer loop $t \geq 1$, we require that

$$\omega_t \eta_t \leq \omega_{t-1}\eta_{t-1} \tag{3.7}$$
$$\omega_t \tau_t \leq \omega_{t-1}(\tau_{t-1} + 1) \tag{3.8}$$
$$\eta_{t-1}\tau_t \geq \theta_t L_{\varepsilon,r}^{\texttt{ADA}} \quad \text{with} \quad \theta_t = \omega_{t-1}/\omega_t \tag{3.9}$$
$$\eta_T(\tau_T + 1) \geq L_{\varepsilon,r}^{\texttt{ADA}} \tag{3.10}$$
$$\gamma^{(t)}\beta^{(t)} \geq \|\nu^t\|^2 \tag{3.11}$$
$$\tilde{\omega}^{(t)}\beta^{(t)} \geq \tilde{\omega}^{(t+1)}\beta^{(t+1)} \tag{3.12}$$
$$\tilde{\omega}^{(t)}\gamma^{(t)} \geq \tilde{\omega}^{(t+1)}\gamma^{(t+1)} \tag{3.13}$$
$$\gamma^{(t)}\beta^{(t)} \geq (\rho^{(t)})^2\|\nu^{t-1}\|^2 \quad \text{with} \quad \rho^{(t+1)} = \tilde{\omega}^{(t)}/\tilde{\omega}^{(t+1)} \tag{3.14}$$

where $\tilde{\omega}^{(t)} := \omega_t/S_t$ denotes the aggregate weights.

Although our algorithm converges for any selection satisfying these requirements, optimized performance follows from particular choices. In particular, our main convergence guarantee below requires knowledge of an upper bound on $L_{\varepsilon,r}^{\texttt{ADA}}$ to set parameters. Some of our convergence guarantee corollaries additionally assume knowledge of positive bounds on the initial distances to a saddle point $D_x \geq \|x^0 - x^\star\|$ and $D_\lambda \geq \|\lambda^0 - \lambda^\star\|$.

As a first result, we establish that a careful setting of stepsizes ensures that the primal iterates

are always bounded and that the dual iterates are bounded if h is L_h-smooth[2]. The parameters of Algorithm 1 below are further parameterized by the choice of two balancing parameters C and Δ.

Proposition 3.3. *Consider any problem of the form* (2.1) *and constants* $\Delta, C, \epsilon, r > 0$*, and suppose Algorithm 1 is run with outer loop stepsizes set as*

$$\tau_t = \frac{t-1}{2}, \;\; \eta_t = \frac{L_{\varepsilon,r}^{\texttt{ADA}}}{\tau_{t+1}}, \;\; \theta_t = \frac{\tau_t}{\tau_{t-1}+1}, \;\; \omega_t = t \;, \tag{3.15}$$

and inner loop stepsizes set as

$$\rho^{(t)} = \tilde{M}_t/\tilde{M}_{t-1}, \quad \beta^{(t)} = C\tilde{M}_t, \quad \gamma^{(t)} = \frac{\tilde{M}_t^2}{\beta^{(t)}} = \frac{\tilde{M}_t}{C} \;, \tag{3.16}$$

with $M_t = \|\nu^t\|$*,* $S_t = \lceil M_t \Delta t \rceil$*,* $\tilde{M}_t = \frac{S_t}{\Delta t}$*. Then*

$$\|x^t - x^\star\|^2 \leq \frac{1}{2L_{\varepsilon,r}^{\texttt{ADA}}} \left[(C/\Delta + 2L_{\varepsilon,r}^{\texttt{ADA}}) \|x^0 - x^\star\|^2 + \frac{1}{C\Delta} \|\lambda^0 - \lambda^\star\|^2 \right] \;. \tag{3.17}$$

Furthermore, if h *is* L_h*-smooth, then for averaged iterates* $\tilde{\lambda}^t$ *computed each loop,*

$$\|\tilde{\lambda}^t - \lambda^\star\|^2 \leq L_h(M\Delta + 1) \left[(C/\Delta + 2L_{\varepsilon,r}^{\texttt{ADA}}) \|x^0 - x^\star\|^2 + \frac{1}{C\Delta} \|\lambda^0 - \lambda^\star\|^2 \right] \;. \tag{3.18}$$

where M *is an upper bound for* $\|\nabla g(x)\|$ *in the neighborhood outlined above* (3.17)*.*

Moreover, under such choices, the following theorem explicitly bounds the number of gradient and proximal oracle calls required to reach any target (ϵ, r)-optimality.

Theorem 3.4. *Consider any problem of the form* (2.1) *with each* g_j *being* L_j*-smooth, and constants* $\Delta, C, \epsilon, r > 0$*. Then Algorithm 1 with stepsizes* (3.15) *and* (3.16) *must find an* (ε, r)*-optimal solution* (1.8) *with complexity bounds*

$$N_{\varepsilon,r} = \sqrt{\frac{(C/\Delta + 2L_{\varepsilon,r}^{\texttt{ADA}}) D_x^2 + 2/(C\Delta)(D_\lambda^2 + r^2)}{\varepsilon}}, \quad P_{\varepsilon,r} = \lceil N_{\varepsilon,r} \rceil + \lceil N_{\varepsilon,r} \rceil^2 \Delta M, \tag{3.19}$$

where M *is an upper bound for* $\|\nabla g(x)\|$ *in the neighborhood outlined in* (3.17)

The following corollaries simplify the above bounds by considering particular choices of Δ, C, and r. The first corollary presents an upper bound in terms of a primal-dual distance while avoiding reliance on knowledge of any upper bounds on initial distances to optimality. The second corollary provides an improved gradient complexity bound depending only on primal distances at the cost of requiring knowledge of upper bounds on the initial primal and dual distances to a saddle point. Our extended theory in Section 4 and Section 5 will focus on generalizing this second, stronger result. The remainder of this section is dedicated to proving these results.

Corollary 3.5. *For any* $\epsilon > 0$*, setting* $C = \sqrt{2}/2$*,* $\Delta = \frac{\sqrt{2}}{4L_{\varepsilon,r}^{\texttt{ADA}}}$ *and* $r = \sqrt{\varepsilon}$*, Algorithm 1 with stepsizes* (3.15) *and* (3.16) *must find an* ε*-optimal solution with complexity bounds*

$$N_{\varepsilon,r} = \mathrm{O}\left(\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}}(D_x^2 + D_\lambda^2 + \varepsilon)}{\varepsilon}} \right), \quad P_{\varepsilon,r} = \mathrm{O}\left(\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}}(D_x^2 + D_\lambda^2 + \varepsilon)}{\varepsilon}} + \frac{M(D_x^2 + D_\lambda^2 + \varepsilon)}{\varepsilon} \right)$$

where M *is an upper bound on* $\|\nabla g(x)\|$ *for all* $x \in B(x^\star, \sqrt{2(D_x^2 + D_\lambda^2)})$

[2]We will abuse notation in the setting of general, nonsmooth h, saying h is $L_h = \infty$-smooth in this limiting case.

Corollary 3.6. *For any* $0 < \epsilon \leq \min\{1, 12L_{\varepsilon,r}^{\texttt{ADA}} D_x^2\}$*, setting* $C = D_\lambda / D_x$*,* $\Delta = C/2L_{\varepsilon,r}^{\texttt{ADA}}$*, and* $r = D_\lambda \sqrt{\varepsilon}$*, Algorithm 1 with stepsizes* (3.15) *and* (3.16) *must find an* ε*-optimal solution with complexity bounds*

$$N_{\varepsilon,r} = \mathrm{O}\left(\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{\varepsilon}}\right), \quad P_{\varepsilon,r} = \mathrm{O}\left(\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{\varepsilon}} + \frac{M D_x D_\lambda}{\varepsilon}\right) ,$$

where M *is an upper bound on* $\|\nabla g(x)\|$ *for all* $x \in B(x^\star, \sqrt{3D_x^2})$*.*

3.5 Analysis of UFCM for Compositions with Smooth Components

Our theory primarily follows from a sequence of three lemmas, which directly extend equivalent results developed for the case of smooth constrained optimization by [42]. In our analysis, note that we select $\hat{g}$, our analytical proxy for $g(x^t)$, differently than the projection choice used by Zhang and Lan. Throughout, we let $L_h \in (0, \infty]$ denote the smoothness constant of h, set to be ∞ if h is nonsmooth, as occurs in the special case of constrained optimization. For each result, we refer to the paralleled proof in their special case. For results requiring generalization, we defer the proofs to Appendix A.1. Our results show that the analysis technique of Zhang and Lan is quite robust, generalizing to compositions, managing new h^* terms, and benefiting from any smoothness in h.

The first lemma provides a useful smoothness bound on the Lagrangian from our Approximate Dualized Aggregate smoothness constant.

Lemma 3.7 (Lemma 2, Zhang and Lan [42])**.** *If each* g_j *is* L_j*-smooth, then*

$$\langle \lambda, U_{g^*}(\nu; \hat{\nu}) \rangle \geq \frac{1}{2L_{\varepsilon,r}^{\texttt{ADA}}} \left\| \sum_{j=1}^m \lambda_j (\nu_j - \hat{\nu}_j) \right\|^2 , \ \forall \lambda \in \Lambda_r, \ \forall \nu, \hat{\nu} \in \{\nabla g(x) : x \in \mathcal{X}\} .$$

Next, we provide a general convergence bound on the Q_x and Q_λ functions associated with the primal and dual variables, extending the result of [42, Equation (4.19)] and proven in Appendix A.1. When h is nonsmooth (i.e., $L_h = +\infty$), the quantity $1/L_h$ below should be interpreted at zero.

Lemma 3.8. *Suppose the stepsizes satisfy* (3.7)-(3.14)*, and let* $z^t := (x^t; \tilde{\lambda}^t, \nu^t)$ *denote the iterates of Algorithm 1. Then* z^t *satisfies the following for any* $z = (x; \lambda, \nu) \in \mathcal{X} \times \Lambda_r \times V$*,*

$$\sum_{t=1}^T \omega_t [Q_x(z^t, z) + Q_\lambda(z^t, z)] + \sum_{t=1}^T \sum_{s=1}^{S_t} \frac{\omega_t}{S_t} \frac{1}{2L_h} \|\lambda_s^{(t)} - \lambda\|^2 + \sum_{t=1}^T \frac{\omega_t \eta_t}{2} \|x^t - x^{t-1}\|^2$$
$$+ \frac{\omega_T \eta_T}{2} \|x^T - x\|^2 - \frac{\omega_1 \eta_1}{2} \|x^0 - x\|^2 \leq \frac{\tilde{\omega}^{(1)}}{2} \left(\gamma^{(1)} \|\lambda_0^{(1)} - \lambda\|^2 + \beta^{(1)} \|y_0^{(1)} - x\|^2 \right) .$$

Lastly, we provide a general convergence bound on the Q_ν function associated with the conjugate variables ν, which requires only mild modifications from the analysis of Zhang and Lan [42, Proposition 2], proven in Appendix A.1.

Lemma 3.9. *Suppose the stepsizes satisfy* (3.7)-(3.14)*. Then* z^t *satisfies the following for any*

$z = (x; \lambda, \nu) \in \mathcal{X} \times \Lambda_r \times V$,

$$\begin{aligned}\sum_{t=1}^{T} \omega_t \left[Q_\nu(z^t, z)\right] \leq &- \left[\omega_T(\tau_T + 1)\left(\sum_{j=1}^{m} \lambda_j U_{g_j^*}(\nu_j; \nu_j^T)\right) - \omega_T \left\langle \sum_{j=1}^{m} \lambda_j(\nu_j - \nu_j^T), x^T - x^{T-1} \right\rangle\right] \\ &- \sum_{t=2}^{T} \left[\omega_t \tau_t \left(\sum_{j=1}^{m} \lambda_j U_{g_j^*}(\nu_j^t; \nu_j^{t-1})\right) - \omega_{t-1} \left\langle \sum_{j=1}^{m} \lambda_j(\nu_j^{t-1} - \nu_j^t), (x^{t-1} - x^{t-2}) \right\rangle\right] \\ &+ \omega_1 \tau_1 \left\langle \lambda, U_{g^*}(\nu, \nu^0) \right\rangle \ .\end{aligned}$$

Combining these three lemmas gives a single convergence result for the entire gap function Q. This result looks nearly identical in form to Proposition 2 from [42] and is proven in Appendix A.1. From this proposition, we then prove our claimed compactness and convergence guarantees in Proposition 3.3 and Theorem 3.4.

Proposition 3.10. *Consider any problem of the form* (2.1) *with stepsizes satisfying* (3.7)-(3.14). *Then for any* $z = (x; \lambda, \nu) \in \mathcal{X} \times \Lambda_r \times V$,

$$\begin{aligned}&\sum_{t=1}^{T} \omega_t Q(z^t, z) + \sum_{t=1}^{T} \sum_{s=1}^{S_t} \frac{\omega_t}{S_t} \frac{1}{2L_h} \|\lambda_s^{(t)} - \lambda\|^2 + \frac{\omega_T \eta_T}{2} \|x^T - x\|^2 \\ &\leq \frac{\tilde{\omega}^{(1)} \beta^{(1)} + \omega_1 \eta_1}{2} \|x^0 - x\|^2 + \frac{\tilde{\omega}^{(1)} \gamma^{(1)}}{2} \|\lambda^0 - \lambda\|^2 + \omega_1 \tau_1 \left\langle \lambda, U_{g^*}(\nu, \nu^0) \right\rangle \ .\end{aligned}$$

Proof of Proposition 3.3. First, we claim the proposed stepsizes in (3.15) and (3.16) satisfy the necessary conditions (3.7)-(3.14) for the preceding proposition and lemmas to apply. Each of these conditions can be directly checked: See [42, Theorem 5] for equivalent verifications in the simplified setting of constrained optimization, only differing in that we consider a generic C rather than fixing $C = \frac{\|\lambda^\star\| + r}{\|x^0 - x^\star\|}$ in our choice of $\beta^{(t)} = C\tilde{M}_t$.

Note that from Proposition 3.10, the assumption that $L_h \in (0, \infty]$, and the fact that $\tau_1 = 0$,

$$\sum_{t=1}^{T} \omega_t Q(z^t, z) + \frac{\omega_T \eta_T}{2} \|x^T - x\|^2 \leq \frac{\tilde{\omega}^{(1)} \beta^{(1)} + \omega_1 \eta_1}{2} \|x^0 - x\|^2 + \frac{\tilde{\omega}^{(1)} \gamma^{(1)}}{2} \|\lambda^0 - \lambda\|^2 \ . \tag{3.20}$$

Furthermore, since this holds for all $z \in \mathcal{X} \times \Lambda_r \times V$, we consider the saddle point $z^\star$. As a saddle point, $Q(z^t, z^\star) \geq 0$. Using the stepsize conditions (3.15) and (3.16),

$$\omega_1 = 1, \quad \tilde{\omega}^{(1)} = \frac{1}{S_1}, \quad \eta_1 = 2L_{\varepsilon,r}^{\texttt{ADA}}, \quad \beta^{(1)} = \frac{CS_1}{\Delta}, \quad \gamma^{(1)} = \frac{\tilde{M}_1^2}{\beta^{(1)}} = \frac{S_1}{C\Delta} \ , \tag{3.21}$$

gives the claimed bound on x^T. Therefore, for $t \geq 1$, each x^t lies in the desired bounded neighborhood around $x^\star$. Note that $\underline{x}^t \in \text{conv}(x^0, \ldots, x^{t-1})$, so $\underline{x}^t$ is in the same neighborhood. As a result, $M_t = \|\nabla g(\underline{x}^t)\|$ is bounded uniformly by M.

For the dual iterates, Proposition 3.10 and the nonnegative of $Q(z^t; z^\star)$ and the norm ensures

$$\sum_{t=1}^{T} \sum_{s=1}^{S_t} \frac{\omega_t}{S_t} \frac{1}{2L_h} \|\lambda_s^{(t)} - \lambda^\star\|^2 \leq \frac{\tilde{\omega}^{(1)} \beta^{(1)} + \omega_1 \eta_1}{2} \|x^0 - x^\star\|^2 + \frac{\tilde{\omega}^{(1)} \gamma^{(1)}}{2} \|\lambda^0 - \lambda^\star\|^2 \ .$$

For any $s, t \geq 1$, one can bound

$$\|\lambda_s^{(t)} - \lambda^\star\|^2 \leq \frac{2L_h S_t}{\omega_t} \left[\frac{\tilde{\omega}^{(1)} \beta^{(1)} + \omega_1 \eta_1}{2} \|x^0 - x^\star\|^2 + \frac{\tilde{\omega}^{(1)} \gamma^{(1)}}{2} \|\lambda^0 - \lambda^\star\|^2\right] \ .$$

Utilizing the values in (3.21), bounding M_t above by M, S_t/ω_t above by $M\Delta+1$ for $t\geq 1$, and noting $\tilde{\lambda}^t$ lies in the convex hull of $\{\lambda_s^{(t)}\}$ yields the claimed bound on λ^t.

Proof of Theorem 3.4. Let $\bar{z}^T=(\bar{x}^T;\bar{\lambda}^T,\bar{\nu}^T)$ where

$$\bar{x}^T=\sum_{t=1}^T\omega_t x_t/\sum_{t=1}^T\omega_t \quad \bar{\lambda}^T=\sum_{t=1}^T\omega_t\tilde{\lambda}^t/\sum_{t=1}^T\omega_t, \quad \bar{\nu}_j^T=\begin{cases}\nabla g_j(x^0) & \text{if } \tilde{\lambda}_j^t=0 \text{ for all } t,\\ \sum_{t=1}^T\omega_t\tilde{\lambda}_j^t\nu_j^t/\sum_{t=1}^T\omega_t\tilde{\lambda}_j^t & \text{otherwise.}\end{cases} \tag{3.22}$$

Consequently,

$$\begin{aligned}\sum_{t=1}^T\omega_t\mathcal{L}(x;\tilde{\lambda}^t,\nu^t)&=\sum_{t=1}^T\omega_t\left[u(x)+\left\langle\tilde{\lambda}^t,\nu^t x-g^*(\nu^t)\right\rangle-h^*(\tilde{\lambda}^t)\right]\\&=\sum_{t=1}^T\omega_t u(x)+\sum_{t=1}^T\omega_t\sum_{j=1}^m\tilde{\lambda}_j^t(\left\langle\nu_j^t,x\right\rangle-g_j^*(\nu_j^t))-\sum_{t=1}^T\omega_t h^*(\tilde{\lambda}^t)\\&\leq\left(\sum_{t=1}^T\omega_t\right)u(x)+\left(\sum_{t=1}^T\omega_t\right)\left[\sum_{j=1}^m\bar{\lambda}_j^T\left(\left\langle\bar{\nu}_j^T,x\right\rangle-g_j^*(\bar{\nu}_j^T)\right)-h^*(\bar{\lambda}^T)\right]\\&=\left(\sum_{t=1}^T\omega_t\right)\mathcal{L}(x;\bar{\lambda}^T,\bar{\nu}^T)\ ,\end{aligned}$$

where the inequality follows from Jensen's inequality. Similarly, Jensen's inequality ensures that

$$\left(\sum_{t=1}^T\omega_t\right)\mathcal{L}(\bar{x}^T;\lambda,\nu)\leq\sum_{t=1}^T\omega_t\mathcal{L}(x^t;\lambda,\nu)\ .$$

Therefore, we have for all $z\in\mathcal{Z}$

$$\begin{aligned}\left(\sum_{t=1}^T\omega_t\right)Q(\bar{z}^T,z)&=\left(\sum_{t=1}^T\omega_t\right)\left[\mathcal{L}(\bar{x}^T;\lambda,\nu)-\mathcal{L}(x;\bar{\lambda}^T,\bar{\nu}^T)\right]\\&\leq\sum_{t=1}^T\omega_t\left[\mathcal{L}(x^t;\lambda,\nu)-\mathcal{L}(x;\tilde{\lambda}^t,\nu^t)\right]=\sum_{t=1}^T\omega_t Q(z^t,z)\ .\end{aligned} \tag{3.23}$$

Utilizing the above inequality, the bound demonstrated in (3.20), the nonnegativity of the norm, our distance bounds, as well as the triangle inequality,

$$\left(\sum_{t=1}^T\omega_t\right)Q(\bar{z}^T,(x^\star;\lambda,\nu))\leq\frac{\tilde{\omega}^{(1)}\beta^{(1)}+\omega_1\eta_1}{2}D_x^2+\tilde{\omega}^{(1)}\gamma^{(1)}(D_\lambda^2+r^2),\ \text{for all } \lambda\in\Lambda_r,\ \nu\in V\ .$$

Finally, we bound $\sum_{t=1}^T\omega_t$ by $T^2/2$ and substitute the values from (3.21) into the above expression. Considering Lemma 3.2, it suffices to bound the above by ε. Since each outer loop of UFCM computes only one gradient of g_j, $N_{\varepsilon,r}=T$ where $T>0$ solves

$$\frac{(C/\Delta+2L_{\varepsilon,r}^{\texttt{ADA}})D_x^2+2/(C\Delta)(D_\lambda^2+r^2)}{T^2}=\varepsilon\ ,$$

resulting in the complexity bound for $N_{\varepsilon,r}$. Noting that each inner loop performs only one proximal step on u and h^*,

$$P_{\varepsilon,r}\leq\sum_{t=1}^{\lceil N_{\varepsilon,r}\rceil}S_t=\sum_{t=1}^{\lceil N_{\varepsilon,r}\rceil}\lceil M_t\Delta t\rceil\leq\sum_{t=1}^{\lceil N_{\varepsilon,r}\rceil}(1+M\Delta t)\leq(\lceil N_{\varepsilon,r}\rceil+\lceil N_{\varepsilon,r}\rceil^2M\Delta)\ . \tag{3.24}$$

4 Compositions with Heterogeneously Hölder Smooth Components

Our convergence theory for problems with smooth components developed so far extends to instances where each g_j has Hölder continuous gradient with individual exponents. Recall, we say a function f is (L, p)-Hölder smooth with $L \geq 0$ and $p \in [0, 1]$ if the function satisfies

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|^p, \ \forall x, y \in \mathrm{dom}(f) \ .$$

When $p = 1$, we recover standard L-smoothness, and when $p = 0$, the function f is Lipschitz. Therefore, Hölder smoothness lets one interpolate between smooth and nonsmooth functions.

4.1 A Universal Definition of the Approximate Dualized Aggregate Smoothness

The key result facilitating the design of methods universally applicable to Hölder smooth problems is proven by [30, Lemma 1], previously introduced here as Lemma 2.3. We further utilize the subsequent cocoercive extension, introduced here as Lemma 2.4, which was proven by [25, Lemma 1]. These results show, for any fixed tolerance $\delta > 0$, there exists a constant $L_\delta = \left[\frac{1-p}{1+p}\frac{1}{\delta}\right]^{\frac{1-p}{1+p}} L^{\frac{2}{1+p}}$ such that the standard quadratic upper bound inequality or cocoercivity inequality of smooth convex functions hold (up to δ) for any (L, p)-Hölder smooth function. Supposing each g_j is (L_j, p_j)-Hölder smooth, define a general smoothness constant for fixed tolerance $\delta > 0$ as

$$L_{\delta,r} := \sum_{j=1}^{m} \left[\left[\frac{1-p_j}{1+p_j} \cdot \frac{m}{\delta} \right]^{\frac{1-p_j}{1+p_j}} \left[(\lambda_j^\star + r) \cdot L_j \right]^{\frac{2}{1+p_j}} \right] \ . \tag{4.1}$$

Noting each $\lambda_j g_j$ is $(\lambda_j L_j, p_j)$-Hölder smooth, this constant is large enough to ensure that Lemmas 2.3 and 2.4 apply to each component with tolerance δ/m. Summing these m components, the ideal dualized problem (3.5) arising if one knew the optimal dual multipliers is approximated within tolerance δ. We utilize this general constant to motivate our unifying theory.

Our universal definition for the Approximate Dualized Aggregate smoothness constant $L_{\varepsilon,r}^{\texttt{ADA}}$, generalizing the smooth case previously defined in (3.6), then follows from careful selection of this δ tolerance. To achieve optimal convergence guarantees, we require the following implicit choice for the definition of δ. Given an initialization $D_x \geq \|x^0 - x^\star\|$ and choices of $\epsilon, r > 0$, we define the Approximate Dualized Aggregate smoothness constant as the unique positive root to the following equation

$$L_{\varepsilon,r}^{\texttt{ADA}} := \left\{ L^{\texttt{ADA}} > 0 : L^{\texttt{ADA}} = \sum_{j=1}^{m} \left[\frac{1-p_j}{1+p_j} \cdot \frac{m\sqrt{L^{\texttt{ADA}}}}{\varepsilon} \cdot \frac{2\sqrt{6}D_x}{\sqrt{\varepsilon}} \right]^{\frac{1-p_j}{1+p_j}} \left[(\lambda_j^\star + r) L_j \right]^{\frac{2}{1+p_j}} \right\} \ . \tag{4.2}$$

We note that this value is precisely $L_{\delta,r}$ (4.1) with specialized $\delta = \varepsilon/\sqrt{24 L_{\varepsilon,r}^{\texttt{ADA}} D_x^2/\varepsilon}$. As each p_j tends to one, the associated coefficient tends to one, becoming independent on ε. As all p_j tend to one, the above sum defining $L_{\varepsilon,r}^{\texttt{ADA}}$ tends to $\sum_{j=1}^m (\lambda_j^\star + r) L_j$, recovering our previous definition (3.6) as a special case.

Lemma 4.1. *The Approximate Dualized Aggregate smoothness constant $L_{\varepsilon,r}^{\texttt{ADA}}$ as defined in* (4.2) *is nonincreasing with respect to ε.*

Proof. Consider $\varepsilon' \geq \varepsilon > 0$. Rearranging the definitions of $L_{\varepsilon,r}^{\texttt{ADA}}$ and $L_{\varepsilon',r}^{\texttt{ADA}}$ ensure that

$$\sum_{j=1}^{m} C_j (L_{\varepsilon,r}^{\texttt{ADA}})^{-\frac{1+3p_j}{2(1+p_j)}} \varepsilon^{-\frac{3-3p_j}{2+2p_j}} = 1, \quad \sum_{j=1}^{m} C_j (L_{\varepsilon',r}^{\texttt{ADA}})^{-\frac{1+3p_j}{2(1+p_j)}} (\varepsilon')^{-\frac{3-3p_j}{2+2p_j}} = 1 \ ,$$

with $C_j = \left[\frac{1-p_j}{1+p_j} \cdot 2\sqrt{6} m D_x\right]^{\frac{1-p_j}{1+p_j}} \left[(\lambda_j^\star + r) L_j\right]^{\frac{2}{1+p_j}}$. Since each $p_j \in [0,1]$, it follows that $(\varepsilon')^{-\frac{3-3p_j}{2+2p_j}} \leq (\varepsilon)^{-\frac{3-3p_j}{2+2p_j}}$, and for the above sums to equal 1, it must hold that the positive solution $L_{\varepsilon',r}^{\texttt{ADA}} \leq L_{\varepsilon,r}^{\texttt{ADA}}$. □

4.2 Guarantees for Composite Optimization with Heterogeneous Components

Importantly, note that we are not making any modifications to UFCM in this section. The algorithm does not require knowledge of the implicitly defined δ value or any (L_j, p_j) pairs. They are for analysis only. Instead, the UFCM algorithm only relies on an estimate of $L_{\varepsilon,r}^{\texttt{ADA}}$, which could be guessed via a geometric parameter schedule without attempting to approximate δ or any of the (L_j, p_j) pairs. (See Remark 4.6.)

Next, we present our convergence theory, further justifying the choice of the implicit constant definition (4.2). Our theory provides guarantees for any choice of δ and approximate smoothness constant $L_{\delta,r}$. However, this choice $L_{\varepsilon,r}^{\texttt{ADA}}$ optimizes the strength of our guarantee over all δ. We first generalize Proposition 3.3, ensuring the iterates of UFCM stay bounded in this heterogeneously smooth setting, accounting for a slightly larger radius due to the additive error term. We defer the proof to Appendix A.2.

Proposition 4.2. *Consider any problem of the form* (2.1) *and constants* $\Delta, C, \epsilon, r > 0$*, and suppose Algorithm 1 is run with outer loop stepsizes* (3.15) *and inner loop stepsizes* (3.16)*, then for all* $t \leq \sqrt{\frac{24 L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{\varepsilon}}$*,*

$$\|x^t - x^\star\|^2 \leq \frac{1}{2L_{\varepsilon,r}^{\texttt{ADA}}} \left[(C/\Delta + 50 L_{\varepsilon,r}^{\texttt{ADA}}) D_x^2 + \frac{1}{C\Delta} D_\lambda^2 \right] \ . \tag{4.3}$$

Furthermore, if h *is* L_h*-smooth, then for averaged iterates* $\tilde{\lambda}^t$ *computed each loop,*

$$\|\tilde{\lambda}^t - \lambda^\star\|^2 \leq L_h (M\Delta + 1) \left[(C/\Delta + 50 L_{\varepsilon,r}^{\texttt{ADA}}) D_x^2 + \frac{1}{C\Delta} D_\lambda^2 \right] \ . \tag{4.4}$$

where M *is an upper bound for* $\|\nabla g(x)\|$ *in the neighborhood outlined above* (4.3)*.*

Theorem 4.3. *Consider any problem of the form* (2.1) *with each* g_j *being* (L_j, p_j)*-Hölder smooth, and constants* $\Delta, C, \epsilon, r > 0$*. Then Algorithm 1 with stepsizes* (3.15) *and* (3.16) *must find an* (ε, r)*-optimal solution* (1.8) *with complexity bounds*

$$N_{\varepsilon,r} = \sqrt{\frac{(2C/\Delta + 4L_{\varepsilon,r}^{\texttt{ADA}}) D_x^2 + 4/(C\Delta)(D_\lambda^2 + r^2)}{\varepsilon}}, \quad P_{\varepsilon,r} = \lceil N_{\varepsilon,r} \rceil + \lceil N_{\varepsilon,r} \rceil^2 \Delta M, \tag{4.5}$$

where M *is an upper bound for* $\|\nabla g(x)\|$ *in the neighborhood outlined in* (4.3)*.*

For target accuracy $0 < \varepsilon \leq \min\{1, 24 L_{\varepsilon,r}^{\texttt{ADA}} D_x^2\}$ *with* $r = D_\lambda \sqrt{\varepsilon}$*, and setting* $C = D_\lambda / D_x$ *and* $\Delta = C/(2L_{\varepsilon,r}^{\texttt{ADA}})$*, these bounds simplify to*

$$N_{\varepsilon,r} = \sqrt{\frac{24 L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{\varepsilon}}, \quad P_{\varepsilon,r} = \sqrt{\frac{24 L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{\varepsilon}} + \frac{48 M D_x D_\lambda}{\varepsilon} + 1 \ , \tag{4.6}$$

where M *is an upper bound on* $\|\nabla g(x)\|$ *for all* $x \in B(x^\star, \sqrt{27 D_x^2})$*.*

The preceding theorem demonstrates how the complicated nature of heterogeneous optimization can be simplified to look analogous to the standard accelerated rate of unconstrained smooth optimization. Of course, in general, this rate is not $\mathrm{O}(1/\sqrt{\varepsilon})$ as $L_{\varepsilon,r}^{\mathtt{ADA}}$ may be non-constant in ε. The aggregating parameter $L_{\varepsilon,r}^{\mathtt{ADA}}$ provides the key mechanism to provide a single unifying, universal guarantee.

Further, $L_{\varepsilon,r}^{\mathtt{ADA}}$ serves as a universal tool to recover optimal rates in terms of gradient oracle complexity for known problem classes. First note that Theorem 4.3 recovers the optimal rates from smooth compositions from Section 3 within a factor of two as our more general definition of $L_{\varepsilon,r}^{\mathtt{ADA}}$ reduces to the previous definition (3.6) when $p_j = 1$. The following corollaries demonstrate $L_{\varepsilon,r}^{\mathtt{ADA}}$'s ability to recover optimal results in terms of gradient oracle complexity from the literature for minimizing a single Hölder smooth function [28], rates for a heterogeneous sum of Hölder smooth terms [15], and rates for smooth constrained optimizations [42].

Corollary 4.4. *Consider minimizing $g_0(x) + u(x)$ where g_0 is (L,p)-Hölder smooth with initial distance bound $D_x \geq \|x^0 - x^\star\|$, initialization $\lambda^0 = 1$, and target accuracy $0 < \varepsilon \leq \min\{1, 2\sqrt{6}LD_x^{1+p}\}$. Then Algorithm 1 finds an ε-optimal solution with complexity bounds*

$$N_{\varepsilon,r} = P_{\varepsilon,r} = \mathrm{O}\left(\left(\frac{L}{\varepsilon}\right)^{\frac{2}{1+3p}} D_x^{\frac{2+2p}{1+3p}}\right) .$$

Proof. For minimizing a single function, $h(z) = z$, and $\lambda^0 = \lambda^\star = 1$. Therefore, D_λ and r can be arbitrarily small, as well as $\Delta = D_\lambda/D_x$. Recall $S_t = \lceil M_t \Delta t \rceil$, so we set $S_t = 1$ for all t, and each inner loop of UFCM only computes a single proximal step on u and h^*. Therefore, $N_{\varepsilon,r} = P_{\varepsilon,r}$.

We now focus on the gradient oracle complexity. It is straightforward to check that our hypothesis enforces $\varepsilon \leq 2\sqrt{6}LD_x^{1+p} \leq L\left(\frac{1-p}{1+p}\right)^{\frac{1-p}{2}} (2\sqrt{6}D_x)^{1+p} \leq 24 L_{\varepsilon,r}^{\mathtt{ADA}} D_x^2$, which allows us to apply Theorem 4.3. Considering our Approximate Dualized Aggregate constant yields

$$L_{\varepsilon,r}^{\mathtt{ADA}} = (1+r)^{\frac{4}{1+3p}} \left[\frac{1-p}{1+p} \cdot \frac{2\sqrt{6}D_x}{\varepsilon\sqrt{\varepsilon}}\right]^{\frac{2-2p}{1+3p}} L^{\frac{4}{1+3p}} .$$

We then conclude

$$N_{\varepsilon,r} = \mathrm{O}\left(\sqrt{\frac{L_{\varepsilon,r}^{\mathtt{ADA}} D_x^2}{\varepsilon}}\right) = \mathrm{O}\left(\sqrt{\left(\frac{D_x}{\varepsilon\sqrt{\varepsilon}}\right)^{\frac{2-2p}{1+3p}} \frac{L^{\frac{4}{1+3p}} D_x^2}{\varepsilon}}\right) = \mathrm{O}\left(\left(\frac{L}{\varepsilon}\right)^{\frac{2}{1+3p}} D_x^{\frac{2+2p}{1+3p}}\right)$$

where the first equality considers (4.6), the second substitutes $L_{\varepsilon,r}^{\mathtt{ADA}}$, and the third simplifies to recover (1.5). □

Corollary 4.5. *In the setting of Theorem 4.3, Algorithm 1 finds an ε-optimal solution with complexity bounds $N_{\varepsilon,r}$ and $P_{\varepsilon,r} = \lceil N_{\varepsilon,r} \rceil + \lceil N_{\varepsilon,r} \rceil^2 M\Delta$ for $N_{\varepsilon,r} := T > 0$ which solves the following*

$$\sum_{j=1}^{m} c_j \frac{((\lambda_j^\star + r)L_j)^{\frac{2}{1+p_j}} D_x^2}{\varepsilon^{\frac{2}{1+p_j}}} T^{-\frac{1+3p_j}{1+p_j}} = 1 , \tag{4.7}$$

with $c_j = 24\left[\frac{1-p_j}{1+p_j} m\right]^{\frac{1-p_j}{1+p_j}}$. Consequently, the convergence rate is at most the sum of the rates of individual terms (1.5), weighted by the appropriate multiplier,

$$N_{\varepsilon,r} = \mathrm{O}\left(\sum_{j=1}^{m} K_{SM}(\varepsilon, (\lambda_j^\star + r)L_j, p_j, D_x)\right) . \tag{4.8}$$

Proof. Considering the gradient oracle complexity bound $N_{\varepsilon,r} = \sqrt{24L_{\varepsilon,r}^{\texttt{ADA}} D_x^2/\varepsilon}$, we substitute $L_{\varepsilon,r}^{\texttt{ADA}} = N_{\varepsilon,r}^2 \cdot \varepsilon/(24D_x^2)$ into the definition (4.2) giving

$$\frac{T^2\varepsilon}{24D_x^2} = \sum_{j=1}^{m} \left[\frac{1-p_j}{1+p_j} \cdot \frac{2\sqrt{6}m\sqrt{\frac{T^2\varepsilon}{24D_x^2}}D_x^2}{\varepsilon^{3/2}} \right]^{\frac{1-p_j}{1+p_j}} \left((\lambda_j^\star + r)L_j\right)^{\frac{2}{1+p_j}} .$$

Rearranging the expression above yields (4.7), which is nonincreasing in T. Therefore, to prove (4.8), it suffices to bound each summand of (4.7) by $1/m$. We then consider solving

$$c_j \frac{\left((\lambda_j^\star + r)L_j\right)^{\frac{2}{1+p_j}} D_x^2}{\varepsilon^{\frac{2}{1+p_j}}} T_j^{-\frac{1+3p_j}{1+p_j}} = \frac{1}{m}$$

for T_j. The result recovers (1.5) component-wise

$$T_j = \mathrm{O}\left(\left(\frac{(\lambda_j^\star + r)L_j}{\varepsilon} \right)^{\frac{2}{1+3p_j}} D_x^{\frac{2+2p_j}{1+3p_j}} \right) = \mathrm{O}\left(K_{SM}(\varepsilon, (\lambda_j^\star + r)L_j, p_j, D_x) \right).$$

Bounding $T \leq \max_j T_j \leq \sum_{j=1}^m T_j$, yields (4.8). □

Fixing $h(z) = \sum_{j=1}^m z_j$, each $\lambda_j^\star = 1$, and this second corollary recovers the results for heterogeneous sums of Hölder smooth terms of [15, Theorem 1.1 and 1.3] when we initialize $\lambda_j^0 = 1$. For constrained optimization, with h as the nonpositive indicator function and when each g_j is (L_j, p)-Hölder smooth with common exponent, this second corollary recovers the $\mathrm{O}(1/\varepsilon^{2/(1+3p)})$ results of [8, Corollary 2.3] as a special case.

Remark 4.6. *We note that one may tradeoff knowledge of $L_{\varepsilon,r}^{\texttt{ADA}}$ for knowledge of distance bounds D_x and D_λ. Considering the sequence $L_k = 2^0, 2^1, 2^2, ...,$ as well as the setting of Theorem 4.3 one may run UFCM for $N_k = \sqrt{\frac{24L_kD_x^2}{\varepsilon}}$ outer loop iterations. Our theory then guarantees that once $k = \max\{\lceil \log_2(L_{\varepsilon,r}^{\texttt{ADA}}) \rceil, 0\}$, an ε-optimal solution has been constructed, taking at most $\mathrm{O}\left(\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{\varepsilon}}\right)$ gradient oracle calls total. Note that neither UFCM nor this modified scheme possesses stopping criteria certifying that an ϵ-optimal solution has been found without additional problem knowledge. So these guarantees are theoretical.*

4.3 Analysis of UFCM for Compositions with Heterogeneous Components

The same process of analysis presented in Section 3 extends to provide guarantees for UFCM given any heterogeneously Hölder smooth components by carefully accounting for the additive errors incurred by using Nesterov-style inequalities. Lemma 3.7, Lemma 3.9, and Proposition 3.10 generalize to this setting as follows. For many of these results, the proof is redundant with prior work except for tracking an additional constant term through the developed inequalities. Below, we present these key results with proofs of Lemmas 4.7 and 4.8 deferred to Appendix A.2 for the sake of completeness.

Lemma 4.7. *If each g_j is (L_j, p_j)-Hölder smooth, then for any fixed $\delta > 0$,*

$$\langle \lambda, U_{g^*}(\nu; \hat{\nu}) \rangle \geq \frac{1}{2L_{\delta,r}} \left\| \sum_{j=1}^m \lambda_j (\nu_j - \hat{\nu_j}) \right\|^2 - \frac{\delta}{2}, \ \forall \lambda \in \Lambda_r, \ \forall \nu, \hat{\nu} \in \{\nabla g(x) : x \in \mathcal{X}\} .$$

The following results utilize the general smoothness constant $L_{\delta,r}$ in both the analysis and the appropriate parameters for completeness. However, we recall that the Approximate Aggregate Smoothness constant $L^{\texttt{ADA}}_{\varepsilon,r}$ used in Theorem 4.3 is specialized with $\delta = \varepsilon/\sqrt{24L^{\texttt{ADA}}_{\varepsilon,r}D_x^2/\varepsilon}$.

Lemma 4.8. *Suppose the stepsizes satisfy* (3.7)-(3.14)*, and let* $z^t := (x^t; \tilde{\lambda}^t, \nu^t)$ *denote the iterates of Algorithm 1. Then* z^t *satisfy the following for any* $z = (x; \lambda, \nu) \in \mathcal{X} \times \Lambda_r \times V$ *and any fixed* $\delta > 0$

$$\sum_{t=1}^{T} \omega_t \left[Q_\nu(z^t, z)\right] \leq \frac{\omega_T L_{\delta,r}}{2(\tau_T+1)}\|x^T - x^{T-1}\|^2 + \sum_{t=1}^{T-1} \frac{\omega_t \theta_{t+1} L_{\delta,r}}{2\tau_{t+1}}\|x^t - x^{t-1}\|^2$$
$$+ \omega_1\tau_1 \left\langle \lambda, U_{g^*}(\nu;\nu^0)\right\rangle + \frac{\delta}{2}\left[\omega_T(\tau_T+1) + \sum_{t=2}^{T}\omega_t\tau_t\right], \quad \forall (x;\lambda,\nu) \in \mathcal{X}\times\Lambda_r\times V \ .$$

The following proposition is the direct analog of Proposition 3.10 in the heterogeneously smooth setting. The proof is analogous, applying the above two lemmas instead of Lemmas 3.7 and 3.9, noting the small additive dependence on tolerance δ.

Proposition 4.9. *Consider any problem of the form* (2.1) *with stepsizes satisfying* (3.7)-(3.14)*. Then for any* $z = (x;\lambda,\nu) \in \mathcal{X}\times\Lambda_r\times V$ *and any fixed* $\delta > 0$*,*

$$\sum_{t=1}^{T}\omega_t Q(z^t, z) + \sum_{t=1}^{T}\sum_{s=1}^{S_t}\frac{\omega_t}{S_t}\frac{1}{2L_h}\|\lambda_s^{(t)} - \lambda\|^2 + \frac{\omega_T\eta_T}{2}\|x^T - x\|^2$$
$$\leq \frac{\tilde{\omega}^{(1)}\beta^{(1)} + \omega_1\eta_1}{2}\|x^0 - x\|^2 + \frac{\tilde{\omega}^{(1)}\gamma^{(1)}}{2}\|\lambda^0 - \lambda\|^2 + \omega_1\tau_1\left\langle \lambda, U_{g^*}(\nu;\nu^0)\right\rangle$$
$$+ \frac{\delta}{2}\left[\omega_T(\tau_T+1) + \sum_{t=2}^{T}\omega_t\tau_t\right], \quad \forall (x;\lambda,\nu) \in \mathcal{X}\times\Lambda_r\times V \ .$$

Proof of Theorem 4.3. We first note that as $\omega_t = t$, $\tau_t = \frac{t-1}{2}$, we can rewrite

$$\left[\omega_T(\tau_T+1) + \sum_{t=2}^{T}\omega_t\tau_t\right] = \frac{T^3+3T^2+2T}{6}, \quad \sum_{t=1}^{T}\omega_t = \frac{T(T+1)}{2} \ .$$

Then for all $\lambda \in \Lambda_r$ and $\nu \in V$, considering Proposition 4.9, Jensen's inequality (3.23), and the particular stepsizes (3.21), we can bound

$$Q(\bar{z}^T, (x^\star;\lambda,\nu)) \leq \frac{(C/\Delta + 2L_{\delta,r})\|x^0 - x^\star\|^2 + 2/(C\Delta)(\|\lambda^0-\lambda^\star\|^2 + r^2)}{T(T+1)} + \frac{\delta(T+2)}{6} \ . \tag{4.9}$$

Recall that $L_{\delta,r}$ depends on the value δ, so we optimize the above bound with respect δ to achieve a universally optimal rate. We fix $T, \varepsilon > 0$.

Setting $\delta = \frac{\varepsilon}{T}$, and bounding $T + 2 \leq 3T$, we obtain the following inequality derived from (4.9)

$$Q(\bar{z}^T, (x^\star;\lambda,\nu)) \leq \frac{(C/\Delta + 2L_{\varepsilon/T,r})\|x^0 - x^\star\|^2 + 2/(C\Delta)(\|\lambda^0-\lambda^\star\|^2 + r^2)}{T^2} + \frac{\varepsilon}{2} \ . \tag{4.10}$$

Our choices of C, Δ, and $r = D_\lambda\sqrt{\varepsilon}$ simplify the expression as for any $T \geq N_{\varepsilon,r} = \sqrt{24L^{\texttt{ADA}}_{\varepsilon,r}D_x^2/\varepsilon}$ one has $Q(\bar{z}^T, (x^\star;\lambda,\nu)) \leq \varepsilon$ over all $\lambda \in \Lambda_r$ and $\nu \in V$. Applying Lemma 3.2, this ensures (ϵ, r)-optimality. Furthermore, when $T = N_{\varepsilon,r}$, then $L_{\varepsilon/N_{\varepsilon,r},r}$ precisely recovers the definition of $L^{\texttt{ADA}}_{\varepsilon,r}$ in (4.2).

The claimed proximal step complexity follows from the general formula (3.24). Since $\varepsilon \leq 24L_{\varepsilon,r}^{\texttt{ADA}} D_x^2$, it holds that $N_{\varepsilon,r} \geq 1$. Therefore, we bound $\lceil N_{\varepsilon,r} \rceil \leq N_{\varepsilon,r} + 1 \leq 2N_{\varepsilon,r}$, which results in

$$P_{\varepsilon,r} = \lceil N_{\varepsilon,r} \rceil + \lceil N_{\varepsilon,r} \rceil^2 \Delta M \leq \sqrt{\frac{24L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{\varepsilon}} + \frac{48MD_xD_\lambda}{\varepsilon} + 1 \ ,$$

where we recall $\Delta = D_\lambda/(2D_x L_{\varepsilon,r}^{\texttt{ADA}})$.

5 Growth Bounds and Restarting

We utilize a simple restarting scheme given initial distance bounds. Primal-dual algorithms have exhibited great success from restarting when the respective gap function possesses certain growth conditions [1, 12, 27]. This algorithm, denoted R-UFCM can then achieve linear convergence in terms of gradient oracle calls when the components are smooth and strongly convex, and the proximal step complexity can achieve linear convergence rates when h is sufficiently smooth. Recall from (1.4) that a function f is (μ, q)-uniformly convex if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{q+1} \|y - x\|^{q+1}, \quad \forall x, y \in \mathrm{dom}(f) \ .$$

When $q = 1$, we recover the notion of μ-strong convexity. As $q \to \infty$, these functions are simply convex. Similarly to Hölder smoothness, we can interpolate between the level of convexity. If g_j is also (L_j, p_j)-Hölder smooth, then the following symmetric, two-sided bound holds

$$g_j(x) + \langle \nabla g_j(x), y - x \rangle + \frac{\mu_j}{q_j+1} \|y - x\|^{q_j+1} \leq g_j(y) \leq g_j(x) + \langle \nabla g_j(x), y - x \rangle + \frac{L_j}{p_j+1} \|y - x\|^{p_j+1} \ .$$

5.1 Growth Structure

The uniform convexity of each g_j can be combined together to ensure a growth condition on the gap function. This perspective plays a central role in our analysis, as it does in most restarted analyses.

Definition 5.1. *Given monotone nondecreasing, convex functions $G_x, G_\lambda \colon \mathbb{R}_+ \to \mathbb{R}_+$, we say that the gap function possesses (G_x, G_λ)-growth if for any $z = (x; \lambda, \nu)$ and $\hat{z} = (x^\star; \lambda^\star, \nabla g(x))$*

$$G_x(\|x - x^\star\|) + G_\lambda(\|\lambda - \lambda^\star\|) \leq Q(z, \hat{z}) \ .$$

As additional structure, the growth functions considered herein will always have $G_x(0) = 0$, $G_\lambda(0) = 0$, and both G_x and G_λ differentiable. The following lemma gives the explicit growth condition when the component functions g_j exhibit varying uniform convexity and h is L_h-smooth.

Lemma 5.2. *Suppose component functions g_j are (μ_j, q_j)-uniformly convex and h is L_h-smooth. Then the gap function possesses G_x, G_λ growth where $G_x(t) = \sum_{j=1}^m \lambda_j^\star \frac{\mu_j}{q_j+1} |t|^{(q_j+1)}$ and $G_\lambda(t) = \frac{1}{2L_h} t^2$. Therefore, for any $z = (x, \lambda, \nu)$ and $\hat{z} = (x^\star, \lambda^\star, \nabla g(x))$*

$$Q(z, \hat{z}) \geq \sum_{j=1}^m \lambda_j^\star \frac{\mu_j}{q_j+1} \|x - x^\star\|^{q_j+1} + \frac{1}{2L_h} \|\lambda - \lambda^\star\|^2 \ .$$

Proof. From the optimality of $x^\star$, $\left\langle \nabla u(x^\star) + \sum_{j=1}^m \lambda_j^\star \nabla g_j(x^\star), x - x^\star \right\rangle \geq 0$. Note that since h is L_h-smooth, h^* is $1/L_h$-strongly convex. Thus,

$$\begin{aligned}
Q(z,\hat{z}) &= \mathcal{L}(x;\lambda^\star,\nabla g(x)) - \mathcal{L}(x^\star;\lambda^\star,\nu^\star) + \mathcal{L}(x^\star;\lambda^\star,\nu^\star) - \mathcal{L}(x^\star;\lambda,\nu) \\
&\geq u(x) + \langle \lambda^\star, g(x)\rangle - h^*(\lambda^\star) - [u(x^\star) + \langle \lambda^\star, g(x^\star)\rangle - h^*(\lambda^\star)] \\
&\quad + \langle \lambda^\star, \nu^\star x^\star - g^*(\nu^\star)\rangle - h^*(\lambda^\star) - [\langle \lambda, \nu x^\star - g^*(\nu)\rangle - h^*(\lambda)] \\
&\quad - \left\langle \nabla u(x^\star) + \sum_{j=1}^m \lambda_j^\star \nabla g_j(x^\star), x - x^\star \right\rangle \\
&\geq [u(x) - u(x^\star) - \langle \nabla u(x^\star), x - x^\star\rangle] + \sum_{j=1}^m \lambda_j^\star \left[g(x) - g(x^\star) - \langle \nabla g(x^\star), x - x^\star\rangle\right] \\
&\quad + h^*(\lambda) - h^*(\lambda^\star) - \langle \lambda - \lambda^\star, g(x^\star)\rangle \\
&\geq \sum_{j=1}^m \lambda_j^\star \frac{\mu_j}{q_j+1}\|x - x^\star\|^{q_j+1} + \frac{1}{2L_h}\|\lambda - \lambda^\star\|^2 \ ,
\end{aligned}$$

where the first equality expands the gap function, the following inequality applies Fenchel-Young and subtracts the nonnegative inner product outlined above, the next inequality regroups terms and applies Fenchel-Young once again, and the final inequality comes directly from the convexity of u, the uniform convexity of g_j, and the strong convexity of h^*. □

5.2 An Approximate Dualized Aggregate Convexity $\mu_\varepsilon^{\texttt{ADA}}$

We now have the necessary tools to define the lower bounding curvature for the composite problem into a single value $\mu_\varepsilon^{\texttt{ADA}}$, generalizing the growth bound strong convexity yields. For (μ_j, q_j)-uniformly convex components g_j and target accuracy $\varepsilon > 0$, we define the Approximate Dualized Aggregate convexity constant implicitly as the unique positive solution to the following equation

$$\mu_\varepsilon^{\texttt{ADA}} := \left\{ \mu^{\texttt{ADA}} > 0 : \frac{\mu^{\texttt{ADA}}}{2} = \sum_{j=1}^m \lambda_j^\star \frac{\mu_j}{q_j+1} (\varepsilon/\mu^{\texttt{ADA}})^{\frac{q_j-1}{2}} \right\} \ . \tag{5.1}$$

Note when $q_j = 1$, the coefficient becomes independent of ε. If all $q_j = 1$, the $\mu_\varepsilon^{\texttt{ADA}}$ simply totals the $\lambda_j^\star$-weighted strong convexity constants. More generally, $\mu_\varepsilon^{\texttt{ADA}}$ aggregates the lower curvature of each component, weighted by the appropriate dual multiplier. This quantity can further be viewed as an approximation of strong convexity as shown in Lemma 5.4 below.

Lemma 5.3. *The Approximate Dualized Aggregate convexity constant $\mu_\varepsilon^{\texttt{ADA}}$ as defined in* (5.1) *is nondecreasing with respect to ε.*

Proof. Consider $\varepsilon' \geq \varepsilon > 0$. Rearranging the definitions of $\mu_\varepsilon^{\texttt{ADA}}$ and $\mu_{\varepsilon'}^{\texttt{ADA}}$ ensure that

$$\sum_{j=1}^m \lambda_j^\star \frac{\mu_j}{q_j+1} (\mu_\varepsilon^{\texttt{ADA}})^{-\frac{q_j+1}{2}} \varepsilon^{\frac{q_j-1}{2}} = 1, \quad \sum_{j=1}^m \lambda_j^\star \frac{\mu_j}{q_j+1} (\mu_{\varepsilon'}^{\texttt{ADA}})^{-\frac{q_j+1}{2}} (\varepsilon')^{\frac{q_j-1}{2}} = 1 \ .$$

Since each $q_j \geq 1$, it follows that $(\varepsilon')^{(q_j-1)/2} \geq (\varepsilon)^{(q_j-1)/2}$, and for the above sums to equal one, it must hold that the positive solution $\mu_{\varepsilon'}^{\texttt{ADA}} \geq \mu_\varepsilon^{\texttt{ADA}}$. □

Lemma 5.4. *Suppose the components g_j are (μ_j, q_j)-uniformly convex and h is L_h-smooth. Then for any $z = (x, \lambda, \nu)$ and $\hat{z} = (x^\star, \lambda^\star, \nabla g(x))$,*

$$Q(z,\hat{z}) \geq \frac{\mu_\varepsilon^{\texttt{ADA}}}{2}\|x - x^\star\|^2 + \frac{1}{2L_h}\|\lambda - \lambda^\star\|^2 - \frac{\varepsilon}{2} \ .$$

Proof. Considering the result of Lemma 5.2, it suffices to bound

$$G_x(\|x - x^\star\|) \geq \frac{\mu_\varepsilon^{\texttt{ADA}}}{2}\|x - x^\star\|^2 - \frac{\varepsilon}{2}, \quad \forall x \in \mathcal{X} \ .$$

Note that $\mu_\varepsilon^{\texttt{ADA}} = \varepsilon/(G_x^{-1}(\varepsilon/2))^2$. Since $G_x(t)$ is differentiable and positive for all $t > 0$ and $\mu_\varepsilon^{\texttt{ADA}}$ is nondecreasing in ε, it follows that

$$\frac{\partial \mu_\varepsilon^{\texttt{ADA}}}{\partial \varepsilon} = \frac{(G_x^{-1}(\varepsilon/2))^2 - \frac{\varepsilon G_x^{-1}(\varepsilon/2)}{G_x'(G_x^{-1}(\varepsilon/2))}}{(G_x^{-1}(\varepsilon/2))^4} \geq 0 \Longrightarrow G_x'(G_x^{-1}(\varepsilon/2)) \geq \frac{\varepsilon}{G_x^{-1}(\varepsilon/2)} = \mu_\varepsilon^{\texttt{ADA}} G_x^{-1}(\varepsilon/2) \ .$$

Therefore, by the monotonicity and nonnegativity of G_x, for any $t \geq G_x^{-1}(\varepsilon/2)$, it holds that $G_x'(t) \geq \mu_\varepsilon^{\texttt{ADA}} t$.

We first consider the case where $G_x(\|x - x^\star\|) \geq \varepsilon/2$. We again note that by monotonicity and nonegativity of G_x, it holds that for any $x \in \mathcal{X}$,

$$\begin{aligned} G_x(\|x - x^*\|) = \int_0^{\|x-x^\star\|} G_x'(t)\mathrm{dt} &\geq \int_{G_x^{-1}(\varepsilon/2)}^{\|x-x^\star\|} \mu_\varepsilon^{\texttt{ADA}} t\mathrm{dt} + \int_0^{G_x^{-1}(\varepsilon/2)} G_x'(t)\mathrm{dt} \\ &= \int_0^{\|x-x^\star\|} \mu_\varepsilon^{\texttt{ADA}} t\mathrm{dt} - \int_0^{G_x^{-1}(\varepsilon/2)} \mu_\varepsilon^{\texttt{ADA}} t\mathrm{dt} + \int_0^{G_x^{-1}(\varepsilon/2)} G_x'(t)\mathrm{dt} \\ &= \frac{\mu_\varepsilon^{\texttt{ADA}}}{2}\|x - x^\star\|^2 \ . \end{aligned}$$

Now we consider the case where $G_x(\|x - x^\star\|) < \varepsilon/2$. Since $\mu_\varepsilon^{\texttt{ADA}} = \varepsilon/(G_x^{-1}(\varepsilon/2))^2$, we note that

$$\frac{\mu_\varepsilon^{\texttt{ADA}}}{2}\|x - x^\star\|^2 = \frac{\varepsilon}{2(G_x^{-1}(\varepsilon/2))^2}\|x - x^\star\|^2 < \frac{\varepsilon}{2} \ ,$$

which implies that

$$G_x(\|x - x^\star\|) \geq 0 > \frac{\mu_\varepsilon^{\texttt{ADA}}}{2}\|x - x^\star\|^2 - \frac{\varepsilon}{2} \ . \qquad \square$$

5.3 A Further Universalized Approximate Dualized Aggregate Smoothness

Recall that the previous definition in (4.2) for the Approximate Dualized Aggregate smoothness constant $L_{\varepsilon,r}^{\texttt{ADA}}$ depended on ε and distance bound D_x. In order to recover (1.5) through Theorem 4.3, this dependence was a necessity. When the components possess uniform convexity in addition to Hölder smoothness, one can further leverage the Approximate Dualized Aggregate Convexity $\mu_\varepsilon^{\texttt{ADA}}$. In its full generality, we define $L_{\varepsilon,r}^{\texttt{ADA}}$ to be the unique positive solution to the following equation

$$L_{\varepsilon,r}^{\texttt{ADA}} := \left\{ L^{\texttt{ADA}} > 0 : L^{\texttt{ADA}} = \sum_{j=1}^m \left[\frac{1-p_j}{1+p_j} \cdot \frac{m\sqrt{L^{\texttt{ADA}}}}{\varepsilon} \cdot \min\left\{ \frac{2\sqrt{6}D_x}{\sqrt{\varepsilon}}, \frac{4\sqrt{6}}{\sqrt{\mu_\varepsilon^{\texttt{ADA}}}} \right\} \right]^{\frac{1-p_j}{1+p_j}} \left[(\lambda_j^\star + r)L_j\right]^{\frac{2}{1+p_j}} \right\} . \tag{5.2}$$

We note that for small enough $\mu_\varepsilon^{\texttt{ADA}}$, the above value recovers (4.2) exactly. Further note that even with this generalization, $L_{\varepsilon,r}^{\texttt{ADA}}$ remains nonincreasing with respect to ε.

Algorithm 2 Restarted Universal Fast Composite Method (R-UFCM)

Input $z^0 \in \mathcal{X} \times \Lambda$, distance bounds D_x and D_λ, target accuracy $\varepsilon > 0$, constants $L_{\varepsilon,r}^{\texttt{ADA}}$ and $\mu_\varepsilon^{\texttt{ADA}}$, and UFCM execution count K

Set $D_x^{(0)}$, $D_\lambda^{(0)}$ and $\{T_k\}$ according to (5.3)
for $k = 0, 1, \ldots, K-1$ **do**
  Run UFCM(z^k, $\lceil T_k \rceil$, $L_{\varepsilon,r}^{\texttt{ADA}}$) returning output $(\bar{x}^{T_k,k}, \bar{\lambda}^{T_k,k})$
  Set $(x^{k+1}, D_x^{(k+1)}) = \begin{cases} (\bar{x}^{T_k,k}, \sqrt{2^{K-k}\varepsilon/\mu_\varepsilon^{\texttt{ADA}}}) & \text{if } \mu_\varepsilon^{\texttt{ADA}} \geq 4\varepsilon/D_x^2 \\ (x^0, D_x) & \text{otherwise} \end{cases}$
  Set $(\lambda^{k+1}, D_\lambda^{(k+1)}) = \begin{cases} (\bar{\lambda}^{T_k,k}, \sqrt{2^{K-k}\varepsilon L_h}) & \text{if } \sqrt{2^{K-k}\varepsilon L_h} \leq D_\lambda \\ (\lambda^0, D_\lambda) & \text{otherwise} \end{cases}$
  Set $z^{k+1} = (x^{k+1}, \lambda^{k+1})$
end for

5.4 Guarantees for Fully Heterogeneous Compositions

Finally, we present our universal theory when each component g_j possesses its own (L_j, p_j)-Hölder smoothness and (μ_j, q_j)-uniform convexity. Algorithmic restarting, as discussed in Section 2.1, is the key to enabling this final improvement in our theory.

Our proposed restarted variant, denoted R-UFCM, sequentially runs K executions of UFCM, each for T_k iterations, restarted at a sequence of initializations $z^k = (x^k, \lambda^k)$ with distance bounds $D_x^{(k)}$ and $D_\lambda^{(k)}$. Using the produced outputs $\bar{x}^{T_k,k}$ and $\bar{\lambda}^{T_k,k}$, the next initialization $z^{k+1} = (x^{k+1}, \lambda^{k+1})$ is determined. The next primal initialization is $\bar{x}^{T_k,k}$ if $\mu_\varepsilon^{\texttt{ADA}} \geq 4\varepsilon/D_x^2$, else x^0 is reused. Similarly, the next dual initialization is $\bar{\lambda}^{T_k,k}$ if $2^{K-k}L_h \leq D_\lambda^2/\varepsilon$, else λ^0 is reused. Algorithm 2 formalizes this process with the following initializations

$$(T_k, D_x^{(0)}) = \begin{cases} \left(\sqrt{\frac{96L_{\varepsilon,r}^{\texttt{ADA}}}{\mu_\varepsilon^{\texttt{ADA}}}}, \sqrt{\frac{2^{K+1}\varepsilon}{\mu_\varepsilon^{\texttt{ADA}}}}\right) & \text{if } \mu_\varepsilon^{\texttt{ADA}} \geq \frac{4\varepsilon}{D_x^2}, \\ \left(\sqrt{\frac{24L_{\varepsilon,r}^{\texttt{ADA}}D_x^2}{2^{K-k-1}\varepsilon}}, D_x\right) & \text{otherwise,} \end{cases} \qquad D_\lambda^{(0)} = \min\left\{D_\lambda, \sqrt{2^{K+1}\varepsilon L_h}\right\} \tag{5.3}$$

Note that when $\mu_\varepsilon^{\texttt{ADA}} \geq 4\varepsilon/D_x^2$, T_k is independent of k.

The following theorem, proven in Section 5.5, establishes our universal convergence theory. We denote the gradient complexity of this restarted method by $N_{\varepsilon,r} := \sum_{k=0}^{K-1} \lceil T_k \rceil$ as R-UFCM computes $\lceil T_k \rceil$ gradients of each g_j in execution k of UFCM. Likewise, we denote the proximal complexity by $P_{\varepsilon,r} := \sum_{k=0}^{K-1} \lceil P_{\varepsilon,r}^{(k)} \rceil$ where $\lceil P_{\varepsilon,r}^{(k)} \rceil$ bounds the number of proximal evaluations of u and h used in the kth execution of UFCM.

Furthermore, we restrict $\varepsilon \in (0, 1]$ sufficiently small such that

$$\sqrt{\frac{24L_{\varepsilon,r}^{\texttt{ADA}}D_x^2}{\varepsilon}} \geq 1, \quad \sqrt{\frac{96L_{\varepsilon,r}^{\texttt{ADA}}}{\mu_\varepsilon^{\texttt{ADA}}}} \geq 1.$$

These restrictions must hold for sufficiently small ε as $\lim_{\varepsilon\to 0^+} L_{\varepsilon,r}^{\texttt{ADA}}/\varepsilon = +\infty$, which holds from Lemma 4.1. Secondly, $\lim_{\varepsilon\to 0^+} L_{\varepsilon,r}^{\texttt{ADA}}/\mu_\varepsilon^{\texttt{ADA}} \geq \max\left\{\lim_{\varepsilon\to 0^+} L_{1,r}^{\texttt{ADA}}/\mu_\varepsilon^{\texttt{ADA}}, \lim_{\varepsilon\to 0^+} L_{\varepsilon,r}^{\texttt{ADA}}/\mu_1^{\texttt{ADA}}\right\} \geq 1$, where the first inequality utilizes Lemmas 4.1 and 5.3, and the second notes that when $p_j = q_j = 1$ for all j, then $L_{\varepsilon,r}^{\texttt{ADA}}$ and $\mu_\varepsilon^{\texttt{ADA}}$ are constant with respect to ε, so $L_{\varepsilon,r}^{\texttt{ADA}} \geq \mu_\varepsilon^{\texttt{ADA}}$, while if any $p_j < 1$ or $q_j > 1$ then the first or second limit diverge to infinity respectively.

For notational ease, we let $\tilde{z}^k = (x^k; \lambda^k, \nabla g(x^k))$, extending each initialization $z^k = (x^k, \lambda^k)$ to include the conjugate variables. In particular $\tilde{z}^0 = (x^0; \lambda^0, \nabla g(x^0))$. We also let $\hat{z}^0 = (x^\star; \lambda^\star, \nabla g(x^0))$, extending the optimal primal-dual pair to include the conjugate variable at the initialization.

Theorem 5.5. *Consider any problem of the form* (2.1) *with each g_j being (L_j, p_j)-Hölder smooth and (μ_j, q_j)-uniformly convex, target accuracy $\varepsilon > 0$ sufficiently small, with $r = D_\lambda \sqrt{\varepsilon}$. Setting $C^{(k)} = D_\lambda^{(k)}/D_x^{(k)}$ and $\Delta^{(k)} = C^{(k)}/(2L_{\varepsilon,r}^{\texttt{ADA}})$, if $K \geq \left\lceil \log_2 \left(\frac{Q(\tilde{z}^0, \hat{z}^0)+\varepsilon}{\varepsilon} \right) \right\rceil$, Algorithm 2 with stepsizes* (3.15) *and* (3.16) *must find an (ε, r)-optimal solution* (1.8)*. If K is within a constant factor of $\left\lceil \log_2 \left(\frac{Q(\tilde{z}^0, \hat{z}^0)+\varepsilon}{\varepsilon} \right) \right\rceil$, this achieves the complexity bounds outlined in Table 1.*

Remark 5.6. *Since $L_{\varepsilon,r}^{\texttt{ADA}}$ is nonincreasing with ε and $\mu_\varepsilon^{\texttt{ADA}}$ is nondecreasing with ε, this bound can be tightened by considering our Approximate Dualized Aggregate constants specialized to the target accuracy sought by each application of UFCM. For each loop, one could run $UFCM(z^k, \lceil T_k \rceil, L_{2^{K-k-1}\varepsilon,r}^{\texttt{ADA}})$ with outer loop iteration count*

$$T_k = \min \left\{ \sqrt{\frac{96 L_{2^{K-k-1}\varepsilon,r}^{\texttt{ADA}}}{\mu_{2^{K-k-1}\varepsilon}^{\texttt{ADA}}}}, \sqrt{\frac{24 L_{2^{K-k-1}\varepsilon,r}^{\texttt{ADA}} D_x^2}{2^{K-k-1}\varepsilon}} \right\} ,$$

instead updating $D_x^{(k+1)}$ with $\sqrt{2^{K-k}\varepsilon/\mu_{2^{K-k}\varepsilon}^{\texttt{ADA}}}$ whenever $\mu_{2^{K-k}\varepsilon}^{\texttt{ADA}} \geq 2^{K-k+2}\varepsilon/D_x^2$ and $D_\lambda^{(k+1)}$ with $\sqrt{2^{K-k}\varepsilon L_h}$ whenever $L_h \leq D_\lambda^2/(2^{K-k}\varepsilon)$. Consequently, one can derive guarantees

$$N_{\varepsilon,r} = \mathrm{O}\left(\sum_{n=0}^{K-1} \sqrt{\frac{L_{2^{K-k-1}\varepsilon,r}^{\texttt{ADA}}}{\mu_{2^{K-k-1}\varepsilon}^{\texttt{ADA}}}} \right), \quad P_{\varepsilon,r} = \mathrm{O}\left(\sum_{n=0}^{K-1} \sqrt{\frac{(L_{2^{K-k-1}\varepsilon,r}^{\texttt{ADA}} + M^2 L_h)}{\mu_{2^{K-k-1}\varepsilon}^{\texttt{ADA}}}} \right)$$

which avoids additional multiplicative log terms if the sums above total up geometrically.

Corollary 5.7. *Consider minimizing $F(x) = g_0(x) + u(x)$ where g_0 is (L, p)-Hölder smooth and (μ, q)-uniformly convex function with $D_x \geq \|x^0 - x^\star\|$, initialization $\lambda^0 = 1$, and any target accuracy $0 < \varepsilon \leq \min \left\{ 1, 2\sqrt{6} L D_x^{1+p}, \frac{2\mu}{1+q} \left(\frac{D_x}{2} \right)^{1+q}, 4 \left(\frac{1+q}{2} \right)^{\frac{2}{3q+1}} L^{\frac{2(q+1)}{(3q+1)(1+p)}} \right\}$. Then Algorithm 2 recovers* (1.6)*:*

$$N_{\varepsilon,r} = P_{\varepsilon,r} = K_{UC}(\varepsilon, L, p, \mu, q) = \begin{cases} \mathrm{O}\left(\left(\frac{L^{1+q}}{\mu^{1+p} \varepsilon^{q-p}} \right)^{\frac{2}{(1+3p)(1+q)}} \right) & \text{if } q > p \ , \\ \mathrm{O}\left(\left(\frac{L^{1+q}}{\mu^{1+p}} \right)^{\frac{2}{(1+3p)(1+q)}} \log \left(\frac{F(x^0) - F^\star}{\varepsilon} \right) \right) & \text{if } q = p \end{cases}$$

up to logarithmic factors[3]*.*

Proof. By hypothesis, ε is sufficiently small to apply Theorem 5.5. Noting our Approximate Dualized Aggregate constants equal

$$L_{\varepsilon,r}^{\texttt{ADA}} = (1+r)^{\frac{4}{1+3p}} \left[\frac{1-p}{1+p} \cdot \frac{4\sqrt{6}}{\varepsilon \sqrt{\mu_\varepsilon^{\texttt{ADA}}}} \right]^{\frac{2-2p}{1+3p}} L^{\frac{4}{1+3p}} \quad \text{and} \quad \mu_\varepsilon^{\texttt{ADA}} = \left(\frac{2\mu}{1+q} \right)^{\frac{2}{1+q}} \varepsilon^{\frac{q-1}{q+1}} ,$$

[3]Using the modification discussed in Remark 5.6, one can recover the optimal rate without incurring log factors.

we conclude

$$N_{\varepsilon,r} = \tilde{\mathcal{O}}\left(\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}}}{\mu_{\varepsilon}^{\texttt{ADA}}}}\right) = \tilde{\mathcal{O}}\left(\frac{(\varepsilon\sqrt{\mu_{\varepsilon}^{\texttt{ADA}}})^{-\frac{1-p}{1+3p}} L^{\frac{2}{1+3p}}}{\sqrt{\mu_{\varepsilon}^{\texttt{ADA}}}}\right) = \tilde{\mathcal{O}}\left(\left(\frac{L^{1+q}}{\mu^{1+p}\varepsilon^{q-p}}\right)^{\frac{2}{(1+3p)(1+q)}}\right) ,$$

where the first equality considers Theorem 5.5, the second equality substitutes $L_{\varepsilon,r}^{\texttt{ADA}}$, and the last equality then substitutes $\mu_{\varepsilon}^{\texttt{ADA}}$ and simplifies to recover (1.6).

Since $\lambda^0 = \lambda^\star$, we can make Δ arbitrarily small, so $S_t = 1$ for each t. Therefore, the resulting proximal complexity equals the gradient oracle complexity. □

Corollary 5.8. *For any problem of the form* (2.1)*, target accuracy $\varepsilon > 0$ sufficiently small, with $r = D_\lambda\sqrt{\varepsilon}$, suppose each g_j is (L_j, p_j)-smooth and (μ_j, q_j)-uniformly convex. Algorithm 2, with stepsizes* (3.15) *and* (3.16)*, with choices of $C^{(k)} = D_\lambda^{(k)}/D_x^{(k)}$ and $\Delta^{(k)} = C^{(k)}/(2L_{\varepsilon,r}^{\texttt{ADA}})$ must find an (ε, r)-optimal solution with oracle complexity bound*

$$N_{\varepsilon,r} = \tilde{\mathcal{O}}\left(\sum_{j=1}^m K_{UC}(\varepsilon, (\lambda_j^\star + r)L_j, p_j, \mu_{\varepsilon}^{\texttt{ADA}}, 1)\right) .$$

Proof. Since $\varepsilon \leq D_x^2\mu_{\varepsilon}^{\texttt{ADA}}/4$, it holds that after rearrangement of the definition in (5.2), $L_{\varepsilon,r}^{\texttt{ADA}}$ is the unique positive root to

$$\sum_{j=1}^m (L_{\varepsilon,r}^{\texttt{ADA}})^{-\frac{1+3p_j}{2(1+p_j)}} \left[\frac{1-p_j}{1+p_j}\cdot\frac{4\sqrt{6}m}{\varepsilon\sqrt{\mu_{\varepsilon}^{\texttt{ADA}}}}\right]^{\frac{1-p_j}{1+p_j}} \left[(\lambda_j^\star + r)L_j\right]^{\frac{2}{1+p_j}} = 1 .$$

Similar to proving Corollary 4.5, we can bound $\sqrt{L_{\varepsilon,r}^{\texttt{ADA}}} \leq \sum_{j=1}^m \sqrt{L_{j,\varepsilon,r}^{\texttt{ADA}}}$ where $L_{j,\varepsilon,r}^{\texttt{ADA}}$ solves componentwise as the unique positive root to the following equation

$$(L_{j,\varepsilon,r}^{\texttt{ADA}})^{-\frac{1+3p_j}{2(1+p_j)}} \left[\frac{1-p_j}{1+p_j}\cdot\frac{4\sqrt{6}m}{\varepsilon\sqrt{\mu_{\varepsilon}^{\texttt{ADA}}}}\right]^{\frac{1-p_j}{1+p_j}} \left[(\lambda_j^\star + r)L_j\right]^{\frac{2}{1+p_j}} = \frac{1}{m} .$$

We can then conclude

$$\sqrt{L_{\varepsilon,r}^{\texttt{ADA}}} \leq \sum_{j=1}^m m^{\frac{2}{1+3p_j}} \left[\frac{1-p_j}{1+p_j}\cdot\frac{4\sqrt{6}}{\varepsilon\sqrt{\mu_{\varepsilon}^{\texttt{ADA}}}}\right]^{\frac{1-p_j}{1+3p_j}} \left[(\lambda_j^\star + r)L_j\right]^{\frac{2}{1+3p_j}} . \tag{5.4}$$

Finally, it holds that

$$N_{\varepsilon,r} = \tilde{\mathcal{O}}\left(\sum_{j=1}^m \frac{\left(\varepsilon\sqrt{\mu_{\varepsilon}^{\texttt{ADA}}}\right)^{-\frac{1-p_j}{1+3p_j}} \left[(\lambda_j^\star + r)L_j\right]^{\frac{2}{1+3p_j}}}{\sqrt{\mu_{\varepsilon}^{\texttt{ADA}}}}\right) = \tilde{\mathcal{O}}\left(\sum_{j=1}^m \left(\frac{\left[(\lambda_j^\star + r)L_j\right]^2}{(\mu_{\varepsilon}^{\texttt{ADA}})^{1+p_j}\varepsilon^{1-p_j}}\right)^{\frac{1}{(1+3p_j)}}\right) ,$$

where the first equality considers the result from Theorem 5.5 and substitutes the upper bound on $L_{\varepsilon,r}^{\texttt{ADA}}$ in (5.4), and the second equality simplifies to yield the desired result. □

Fixing $h(z) = \sum_{j=1}^m z_j$, each $\lambda_j^\star = 1$, and bounding $\mu_{\varepsilon}^{\texttt{ADA}}$ by only considering a single component in its sum recovers the results for heterogeneous sums of Hölder smooth terms of [15, Theorem 1.2]. When each g_j is smooth and h is a nonpositive indicator function, this second corollary recovers the results of [42, Theorem 6] by lower bounding $\mu_{\varepsilon}^{\texttt{ADA}}$ by the μ_0-strong convexity of g_0 (see the concluding Section 5.6 for further consideration of this special case).

5.5 Analysis of R-UFCM (Proof of Theorem 5.5)

Recall our analysis only depends on the (L_j, p_j)-Hölder smooth and (μ_j, q_j)-uniformly convex of g_j through our analysis through the universal constants $L_{\varepsilon,r}^{\texttt{ADA}}$ and $\mu_{\varepsilon}^{\texttt{ADA}}$ defined in (5.2) and (5.1). Let $\lambda \in \Lambda_r$ and $\nu \in V$. Below, we inductively prove that in all four of the cases in Table 1 (determined by whether $\mu_{\varepsilon}^{\texttt{ADA}} < 4\varepsilon/D_x^2$ and whether $L_h > D_\lambda^2/\varepsilon$) the following are maintained at each outer iteration of the restarted method $k = 0, 1, \ldots, K-1$

$$D_x^{(k)} \geq \|x^k - x^\star\|, \quad D_\lambda^{(k)} \geq \|\lambda^k - \lambda^\star\|, \quad Q((\bar{x}^{T_k,k}; \bar{\lambda}^{T_k,k}, \bar{\nu}^{T_k,k}), (x^\star; \lambda, \nu)) \leq 2^{K-k-1}\varepsilon \ ,$$

where we recall $(\bar{x}^{T_k,k}; \bar{\lambda}^{T_k,k}, \bar{\nu}^{T_k,k})$ from our averaging scheme (3.22). By definition and application of Lemma 5.4, $D_x^{(0)} \geq \|x^0 - x^\star\|$ and $D_\lambda^{(0)} \geq \|\lambda^0 - \lambda^\star\|$ both hold at $k = 0$, regardless of the relative sizes of $\mu_{\varepsilon}^{\texttt{ADA}}$ and L_h. Our inductive proof proceeds by first establishing that

$$D_x^{(k)} \geq \|x^k - x^\star\|, \quad D_\lambda^{(k)} \geq \|\lambda^k - \lambda^\star\| \implies Q((\bar{x}^{T_k,k}; \bar{\lambda}^{T_k,k}, \bar{\nu}^{T_k,k}), (x^\star; \lambda, \nu)) \leq 2^{K-k-1}\varepsilon \tag{5.5}$$

for each k. The key result to this end is that $Q((\bar{x}^{T_k,k}; \bar{\lambda}^{T_k,k}, \bar{\nu}^{T_k,k}), (x^\star; \lambda, \nu)) \leq 2^{K-k-1}\varepsilon$ if

$$T_k \geq \sqrt{\frac{24 L_{\varepsilon,r}^{\texttt{ADA}} (D_x^{(k)})^2}{2^{K-k-1}\varepsilon}}$$

by Theorem 4.3. Hence, we just need to verify our choice of T_k satisfies this inequality in each case. Then, to complete the induction, we establish

$$Q((\bar{x}^{T_k,k}; \bar{\lambda}^{T_k,k}, \bar{\nu}^{T_k,k}), (x^\star; \lambda, \nu)) \leq 2^{K-k-1}\varepsilon \implies D_x^{(k+1)} \geq \|x^{k+1} - x^\star\|, \quad D_\lambda^{(k+1)} \geq \|\lambda^{k+1} - \lambda^\star\| \ . \tag{5.6}$$

The key result to this end is the growth condition from Lemma 5.2, which guarantees that

$$G_x(\|\bar{x}^{T_k,k} - x^\star\|) + G_\lambda(\|\bar{\lambda}^{T_k,k} - \lambda^\star\|) \leq Q((\bar{x}^{T_k,k}; \bar{\lambda}^{T_k,k}, \bar{\nu}^{T_k,k}), (x^\star; \lambda^\star, \nabla g(\bar{x}^{T_k,k}))) \leq 2^{K-k-1}\varepsilon \ .$$

The remainder of this proof verifies the implications (5.5) and (5.6) and calculates the total gradient and proximal complexity in each case of Table 1. Finally, we deduce that

$$Q((\bar{x}^{T_{K-1},K-1}; \bar{\lambda}^{T_{K-1},K-1}, \bar{\nu}^{T_{K-1},K-1}), (x^\star, \hat{\lambda}, \nabla g(\bar{x}^{T_{K-1},K-1}))) \leq \varepsilon$$

and apply Lemma 3.2 to conclude that $\bar{x}^{T_{K-1},K-1}$ is (ε, r)-optimal.

Case 1: Suppose $\mu_{\varepsilon}^{\texttt{ADA}} < 4\varepsilon/D_x^2$. Observe the first needed implication for our induction (5.5) is immediate from Theorem 4.3 as

$$T_k = \sqrt{\frac{24 L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{2^{K-k-1}\varepsilon}} \ .$$

The gradient complexity follows from geometrically summing this quantity and bounding $K < \infty$, so

$$\sum_{k=0}^{K-1} \lceil T_k \rceil \leq \sum_{k=0}^{K-1} 1 + \sqrt{\frac{24 L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{2^{K-k-1}\varepsilon}} \leq K + \sum_{j=0}^{\infty} \sqrt{\frac{24 L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{2^j \varepsilon}} = \sqrt{\frac{(144 + 96\sqrt{2}) L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{\varepsilon}} + K \ .$$

Next, we verify the second needed implication (5.6). The primal bound is vacuously the case since the primal initialization is constant, so $x^k = x^0$ for each $k = 0, \ldots, K-1$ and

$$D_x^{(k)} = D_x \geq \|x^0 - x^\star\| = \|x^k - x^\star\| \ .$$

To derive the dual distance bound, we consider the two cases of dual restarting.

Case 1a: Suppose $L_h > D_\lambda^2/\varepsilon$. In this setting, the dual variable does not reinitialize each iteration and $D_\lambda^{(k)} = D_\lambda \geq \|\lambda^0 - \lambda^\star\| = \|\lambda^k - \lambda^\star\|$, completing the proof of (5.6). Observe that since $\Delta^{(k)} = D_\lambda/(2D_x L_{\varepsilon,r}^{\texttt{ADA}})$ as neither variable reinitializes, the number of proximal steps on u and h taken each iteration k of R-UFCM is

$$P_{\varepsilon,r}^{(k)} = \lceil T_k \rceil + \lceil T_k \rceil^2 \Delta^{(k)} M \leq 1 + \sqrt{\frac{24 L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{2^{K-k-1}\varepsilon}} + \frac{24 M D_x D_\lambda^{(k)}}{2^{K-k-1}\varepsilon} + \frac{M D_\lambda^{(k)}}{L_{\varepsilon,r}^{\texttt{ADA}} D_x}, \tag{5.7}$$

$$= 1 + \sqrt{\frac{24 L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{2^{K-k-1}\varepsilon}} + \frac{24 M D_x D_\lambda}{2^{K-k-1}\varepsilon} + \frac{M D_\lambda}{L_{\varepsilon,r}^{\texttt{ADA}} D_x}$$

where M bounds $\|\nabla g(x)\|$ for $x \in B(x^\star, \sqrt{27 D_x^2})$, and we use the facts $\lceil T_k \rceil \leq T_k + 1$ and $\lceil T_k \rceil^2 \leq 2T_k^2 + 2$. The total proximal complexity is then at most

$$\sum_{k=0}^{K-1} P_{\varepsilon,r}^{(k)} \leq \sqrt{\frac{(144 + 96\sqrt{2}) L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{\varepsilon}} + \frac{48 M D_x D_\lambda}{\varepsilon} + K\left(1 + \frac{M D_\lambda}{L_{\varepsilon,r}^{\texttt{ADA}} D_x}\right) .$$

Case 1b: Suppose $L_h \leq D_\lambda^2/\varepsilon$. To verify (5.6), we note that for the primary executions of UFCM where $D_\lambda^2 < 2^{K-k+1}\varepsilon L_h$, our initializations maintain $D_\lambda^{(k)} = D_\lambda$ and $\lambda^k = \lambda^0$. So $D_\lambda^{(k)} \geq \|\lambda^k - \lambda^\star\|$. For the subsequent executions, observe that Lemma 5.2 ensures

$$G_\lambda(\|\lambda^{k+1} - \lambda^\star\|) \leq Q((\bar{x}^{T_k,k}; \bar{\lambda}^{T_k,k}, \bar{\nu}^{T_k,k}), (x^\star; \lambda^\star, \nabla g(\bar{x}^{T_k,k}))) \leq 2^{K-k-1}\varepsilon ,$$

where $\lambda^{k+1} \leftarrow \bar{\lambda}^{T_k,k}$ by line 5 of Algorithm 2. We then utilize the growth bound to yield

$$\|\lambda^{k+1} - \lambda^\star\| \leq G_\lambda^{-1}(2^{K-k-1}\varepsilon) = \sqrt{2^{K-k}\varepsilon L_h} = D_\lambda^{(k+1)} ,$$

completing our induction in this case.

From the proximal complexity for application k of UFCM (5.7), we note that

$$P_{\varepsilon,r}^{(k)} \leq 1 + \sqrt{\frac{24 L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{2^{K-k-1}\varepsilon}} + \frac{24 M D_x \sqrt{2^{K-k+1}\varepsilon L_h}}{2^{K-k-1}\varepsilon} + \frac{M D_\lambda}{L_{\varepsilon,r}^{\texttt{ADA}} D_x}$$

$$\leq 1 + \sqrt{\frac{(48 L_{\varepsilon,r}^{\texttt{ADA}} + 1152 M^2 L_h) D_x^2}{2^{K-k-1}\varepsilon}} + \frac{M D_\lambda}{L_{\varepsilon,r}^{\texttt{ADA}} D_x} ,$$

since $D_\lambda^{(k)} \leq \min\{D_\lambda, \sqrt{2^{K-k+1}\varepsilon L_h}\}$ for each k. The total proximal complexity is then bounded by

$$\sum_{k=0}^{K-1} P_{\varepsilon,r}^{(k)} \leq \sqrt{\frac{(6 + 4\sqrt{2})(48 L_{\varepsilon,r}^{\texttt{ADA}} + 1152 M^2 L_h) D_x^2}{\varepsilon}} + K\left(1 + \frac{M D_\lambda}{L_{\varepsilon,r}^{\texttt{ADA}} D_x}\right) .$$

Case 2: Now suppose $\mu_\varepsilon^{\texttt{ADA}} \geq 4\varepsilon/D_x^2$. Observe that the first step of our induction (5.5) holds immediately after noting $D_x^{(k)} = \sqrt{2^{K-k+1}\varepsilon/\mu_\varepsilon^{\texttt{ADA}}}$ and applying Theorem 4.3 as

$$T_k = \sqrt{\frac{96 L_{\varepsilon,r}^{\texttt{ADA}}}{\mu_\varepsilon^{\texttt{ADA}}}} = \sqrt{\frac{96 L_{\varepsilon,r}^{\texttt{ADA}} \left(D_x^{(k)}\right)^2}{2^{K-k+1}\varepsilon}} .$$

Hence, the total gradient complexity is bounded by

$$\sum_{k=0}^{K-1} \lceil T_k \rceil \leq K \left(1 + \sqrt{\frac{96 L_{\varepsilon,r}^{\texttt{ADA}}}{\mu_\varepsilon^{\texttt{ADA}}}} \right) ,$$

where we bound $\lceil T_k \rceil$ by $T_k + 1$. Note when K is within a constant factor of $\left\lceil \log_2 \left(\frac{Q(\tilde{z}^0, \hat{z}^0) + \varepsilon}{\varepsilon} \right) \right\rceil$, then the gradient complexity is $\mathrm{O}\left(\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}}}{\mu_\varepsilon^{\texttt{ADA}}}} \log\left(\frac{1}{\varepsilon}\right) \right)$.

Next, we verify the second needed implication (5.6), noting that the dual distance bounds $D_\lambda^{(k)} \geq \|\lambda^k - \lambda^\star\|$ have already been shown to hold. Thus, we only need to consider the primal distance bounds. For $k = 0$, our initializations and Lemma 5.4 ensure

$$D_x^{(0)} = \sqrt{\frac{2^{K+1}\varepsilon}{\mu_\varepsilon^{\texttt{ADA}}}} \geq \sqrt{\frac{2(Q(\tilde{z}^0, \hat{z}^0) + \varepsilon)}{\mu_\varepsilon^{\texttt{ADA}}}} \geq \|x^0 - x^\star\| .$$

For $k \geq 1$, Lemma 5.2 ensures that

$$G_x(\|x^{k+1} - x^\star\|) \leq Q((\bar{x}^{T_k,k}; \bar{\lambda}^{T_k,k}, \bar{\nu}^{T_k,k}), (x^\star; \lambda^\star, \nabla g(\bar{x}^{T_k,k}))) \leq 2^{K-k-1}\varepsilon ,$$

where $x^{k+1} \leftarrow \bar{x}^{T_k,k}$ by line 4 of Algorithm 2. This bound implies

$$\|x^{k+1} - x^\star\| \leq G_x^{-1}(2^{K-k-1}\varepsilon) = \sqrt{2^{K-k}\varepsilon / \mu_{2^{K-k}\varepsilon}^{\texttt{ADA}}} \leq \sqrt{2^{K-k}\varepsilon / \mu_\varepsilon^{\texttt{ADA}}} = D_x^{(k+1)} ,$$

where the first equality comes from the characterization that $\mu_\varepsilon^{\texttt{ADA}} = \varepsilon / (G_x^{-1}(\varepsilon/2))^2$ and the last inequality holds as $\mu_\varepsilon^{\texttt{ADA}}$ is a nondecreasing function of ε. Next, we consider the proximal operator complexity.

Case 2a: Suppose $L_h > D_\lambda^2/\varepsilon$. In this case, $D_\lambda^{(k)} = D_\lambda$. The number of proximal steps performed each execution of UFCM is

$$P_{\varepsilon,r}^{(k)} = \lceil T_k \rceil + \lceil T_k \rceil^2 \Delta^{(k)} M \leq 1 + \sqrt{\frac{96 L_{\varepsilon,r}^{\texttt{ADA}}}{\mu_\varepsilon^{\texttt{ADA}}}} + \frac{192 M D_\lambda}{\sqrt{2^{K-k+1} \mu_\varepsilon^{\texttt{ADA}} \varepsilon}} ,$$

where the first equality uses (3.24), and the inequality substitutes $\Delta^{(k)} = D_\lambda / (2 L_{\varepsilon,r}^{\texttt{ADA}} D_x^{(k)})$ with $D_x^{(k)} = \sqrt{2^{K-k+1}\varepsilon / \mu_\varepsilon^{\texttt{ADA}}}$, further noting that since $\mu_\varepsilon^{\texttt{ADA}} \leq 96 L_{\varepsilon,r}^{\texttt{ADA}}$, $\lceil T_k \rceil \leq T_k + 1 \leq 2T_k$. Therefore, the total proximal complexity is bounded by

$$\sum_{k=0}^{K-1} P_{\varepsilon,r}^{(k)} \leq \frac{(1152 + 786\sqrt{2}) M D_\lambda}{\sqrt{\mu_\varepsilon^{\texttt{ADA}} \varepsilon}} + K \left(1 + \sqrt{\frac{96 L_{\varepsilon,r}^{\texttt{ADA}}}{\mu_\varepsilon^{\texttt{ADA}}}} \right) .$$

Case 2b: Suppose $L_h \leq D_\lambda^2/\varepsilon$. Now $D_\lambda^{(k)} \leq \sqrt{2^{K-k+1}\varepsilon L_h}$, in which case

$$P_{\varepsilon,r}^{(k)} \leq 1 + \sqrt{\frac{96 L_{\varepsilon,r}^{\texttt{ADA}}}{\mu_\varepsilon^{\texttt{ADA}}}} + \frac{192 M \sqrt{L_h}}{\sqrt{\mu_\varepsilon^{\texttt{ADA}}}} \leq 1 + \sqrt{\frac{192 L_{\varepsilon,r}^{\texttt{ADA}} + 73728 M^2 L_h}{\mu_\varepsilon^{\texttt{ADA}}}} .$$

The total proximal complexity is then bounded by

$$\sum_{k=0}^{K-1} P_{\varepsilon,r}^{(k)} \leq K \left(1 + \sqrt{\frac{192 L_{\varepsilon,r}^{\texttt{ADA}} + 73728 M^2 L_h}{\mu_\varepsilon^{\texttt{ADA}}}} \right) .$$

5.6 Application to Functionally Constrained Optimization

We conclude this section considering functionally constrained optimization with strongly convex and smooth components, recovering the linear convergence in terms of first-order oracle calls to g and sublinear convergence in terms of proximal operations analogous to [42]. In this setting, $h(z_0, \dots, z_m) = z_0 + \iota_{z \le 0}(z_1, \dots, z_m)$ with each g_j being μ_j-strongly convex results in constant $\mu_\varepsilon^{\texttt{ADA}} = \mu_0 + \sum_{j=1}^m \lambda_j^\star \mu_j$. (Note that our method and theory also apply more generally, given only Hölder smoothness and uniform convexity, but for the sake of this comparison, we restrict ourselves to considering only smooth and strongly convex constraints.)

Since h is nonsmooth, i.e. $L_h = \infty$, each restarted application of UFCM uses the fixed dual initialization λ^0. However, for small enough ε, the primal variables and distance bounds will update, with $D_x^{(k)} = \sqrt{2^{K-k+1}\varepsilon/\mu_\varepsilon^{\texttt{ADA}}}$. Therefore, Algorithm 2 reaches an ε-optimal solution with complexity bounds

$$N_{\varepsilon,r} = \mathrm{O}\left(\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}}}{\mu_\varepsilon^{\texttt{ADA}}}} \log\left(\frac{Q(\tilde{z}^0, \hat{z}^0)}{\varepsilon}\right)\right), \quad P_{\varepsilon,r} = \mathrm{O}\left(N_{\varepsilon,r} + \frac{D_\lambda M}{\sqrt{\mu_\varepsilon^{\texttt{ADA}}\varepsilon}}\right)$$

In contrast, the ACGD-S method of [42, Corollary 4] has oracle complexities

$$N_{\varepsilon,r} = \mathrm{O}\left(\sqrt{\frac{L(\Lambda_r)}{\mu_0}} \log\left(\frac{\sqrt{L(\Lambda_r)\mu_0} D_x^2}{\varepsilon}\right)\right), \quad P_{\varepsilon,r} = \mathrm{O}\left(N_{\varepsilon,r} + \frac{d(\Lambda_r)M}{\sqrt{\mu_0\varepsilon}}\right),$$

where $L(\Lambda_r) = \max_{\lambda \in \Lambda_r}\{\sum_{j=1}^m \lambda_j L_j\}$ and $d(\Lambda_r) = \|\lambda^\star\| + r$. In the case where only g_0 is strongly convex, our rate recovers theirs as $\mu_\varepsilon^{\texttt{ADA}} = \mu_0 + \sum_{j=1}^m \lambda_j^\star 0 = \mu_0$. Importantly, our method additionally benefits from strong convexity in the components as $\mu_\varepsilon^{\texttt{ADA}} > \mu_0$ whenever any active constraint is strongly convex (or even just, uniformly convex).

Funding

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

References

[1] Applegate, D., Hinder, O. & Lu, H. et al. (2023) Faster first-order primal-dual methods for linear programming using restarts and sharpness. *Math. Program.*, **201**, 133–184. Available at: `https://doi.org/10.1007/s10107-022-01901-9`

[2] Aybat, N. S., Fallah, A., Gurbuzbalaban, M. & Ozdaglar, A. (2019) A universally optimal multistage accelerated stochastic gradient method. *Adv. Neural Inf. Process. Syst.*, **32**. Available at: `https://proceedings.neurips.cc/paper_files/paper/2019/file/d630553e32ae21fb1a6df39c702d2c5c-Paper.pdf`

[3] Beck, A. (2017) *First-order methods in optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM). Available at: `https://epubs.siam.org/doi/abs/10.1137/1.9781611974997`

[4] Beck, A. & Teboulle, M. (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, **2**, 183–202. Available at: https://doi.org/10.1137/080716542

[5] Beck, A. & Teboulle, M. (2012) Smoothing and first order methods: a unified framework. *SIAM J. Optim.*, **22**, 557–580. Available at: https://doi.org/10.1137/100818327

[6] Ben-Tal, A., El Ghaoui, L. & Nemirovski, A. (2009) *Robust optimization*. Princeton, NJ: Princeton University Press. Available at: https://doi.org/10.1515/9781400831050

[7] Brekelmans, R., Masrani, V., Wood, F., Ver Steeg, G., & Galstyan, A. In *Thirty-seventh International Conference on Machine Learning (ICML 2020)*. Available at: https://proceedings.icml.cc/static/paper_files/icml/2020/2826-Paper.pdf

[8] Deng, Q., Lan, G. & Lin, Z. (2024) Uniformly optimal and parameter-free first-order methods for convex and function-constrained optimization. *arXiv preprint*, arXiv:2412.06319. Available at: https://arxiv.org/abs/2412.06319

[9] Devolder, O., Glineur, F. & Nesterov, Y. (2014). First-order methods of smooth convex optimization with inexact oracle. *Math. Program.* 146, 37–75. Available at: https://doi.org/10.1007/s10107-013-0677-5

[10] Diakonikolas, J. & Guzmán, C. (2024) Optimization on a finer scale: bounded local subgradient variation perspective. *arXiv preprint*, arXiv:2403.16317. Available at: https://arxiv.org/abs/2403.16317

[11] Fenchel, W. (1949) On conjugate convex functions. *Canad. J. Math.*, **1**, 73–77. Available at: https://doi.org/10.4153/CJM-1949-007-x

[12] Fercoq, O. (2023) Quadratic error bound of the smoothed gap and the restarted averaged primal-dual hybrid gradient. *Open J. Math. Optim.*, **4**, 1–34. Available at: https://ojmo.centre-mersenne.org/articles/10.5802/ojmo.26/

[13] Gasnikov, A. V. & Nesterov, Y. (2018) Universal method for stochastic composite optimization problems. *Comput. Math. Math. Phys.*, **58**, 48–64. Available at: https://doi.org/10.1134/S0965542518010050

[14] Ghadimi, S., Lan, G. & Zhang, H. (2019) Generalized uniformly optimal methods for nonlinear programming. *J. Sci. Comput.*, **79**, 1854–1881. Available at: https://doi.org/10.1007/s10915-019-00915-4

[15] Grimmer, B.(2023). On optimal universal first-order methods for minimizing heterogeneous sums. *Optim Lett* 18, 427–445 (2024). Available at: https://doi.org/10.1007/s11590-023-02060-2

[16] Grimmer, B. (2024) Radial duality part I: foundations. *Math. Program.*, **205**, 33–68. Available at: https://doi.org/10.1007/s10107-023-02006-7

[17] Grimmer, B. (2024) Radial duality part II: applications and algorithms. *Math. Program.*, **205**, 69–105. Available at: https://doi.org/10.1007/s10107-023-01974-0

[18] Guigues, V., Liang, J., & Monteiro, R. D. C. (2025) Universal subgradient and proximal bundle methods for convex and strongly convex hybrid composite optimization *arXiv preprint*, arXiv:2407.10073 https://arxiv.org/abs/2407.10073

[19] Ito, M., Lu, Z. & He, C. (2023) A parameter-free conditional gradient method for composite minimization under Hölder condition. *J. Mach. Learn. Res.*, **24**, 1–34. Available at: http://jmlr.org/papers/v24/22-0983.html

[20] Kavis, A., Levy, K. Y., Bach, F. & Cevher, V. (2019) UniXGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. *Advances in Neural Information Processing Systems 32*. Available at: https://proceedings.neurips.cc/paper_files/paper/2019/file/88855547570f7ff053fff7c54e5148cc-Paper.pdf

[21] Lan, G. (2015) Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. *Math. Program.*, **149**, 1–45. Available at: https://doi.org/10.1007/s10107-013-0737-x

[22] Lan, G. (2020) *First-order and stochastic optimization methods for machine learning*. New York, NY: Springer. Available at: https://doi.org/10.1007/978-3-030-39568-1

[23] Lan, G. (2016) Gradient sliding for composite optimization. *Math. Program.*, **159**, 201–235. Available at: https://doi.org/10.1007/s10107-015-0955-5

[24] Lan, G. & Zhou, Y. (2016) Conditional gradient sliding for convex optimization. *SIAM J. Optim.*, **26**, 1379–1409. Available at: https://doi.org/10.1137/140992382

[25] Li, T., Lan, G. (2025). A simple uniformly optimal method without line search for convex optimization. *Math. Program.* Available at: https://doi.org/10.1007/s10107-025-02250-z

[26] Liang, J. & Monteiro, R. D. C. (2023) A unified analysis of a class of proximal bundle methods for solving hybrid convex composite optimization problems. *Math. Oper. Res.*, **49**, 832–855. Available at: https://doi.org/10.1287/moor.2023.1372

[27] Lu, H., Yang, J.(2025). A Practical and Optimal First-Order Method for Large-Scale Convex Quadratic Programming. *Math. Program.* Available at: https://doi.org/10.1007/s10107-025-02241-0

[28] Nemirovski, A. S. & Nesterov, Y. E. (1985) Optimal methods of smooth convex minimization. *USSR Comput. Math. Math. Phys.*, **25**, 21–30. Available at: https://doi.org/10.1016/0041-5553(85)90100-4

[29] Nesterov, Y. (1983) A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, **269**, 543. Available at: https://cir.nii.ac.jp/crid/1370862715914709505

[30] Nesterov, Y. (2014) Universal gradient methods for convex optimization problems. *Math. Program.*, **152**, 381–404. Available at: https://doi.org/10.1007/s10107-014-0790-0

[31] Park, J. (2022) Fast gradient methods for uniformly convex and weakly smooth problems. *Adv. Comput. Math.*, **48**. Available at: https://doi.org/10.1007/s10444-022-09943-5

[32] Polak, E. (1997) Finite min-max and constrained optimization. In *Optimization: algorithms and consistent approximations*, pp. 167–367. Springer, New York. Available at: https://doi.org/10.1007/978-1-4612-0663-7_2

[33] Renegar, J. & Grimmer, B. (2022) A simple nearly optimal restart scheme for speeding up first-order methods. *Found. Comput. Math.*, **22**, 211–256. Available at: https://doi.org/10.1007/s10208-021-09502-2

[34] Rockafellar, R. T. (1996) *Convex analysis*. Princeton, NJ: Princeton University Press. Available at: https://doi.org/10.1137/1013042

[35] Rodomanov, A., Kavis, A., Wu, Y., Antonakopoulos, K., & Cevher, V. Universal Gradient Methods for Stochastic Convex Optimization. *arXiv preprint*, arXiv:2402.03210. Available at: https://arxiv.org/abs/2402.03210

[36] Roulet, V. & d'Aspremont, A. (2020) Sharpness, restart, and acceleration. *SIAM J. Optim.*, **30**, 262–289. Available at: https://doi.org/10.1137/18M1224568

[37] Thekumparampil, K., Jain, P., Netrapalli, P., & Oh, S. Projection Efficient Subgradient Method and Optimal Nonsmooth Frank-Wolfe Method. *Advances in Neural Information Processing Systems*. Available at: https://proceedings.neurips.cc/paper_files/paper/2020/file/8f468c873a32bb0619eaeb2050ba45d1-Paper.pdf.

[38] Wang, N. & Zhang, S. (2024) A gradient complexity analysis for minimizing the sum of strongly convex functions with varying condition numbers. *SIAM J. Optim.*, **34**, 1374–1401. Available at: https://doi.org/10.1137/22M1503646

[39] Yang, T. & Lin, Q. (2018) RSG: Beating subgradient method without smoothness and strong convexity. *J. Mach. Learn. Res.*, **19**, 1–33. Available at: http://jmlr.org/papers/v19/17-016.html

[40] Zhang, Z., Ahmed, S. & Lan, G. (2021) Efficient algorithms for distributionally robust stochastic optimization with discrete scenario support. *SIAM J. Optim.*, **31**, 1690–1721. Available at: https://doi.org/10.1137/19M1290115

[41] Zhang, Z. & Lan, G. (2022a) Optimal algorithms for convex nested stochastic composite optimization. *arXiv preprint*, arXiv:2011.10076. Available at: https://arxiv.org/abs/2011.10076

[42] Zhang, Z. & Lan, G. (2022b) Solving convex smooth function constrained optimization is almost as easy as unconstrained optimization. *arXiv preprint*, arXiv:2210.05807, version 3. Available at: https://arxiv.org/abs/2210.05807

[43] Zhang, Z. & Lan, G. (2023) Optimal methods for convex risk-averse distributed optimization. *SIAM J. Optim.*, **33**, 1518–1557. Available at: https://doi.org/10.1137/22M1485309

[44] Zhou, D., Ma, S., & Yang, J. (2025) AdaBB: Adaptive Barzilai-Borwein Method for Convex Optimization. *Mathematics of Operations Research*. Available at: https://pubsonline.informs.org/doi/abs/10.1287/moor.2024.0510

A Deferred Proofs

A.1 Deferred Proofs for Smooth Composite Analysis

Proof of Lemma 3.8. First we establish a convergence bound on the inner loop for each phase. Fix $t \geq 1$. Since $u(y) + \eta_t \|y - x^{t-1}\|^2/2$ has strong convexity with modulus η_t, the proximal step for y in line 8 of Algorithm 1 satisfies the three point inequality (see [22, Lemma 3.5])

$$\left\langle y_s^{(t)} - x, \tilde{h}^{(t),s} \right\rangle + u(y_s^{(t)}) - u(x) + \frac{\eta_t}{2}\left(\|y_s^{(t)} - x^{t-1}\|^2 - \|x - x^{t-1}\|^2\right) + \frac{1}{2}\left[(\beta^{(t)} + \eta_t)\|x - y_s^{(t)}\|^2 + \beta^{(t)}\|y_s^{(t)} - y_{s-1}^{(t)}\|^2 - \beta^{(t)}\|y_{s-1}^{(t)} - x\|^2\right] \leq 0 \ . \tag{A.1}$$

Recall that

$$\tilde{h}^{(t),s} = \begin{cases} (\nu^t)^\top \lambda_0^{(t)} + \rho^{(t)}(\nu^{t-1})^\top(\lambda_0^{(t)} - \lambda_{-1}^{(t)}) & \text{if } s = 1, \\ (\nu^t)^\top \lambda_{s-1}^{(t)} + (\nu^t)^\top(\lambda_{s-1}^{(t)} - \lambda_{s-2}^{(t)}) & \text{otherwise.} \end{cases}$$

In particular, line 7 of Algorithm 1 ensures that for $s \geq 2$

$$\left\langle y_s^{(t)} - x, \tilde{h}^{(t),s} \right\rangle = \left\langle y_s^{(t)} - x, \sum_{j=1}^m \lambda_{s,j}^{(t)} \nu_j^t \right\rangle - \left\langle y_s^{(t)} - x, \sum_{j=1}^m (\lambda_{s,j}^{(t)} - \lambda_{s-1,j}^{(t)}) \nu_j^t \right\rangle + \left\langle y_s^{(t)} - y_{s-1}^{(t)}, \sum_{j=1}^m (\lambda_{s-1,j}^{(t)} - \lambda_{s-2,j}^{(t)}) \nu_j^t \right\rangle + \left\langle y_{s-1}^{(t)} - x, \sum_{j=1}^m (\lambda_{s-1,j}^{(t)} - \lambda_{s-2,j}^{(t)}) \nu_j^t \right\rangle \ , \tag{A.2}$$

and for $s = 1$

$$\left\langle y_1^{(t)} - x, \tilde{h}^{(t),1} \right\rangle = \left\langle y_1^{(t)} - x, \sum_{j=1}^m \lambda_{1,j}^{(t)} \nu_j^t \right\rangle - \left\langle y_1^{(t)} - x, \sum_{j=1}^m (\lambda_{1,j}^{(t)} - \lambda_{0,j}^{(t)}) \nu_j^t \right\rangle + \rho^{(t)} \left\langle y_1^{(t)} - y_0^{(t)}, \sum_{j=1}^m (\lambda_{0,j}^{(t)} - \lambda_{-1,j}^{(t)}) \nu_j^{t-1} \right\rangle + \rho^{(t)} \left\langle y_0^{(t)} - x, \sum_{j=1}^m (\lambda_{0,j}^{(t)} - \lambda_{-1,j}^{(t)}) \nu_j^{t-1} \right\rangle \ . \tag{A.3}$$

Observe the third term is bounded by Young's inequality ($ab \leq \frac{a^2}{2\varepsilon} + \frac{b^2\varepsilon}{2}$ for all $\varepsilon > 0$) with

$$\begin{aligned} \left\langle y_s^{(t)} - y_{s-1}^{(t)}, \sum_{j=1}^m (\lambda_{s-1,j}^{(t)} - \lambda_{s-2,j}^{(t)}) \nu_j^t \right\rangle &\leq \frac{\beta^{(t)}\|y_s^{(t)} - y_{s-1}^{(t)}\|^2}{2} + \frac{\|\lambda_{s-1}^{(t)} - \lambda_{s-2}^{(t)}\|^2 \|\nu^t\|^2}{2\beta^{(t)}} \ , \\ \left\langle y_1^{(t)} - y_0^{(t)}, \sum_{j=1}^m (\lambda_{0,j}^{(t)} - \lambda_{-1,j}^{(t)}) \nu_j^{t-1} \right\rangle &\leq \frac{\beta^{(t)}\|y_1^{(t)} - y_0^{(t)}\|^2}{2\rho^{(t)}} + \frac{\rho^{(t)}\|\lambda_0^{(t)} - \lambda_{-1}^{(t)}\|^2 \|\nu^{t-1}\|^2}{2\beta^{(t)}} \ , \end{aligned} \tag{A.4}$$

for $s \geq 2$ and $s = 1$ respectively. Moreover, note that when summing over $s = 1, ..., S_t$,

$$\begin{aligned} \sum_{s=1}^{S_t} \mathcal{L}(y_s^{(t)}; \lambda_s^{(t)}, \nu^t) - \mathcal{L}(x; \lambda_s^{(t)}, \nu^t) &= \sum_{s=1}^{S_t} \left\langle y_s^{(t)} - x, \sum_{j=1}^m \lambda_{s,j}^{(t)} \nu_j^t \right\rangle + u(y_s^{(t)}) - u(x), \\ \frac{\beta^{(t)}}{2}\left(\|y_{S_t}^{(t)} - x\|^2 - \|y_0^{(t)} - x\|^2\right) &= \sum_{s=1}^{S_t} \frac{\beta^{(t)}}{2}\left(\|x - y_s^{(t)}\|^2 - \|y_{s-1}^{(t)} - x\|^2\right), \\ \sum_{s=2}^{S_t} \frac{\|\lambda_{s-1}^{(t)} - \lambda_{s-2}^{(t)}\|^2 \|\nu^t\|^2}{2\beta^{(t)}} &\leq \sum_{s=2}^{S_t} \frac{\gamma^{(t)}}{2}\|\lambda_{s-1}^{(t)} - \lambda_{s-2}^{(t)}\|^2, \end{aligned}$$

where the first equality holds by definition, the second holds from telescoping the norm squared terms, and the inequality holds from requirement (3.11). Furthermore, the inner product terms

telescope as well since

$$\begin{aligned}\rho^{(t)}\left\langle y_0^{(t)}-x,\sum_{j=1}^m(\lambda_{0,j}^{(t)}-\lambda_{-1,j}^{(t)})\nu_j^{t-1}\right\rangle-\left\langle y_{S_t}^{(t)}-x,\sum_{j=1}^m(\lambda_{S_t,j}^{(t)}-\lambda_{S_t-1,j}^{(t)})\nu_j^t\right\rangle\\ =\sum_{s=2}^{S_t}\left\langle y_{s-1}^{(t)}-x,\sum_{j=1}^m(\lambda_{s-1,j}^{(t)}-\lambda_{s-2,j}^{(t)})\nu_j^t\right\rangle-\left\langle y_s^{(t)}-x,\sum_{j=1}^m(\lambda_{s,j}^{(t)}-\lambda_{s-1,j}^{(t)})\nu_j^t\right\rangle\\ +\rho^{(t)}\left\langle y_0^{(t)}-x,\sum_{j=1}^m(\lambda_{0,j}^{(t)}-\lambda_{-1,j}^{(t)})\nu_j^{t-1}\right\rangle-\left\langle y_1^{(t)}-x,\sum_{j=1}^m(\lambda_{1,j}^{(t)}-\lambda_{0,j}^{(t)})\nu_j^t\right\rangle,\end{aligned}$$

where the substitutions $\lambda_0^{(t+1)}=\lambda_{S_t}^{(t)},\ \lambda_{-1}^{(t+1)}=\lambda_{S_t-1}^{(t)},\ y_0^{(t+1)}=y_{S_t}^{(t)}$ hold by line 11 of Algorithm 1.

We then sum (A.1) over $s=1,\dots,S_t$. After plugging in (A.2) and (A.3), applying the above equalities and inequalities, and considering requirement (3.11), we bound

$$\begin{aligned}&\sum_{s=1}^{S_t}\left(\mathcal{L}(y_s^{(t)};\lambda_s^{(t)},\nu^t)-\mathcal{L}(x;\lambda_s^{(t)},\nu^t)+\frac{\eta_t\|y_s^{(t)}-x^{t-1}\|^2+\eta_t\|y_s^{(t)}-x\|^2-\eta_t\|x-x^{t-1}\|^2}{2}\right)\\ &\quad+\rho^{(t)}\left\langle y_0^{(t)}-x,\sum_{j=1}^m(\lambda_{0,j}^{(t)}-\lambda_{-1,j}^{(t)})\nu_j^{t-1}\right\rangle-\left\langle y_{S_t}^{(t)}-x,\sum_{j=1}^m(\lambda_{S_t,j}^{(t)}-\lambda_{S_t-1,j}^{(t)})\nu_j^t\right\rangle \qquad (A.5)\\ &\quad\le\sum_{s=2}^{S_t}\frac{\gamma^{(t)}}{2}\|\lambda_{s-1}^{(t)}-\lambda_{s-2}^{(t)}\|^2+\frac{(\rho^{(t)})^2\|\nu^{t-1}\|^2}{2\beta^{(t)}}\|\lambda_0^{(t)}-\lambda_{-1}^{(t)}\|^2-\frac{1}{2}\left[\beta^{(t)}\|y_{S_t}^{(t)}-x\|^2-\beta^{(t)}\|y_0^{(t)}-x\|^2\right],\end{aligned}$$

since the terms $\|y_s^{(t)}-y_{s-1}^{(t)}\|^2$ cancel as (A.1) and (A.4) have the same coefficients.

Next we leverage this inner loop bound to derive bounds on the $\lambda_s^{(t)}$ terms. The proximal mapping in Line 9 of Algorithm 1 applies a proximal step to

$$p(\lambda):=\left\langle\lambda,\nu^t(y_s^{(t)}-\underline{x}^t)+g(\underline{x}^t)\right\rangle-h^*(\lambda)\ .$$

From the Fenchel-Young inequality, $\left\langle\lambda,\nu^t y_s^{(t)}-g^*(\nu^t)\right\rangle=\langle\lambda,\nu^t(y_s^t-\underline{x}^t)+g(\underline{x}^t)\rangle$. Therefore, the proximal mapping in line 9, the three point inequality of [22, Lemma 3.5], and L_h-smoothness of h imply the following is nonpositive

$$\mathcal{L}(y_s^{(t)};\lambda,\nu^t)-\mathcal{L}(y_s^{(t)};\lambda_s^{(t)},\nu^t)+\frac{\gamma^{(t)}}{2}\left[\|\lambda-\lambda_s^{(t)}\|^2+\|\lambda_s^{(t)}-\lambda_{s-1}^{(t)}\|^2-\|\lambda-\lambda_{s-1}^{(t)}\|^2\right]+\frac{1}{2L_h}\|\lambda_s^{(t)}-\lambda\|^2\ .$$

Taking the sum over $s=1,\dots,S_t$ and combining with (A.5),

$$\begin{aligned}&\sum_{s=1}^{S_t}\left(\mathcal{L}(y_s^{(t)};\lambda,\nu^t)-\mathcal{L}(x;\lambda_s^{(t)},\nu^t)+\frac{\eta_t\|y_s^{(t)}-x^{t-1}\|^2+\eta_t\|y_s^{(t)}-x\|^2-\eta_t\|x-x^{t-1}\|^2}{2}\right)\\ &\quad+\sum_{s=1}^{S_t}\frac{1}{2L_h}\|\lambda_s^{(t)}-\lambda\|^2+\rho^{(t)}\left\langle y_0^{(t)}-x,\sum_{j=1}^m(\lambda_{0,j}^{(t)}-\lambda_{-1,j}^{(t)})\nu_j^{t-1}\right\rangle-\left\langle y_{S_t}^{(t)}-x,\sum_{j=1}^m(\lambda_{S_t,j}^{(t)}-\lambda_{S_t-1,j}^{(t)})\nu_j^t\right\rangle\\ &\quad\le\frac{\gamma^{(t)}}{2}\|\lambda-\lambda_0^{(t)}\|^2-\frac{\gamma^{(t)}}{2}\left[\|\lambda-\lambda_{S_t}^{(t)}\|^2+\|\lambda_{S_t}^{(t)}-\lambda_{S_t-1}^{(t)}\|^2\right]+\frac{(\rho^{(t)})^2\|\nu^{t-1}\|^2}{2\beta^{(t)}}\|\lambda_0^{(t)}-\lambda_{-1}^{(t)}\|^2\\ &\qquad-\frac{\beta^{(t)}}{2}\left[\|y_{S_t}^{(t)}-x\|^2-\|y_0^{(t)}-x\|^2\right]\ ,\end{aligned}$$

noting that the $\|\lambda - \lambda_s^{(t)}\|^2$ terms telescope and $\gamma^{(t)}/2 \sum_{s=2}^{S_t} \|\lambda_{s-1}^{(t)} - \lambda_{s-2}^{(t)}\|^2$ cancels, leaving only $\gamma^{(t)}/2\|\lambda_{S_t}^{(t)} - \lambda_{S_t-1}\|^2$. Since $\mathcal{L}(y_s^{(t)}; \lambda, \nu^t)$ and $\|y_s^{(t)} - x\|^2$ are convex in $y_s^{(t)}$ and $\mathcal{L}(x; \lambda_s^{(t)}, \nu^t)$ is concave in $\lambda_s^{(t)}$, multiplying by $\tilde{\omega}^{(t)} = \omega_t/S_t$ and considering the averaging scheme in line 14 of Algorithm 1, one can apply Jensen's inequality to derive a bound with respect to x^t and $\tilde{\lambda}^t$ of

$$\begin{aligned}
&\omega_t \left(\mathcal{L}(x^t; \lambda, \nu^t) - \mathcal{L}(x; \tilde{\lambda}^t, \nu^t) + \frac{\eta_t}{2} \left(\|x^t - x^{t-1}\|^2 + \|x^t - x\|^2 - \|x^{t-1} - x\|^2 \right) \right) + \sum_{s=1}^{S_t} \frac{\omega_t \|\lambda - \lambda_s^{(t)}\|^2}{2L_h S_t} \\
&+ \tilde{\omega}^{(t)} \rho^{(t)} \left\langle y_0^{(t)} - x, \sum_{j=1}^{m} (\lambda_{0,j}^{(t)} - \lambda_{-1,j}^{(t)}) \nu_j^{t-1} \right\rangle - \tilde{\omega}^{(t)} \left\langle y_{S_t}^{(t)} - x, \sum_{j=1}^{m} (\lambda_{S_t,j}^{(t)} - \lambda_{S_t-1,j}^{(t)}) \nu_j^{t} \right\rangle \\
&\leq \frac{\tilde{\omega}^{(t)} \gamma^{(t)}}{2} \left(\|\lambda - \lambda_0^{(t)}\|^2 - \|\lambda - \lambda_{S_t}^{(t)}\|^2 - \|\lambda_{S_t}^{(t)} - \lambda_{S_t-1}^{(t)}\|^2 \right) \\
&+ \frac{\tilde{\omega}^{(t)} (\rho^{(t)})^2 \|\nu^{t-1}\|^2}{2\beta^{(t)}} \|\lambda_0^{(t)} - \lambda_{-1}^{(t)}\|^2 - \frac{\tilde{\omega}^{(t)} \beta^{(t)}}{2} \left[\|y_{S_t}^{(t)} - x\|^2 - \|y_0^{(t)} - x\|^2 \right] . \qquad \text{(A.6)}
\end{aligned}$$

Next, we note that

$$\begin{aligned}
Q_x(z^t, z) + Q_\lambda(z^t, z) &= \mathcal{L}(x^t; \lambda, \nu^t) - \mathcal{L}(x; \tilde{\lambda}^t, \nu^t), \\
\frac{\omega_T \eta_T}{2} \|x^T - x\|^2 - \frac{\omega_1 \eta_1}{2} \|x_0 - x\|^2 &\leq \sum_{t=1}^{T} \frac{\omega_t \eta_t}{2} \left(\|x^t - x\|^2 - \|x^{t-1} - x\|^2 \right)
\end{aligned}$$

where the first equality comes from definition and the inequality comes from telescoping along with requirement (3.7). Then by telescoping, we produce the following bounds

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\tilde{\omega}^{(t)} \beta^{(t)}}{2} \left(\|y_0^{(t)} - x\|^2 - \|y_{S_t}^{(t)} - x\|^2 \right) &\leq \frac{\tilde{\omega}^{(1)} \beta^{(1)}}{2} \|y_0^{(1)} - x\|^2 - \frac{\tilde{\omega}^{(T)} \beta^{(T)}}{2} \|y_{S_T}^{(T)} - x\|^2 \\
\sum_{t=1}^{T} \frac{\tilde{\omega}^{(t)} \gamma^{(t)}}{2} \left(\|\lambda - \lambda_0^{(t)}\|^2 - \|\lambda - \lambda_{S_t}^{(t)}\|^2 \right) &\leq \frac{\tilde{\omega}^{(1)} \gamma^{(1)}}{2} \|\lambda - \lambda_0^{(1)}\|^2 - \frac{\tilde{\omega}^{(T)} \gamma^{(T)}}{2} \|\lambda - \lambda_{S_T}^{(T)}\|^2 \\
\sum_{t=1}^{T} -\frac{\tilde{\omega}^{(t)} \gamma^{(t)}}{2} \|\lambda_{S_t}^{(t)} - \lambda_{S_t-1}^{(t)}\|^2 + \frac{\tilde{\omega}^{(t)} (\rho^{(t)})^2 \|\nu^{t-1}\|^2}{2\beta^{(t)}} \|\lambda_0^{(t)} - \lambda_{-1}^{(t)}\|^2 & \\
\leq \sum_{t=1}^{T} \frac{\tilde{\omega}^{(t)} \gamma^{(t)}}{2} \left(\|\lambda_0^{(t)} - \lambda_{-1}^{(t)}\|^2 - \|\lambda_{S_t}^{(t)} - \lambda_{S_t-1}^{(t)}\|^2 \right) & \\
\leq \frac{\tilde{\omega}^{(1)} \gamma^{(1)}}{2} \|\lambda_0^{(1)} - \lambda_{-1}^{(1)}\|^2 - \frac{\tilde{\omega}^{(T)} \gamma^{(T)}}{2} \|\lambda_{S_T}^{(T)} - \lambda_{S_T-1}^{(T)}\|^2 &
\end{aligned}$$

where the first two inequalities apply (3.12) and (3.13) respectively, while the following bound applies requirement (3.14), line 11 of Algorithm 1, and uses (3.13) to telescope.

Finally for $t \geq 2$, by line 11 of Algorithm 1 and the second condition of (3.14),

$$\begin{aligned}
&\tilde{\omega}^{(t)} \rho^{(t)} \left\langle y_0^{(t)} - x, \sum_{j=1}^{m} (\lambda_{0,j}^{(t)} - \lambda_{-1,j}^{(t)}) \nu_j^{t-1} \right\rangle - \tilde{\omega}^{(t)} \left\langle y_{S_t}^{(t)} - x, \sum_{j=1}^{m} (\lambda_{S_t,j}^{(t)} - \lambda_{S_t-1,j}^{(t)}) \nu_j^{t} \right\rangle = \\
&\tilde{\omega}^{(t-1)} \left\langle y_{S_{t-1}}^{(t-1)} - x, \sum_{j=1}^{m} (\lambda_{S_{t-1},j}^{(t-1)} - \lambda_{S_{t-1}-1,j}^{(t-1)}) \nu_j^{t-1} \right\rangle - \tilde{\omega}^{(t)} \left\langle y_{S_t}^{(t)} - x, \sum_{j=1}^{m} (\lambda_{S_t,j}^{(t)} - \lambda_{S_t-1,j}^{(t)}) \nu_j^{t} \right\rangle .
\end{aligned}$$

Since $\lambda_{-1}^{(1)} = \lambda_0^1$ by initialization, summing the inner product terms over $t = 1, ..., T$ results in $-\tilde{\omega}^{(T)} \left\langle y_{S_T}^{(T)} - x, \sum_{j=1}^m (\lambda_{S_T,j}^{(T)} - \lambda_{S_T-1,j}^{(T)}) \nu_j^T \right\rangle$. Rearranging, applying Young's inequality, using requirement (3.11), and combining with the bounds outlined above results in

$$\sum_{t=1}^T \omega_t [Q_x(z^t, z) + Q_\lambda(z^t, z)] + \sum_{t=1}^T \sum_{s=1}^{S_t} \frac{\omega_t}{S_t} \frac{1}{2L_h} \|\lambda - \lambda_s^{(t)}\|^2 + \sum_{t=1}^T \frac{\omega_t \eta_t}{2} \|x^t - x^{t-1}\|^2$$
$$+ \frac{\omega_T \eta_T}{2} \|x^T - x\|^2 - \frac{\omega_1 \eta_1}{2} \|x^0 - x\|^2 \leq \frac{\tilde{\omega}^{(1)}}{2} \left(\gamma^{(1)} \|\lambda_0^{(1)} - \lambda\|^2 + \beta^{(1)} \|y_0^{(1)} - x\|^2 \right) .$$

Proof of Lemma 3.9. Recall that $\tilde{x}^t = x^{t-1} + \theta_t (x^{t-1} - x^{t-2})$. Thus,

$$\left\langle \nu_j - \nu_j^t, (\tilde{x}^t - x^t) \right\rangle = - \left\langle \nu_j - \nu_j^t, (x^t - x^{t-1}) \right\rangle + \theta_t \left\langle \nu_j - \nu_j^{t-1}, (x^{t-1} - x^{t-2}) \right\rangle$$
$$+ \theta_t \left\langle \nu_j^{t-1} - \nu_j^t, (x^{t-1} - x^{t-2}) \right\rangle .$$

Using [22, Lemma 3.5] and that g_j^* is strongly convex with modulus 1 with respect to the Bregman divergence $U_{g_j^*}$, the proximal mapping

$$\nu_j^t \leftarrow \operatorname*{argmax}_{\nu_j \in V_j} \left\langle \nu_j, \tilde{x}^t \right\rangle - g_j^*(\nu_j) - \tau_t U_{g_j^*}(\nu, \nu^{t-1}) \quad \forall j \in \{1, \ldots, m\} ,$$

which is equivalent to line 4 of Algorithm 1, satisfies $j \in \{1, \ldots, m\}$

$$\left\langle \nu_j - \nu_j^t, x^t \right\rangle + \left\langle \nu_j - \nu_j^t, \tilde{x}^t - x^t \right\rangle + g_j^*(\nu_j^t) - g_j^*(\nu_j)$$
$$\leq \tau_t U_{g_j^*}(\nu_j; \nu_j^{t-1}) - (\tau_t + 1) U_{g_j^*}(\nu_j; \nu_j^t) - \tau_t U_{g_j^*}(\nu_j^t; \nu_j^{t-1}) .$$

Summing over $t = 1, \ldots, T$ with weights ω_t yields

$$\sum_{t=1}^T \omega_t \left[\left\langle \nu_j - \nu_j^t, x^t \right\rangle + g_j^*(\nu_j^t) - g_j^*(\nu_j) \right]$$
$$+ \sum_{t=1}^T -\omega_t \left\langle \nu_j - \nu_j^t, (x^t - x^{t-1}) \right\rangle + \sum_{t=1}^T \omega_t \theta_t \left\langle \nu_j - \nu_j^{t-1}, (x^{t-1} - x^{t-2}) \right\rangle$$
$$+ \sum_{t=1}^T \omega_t \theta_t \left\langle \nu_j^{t-1} - \nu_j^t, (x^{t-1} - x^{t-2}) \right\rangle$$
$$\leq \sum_{t=1}^T \omega_t \left[\tau_t U_{g_j^*}(\nu_j; \nu_j^{t-1}) - (\tau_t + 1) U_{g_j^*}(\nu_j; \nu_j^t) - \tau_t U_{g_j^*}(\nu_j^t; \nu_j^{t-1}) \right] .$$

Applying requirements (3.8) and (3.9), one can conclude that

$$\sum_{t=1}^T \omega_t \left[\left\langle \nu_j - \nu_j^t, x^t \right\rangle + g_j^*(\nu_j^t) - g_j^*(\nu_j) \right] \leq - \left[\omega_T (\tau_T + 1) U_{g_j^*}(\nu_j; \nu_j^T) - \omega_T \left\langle \nu_j - \nu_j^T, x^T - x^{T-1} \right\rangle \right]$$
$$- \left[\sum_{t=1}^T \omega_t \tau_t U_{g_j^*}(\nu_j^t; \nu_j^{t-1}) - \omega_{t-1} \left\langle \nu_j^{t-1} - \nu_j^t, (x^{t-1} - x^{t-2}) \right\rangle \right] + \omega_1 \tau_1 U_{g_j^*}(\nu_j, \nu_j^0) .$$

Taking the sum over $j = 1, \ldots, m$ with weights λ_j yields the desired result.

Proof of Proposition 3.10. From Lemma 3.7, any $\nu \in V$ satisfies

$$\sum_{j=1}^{m} \lambda_j U_{g_j^*}(\nu; \nu_j^T) \geq \frac{1}{2L_{\varepsilon,r}^{\texttt{ADA}}} \left\| \sum_{j=1}^{m} \lambda_j (\nu_j - \nu_j^T) \right\|^2 .$$

Therefore, by Lemma 3.9 and requirement (3.9),

$$\begin{aligned} \sum_{t=1}^{T} \omega_t \left[Q_\nu(z^t, z) \right] \leq & \left[\omega_T \left\langle \sum_{j=1}^{m} \lambda_j (\nu_j - \nu_j^T), x^T - x^{T-1} \right\rangle - \frac{\omega_T(\tau_T + 1)}{2L_{\varepsilon,r}^{\texttt{ADA}}} \left\| \sum_{j=1}^{m} \lambda_j (\nu_j - \nu_j^T) \right\|^2 \right] \\ & + \sum_{t=2}^{T} \left[\omega_{t-1} \left\langle \sum_{j=1}^{m} \lambda_j (\nu_j^{t-1} - \nu_j^t), (x^{t-1} - x^{t-2}) \right\rangle - \frac{\omega_{t-1}\tau_t}{2\theta_t L_{\varepsilon,r}^{\texttt{ADA}}} \left\| \sum_{j=1}^{m} \lambda_j (\nu_j^t - \nu_j^{t-1}) \right\|^2 \right] \\ & + \omega_1 \tau_1 \left(\sum_{j=1}^{m} \lambda_j U_{g_j^*}(\nu_j, \nu_j^0) \right) . \end{aligned} \tag{A.7}$$

Applying Young's inequality to the inner product yields

$$\sum_{t=1}^{T} \omega_t \left[Q_\nu(z^t, z) \right] \leq \frac{\omega_T L_{\varepsilon,r}^{\texttt{ADA}}}{2(\tau_T + 1)} \|x^T - x^{T-1}\|^2 + \sum_{t=1}^{T-1} \frac{\omega_t \theta_{t+1} L_{\varepsilon,r}^{\texttt{ADA}}}{2\tau_{t+1}} \|x^t - x^{t-1}\|^2 + \omega_1 \tau_1 \left\langle \lambda, U_{g^*}(\nu_j, \nu_j^0) \right\rangle . \tag{A.8}$$

Utilizing the stepsize conditions (3.9) and (3.10), combining with Lemma 3.8, and the fact that $y_0^{(1)} = x^0$ and $\lambda_0^{(1)} = \lambda^0$, achieves the desired bound.

A.2 Proofs Deferred for Heterogeneously Smooth Composite Analysis

Proof of Proposition 4.2. Similar to the proof of Proposition 3.3, we note the bound from Proposition 4.9. Considering $\tau_1 = 0$ and the nonnegativity of L_h, the associated norm, and $Q(z^t, z^\star)$, it holds for all $T \geq 1$,

$$\frac{\omega_T \eta_T}{2} \|x^T - x^\star\|^2 \leq \frac{\tilde{\omega}^{(1)} \beta^{(1)} + \omega_1 \eta_1}{2} \|x^0 - x^\star\|^2 + \frac{\tilde{\omega}^{(1)} \gamma^{(1)}}{2} \|\lambda^0 - \lambda^\star\|^2 + \frac{\delta}{2} \left[\omega_T(\tau_T + 1) + \sum_{t=2}^{T} \omega_t \tau_t \right] .$$

Using the stepsize conditions (3.15) and (3.16) and our distance bounds, and letting $N_\varepsilon = \sqrt{\frac{24 L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{\varepsilon}}$ then

$$\|x^T - x^\star\|^2 \leq \frac{1}{2L_{\varepsilon,r}^{\texttt{ADA}}} \left[(C/\Delta + 2L_{\varepsilon,r}^{\texttt{ADA}}) D_x^2 + \frac{1}{C\Delta} D_\lambda^2 + \frac{12 L_{\varepsilon,r}^{\texttt{ADA}} D_x^2}{N_\varepsilon^3} \left(\frac{T^3 + 3T^2 + 2T}{6} \right) \right], \quad \forall T \geq 1,$$

where we bounded $\left[\omega_T(\tau_T + 1) + \sum_{t=2}^{T} \omega_t \tau_t \right] \leq \frac{T^3 + 3T^2 + 2T}{6}$ with specialized $\delta = \varepsilon / N_\varepsilon$. Finally considering particular iterate $t \geq 1$, we can further bound the rightmost product by $48 L_{\varepsilon,r}^{\texttt{ADA}} D_x^2$ for any $t \leq \lceil N_\varepsilon \rceil$.

The remainder of the proof follows analogously to the proof of Proposition 3.3.

Proof of Lemma 4.7. Let $\lambda \in \Lambda_r$ and $\bar{g}(x) = \sum_{j=1}^m \lambda_j g_j$. Moreover, consider any $\delta > 0$ and

$$L_\lambda \geq \sum_{j=1}^m \left[\left[\frac{1-p_j}{1+p_j} \cdot \frac{m}{\delta} \right]^{\frac{1-p_j}{1+p_j}} \lambda_j^{\frac{2}{1+p_j}} L_j^{\frac{2}{1+p_j}} \right] .$$

Since $\nabla \bar{g}(y) = \sum_{j=1}^m \lambda_j \nabla g_j(y)$ for all $y \in X$, letting $\nu = \nabla g(x)$ and $\hat{\nu} = \nabla g(\hat{x})$, one has that

$$\langle \lambda, U_{g^*}(\nu; \hat{\nu}) \rangle = \sum_{j=1}^m \lambda_j U_{g_j^*}(\nu_j; \hat{\nu}_j) = \sum_{j=1}^m \lambda_j U_{g_j}(\hat{x}; x) = U_{\bar{g}}(\hat{x}; x) \geq \frac{1}{2L_\lambda} \| \sum_{j=1}^m \lambda_j (\nu_j - \hat{\nu}_j) \|^2 - \frac{\delta}{2}$$

where the second equality follows from [7, Appendix A.2] and the inequality from using the cocoercivity condition in Lemma 2.4. Noting in (4.1), $L_{\delta,r} \geq L_\lambda$ for all $\lambda \in \Lambda_r$ gives the desired bound.

Proof of Lemma 4.8. This result follows analogously to deriving (A.7) in the proof of Proposition 3.10, with the modification of noting that by Lemma 4.7, for any $\nu \in V$ and $\delta > 0$,

$$\sum_{j=1}^m \lambda_j U_{g_j^*}(\nu; \nu_j^T) \geq \frac{1}{2L_{\delta,r}} \left\| \sum_{j=1}^m \lambda_j (\nu_j - \nu_j^T) \right\|^2 - \frac{\delta}{2} .$$

By Lemma 3.9 and requirement (3.9), with the above substitution

$$\begin{aligned} \sum_{t=1}^T \omega_t \left[Q_\nu(z^t, z) \right] \leq \omega_T(\tau_T + 1) & \left[\frac{\delta}{2} - \frac{1}{2L_{\delta,r}} \left\| \sum_{j=1}^m \lambda_j (\nu_j - \nu_j^T) \right\|^2 \right] \\ & + \omega_T \left\langle \sum_{j=1}^m \lambda_j (\nu_j - \nu_j^T), x^T - x^{T-1} \right\rangle + \sum_{t=2}^T \Bigg[\omega_t \tau_t \left(\frac{\delta}{2} - \frac{1}{2L_{\delta,r}} \left\| \sum_{j=1}^m \lambda_j (\nu_j^t - \nu_j^{t-1}) \right\|^2 \right) \\ & + \omega_{t-1} \left\langle \sum_{j=1}^m \lambda_j (\nu_j^{t-1} - \nu_j^t), (x^{t-1} - x^{t-2}) \right\rangle \Bigg] + \omega_1 \tau_1 \left(\sum_{j=1}^m \lambda_j U_{g_j^*}(\nu_j, \nu_j^0) \right) . \end{aligned}$$

Rearranging and applying Young's inequality (analogous to (A.8)) concludes the proof.