On Optimal Universal First-Order Methods for Minimizing Heterogeneous Sums

Benjamin Grimmer*

Abstract

This work considers minimizing a convex sum of functions, each with potentially different structure ranging from nonsmooth to smooth, Lipschitz to non-Lipschitz. Nesterov's universal fast gradient method [1] provides an optimal black-box first-order method for minimizing a single function that takes advantage of any continuity structure present without requiring prior knowledge. In this paper, we show that this landmark method (without modification) further adapts to heterogeneous sums. For example, it minimizes the sum of a nonsmooth M-Lipschitz function and an L-smooth function at a rate of $O(M^2/\epsilon^2 + \sqrt{L/\epsilon})$ without knowledge of M, L, or even that the objective was a sum of two terms. This rate is precisely the sum of the optimal convergence rates for each term's individual complexity class. More generally, we show that sums of varied Hölder smooth functions introduce no new complexities and require at most as many iterations as is needed for minimizing each summand separately. Extensions to strongly convex and Hölder growth settings as well as simple matching lower bounds are also provided.

1 Introduction

In this paper, we are interested in approximately solving convex optimization problems of the form

$$p_* = \min_{x \in Q} F(x) + \Psi(x) \tag{1.1}$$

where $F(x) = \sum_{j \in \mathcal{J}} f_j(x)$ is given by a sum of functions f_j each individually possessing some standard structure (Lipschitz continuity, smoothness, or more generally Hölder smoothness). This sum is heterogeneous in that it is a composition of several terms ranging from smooth to nonsmooth, Lipschitz to non-Lipschitz. Note typically F will not possess any of the structure held by its components. We consider methods given a (sub)gradient oracle for F, seeking an approximate $\epsilon > 0$-minimizer satisfying $F(y) + \Psi(y) - p_* \leq \epsilon$.

In the landmark paper [1], Nesterov introduced the Universal Fast Gradient Method (`UFGM`), which we review in Section 2. This method was analyzed for minimizing $F(x) + \Psi(x)$ where $F = f$ is a single (M, v)-Hölder smooth function, defined for $M \geq 0$ and $v \in [0, 1]$ as

$$\|\nabla f(x) - \nabla f(y)\|_* \leq M\|x - y\|^v, \quad x, y \in Q \ . \tag{1.2}$$

Note $(L, 1)$-Hölder smoothness corresponds to the typical smooth optimization assumption of an L-Lipschitz gradient, and $(M, 0)$ corresponds to the typical nonsmooth optimization assumption of having an M-Lipschitz objective function. `UFGM`'s iterates y_k are all ϵ-minimizers once

$$k \geq 2^{\frac{3+5v}{1+3v}} \left(\frac{M}{\epsilon}\right)^{\frac{2}{1+3v}} \xi(x_0, x^*)^{\frac{1+v}{1+3v}} \ , \tag{1.3}$$

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

where $x^* \in \operatorname{argmin}_{x \in Q} F(x) + \Psi(x)$ and $\xi(x_0, x^*)$ is a Bregman divergence, formally defined in (2.1), measuring the initial distance from optimality. This rate cannot be improved upon with respect to all three of $M, \epsilon, \xi(x_0, x^*)$ for any Hölder exponent $v \in [0, 1]$ [2]. Moreover, `UFGM` does not require knowledge of any of these parameters to be run. Hence this algorithm is "optimal" and "universal".

In this paper, we extend this optimal, universal theory to apply to generic sums of Hölder smooth functions of the form (1.1). As a simple first example, consider $F(x) = \frac{1}{2}(|x| + x^2)$, which is not Hölder smooth for any (M, v) despite being the sum of $(1, 0)$ and $(1, 1)$-Hölder smooth functions. Our main result finds that the performance of `UFGM` on such a sum is simply given by summing up all the individual convergence rates (1.3). Hence the method's behavior can be understood as the superposition of each summand's individual Hölder smooth setting. This is stated below and proven in Section 4.

Theorem 1.1. *For any convex $F(x) = \sum_{j \in \mathcal{J}} f_j(x)$ where each f_j is (M_j, v_j)-Hölder smooth* (1.2) *and target accuracy $\epsilon > 0$, the iterates y_k of the* `UFGM` *are ϵ-minimizers of* (1.1) *for all*

$$k \geq \sum_{j \in \mathcal{J}} \left[c_j \left(\frac{M_j}{\epsilon} \right)^{\frac{2}{1+3v_j}} \xi(x_0, x^*)^{\frac{1+v_j}{1+3v_j}} \right] . \tag{1.4}$$

where the coefficient $c_j = \frac{1+3v_j}{1+v_j} 2^{\frac{2+2v_j}{1+3v_j}} |\mathcal{J}|^{\frac{1-v_j}{1+3v_j}}$ for each $j \in \mathcal{J}$ and x^* minimizes (1.1).

Importantly, we do not modify the method at all, only providing it a subgradient oracle for the overall summation F. Consequently, we arrive at a stronger statement of `UFGM`'s universality, it adapts to sums of structured functions without knowledge of their types or even the number of summands. Note when there is more than one term in the sum, the number of summands appears in the coefficients c_j of our theorem. Section 1.2 discusses the necessity of a dependence on $|\mathcal{J}|$.

Many optimization problems take the form of minimizing a nonsmooth Lipschitz function f_1 plus a smooth function f_2 (see the example applications in Section 3). For such applications where f_1 is $(M, 0)$-Hölder smooth and f_2 is $(L, 1)$-Hölder smooth, Theorem 1.1 guarantees `UFGM` has y_k as an ϵ-minimizer of $F(x) = f_1(x) + f_2(x)$ for all

$$k \geq 8 \left(\frac{M}{\epsilon} \right)^2 \xi(x_0, x^*) + 4 \sqrt{\frac{L \xi(x_0, x^*)}{\epsilon}} . \tag{1.5}$$

Up to small constants, this convergence rate is the sum of the optimal rate for nonsmooth M-Lipschitz minimization plus the optimal rate for L-smooth minimization.

The optimality of this rate in the Euclidean setting (where $\xi(x, y) = \frac{1}{2}\|x - y\|^2$) follows immediately from the known lower bounds in each individual Hölder-smooth setting. Restricting to methods satisfying the gradient span model typical to first-order method [3] gives the following matching lower bound.

Theorem 1.2. *Consider any $\epsilon > 0$, $\{(M_j, v_j)\}_{j \in \mathcal{J}}$ and $R > 0$ and any algorithm generating iterates satisfying $x_{k+1} \in \operatorname{span}(x_0, \nabla F(x_0), \ldots \nabla F(x_k))$. There exists convex, (M_j, v_j)-Hölder smooth functions f_j and x_0 with $\frac{1}{2}\|x_0 - x^*\|^2 \leq R$ such that computing an ϵ-minimizer x_k of $F(x) = \sum f_j(x)$ requires at least*

$$k \geq \max_{j \in \mathcal{J}} \left[c'_j \left(\frac{M_j}{\epsilon} \right)^{\frac{2}{1+3v_j}} R^{\frac{1+v_j}{1+3v_j}} \right] \geq \sum_{j \in \mathcal{J}} \left[\frac{c'_j}{|\mathcal{J}|} \left(\frac{M_j}{\epsilon} \right)^{\frac{2}{1+3v_j}} R^{\frac{1+v_j}{1+3v_j}} \right]$$

gradient oracle evaluations where the universal coefficients $c'_j > 0$ depend only on v_j.

Proof. Let c_j denote the coefficient associated with each Hölder smooth lower bound in [2] and let j_* attain $\max_{j\in\mathcal{J}}\left[c_j'\left(\frac{M_j}{\epsilon}\right)^{\frac{2}{1+3v_j}}R^{\frac{1+v_j}{1+3v_j}}\right]$. Then the first inequality holds by selecting f_{j^*} as the lower bounding (M_{j^*}, v_{j^*}) Hölder smooth instance of [2] and all other $f_j = 0$, which is trivially (M_j, v_j)-Hölder smooth. The second inequality trivially bounds the maximum by an average. □

Together these two theorems provide the principle takeaway of this work:

> *Summations of known problem settings introduce no new difficulties. Instead, existing universal methods have the optimal complexity of the sum of each setting's complexity.*

Outline. In the remainder of this introduction, we discuss extensions of our main result to strongly convex problems (and more generally growth/error bounded settings), scaling with respect to $|\mathcal{J}|$, and the importance of universal, blackbox results on such heterogeneous sums. Section 2 briefly introduces Nesterov's universal fast gradient method. Then in Section 3, we discuss applications and simple numerics showing a transition from fast smooth convergence to slow nonsmooth convergence as the dominant term in (1.5) changes. Finally, Section 4 proves our main theorems.

1.1 Improved Convergence Guarantees Under Hölder Growth Bounds

Many works [4–7] have shown improved convergence guarantees whenever a growth bound

$$F(x) + \Psi(x) - p_* \geq \mu\xi(x, x^*)^{p/2} \tag{1.6}$$

holds, which we refer to as (μ, p)-Hölder growth. These conditions are closely related to the Kurdyka-Łojasiewicz condition [8], which are widespread, holding for generic subanalytic functions [9, 10] and nonsmooth subanalytic convex functions [11].

In the setting of Euclidean distances, the recent work of [6] showed for any (M, v)-Hölder smooth function with (μ, p)-Hölder growth, a restarted variant of `UFGM` finds an ϵ-minimizer within

$$\begin{cases} O\left(\frac{M^{\frac{2}{1+3v}}}{\mu^{\frac{2(1+v)}{p(1+3v)}}\epsilon^{\frac{2(p-1-v)}{p(1+3v)}}}\right) & \text{if } v < p-1 \\ O\left(\left(\frac{M}{\mu}\right)^{\frac{2}{1+3v}}\log(1/\epsilon)\right) & \text{if } v = p-1 \end{cases} \tag{1.7}$$

iterations. Our analysis directly extends this showing a convergence rate for a sum of (M_j, v_j)-Hölder smooth functions, which satisfies (μ, p)-Hölder growth, equal to the sum of the individual M_j, v_j, μ, p rates of (1.7). As our focus is not on the details of restarting schemes, we analyze a simple restarted method (`R-UFGM` defined in Algorithm 2) which assumes knowledge of the optimal objective value.

Theorem 1.3. *For any convex $F(x) = \sum_{j\in\mathcal{J}} f_j(x)$ satisfying (μ, p)-Hölder growth* (1.6) *where each f_j is (M_j, v_j)-Hölder smooth* (1.2) *and target accuracy $\tilde{\epsilon} > 0$,* `R-UFGM` *finds an $\tilde{\epsilon}$-minimizers of* (1.1) *after at most*

$$\sum_{j\in\mathcal{J}}\left[\left(c_j''\min\left\{\frac{2^{\frac{2(p-1-v_j)}{p(1+3v_j)}}}{2^{\frac{2(p-1-v_j)}{p(1+3v_j)}}-1}, \frac{N}{2^{\frac{2(p-1-v_j)}{p(1+3v_j)}}}\right\}\right)\frac{M_j^{\frac{2}{1+3v_j}}}{\mu^{\frac{2(1+v_j)}{p(1+3v_j)}}\tilde{\epsilon}^{\frac{2(p-1-v_j)}{p(1+3v_j)}}}\right] + N \tag{1.8}$$

iterations where $N = \lceil\log_2(F(z_0) + \Psi(z_0) - p_*)/\tilde{\epsilon})\rceil$ and $c_j'' = \frac{1+3v_j}{1+v_j}2^{\frac{(v_j-1)(p-2)}{p(1+3v_j)}}|\mathcal{J}|^{\frac{1-v_j}{1+3v_j}}$.

1.2 Improved Convergence Guarantees with respect to $|\mathcal{J}|$

For any fixed number of summands, Theorems 1.1 and 1.2 agree up to their constant coefficients c_j and $c_j'/|\mathcal{J}|$. Hence the fast universal method optimally adapts to any fixed sum structure. However, the dependence on the number of summands can be improved as its power does not agree between our upper and lower bounds. We conjecture the following optimal dependence on $|\mathcal{J}|$.

Conjecture 1.1. *The optimal first-order oracle complexity for minimizing a convex sum $\sum_{j\in\mathcal{J}} f_j(x)$ of (M_j, v_j)-Hölder smooth functions to a target accuracy $\epsilon > 0$ given $\frac{1}{2}\|x_0 - x^*\|^2 \leq R$ is*

$$\sum_{j\in\mathcal{J}} \left[\bar{c}_j \left(\frac{M_j}{\epsilon}\right)^{\frac{2}{1+3v_j}} R^{\frac{1+v_j}{1+3v_j}}\right]$$

where $\bar{c}_j = \Theta(|\mathcal{J}|^{\frac{1-3v_j}{1+3v_j}})$, depending only on universal constants and v_j.

To motivate this conjecture and the necessity of a dependence on $|\mathcal{J}|$ consider the following setting: Given a (M, v)-Hölder-smooth function f to minimize, the optimal convergence rate is given by (1.3). For any $|\mathcal{J}|$, we can write this problem as minimizing a sum of $(M/|\mathcal{J}|, v)$-Hölder smooth functions given by $\sum_{j\in\mathcal{J}}[f(x)/|\mathcal{J}|]$. The optimal convergence rate for each summand here is on the order of $(M/|\mathcal{J}|\epsilon)^{2/(1+3v)})\xi(x_0, x^*)^{(1+v)/(1+3v)}$. Then for the sum these individual rates to match the optimal convergence guarantee without rewriting f as a sum, the coefficient $\bar{c}$ when summing the individual weights must depend on $|\mathcal{J}|$ as

$$\sum_{j\in\mathcal{J}} \left[\bar{c}\left(\frac{M}{|\mathcal{J}|\epsilon}\right)^{\frac{2}{1+3v}} \xi(x_0, x^*)^{\frac{1+v}{1+3v}}\right] = \Theta\left(\left(\frac{M}{\epsilon}\right)^{\frac{2}{1+3v}} \xi(x_0, x^*)^{\frac{1+v}{1+3v}}\right) \iff \bar{c} = \Theta\left(|\mathcal{J}|^{\frac{1-3v}{1+3v}}\right) .$$

As a direction toward tightening this gap in our theory, we derive the following implicitly defined convergence guarantee for `UFGM` in Section 4.2.

Theorem 1.4. *For any convex $F(x) = \sum_{j\in\mathcal{J}} f_j(x)$ where each f_j is (M_j, v_j)-Hölder smooth (1.2) and target accuracy $\epsilon > 0$, the iterates y_k of the `UFGM` are ϵ-minimizers of (1.1) for all $k \geq 5K$ where K is the unique positive solution to*

$$\sum_{j\in\mathcal{J}} \frac{|\mathcal{J}|^{\frac{1-v_j}{1+v_j}} M_j^{\frac{2}{1+v_j}} \xi(x_0, x^*)}{\epsilon^{\frac{2}{1+v_j}}} K^{\frac{-(1+3v_j)}{1+v_j}} = 1 .$$

A short calculation[1] shows this maintains the optimal guarantee for the motivating example above as $|\mathcal{J}|$ grows (unlike Theorem 1.1).

[1]Namely, supposing all $(M_j, v_j) = (M/|\mathcal{J}|, v)$, the defining equation for K simplifies to

$$|\mathcal{J}| \left(\frac{|\mathcal{J}|^{\frac{1-v}{1+v}} (M/|\mathcal{J}|)^{\frac{2}{1+v}} \xi(x_0, x^*)}{\epsilon^{\frac{2}{1+v}}} K^{\frac{-1-3v}{1+v}}\right) = \frac{M^{\frac{2}{1+v}} \xi(x_0, x^*)}{\epsilon^{\frac{2}{1+v}}} K^{\frac{-1-3v}{1+v}} = 1 ,$$

solved by $K = \Theta\left(\left(\frac{M}{\epsilon}\right)^{\frac{2}{1+3v}} \xi(x_0, x^*)^{\frac{1+v}{1+3v}}\right)$, matching the optimal rate for (M, v)-Hölder smooth minimization.

1.3 Related Works

Importance of Universal, Blackbox Guarantees An algorithm is universal if it applies across a range of different problem structures (e.g., different levels of Hölder-smoothness or the existence of different growth conditions). The universal fast gradient method of Nesterov is one such algorithm, applying to a generic Hölder smooth objective $f + \Psi$, only needing access to function and first-order evaluations $(f(x_k), \nabla f(x_k))$ and a target accuracy $\epsilon > 0$. A few varied examples of other universal optimization methods and analysis: bundle methods [12, 13], solving stochastic variational inequalities [14], and Newton's method [15].

Typically universal methods are adaptive or blackbox, meaning they do not require the input of constants related to whatever problem structures exist. Adaptivity is of real practical importance, classically motivating in linesearching and trust-region methodologies [16]. The restarting schemes [6, 7] adapt to whatever Hölder growth exists. Nesterov's universal methods adapt to whatever Hölder smoothness exists, learning an inexact smoothness constant L_k over time. Moreover, from our analysis, it adapts to sums of Hölder smooth terms without knowledge of how $F(x) = \sum_{j \in \mathcal{J}} f_j(x)$ can be written as a sum (i.e., the number of terms and Hölder smoothness of each) is used.

As a benefit of this adaptivity, if many such formulations exist, the universal method will converge as fast as the infimum of (1.4) or (1.8) over all such formulations. For example, this offers the following improvement for minimizing a μ-strongly convex function f over a compact domain D. Let $M(f, D)$ denote the Lipschitz constant of f on D. Then the classic convergence guarantee gives a rate of

$$O\left(\frac{M(f,D)^2}{\mu\epsilon}\right) .$$

However, we can write f as the sum of two convex functions $(f(x) - \frac{\mu}{2}\|x\|^2) + \frac{\mu}{2}\|x\|^2$. Then Theorem 1.3 gives the following rate with potentially much smaller Lipschitz constant

$$O\left(\frac{M(f - \frac{\mu}{2}\|\cdot\|^2, D)^2}{\mu\epsilon} + \log\left(\frac{f(x_0) - \inf_{x \in D} f(x)}{\epsilon}\right)\right) .$$

Importance of Heterogeneous Summations Minimizing finite sums have attracted substantial interest, typically assuming a common structure among all the summands. Variance reduced approaches give a tractable stochastic approach when the number of terms is large. However, such methods do not fit within our blackbox model as they rely on knowing the structure of F. Arjevani et al. [17] provide a customized approach to minimizing sums without using indexing information.

The recent work of Wang and Zhang [18] is motivated similarly to us. They consider minimizing heterogeneous sums of L_i-smooth μ_i-strongly convex functions given a gradient oracle for individual terms. They show a method using variance reduction attains the optimal rate of $O\left(m + \frac{\sum_i \sqrt{L_i}}{\sqrt{\sum_i \mu_i}} \log(1/\epsilon)\right)$ individual function evaluations. This is the same rate with respect to L_i and μ_i given by applying our Theorem 1.3 as such a sum has $(\sum \mu_i, 2)$-Hölder growth[2].

Several past works have considered sums of smooth and nonsmooth but Lipschitz terms. The optimal rate (1.5) in this setting were shown by [19–21] for several dual averaging methods. A normalized subgradient method was analyzed in [22] that converges at the sum of (sub)gradient descent's suboptimal convergence rates. In part, this work aims to follow up on these ideas.

[2]Note this comparison is only superficial as we assume a more expensive oracle giving (sub)gradients of the whole objective. Given our oracle, considering all the terms as one $L = \sum L_i$-smooth term gives a faster $O\left(\frac{\sqrt{\sum_i L_i}}{\sqrt{\sum_i \mu_i}} \log(1/\epsilon)\right)$.

Section 3 discusses two applications where heterogeneous sums naturally occur (maximum likelihood estimation over heterogeneous data sources and support vector machine training). Since our guarantees are universal and optimal, any further improvements on these problems would require customized algorithms with knowledge of the particular problem structure.

2 Preliminaries and the Universal Fast Gradient Method

Notationally, we closely follow [1] to ease the development of our analysis. We assume access to a first-order oracle for the summation $F(x) = \sum_{j\in\mathcal{J}} f_j(x)\colon \mathbb{R}^d \to \mathbb{R}$: for any $x \in Q$, we will utilize

$$x \mapsto (F(x), \nabla F(x))$$

where $\nabla F(x)$ is a subgradient of F at x. (Note a subgradient rather than gradient oracle is needed here since $(M, 0)$-Hölder smoothness only corresponds to Lipschitz continuity of the objective function, rather than some continuity of the gradient.) By the sum rule of subgradient calculus, this could be implemented as a sum over each summand $(F(x), \nabla F(x)) = \sum_{j\in\mathcal{J}}(f_j(x), \nabla f_j(x))$.

For any convex $d(x)$ satisfying the following strong convexity condition (with parameter one)

$$d(y) \geq d(x) + \langle \nabla d(x), y - x\rangle + \frac{1}{2}\|y - x\|^2, \qquad x, y \in \text{rint } Q \ ,$$

we consider the associated *Bregman distance* (or divergence)

$$\xi(x, y) = d(y) - (d(x) + \langle \nabla d(x), y - x\rangle) \ . \tag{2.1}$$

When $d(x) = \frac{1}{2}\|x\|^2$, the Bregman divergence recovers the classic Euclidean distance $\frac{1}{2}\|y - x\|^2$.

For any Bregman distance, we denote the *Bregman mapping* given functions F, Ψ and a set Q as

$$\mathcal{B}_M(x) = \operatorname*{argmin}_{y\in Q} \{\Psi_M(x, y) := F(x) + \langle \nabla F(x), y - x\rangle + M\xi(x, y) + \Phi(y)\} \ .$$

We assume this operation can be computed either in closed form or by some efficient subroutine. This amounts to requiring the constraints Q and generic function Φ are sufficiently simple. Based on this operation, given any target accuracy $\epsilon > 0$, the Universal Fast Gradient Method iterates as defined in Algorithm 1.

The key lemma behind Nesterov [1]'s analysis of universal methods across these different Hölder-smooth settings is the following unifying condition. In essence, for any level of smoothness, the gradient of f can be viewed as an inexact oracle for the case of smooth ($v = 1$) optimization, where gradients provide quadratic upper bounds.

Lemma 2.1 (Nesterov [1], Lemma 2)**.** *Suppose F is (M, v)-Hölder smooth* (1.2)*. Then for any $\delta > 0$ and*

$$L \geq \left[\frac{1 - v}{1 + v} \cdot \frac{1}{\delta}\right]^{\frac{1-v}{1+v}} M^{\frac{2}{1+v}} \ ,$$

we have

$$F(y) \leq F(x) + \langle \nabla F(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 + \frac{\delta}{2}, \quad x, y \in Q \ .$$

This lemma ensures that for large enough i_k the condition (2.2) is satisfied. In Section 4, we generalize this result to similarly apply to sums of Hölder smooth functions (see Lemma 4.1). As a result, the iterates of `UFGM` are well-defined in our more general setting of interest.

On average (amortized), `UFGM` uses at most four first-order oracle evaluations $(F(x_k), \nabla F(x_k))$ per iteration. Stopping criteria and the details of this amortized analysis are given in Nesterov's original development [1] and we refer any interested reader there.

Algorithm 1 Universal Fast Gradient Method (`UFGM`) of Nesterov [1]

1: **Initialization**: Choose $x_0 \in Q$, $\epsilon > 0$, $L_0 > 0$. Define $y_0 = x_0$, $A_0 = 0$, $\phi_0(x) = \xi(x_0, x)$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Find $v_k = \operatorname{argmin}_{x \in Q} \phi_k(x)$.
4: Find the smallest integer $i_k \geq 0$ such that the definitions

$$\begin{aligned} a_{k+1,i_k}^2 &= \frac{1}{2^{i_k} L_k}(A_k + a_{k+1,i_k}),\ a_{k+1,i_k} > 0 \\ A_{k+1,i_k} &= A_k + a_{k+1,i_k} \\ \tau_{k,i_k} &= \frac{a_{k+1,i_k}}{A_{k+1,i_k}} \\ x_{k+1,i_k} &= \tau_{k,i_k} v_k + (1 - \tau_{k,i_k}) y_k \\ \hat{x}_{k+1,i_k} &= \operatorname*{argmin}_{y \in Q} \{\xi(v_k, y) + a_{k+1,i_k}[\langle \nabla F(x_{k+1,i_k}), y\rangle + \Psi(y)]\} \\ y_{k+1,i_k} &= \tau_{k,i_k} \hat{x}_{k+1,i_k} + (1 - \tau_{k,i_k}) y_k \end{aligned}$$

satisfy

$$\begin{aligned} F(y_{k+1,i_k}) \leq F(x_{k+1,i_k}) &+ \langle \nabla F(x_{k+1,i_k}), y_{k+1,i_k} - x_{k+1,i_k}\rangle \\ &+ 2^{i_k - 1} L_k \|y_{k+1,i_k} - x_{k+1,i_k}\|^2 + \frac{\epsilon \tau_{k,i_k}}{2}. \end{aligned} \tag{2.2}$$

5: Set $x_{k+1} = x_{k+1,i_k}$, $y_{k+1} = y_{k+1,i_k}$, $a_{k+1} = a_{k+1,i_k}$, $\tau_k = \tau_{k,i_k}$ and define

$$\begin{aligned} A_{k+1} &= A_k + a_{k+1} \\ L_{k+1} &= 2^{i_k - 1} L_k \\ \phi_{k+1}(x) &= \phi_k(x) + a_{k+1}[F(x_{k+1}) + \langle \nabla F(x_{k+1}), x - x_{k+1}\rangle + \Psi(x)] \end{aligned}$$

6: **end for**

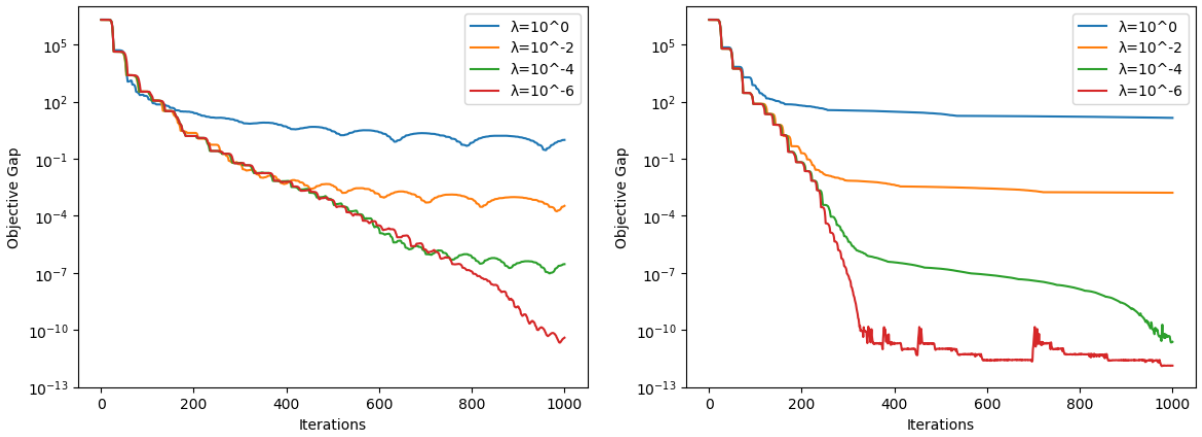
3 Motivating Applications and Numerics

Here we consider two applications with heterogeneous sums for the objective function. Simple numerics are conducted showing the universal fast gradient method converges in much the same fashion as our theory predicts.

Mixtures of Maximum Likelihoods Models Given observed features $A \in \mathbb{R}^{n \times d}$ and labels $b \in \mathbb{R}^n$, ℓ_p regression fits a model x by computing the maximum log-likelihood estimator via

$$\min_{x \in Q} \|Ax - b\|_p^p$$

where $Q \subseteq \mathbb{R}^d$ reflects prior knowledge or regularization (for example, imposing nonnegativity $x \geq 0$ or seeking sparsity $\|x\|_1 \leq \delta$). When $p = 2$, this corresponds to the measurements with Gaussian noise. More generally, $p = 1$ corresponds to Laplacian noise, and all other values of p correspond to a generalized Gaussian distribution. When $p < 2$, these models allow for heavier tails distributions. The improved performance of estimators outside of Gaussian settings $p \neq 2$ is well-documented



(a) `UFGM` (b) `R-UFGM`

Figure 1: Applying `UFGM` and `R-UFGM` to $F(x) = \frac{1}{2}\|A_1x - b_1\|_2^2 + \lambda\|A_2x - b_2\|_1$ with $x_0 = 0, L_0 = 1, \epsilon = 10^{-9}$, standard normal $A_i \in \mathbb{R}^{2000\times1000}, x^* \in \mathbb{R}^{1000}$, and $b_i = A_ix^*$ for various values of $\lambda > 0$. Note absolutely no parameter tuning was done as the method is universal.

(see [23–26]), although this requires careful analysis of the (potentially heavy-tailed) errors to identify an suitable value of p.

Suppose data sets (A_j, b_j) from independent sources $j \in \{1 \dots J\}$ are aggregated, each with its own, different source of errors with log-likelihoods proportional to $-\|A_jx - b_j\|_{p_j}^{p_j}$ for some $p_j \in [1, 2]$. Then the maximum likelihood estimator given all the J data sets is given by

$$\min_{x\in Q} F(x) = \sum_{j=1}^{J} \|A_jx - b_j\|_{p_j}^{p_j} .$$

Observing that each $\|A_jx - b_j\|_{p_j}^{p_j}$ is Hölder smooth with exponent $p_j - 1$, we conclude `UFGM` can compute the maximum likelihood estimator for aggregated independent data sets in time proportional to that of computing separate maximum likelihood estimators for each data set.

Figure 1 applies `UFGM` and `R-UFGM` to $F(x) = \frac{1}{2}\|A_1x - b_1\|_2^2 + \lambda\|A_2x - b_2\|_1$ with varied $\lambda > 0$. In all cases, we see fast convergence early on as $\epsilon \approx F(x) - p_*$ is large and so the smooth convergence rate dominates the nonsmooth component $O(1/\sqrt{\epsilon}) >> O(\lambda^2/\epsilon^2)$. For each problem instance, the method suddenly slows down once the nonsmooth rate dominates (around height $\epsilon \approx \lambda^{4/3}$). Noting $F(\cdot)$ here is $\lambda_{min}(A_1^TA_1)$-strongly convex, we see the speedup from restarting predicted by Theorem 1.3.

Support Vector Machines Consider the following unconstrained formulation of training a support vector machine (SVM) given n data points $(x_i, y_i) \in \mathbb{R}^d \times \{\pm1\}$ and a scalar $\lambda > 0$

$$\min_{w\in\mathbb{R}^d} F(w) = \sum_{i=1}^{n} \max\{0, 1 - y_i \cdot x_i^T w\} + \frac{\lambda}{2}\|w\|_2^2.$$

Note that this objective F is λ-strongly convex and the sum of an M-Lipschitz, nonsmooth function and a function with λ-Lipschitz gradient. Alas, their sum is neither smooth nor Lipschitz globally. Previous works have overcome this limitation in several customized ways. Deriving some bound on the iterates is often possible and then a Lipschitz constant from that compact region can be

used [27]. Alternatively, the analysis of [22, Corollary 12] showed a normalized subgradient method converges at a rate of $O(M^2/\lambda\epsilon + \sqrt{\lambda\|x_0 - x^*\|^2/\epsilon})$.

Our theory improves upon these results. Theorem 1.3 ensures R-UFGM finds an ϵ-minimizer within

$$\frac{32M^2}{\lambda\epsilon} + 4\log\left(\frac{F(w_0) - \inf F}{\epsilon}\right)$$

iterations. The dual averaging methods of [20] also attain this fast rate (although their methods need to know several problem constants to be applied). Such fast algorithms are very resilient to the choice of the initial w_0 anywhere in $\mathbb{R}^d$ as it only appears in the logarithmic term. In contrast, the works mentioned in the paragraph above require w_0 to be limited to a bounding ball or have the dependence w_0 decay at a sublinear rate.

4 Convergence Theory

Our analysis of UFGM closely follows the form of Nesterov's original analysis when considering the minimization of a single Hölder smooth function. The primary difference in deriving our more general convergence rates comes from how we estimate the coefficients A_k. A simple extension of Lemma 2.1 is given below, showing subgradient evaluations of the sum $F(x) = \sum_{j\in\mathcal{J}} f_j(x)$ can be viewed as inexact gradient evaluations yielding quadratic upper bounds with constant depending on a combination of the Hölder-smoothness of each f_j.

Lemma 4.1. *Suppose $F(x) = \sum_{j\in\mathcal{J}} f_j(x)$ where each f_j is (M_j, v_j)-Hölder smooth (1.2). Then for any $\delta > 0$ and*

$$L \geq \sum_{j\in\mathcal{J}} \left[\left[\frac{1-v_j}{1+v_j}\cdot\frac{|\mathcal{J}|}{\delta}\right]^{\frac{1-v_j}{1+v_j}} M_j^{\frac{2}{1+v_j}}\right],$$

we have

$$F(y) \leq F(x) + \langle\nabla F(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 + \frac{\delta}{2}, \quad x, y \in Q\ .$$

Proof. For each (M_j, v_j)-Hölder smooth function f_j, $\delta > 0$ and $L_j \geq \left[\frac{1-v}{1+v}\cdot\frac{|\mathcal{J}|}{\delta}\right]^{\frac{1-v}{1+v}} M^{\frac{2}{1+v}}$, Lemma 2.1 ensures

$$f_j(y) \leq f_j(x) + \langle\nabla f_j(x), y - x\rangle + \frac{L_j}{2}\|y - x\|^2 + \frac{\delta}{2|\mathcal{J}|}, \quad x, y \in Q\ .$$

Note that the sum rule allows us to decompose the subgradient $\nabla F(x) = \sum_{j\in\mathcal{J}} \nabla f_j(x)$ as a sum of subgradients of each of its component. Then summing this over all j gives the claim. □

This lemma ensures the backtracking search in line 4 of Algorithm 1 always terminates and will further allow us to bound the rate A_k grows. Equip with this inexact oracle result, deriving convergence guarantees for the Nesterov's universal fast gradient method follows nearly from the proof of Theorem 3 in [1], complicated by a more difficult recurrence relation arising at the end of the argument. We show that the convergence of UFGM on sums of Hölder smooth functions is controlled by the following recurrence.

Theorem 4.1. *For any convex $F(x) = \sum_{j\in\mathcal{J}} f_j(x)$ where each f_j is (M_j, v_j)-Hölder smooth, all of the iterations of* UFGM *are well-defined. Moreover, for any $k \geq 0$,*

$$A_k\left(F(y_k) + \Psi(y_k) - \frac{\epsilon}{2}\right) \leq \phi_k^* = \min_{x\in Q}\phi_k(x) \tag{4.1}$$

where A_k increases monotonically, satisfying the recurrence relation

$$\sum_{j\in\mathcal{J}}\left[\frac{2|\mathcal{J}|^{\frac{1-v_j}{1+v_j}}M_j^{\frac{2}{1+v_j}}}{\epsilon^{\frac{1-v_j}{1+v_j}}}\frac{(A_{k+1}-A_k)^{\frac{1+3v_j}{1+v_j}}}{A_{k+1}^{\frac{2v_j}{1+v_j}}}\right]\geq 1\ . \tag{4.2}$$

Proof. Our extended Lemma 4.1 establishes the iterates of `UFGM` are well-defined. Then the guarantee (4.1) exactly follows its derivation in [1, Theorem 3]. Furthermore, Lemma 4.1 implies that line 3 of the `UFGM` has

$$2^{i_k}L_k\leq\sum_{j\in\mathcal{J}}\left[2\left(\frac{|\mathcal{J}|A_{k+1}}{\epsilon a_{k+1}}\right)^{\frac{1-v_j}{1+v_j}}M_j^{\frac{2}{1+v_j}}\right].$$

Observing that $a_{k+1}^2/A_{k+1}=1/2^{i_k}L_k$ and $a_{k+1}=A_{k+1}-A_k$ yields the implicit recurrence relation

$$\sum_{j\in\mathcal{J}}\left[\frac{2|\mathcal{J}|^{\frac{1-v_j}{1+v_j}}M_j^{\frac{2}{1+v_j}}}{\epsilon^{\frac{1-v_j}{1+v_j}}}\frac{(A_{k+1}-A_k)^{\frac{1+3v_j}{1+v_j}}}{A_{k+1}^{\frac{2v_j}{1+v_j}}}\right]\geq 1\ . \qquad \square$$

From this theorem, the primary difficulty in bounding the convergence of the fast universal method applied to a sum (and thus the primary difficulty in proving Theorems 1.1 and 1.4) is then in solving this recurrence relation. To do so, we prove two bounds on any sequence satisfying the recurrence (4.2) in the following two subsections. The first bound proven in Lemma 4.2 gives an explicit bound on k in terms of A_k, which when combined with (4.1) gives the explicit bound on the accuracy of `UFGM` of Theorem 1.1. Our second bound proven in Lemma 4.3 gives an implicit bound for A_k based on the solution of a related nonlinear equation. In turn, this yields the improved (although implicit) guarantee of Theorem 1.4.

4.1 Proof of Explicit Convergence Guarantee (Theorem 1.1)

From Theorem 4.1, it follows that every y_k will be an ϵ-minimizer of F whenever $A_k\geq 2\xi(x_0,x^*)/\epsilon$. To ensure A_k reaches this needed size, below we show any recurrence satisfying (4.2) has k lower bound a certain summation of powers of A_k.

Lemma 4.2. *Suppose a nonnegative, increasing sequence A_k satisfies*

$$\sum_{j\in\mathcal{J}}\left[\alpha_j\frac{(A_{k+1}-A_k)^{1+q_j}}{A_{k+1}^{q_j}}\right]\geq 1$$

where $\alpha_j>0$ and $q_j\in[0,1]$ are generic constants for each $j\in\mathcal{J}$. Then for any $k\geq 0$,

$$k\leq\sum_{j\in\mathcal{J}}\left[(1+q_j)(\alpha_j A_k)^{\frac{1}{1+q_j}}\right]\ .$$

Proof. Trivially this holds for $k=0$. Inductively suppose the claimed lower bound holds for some k. First suppose some single summand $j'\in\mathcal{J}$ has $\alpha_{j'}\frac{(A_{k+1}-A_k)^{1+q_{j'}}}{A_{k+1}^{q_{j'}}}\geq 1$. Then combining this bound

with the inductive hypothesis gives our inductive step at $k+1$ as

$$\begin{aligned}
k+1 &\leq \sum_{j\in\mathcal{J}} \left[(1+q_j)(\alpha_j A_k)^{\frac{1}{1+q_j}}\right] + \left(\alpha_{j'}\frac{(A_{k+1}-A_k)^{1+q_{j'}}}{A_{k+1}^{q_{j'}}}\right)^{\frac{1}{1+q_{j'}}} \\
&\leq \sum_{j\in\mathcal{J}\setminus\{j'\}} \left[(1+q_j)(\alpha_j A_{k+1})^{\frac{1}{1+q_j}}\right] + (1+q_{j'})(\alpha_{j'}A_k)^{\frac{1}{1+q_{j'}}} + \alpha_{j'}^{\frac{1}{1+q_{j'}}}\frac{A_{k+1}-A_k}{A_{k+1}^{\frac{q_{j'}}{1+q_{j'}}}} \\
&\leq \sum_{j\in\mathcal{J}\setminus\{j'\}} \left[(1+q_j)(\alpha_j A_{k+1})^{\frac{1}{1+q_j}}\right] + (1+q_{j'})(\alpha_{j'}A_{k+1})^{\frac{1}{1+q_{j'}}}
\end{aligned}$$

where the first inequality uses the assumption on j' (or rather the $(1+q_{j'})$th root of it), the second uses the monotonicity of A_k on each $j \neq j'$ term, the third uses the concavity of $z^{1/(1+q_{j'})}$ between A_k and A_{k+1}.

Now suppose instead that every $j \in \mathcal{J}$ has $\alpha_j \frac{(A_{k+1}-A_k)^{1+q_j}}{A_{k+1}^{q_j}} < 1$. Then combining our inductive hypothesis with the given recurrence relation gives our inductive step at $k+1$ as

$$\begin{aligned}
k+1 &\leq \sum_{j\in\mathcal{J}} \left[(1+q_j)(\alpha_j A_k)^{\frac{1}{1+q_j}} + \alpha_j\frac{(A_{k+1}-A_k)^{1+q_j}}{A_{k+1}^{q_j}}\right] \\
&\leq \sum_{j\in\mathcal{J}} \left[(1+q_j)(\alpha_j A_k)^{\frac{1}{1+q_j}} + \alpha_j^{\frac{1}{1+q_j}}\frac{A_{k+1}-A_k}{A_{k+1}^{\frac{q_j}{1+q_j}}}\right] \\
&\leq \sum_{j\in\mathcal{J}} \left[(1+q_j)(\alpha_j A_{k+1})^{\frac{1}{1+q_j}}\right]
\end{aligned}$$

where the second inequality uses our assumed bound on every j (and that $z^{1/(1+q_j)} \geq z$ for all $j \geq 0$ and $z \in [0,1]$) and the third uses the concavity of $z^{1/(1+q_j)}$ between A_k and A_{k+1}. □

Lemma 4.2 suffices to complete our proof of Theorem 1.1 as it follows from our recurrence that

$$k \leq \sum_{j\mathcal{J}} \left\lceil \frac{1+3v_j}{1+v_j}\left(\frac{2|\mathcal{J}|^{\frac{1-v_j}{1+v_j}} M_j^{\frac{2}{1+v_j}}}{\epsilon^{\frac{1-v_j}{1+v_j}}} A_k\right)^{\frac{1+v_j}{1+3v_j}}\right\rceil .$$

Since (4.1) ensures all y_k are ϵ-minimizers once $A_k \geq 2\xi(x_0,x^*)/\epsilon$, such a minimizer is found once

$$k \geq \sum_{j\mathcal{J}} \left\lceil \left(\frac{1+3v_j}{1+v_j} 2^{\frac{2+2v_j}{1+3v_j}} |\mathcal{J}|^{\frac{1-v_j}{1+3v_j}}\right)\left(\frac{M_j}{\epsilon}\right)^{\frac{2}{1+3v_j}} \xi(x_0,x^*)^{\frac{1+v_j}{1+3v_j}}\right\rceil .$$

4.2 Proof of Improved Implicit Convergence Guarantee (Theorem 1.4)

Here we improve on the convergence guarantee of Theorem 1.1 by providing a tighter analysis of the recurrence relation (4.2) than Lemma 4.2 provides in the following Lemma 4.3. By applying this lemma in the place of Lemma 4.2, Theorem 1.4 immediately follows.

Lemma 4.3. *Suppose a nonnegative, increasing sequence A_k satisfies*

$$\sum_{j\in\mathcal{J}}\left[\alpha_j\frac{(A_{k+1}-A_k)^{1+q_j}}{A_{k+1}^{q_j}}\right]\geq 1$$

where $\alpha_j>0$ and $q_j\in[0,1]$ are generic constants for each $j\in\mathcal{J}$. Then for any $\Delta>0$, $A_k\geq\Delta$ for all $k\geq 5C$ where C is the unique positive root of the equation

$$\sum_{j\in\mathcal{J}}\alpha_j\Delta C^{-(1+q_j)}-1=0\ .$$

Proof. First, observe that since $A_0\geq 0$, it follows that $A_1\geq 1/\sum\alpha_j$. Hence without loss of generality, we can assume $\Delta>1/\sum\alpha_j$ (as the result is immediate otherwise since $A_k\geq A_1>\Delta$).

For any $\delta\geq 1/\sum\alpha_j$, let $C(\delta)>1$ denote the unique positive root of

$$\sum_{j\in\mathcal{J}}\alpha_j\delta C^{-(1+q_j)}-1=0\ .$$

Uniqueness of the positive solution $C(\delta)$ follows from the fact that the function $C\mapsto\sum\alpha_j\delta C^{-1-q_j}-1$ is strictly decreasing for $C>0$ and approaches -1 as $C\to\infty$. Existence of a solution $C(\delta)>1$ follows as this function equals $\sum\alpha_j\delta-1>0$ at $C=1$. As a final useful property of $C(\delta)$, observe that for any $\lambda\leq 1$,

$$\lambda C(\delta)\leq C(\lambda\delta)\leq\sqrt{\lambda}C(\delta)\ . \tag{4.3}$$

Let $N:=\lceil\log_2(\Delta/A_1)\rceil$, which ensures $2^{N-1}A_1\leq\Delta\leq 2^NA_1$. Then we prove the lemma by showing that for any $n\in\{0,\dots,N-1\}$, at most

$$\log(2)C(2^{n+1}A_1) \tag{4.4}$$

many different values of k have A_k in the interval $[2^n,2^{n+1}]/\sum\alpha_j$. Summing this up from $n=0$ to $N-1$ gives the claim the number of steps before $A_k\geq 2^NA_1\geq\Delta$ is at most

$$\sum_{n=0}^{N-1}\log(2)C(2^{n+1}A_1)\leq\sum_{n=0}^{N-1}\log(2)C(2^NA_1)\sqrt{2}^{n+1-N}\leq\frac{\sqrt{2}\log(2)C(2^NA_1)}{(\sqrt{2}-1)}<5C(\Delta)$$

where the first inequality uses (4.3), the second upper bounds this sum by a geometric series, and the third again uses (4.3) (and then rounds the coefficient $2\sqrt{2}\log(2)/(\sqrt{2}-1)$ up to 5 for simplicity).

Now we complete the proof by showing the claimed bound (4.4) for any $n\geq 1$. Note for any $A_{k+1}\leq 2^{n+1}A_1$, the given recurrence relation implies that

$$\sum_{j\in\mathcal{J}}\left[\alpha_j2^{n+1}A_1\left(\frac{A_{k+1}-A_k}{A_{k+1}}\right)^{1+q_j}\right]\geq 1\ .$$

Hence $(A_{k+1}-A_k)/A_{k+1}$ must be at least $C(2^{n+1}A_1)^{-1}$, or equivalently,

$$A_{k+1}\geq\frac{C(2^{n+1}A_1)}{C(2^{n+1}A_1)-1}A_k\ .$$

Thus the value of A_k grows geometrically within the interval $[2^n,2^{n+1}]/\sum\alpha_j$. As a result, the number of different k with A_k in this interval is at most

$$\log_{C(2^{n+1}A_1)/(C(2^{n+1}A_1)-1)}(2)=\frac{\log(2)}{\log(C(2^{n+1}A_1)/(C(2^{n+1}A_1)-1))}\leq\log(2)C(2^{n+1}A_1))\ .$$

where the inequality utilizes the upper bound $x\geq 1/\log(x/(x-1))$ for all $x\geq 1$. □

4.3 Improved Theory Under Hölder Growth Bounds

For many first-order methods in a range of different settings, faster convergence guarantees are well-known under strong convexity of $F+\Psi$ or more generally under some growth/error bound or KL condition. The most common such setting is μ-strongly convex optimization, which possesses $(\mu, 2)$-Hölder growth. Recall for (M, v)-Hölder smooth optimization satisfying (μ, p)-Hölder growth (1.6), the optimal[3] rate is $O\left(\frac{M^{\frac{2}{1+3v}}}{\mu^{\frac{2(1+v)}{p(1+3v)}}\epsilon^{\frac{2(p-1-v)}{p(1+3v)}}}\right)$ if $v < p-1$ and $O((M/\mu)^{\frac{2}{1+3v}}\log(1/\epsilon))$ if $v = p-1$.

Here we generalize this, showing that whenever a sum of different Hölder smooth terms satisfies (μ, p)-Hölder growth, a variant of the universal fast gradient method produces an ϵ-minimizer after at most the sum of the above Hölder smooth, Hölder growth iteration bounds. For our algorithmic variant, we will assume that the optimal value of (1.1) is known for ease of development. The works [6, 7] have provided more sophisticated restarting approaches, avoiding such assumptions, often at the cost of a logarithmic term in the oracle complexity. Utilizing these schemes would give a wholly parameter-free approach but is beyond the scope of this work.

We consider the simple restarted variant of the `UFGM`, dubbed `R-UFGM`, in Algorithm 2. Such

Algorithm 2 Restarted Universal Fast Gradient Method (`R-UFGM`)

1: **Initialization**: Choose $z_0 \in Q$, $L_0 > 0$. Define $\epsilon_0 = (F(z_0) + \Psi(z_0) - p_*)/2$.
2: **for** $n = 0, 1, 2, \ldots$ **do**
3: Let $y_k^{(n)}$ denote the iterates of `UFGM`(z_n, ϵ_n, L_0) run until $y_k^{(n)}$ is an ϵ_n-minimizer for some k_n.
4: Set $z_{n+1} = y_{k_n,n}$ and $\epsilon_{n+1} = \epsilon_n/2$.
5: **end for**

restarting is well known to speed up the convergence of many first-order methods in the presence of growth bounds. Indeed Theorem 1.3 has the convergence rate with respect to each summand with $v_j + 1 \leq p$ improve to the optimal convergence rate (1.7) for (M_j, v_j)-Hölder smooth, (μ, p)-Hölder growth optimization. Each summand in our convergence rate with $v_j + 1 > p$ decreases superlinearly. These rapidly decaying terms cannot be compared to an isolated setting as no function can have (M, v)-Hölder smoothness and (μ, p)-Hölder growth with $v + 1 > p$, even on a compact domain.

4.3.1 Proof of Improved Hölder Growth Convergence Guarantee (Theorem 1.3)

Observe that the stopping criteria for each run of `UFGM` ensures that

$$F(z_n) + \Psi(z_n) - p_* \leq 2^{-n}(F(z_0) + \Psi(z_0) - p_*) = 2\epsilon_n.$$

It follows that after N restarts, an $\tilde{\epsilon}$-minimizer has been found. First, we bound the number of iterations needed to produce to reach the stopping criteria for each iteration n of `R-UFGM`. We bound

[3] The original source for such a lower bound is difficult to identify in the literature, leaving this optimality somewhat as folklore. The theory of [28] allows these bounds to be concluded from the classic lower bounds without growth.

the number of iterations k_n needed to reach the target accuracy ϵ_n from z_n as

$$\begin{aligned}
k_n - 1 &\leq \sum_{j\in\mathcal{J}} \left[\left(\frac{1+3v_j}{1+v_j} 2^{\frac{1+v_j}{1+3v_j}} |\mathcal{J}|^{\frac{1-v_j}{1+3v_j}} \right) \left(\frac{M_j}{\epsilon_n} \right)^{\frac{2}{1+3v_j}} \xi(z_n, x^*)^{\frac{1+v_j}{1+3v_j}} \right] \\
&\leq \sum_{j\in\mathcal{J}} \left[\left(\frac{1+3v_j}{1+v_j} 2^{\frac{1+v_j}{1+3v_j}} |\mathcal{J}|^{\frac{1-v_j}{1+3v_j}} \right) \left(\frac{M_j}{\epsilon_n} \right)^{\frac{2}{1+3v_j}} \left(\frac{\epsilon_n}{\mu} \right)^{\frac{2(1+v_j)}{p(1+3v_j)}} \right] \\
&= \sum_{j\in\mathcal{J}} \left[\left(\frac{1+3v_j}{1+v_j} 2^{\frac{1+v_j}{1+3v_j}} |\mathcal{J}|^{\frac{1-v_j}{1+3v_j}} \right) \frac{M_j^{\frac{2}{1+3v_j}}}{\mu^{\frac{2(1+v_j)}{p(1+3v_j)}} \epsilon_n^{\frac{2(p-1-v_j)}{p(1+3v_j)}}} \right]
\end{aligned}$$

where the first inequality uses Theorem 1.1 and the second uses the assumed Hölder growth. Then bounding $\epsilon_n \geq 2^{N-n-1}\tilde{\epsilon}$, the total number of steps used by `R-UFGM` is then

$$\begin{aligned}
\sum_{n=0}^{N} k_n &\leq \sum_{n=0}^{N} \sum_{j\in\mathcal{J}} \left[\left(\frac{1+3v_j}{1+v_j} 2^{\frac{1+v_j}{1+3v_j}} |\mathcal{J}|^{\frac{1-v_j}{1+3v_j}} \right) \frac{M_j^{\frac{2}{1+3v_j}}}{\mu^{\frac{2(1+v_j)}{p(1+3v_j)}} \left(2^{N-n-1}\tilde{\epsilon}\right)^{\frac{2(p-1-v_j)}{p(1+3v_j)}}} \right] + N \\
&= \sum_{j\in\mathcal{J}} \left[c_j''' \frac{M_j^{\frac{2}{1+3v_j}}}{\mu^{\frac{2(1+v_j)}{p(1+3v_j)}} \tilde{\epsilon}^{\frac{2(p-1-v_j)}{p(1+3v_j)}}} \sum_{n=0}^{N} \frac{1}{\left(2^{N-n}\right)^{\frac{2(p-1-v_j)}{p(1+3v_j)}}} \right] + N \\
&\leq \sum_{j\in\mathcal{J}} \left[c_j''' \frac{M_j^{\frac{2}{1+3v_j}}}{\mu^{\frac{2(1+v_j)}{p(1+3v_j)}} \tilde{\epsilon}^{\frac{2(p-1-v_j)}{p(1+3v_j)}}} \min\left\{ \sum_{r=0}^{\infty} \left(\frac{1}{2^{\frac{2(p-1-v_j)}{p(1+3v_j)}}} \right)^r, \frac{N}{2^{\frac{2(p-1-v_j)}{p(1+3v_j)}}} \right\} \right] + N \\
&= \sum_{j\in\mathcal{J}} \left[c_j''' \min\left\{ \frac{2^{\frac{2(p-1-v_j)}{p(1+3v_j)}}}{2^{\frac{2(p-1-v_j)}{p(1+3v_j)}} - 1}, \frac{N}{2^{\frac{2(p-1-v_j)}{p(1+3v_j)}}} \right\} \frac{M_j^{\frac{2}{1+3v_j}}}{\mu^{\frac{2(1+v_j)}{p(1+3v_j)}} \tilde{\epsilon}^{\frac{2(p-1-v_j)}{p(1+3v_j)}}} \right] + N \ .
\end{aligned}$$

References

[1] Yurii Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152:381–404, 2015.

[2] A S Nemirovskii and Y E Nesterov. Optimal methods of smooth convex minimization. *USSR Comput. Math. Math. Phys.*, 25(3–4):21–30, jul 1986.

[3] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2004.

[4] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, Oct 2017.

[5] Tianbao Yang and Qihang Lin. Rsg: Beating subgradient method without smoothness and strong convexity. *Journal of Machine Learning Research*, 19(6):1–33, 2018.

[6] Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart, and acceleration. *SIAM Journal on Optimization*, 30(1):262–289, 2020.

[7] James Renegar and Benjamin Grimmer. A simple nearly optimal restart scheme for speeding up first-order methods. *Found. Comput. Math.*, 22(1):211–256, 2022.

[8] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48(3):769–783, 1998.

[9] Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[10] Stanislas Łojasiewicz. Sur la géométrie semi-et sous-analytique. In *Annales de l'institut Fourier*, volume 43, pages 1575–1595, 1993.

[11] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[12] Guanghui Lan. Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. *Math. Program.*, 149(1–2):1–45, feb 2015.

[13] Mateo Díaz and Benjamin Grimmer. Optimal convergence rates for the proximal bundle method, 2021. https://arxiv.org/abs/2105.07874.

[14] Kimon Antonakopoulos, Thomas Pethick, Ali Kavis, Panayotis Mertikopoulos, and Volkan Cevher. Sifting through the noise: Universal first-order methods for stochastic variational inequalities. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 13099–13111. Curran Associates, Inc., 2021.

[15] Nikita Doikov, Konstantin Mishchenko, and Yurii Nesterov. Super-universal regularized newton method, 2022. https://arxiv.org/abs/2208.05888.

[16] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, 2e edition, 2006.

[17] Yossi Arjevani, Amit Daniely, Stefanie Jegelka, and Hongzhou Lin. On the complexity of minimizing convex finite sums without using the indices of the individual functions, 2020. https://arxiv.org/abs/2002.03273.

[18] Nuozhou Wang and Shuzhong Zhang. A gradient complexity analysis for minimizing the sum of strongly convex functions with varying condition numbers, 2022. https://arxiv.org/abs/2208.06524.

[19] Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[20] Xi Chen, Qihang Lin, and Javier Pena. Optimal regularized dual averaging methods for stochastic optimization. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[21] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[22] Benjamin Grimmer. Convergence rates for deterministic and stochastic subgradient methods without lipschitz continuity. *SIAM Journal on Optimization*, 29(2):1350–1365, 2019.

[23] Hans Nyquist. *Recent studies on Lp-norm estimation*. PhD thesis, Umeå universitet, 1980.

[24] Arthur Money, John F. Affleck-Graves, M. L. Hart, and G. D. I. Barr. The linear regression model: Lp norm estimation and the choice of p. *Communications in Statistics - Simulation and Computation*, 11:89–109, 1982.

[25] Alejandro Llorente and Alberto Suárez. Critical sample size for the lp-norm estimator in linear regression models. *2013 Winter Simulations Conference (WSC)*, pages 1047–1056, 2013.

[26] Bradley Efron. Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81(395):709–721, 1986.

[27] Simon Lacoste-Julien, Mark Schmidt, and Francis R. Bach. A simpler approach to obtaining an o(1/t) convergence rate for the projected stochastic subgradient method. *CoRR*, abs/1212.2002, 2012.

[28] Benjamin Grimmer. General holder smooth convergence rates follow from specialized rates assuming growth bounds, 2021. https://arxiv.org/abs/2104.10196.