Universal Gradient Methods for Stochastic Convex Optimization

Anton Rodomanov [1] Ali Kavis [2] Yongtao Wu [3] Kimon Antonakopoulos [3] Volkan Cevher [3]

Abstract

We develop universal gradient methods for Stochastic Convex Optimization (SCO). Our algorithms automatically adapt not only to the oracle's noise but also to the Hölder smoothness of the objective function without a priori knowledge of the particular setting. The key ingredient is a novel strategy for adjusting step-size coefficients in the Stochastic Gradient Method (SGD). Unlike AdaGrad, which accumulates gradient norms, our Universal Gradient Method accumulates appropriate combinations of gradient- and iterate differences. The resulting algorithm has state-of-the-art worst-case convergence rate guarantees for the entire Hölder class including, in particular, both nonsmooth functions and those with Lipschitz continuous gradient. We also present the Universal Fast Gradient Method for SCO enjoying optimal efficiency estimates.

1. Introduction

Motivation. The complexity of modern machine learning problems makes it difficult to estimate their mathematical properties, let alone characterize them accurately. The problems thus demand sophisticated solutions which are robust to possible variations in the parameters. One therefore needs algorithms that can work simultaneously under multiple assumptions while implicitly adapting to the parameters of the problem. The sheer scale of modern problems also raises efficiency concerns, which paves the way for the stochastic methods leveraging randomized computations.

In this paper, we study the convex optimization problem

$$\min_{x \in \operatorname{dom} \psi} \bigl[F(x) := f(x) + \psi(x) \bigr], \tag{1}$$

where f is the main (difficult) part of the problem, and ψ is a *simple* convex function (e.g., indicator of a set). Furthermore, we assume that f can be queried only via an unbiased *stochastic gradient oracle* with (unknown) variance σ^2.

Optimization algorithms are typically designed for a particular problem class and tailored to its properties. Two standard classes are *nonsmooth* (f is Lipschitz continuous) and *smooth* (f has Lipschitz gradient). It is common for the problem class to dictate the selection of algorithm's parameters to ensure the optimal convergence.

However, in practice, every specific problem typically belongs to multiple problem classes at the same time, and it is usually very difficult (if not impossible) to say in advance which particular class better fits our problem. To address this issue, we need *universal methods* that can automatically adjust to the "correct" problem class when applied to a concrete problem instance given to them.

The important example of such algorithms is given by Universal Gradient Methods (UGMs) of (Nesterov, 2015). These methods are capable of solving the more general class of *Hölder-smooth* problems:

$$\|\nabla f(x) - \nabla f(y)\| \le L_\nu \|x - y\|^\nu, \quad \forall x, y \in \operatorname{dom} \psi,$$

which continuously connects nonsmooth problems ($\nu = 0$) with the smooth ones ($\nu = 1$). To achieve universality, UGMs use a special line-search procedure which automatically selects an appropriate step size for any possible Hölder exponent and the corresponding Hölder constant, without knowing these parameters. As a result, the methods automatically adjust to the best possible problem class. However, UGMs require exact computations of gradients ($\sigma = 0$).

The extension of UGMs to stochastic optimization has been a challenging *open problem*. The desired algorithms should automatically adjust not only to the Hölder smoothness of the objective function, but also to the oracle's noise.

In this paper, we address this open problem and provide a solution to it. We design line-search-free variants of UGMs which automatically adapt to: *(i)* Hölder exponent ν, *(ii)* Hölder constant L_ν, *(iii)* variance of the stochastic oracle σ, without having the prior knowledge of neither the problem class nor the nature of the gradient information.

[1] CISPA Helmholtz Center for Information Security, Saarbrücken, Germany [2] Institute for Foundations of Machine Learning (IFML), UT Austin, Texas, USA [3] Laboratory for Information and Inference Systems (LIONS), EPFL, Lausanne, Switzerland. Correspondence to: Anton Rodomanov <anton.rodomanov@cispa.de>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Contributions. We develop new Universal Gradient Methods (UGMs) for problem (1), which are robust to the stochastic noise in gradient computations. To achieve that, we assume the knowledge of a certain upper bound D on the diameter of the feasible set $\operatorname{dom}\psi$ (or, somewhat equivalently, the distance from the initial point to the solution), which is a common assumption in a variety of other adaptive algorithms for Stochastic Convex Optimization (SCO).

Our main contributions can be summarized as follows:

1. We first rethink (in Section 3) the theoretical analysis of the line-search-based UGM for deterministic optimization, and identify a simple mechanism to remove the line search from this algorithm while retaining the same worst-case efficiency estimates. The key element is a *novel strategy for adjusting step-size coefficients* based on the idea of balancing the two error terms appearing in the convergence analysis.
2. We then show (in Section 4) that our techniques can easily be extended to stochastic optimization problems. The only essential change that we need to make is to replace the Bregman distance for the objective function, appearing in the formula for the step-size, with the stochastic version of the symmetrized Bregman distance involving gradient- and iterate differences. The resulting Universal Stochastic Gradient Method requires at most $O\bigl(\inf_{\nu\in[0,1]}[\frac{L_\nu}{\epsilon}]^{2/(1+\nu)}D^2+\frac{\sigma^2D^2}{\epsilon^2}\bigr)$ stochastic oracle calls to reach ϵ-accuracy in terms of the expected function residual (Theorem 4.2).
3. Finally, we present (in Section 5) the Universal Stochastic Fast Gradient Method enjoying the worst-case optimal efficiency of $O\bigl(\inf_{\nu\in[0,1]}[\frac{L_\nu D^{1+\nu}}{\epsilon}]^{2/(1+3\nu)}+\frac{\sigma^2D^2}{\epsilon^2}\bigr)$ oracle calls (Theorem 5.1).

Note that all our methods are agnostic to the smoothness exponent ν, smoothness constant L_ν and the noise level σ. To our knowledge, this is the first work proposing algorithms with such characteristics.

Related work. Pioneered by the AdaGrad algorithm Duchi et al. (2011); McMahan & Streeter (2010), *adaptive methods* have been at the forefront of training machine learning models. AdaGrad accumulates the sequence of observed gradient norms to construct a decreasing step size. This construction enables data-adaptive regret bounds and has many useful properties. Following the success of the AdaGrad, several methods have been proposed (Kingma & Ba, 2015; Tieleman & Hinton, 2012; Rakhlin & Sridharan, 2013; Reddi et al., 2018). Levy et al. (2018) proposed the first accelerated algorithm with data-adaptive step-size without the knowledge of Lipschitz constant and the variance bound. They prove convergence results for nonsmooth and smooth objectives in the presence of stochastic noise. These results are further refined and extended by Kavis et al. (2019); Joulani et al. (2020); Ene et al. (2021). Despite the significant interest in the adaptation to smoothness and noise, existing methods are not known to handle Hölder-smooth objectives.

Another popular type of adaptive methods is known as *parameter-free*. This direction is very interesting but somewhat orthogonal to ours. Parameter-free algorithms have been studied for over a decade in online learning (McMahan & Streeter, 2012; Orabona, 2014; Cutkosky & Boahen, 2017; Cutkosky & Orabona, 2018; Jacobsen & Cutkosky, 2023; Mhammedi & Koolen, 2020). They are usually endowed with appropriate mechanisms to achieve efficiency bounds that are almost insensitive (typically, with logarithmic dependency) to the error of estimating certain problem parameters, such as the diameter of the feasible set (Carmon & Hinder, 2022; Ivgi et al., 2023; Defazio & Mishchenko, 2023; Khaled et al., 2023; Mishchenko & Defazio, 2023). However, these methods typically consider the extreme cases of the Hölder class.

Within the context of online learning, there exists an independent notion of universality such that the algorithms adapt unknowingly to the degrees and types of convexity. The goal is designing algorithms that achieve, up to logarithmic factors, optimal regret bounds simultaneously for convex, strongly convex and exponentially concave functions (Van Erven & Koolen, 2016; Wang et al., 2020; Zhang et al., 2022; Yan et al., 2023). The associated design and proof techniques are not transferable to our setup as we focus on the degree of smoothness while the aforementioned works study degrees of convexity.

The first UGM for deterministic optimization, including the Fast UGM with optimal worst-case oracle complexity, was proposed in (Nesterov, 2015). The corresponding methods achieve the adaptation to Hölder smoothness by the means of line search but must set the target accuracy a priori. A possible extension of these algorithms to stochastic optimization was considered in (Gasnikov & Nesterov, 2018). They proposed an accelerated gradient method for stochastic optimization problems, which adapts to the Hölder characteristics of the objective using line search combined with mini-batching. However, this method additionally relies on the knowledge of the oracle's variance to correctly set up the size of the mini-batch at each iteration, and therefore cannot be considered adaptable to the noise level.

More recently, Li & Lan (2023) studied the same problem but with the deterministic oracle, and designed a line-search-free universal method that estimates local smoothness in the sense of Malitsky & Mishchenko (2020; 2023). Their step-size formula shares some similarities to ours, and does not require any (artificial) bounds on the diameter of the feasible set. However, they only consider exact gradient computations, and it is unknown whether their construction can be extended to stochastic problems. On a related note, Orabona (2023) showed that, in the deterministic case, both AdaGrad and the normalized gradient method (Nesterov, 2018, Section 3.2.3) automatically adapt to Hölder smoothness. Although the corresponding proof for AdaGrad could be extended to the stochastic setting at the expense of additional assumptions, the same type of argument cannot be trivially applied to the accelerated method.

2. Preliminaries

2.1. Notation

In this text, we work in the space $\mathbb{R}^n$ equipped with the standard inner product $\langle \cdot, \cdot \rangle$ and the certain Euclidean norm:

$$\|x\| := \langle Bx, x \rangle^{1/2}, \qquad x \in \mathbb{R}^n, \tag{2}$$

where $B \in \mathbb{S}^n_{++}$ is a sufficiently simple symmetric positive definite matrix (e.g., the identity or a diagonal one). The corresponding dual norm is defined in the standard way:

$$\|s\|_* := \max_{\|x\|=1} \langle s, x \rangle = \langle s, B^{-1} s \rangle^{1/2}, \qquad s \in \mathbb{R}^n. \tag{3}$$

Thus, for any $s, x \in \mathbb{R}^n$, we have the Cauchy–Schwarz inequality $|\langle s, x \rangle| \leq \|s\|_* \|x\|$.

For a convex function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, by $\operatorname{dom} f := \{x \in \mathbb{R}^n : f(x) < +\infty\}$, we denote its *effective domain*. The subdifferential of f at a point $x \in \operatorname{dom} f$ is denoted by $\partial f(x)$. For any two points $x, y \in \operatorname{dom} f$ and any $g \in \partial f(x)$, we define the *Bregman distance* generated by f as

$$\beta_f^g(x, y) := f(y) - f(x) - \langle g, y - x \rangle \quad (\geq 0), \tag{4}$$

In the case when there is no ambiguity with the subgradient g, we use a simpler notation $\beta_f(x, y)$.

For any $t \in \mathbb{R}$, by $[t]_+ := \max\{t, 0\}$, we denote its positive part. For random variables X and ξ, by $\mathbb{E}_\xi[X]$ and $\mathbb{E}[X]$, we denote the expectation of X w.r.t. to ξ, and the full expectation of X, respectively.

2.2. Problem Setting

In this paper, we study the following optimization problem:

$$F^* := \min_{x \in \operatorname{dom} \psi} \bigl[F(x) := f(x) + \psi(x) \bigr], \tag{5}$$

where $\psi\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a sufficiently *simple* proper closed convex function, and $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a closed convex function which is finite and subdifferentiable over an open set containing $\operatorname{dom} \psi$.

Our main assumption on problem (5) is the *boundedness of the feasible set* $\operatorname{dom} \psi$.

Assumption 2.1. *For problem* (5)*, there is $D > 0$ such that $\|x - y\| \leq D$ for all $x, y \in \operatorname{dom} \psi$.*

In what follows, we assume that the diameter D is known. This will be the only parameter in our methods. Note that Assumption 2.1 guarantees that the problem (5) has a solution (since the objective function F is proper and closed). An important example that satisfies this assumption is when ψ is the indicator function of a certain compact set $Q \subseteq \mathbb{R}^n$: $\psi(x) := 0$ if $x \in Q$, and $\psi(x) := +\infty$ if $x \notin Q$.

By calling ψ simple, we mean the following standard assumption: for any $c \in \mathbb{R}^n$, $\bar{x} \in \operatorname{dom} \psi$, and $H \geq 0$, we can efficiently compute a solution to the following subproblem: $\min_{x \in \operatorname{dom} \psi} \{\langle c, x \rangle + \frac{H}{2} \|x - \bar{x}\|^2 + \psi(x)\}$. For instance, when ψ is the indicator function of a compact set Q, it corresponds to finding a Euclidean projection onto Q (or minimizing a linear function over Q if $H = 0$; in this case, we allow for an arbitrary solution of the subproblem).

To characterize the smoothness of f in problem (5), let us introduce, for each $\nu \in [0, 1]$, the *Hölder constant*:

$$L_\nu := \sup_{\substack{x, y \in \operatorname{dom} \psi,\ x \neq y, \\ g(x) \in \partial f(x),\ g(y) \in \partial f(y)}} \frac{\|g(x) - g(y)\|_*}{\|x - y\|^\nu}. \tag{6}$$

Of course, for certain values of the exponent $\nu \in [0, 1]$, it may happen that $L_\nu = +\infty$. However, we assume that there exists (at least one) exponent for which the corresponding Hölder constant is finite.

Assumption 2.2. *For problem* (5) *and L_ν given by* (6)*, there exists $\nu \in [0, 1]$ such that $L_\nu < +\infty$.*

The case $L_0 < +\infty$ corresponds to the situation when f has *bounded variation of subgradients*: for all $x, y \in \operatorname{dom} \psi$, and all $g(x) \in \partial f(x)$, $g(y) \in \partial f(y)$, it holds that $\|g(x) - g(y)\|_* \leq L_0$. If f has bounded subgradients over $\operatorname{dom} \psi$, i.e., there exists $L_0' \geq 0$ such that $\|g(x)\|_* \leq L_0'$ for all $x \in \operatorname{dom} \psi$ and all $g(x) \in \partial f(x)$, then $L_0 \leq 2L_0'$. On the other hand, if $L_\nu < +\infty$ for some $\nu \in (0, 1]$, then f is actually differentiable over $\operatorname{dom} \psi$, and, for all $x, y \in \operatorname{dom} \psi$, it holds $\|\nabla f(x) - \nabla f(y)\|_* \leq L_\nu \|x - y\|^\nu$. The case $L_1 < +\infty$ corresponds to the *Lipschitz gradient*.

One simple example of the convex function with Hölder (sub)gradients is the p-th power of the ℓ_p-norm of the residual for the system of linear equations, $f(x) = \frac{1}{m} \sum_{i=1}^m |\langle a_i, x \rangle - b_i|^p$ with $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$, $p \in [1, 2]$, which generalizes the classical least-squares loss (corresponding to $p = 2$); this function is Hölder smooth with $\nu = p - 1$ but not Lipschitz smooth (unless $p = 2$). Another simple example is the similar residual but for linear inequalities, $f(x) = \frac{1}{m} \sum_{i=1}^m [\langle a_i, x \rangle - b_i]_+^p$, which is the smooth counterpart of the classical loss function used by the Support Vector Machines (SVMs). More generally, there is a duality relationship between Hölder smoothness and uniform convexity: if f_* is a uniformly convex function[1] of degree $q \geq 2$ with parameter $\sigma_q > 0$, then its Fenchel dual f is Hölder smooth with $\nu = \frac{1}{q-1}$ and $L_\nu \leq \left(\frac{1}{\sigma_q}\right)^{\frac{1}{q-1}}$ (see, e.g., Lemma 1 in (Nesterov, 2015)), and vice versa; in particular[2], for any convex function f_*, the function $f(x) = \max_{s \in \operatorname{dom} f_*} [\langle s, x \rangle - f_*(s) - \frac{\sigma_q}{q} \|s\|_*^q]$ is Hölder smooth with $\nu = \frac{1}{q-1}$ and $L_\nu \leq \left(\frac{2^{q-2}}{\sigma_q}\right)^{\frac{1}{q-1}}$.

It is not difficult to see from (6) that, under Assumption 2.1, for any $0 \leq \nu_1 \leq \nu_2 \leq 1$, we have the following monotonicity relation: $L_{\nu_1} D^{\nu_1} \leq L_{\nu_2} D^{\nu_2}$. (This is the consequence of the fact that τ^p is increasing in $p > 0$ for any fixed $\tau = \frac{D}{\|x-y\|} \geq 1$.) In particular, if $L_{\nu'} < +\infty$ for some $\nu' \in [0, 1]$, then $L_\nu < +\infty$ for all $\nu \in [0, \nu']$.

One standard and important consequence of (6) is that, for any $\nu \in [0, 1]$ (such that $L_\nu < +\infty$), and all $x, y \in \operatorname{dom} \psi$ and all $g \in \partial f(x)$, we have the following upper bound on the Bregman distance of the function f:

$$\beta_f^g(x, y) \leq \frac{L_\nu}{1+\nu} \|x - y\|^{1+\nu}. \qquad (7)$$

Our goal in this paper is to present numerical methods for solving (5) that are *universal*: they can automatically adapt to the actual level of smoothness of the function f without knowing neither the Hölder exponent ν, nor the corresponding Hölder constant L_ν.

3. Universal Line-Search-Free Gradient Method

3.1. Main Idea

To explain the main idea behind our construction of adaptive step-size coefficients, let us consider the usual (Composite) Gradient Method for solving problem (5):

$$x_{k+1} = \operatorname*{argmin}_{x \in \operatorname{dom} \psi} \Bigl\{ \langle f'(x_k), x \rangle + \psi(x) + \frac{H_k}{2} \|x - x_k\|^2 \Bigr\}, \quad (8)$$

assuming we can compute the exact (sub)gradient $f'(x_k) \in \partial f(x_k)$ at each iteration $k \geq 0$ (i.e., the oracle is deterministic). The question is how to choose the step-size coefficients H_k at each iteration to ensure that the algorithm properly works for any possible Hölder exponent ν and the corresponding coefficient L_ν without explicitly using these constants in the method.

The standard convergence analysis of method (8) uses the following central inequality (for $r_{k+1} := \|x_{k+1} - x_k\|$ and $d_k := \|x_k - x^*\|$ with x^* being a solution of (5)):

$$f(x_k) + \langle f'(x_k), x_{k+1} - x_k \rangle + \psi(x_{k+1}) + \frac{H_k}{2} r_{k+1}^2 + \frac{H_k}{2} d_{k+1}^2$$
$$\leq f(x_k) + \langle f'(x_k), x^* - x_k \rangle + \psi(x^*) + \frac{H_k}{2} d_k^2,$$

which is a simple consequence of the strong convexity of the objective function in the auxiliary subproblem (8) (c.f. Lemma E.2). Rewriting now $f(x_k) + \langle f'(x_k), x_{k+1} - x_k \rangle = f(x_{k+1}) - \beta_f(x_k, x_{k+1})$ using the Bregman distance, and estimating $f(x_k) + \langle f'(x_k), x^* - x_k \rangle + \psi(x^*) \leq f(x^*) + \psi(x^*) = F^*$ using the convexity of f, we get

$$F(x_{k+1}) - F^* + \frac{H_k}{2} d_{k+1}^2 \leq \frac{H_k}{2} d_k^2 + \beta_{k+1} - \frac{H_k}{2} r_{k+1}^2, \quad (9)$$

where $\beta_{k+1} := f(x_{k+1}) - f(x_k) - \langle f'(x_k), x_{k+1} - x_k \rangle$.

3.1.1. Line-Search Approach

The standard approach to proceed, pioneered by (Nesterov, 2015), is to choose the coefficient H_k in such a way that the error term $\beta_{k+1} - \frac{H_k}{2} r_{k+1}^2$ in (9) is sufficiently small:

$$\Delta_k := \beta_{k+1} - \frac{H_k}{2} r_{k+1}^2 \leq \frac{\epsilon}{2} \qquad (10)$$

(for a certain fixed $\epsilon > 0$), and then divide both sides of (9) by H_k to get a telescopic recurrence:

$$\frac{1}{H_k} [F(x_{k+1}) - F^*] + \frac{1}{2} d_{k+1}^2 \leq \frac{1}{2} d_k^2 + \frac{\epsilon}{2 H_k}.$$

Telescoping and dividing by $S_k := \sum_{i=0}^{k-1} \frac{1}{H_i}$, we get

$$F(x_k^*) - F^* \leq \frac{D_0^2}{2 S_k} + \frac{\epsilon}{2} \leq \frac{H_k^* D_0^2}{2k} + \frac{\epsilon}{2}, \qquad (11)$$

where $D_0 := d_0 = \|x_0 - x^*\|$, $H_k^* := \max_{0 \leq i \leq k-1} H_i$, and x_k^* is the "best" iterate:

$$x_k^* := \operatorname{argmin}\bigl\{ f(x) : x \in \{x_1, \ldots, x_k\} \bigr\}. \qquad (12)$$

(Alternatively, one could also define x_k^* as the average of x_i with weights $\frac{1}{H_i}$.) This gives us the convergence of the function residual to ϵ, provided that H_k^* is reasonably bounded from above (e.g., by a constant).

To ensure that (10) is satisfied for a sufficiently large H_k and estimate the corresponding H_k^*, we start with the observation that, by (7), $\beta_{k+1} \leq \frac{L_\nu}{1+\nu} r_{k+1}^{1+\nu}$ for any $\nu \in [0, 1]$, and hence

$$\Delta_k \leq \frac{L_\nu}{1+\nu} r_{k+1}^{1+\nu} - \frac{H_k}{2} r_{k+1}^2 \leq \frac{(1-\nu) L_\nu^{2/(1-\nu)}}{2(1+\nu) H_k^{(1+\nu)/(1-\nu)}}. \qquad (13)$$

(The final inequality follows by maximizing the expression in r_{k+1}, see Lemma E.3; for $\nu = 1$, the right-hand side

[1] This means that $\langle f_*'(x) - f_*'(y), x - y \rangle \geq \sigma_q \|x - y\|^q$ for all x, y and all $f_*'(x) \in \partial f_*(x)$, $f_*'(y) \in \partial f_*(y)$.

[2] Here we use the standard fact that $\frac{1}{q} \|\cdot\|_*^q$ is a uniformly convex function of degree q with parameter $\frac{1}{2^{q-2}}$, (see, e.g., Lemma 4 in (Nesterov, 2008)).

should be understood as 0 if $H_k \geq L_\nu$ and $+\infty$ otherwise.) Making the right-hand side of the above display $\leq \frac{\epsilon}{2}$, we see that (10) is satisfied whenever $H_k \geq \bar{H}_\nu$, where

$$\bar{H}_\nu \coloneqq L_\nu^{2/(1+\nu)} \Big[\frac{1-\nu}{(1+\nu)\epsilon}\Big]^{(1-\nu)/(1+\nu)}.$$

Notice that $\nu = 1$ implies $\bar{H}_\nu \geq L_\nu$. Since we do not know the best (= smallest) possible value of L_ν over all ν, we cannot simply set $H_k = \bar{H}_* \coloneqq \inf_{\nu \in [0,1]} \bar{H}_\nu$. However, we can let the line search estimate this value for us: at each iteration, we start with a certain initial guess H_k' for H_k and then repeatedly double this value until condition (10) is satisfied (note that x_{k+1} depends on H_k and thus needs to be recomputed at every iteration of the line search procedure). Provided that the initial guess H_0' at the very first iteration is not sufficiently large, e.g., $H_0' \leq \bar{H}_*$ and that H_{k+1}' is chosen appropriately (e.g, $H_{k+1}' = \frac{1}{2} H_k$), we then have the guarantee that H_k computed by the line search does not significantly exceed the "right" value: $H_k \leq 2\bar{H}_*$; furthermore, the total number of line search iterations across all iterations of the algorithm is reasonably bounded. Substituting now this bound on H_k into (11) and looking at the number of iterations k that one needs to make $\frac{H_k^* D_0^2}{k} \leq \epsilon$, we see that the outlined above line-search method needs

$$O\bigg(\inf_{\nu \in [0,1]} \frac{\bar{H}_\nu D_0^2}{\epsilon}\bigg) = O\bigg(\inf_{\nu \in [0,1]} \Big[\frac{L_\nu}{\epsilon}\Big]^{2/(1+\nu)} D_0^2\bigg) \tag{14}$$

iterations to reach $F(x_k^*) - F^* \leq \epsilon$.

3.1.2. Our Idea: How to Avoid Line Search

The problem with the line-search approach, which makes it difficult to extend the corresponding reasoning to the stochastic case, is that it creates a dependency (correlation) between x_k and H_k. This does not allow us to use the unbiasedness of the stochastic gradient oracle once we divide (the stochastic counterpart of) (9) by H_k (see Section 4.2).

However, we can follow a different approach to convert (9) into a telescopic recurrence. Specifically, let us replace the coefficient H_k in the left-hand side of (9) with H_{k+1}:

$$\begin{aligned} F(x_{k+1}) - F^* + \frac{H_{k+1}}{2} d_{k+1}^2 - \frac{H_k}{2} d_k^2 \\ \leq \beta_{k+1} - \frac{H_k}{2} r_{k+1}^2 + \frac{1}{2}(H_{k+1} - H_k) d_{k+1}^2. \end{aligned}$$

As we can see, such an operation may introduce an additional error term $\frac{1}{2}(H_{k+1} - H_k) d_{k+1}^2$ if we plan to increase our step-size coefficient: $H_k \leq H_{k+1}$ (which is a natural thing to do if it is currently too small making the other error term $\beta_{k+1} - \frac{H_k}{2} r_{k+1}^2$ too large). Nevertheless, using Assumption 2.1, we can easily control this additional error term and make it telescopic:

$$\begin{aligned} F(x_{k+1}) - F^* + \frac{H_{k+1}}{2} d_{k+1}^2 - \frac{H_k}{2} d_k^2 \\ \leq \beta_{k+1} - \frac{H_k}{2} r_{k+1}^2 + \frac{1}{2}(H_{k+1} - H_k) D^2. \end{aligned} \tag{15}$$

Our main idea now is to choose the next coefficient H_{k+1} so that the two error terms are balanced:

$$\frac{1}{2}(H_{k+1} - H_k) D^2 = \Big[\beta_{k+1} - \frac{H_k}{2} r_{k+1}^2\Big]_+, \tag{16}$$

where we additionally put the positive part $[\cdot]_+$ to respect the monotonicity relation $H_k \leq H_{k+1}$. Recall that β_{k+1} and r_{k+1} depend only on x_k and x_{k+1} (which themselves depend on H_{k-1} and H_k, see (8)). Thus, (16) is a simple linear equation for H_{k+1} which does not require any line search for solving it.

Substituting our choice (16) of H_{k+1} into (15) (and using the fact that $\tau \leq [\tau]_+$ for any $\tau \in \mathbb{R}$), we arrive at the following simple telescopic inequality:

$$F(x_{k+1}) - F^* + \frac{H_{k+1}}{2} d_{k+1}^2 \leq \frac{H_k}{2} d_k^2 + (H_{k+1} - H_k) D^2. \tag{17}$$

Telescoping these inequalities, we get

$$F(x_k^*) - F^* \leq \frac{1}{k}\Big[\frac{H_0}{2} d_0^2 + (H_k - H_0) D^2\Big] \leq \frac{H_k D^2}{k}, \tag{18}$$

where x_k^* is the "best" iterate (see (12)). (Alternatively, one could also define $x_k^* = \frac{1}{k}\sum_{i=1}^k x_i$.)

The main question is how fast the coefficient H_k grows. Following exactly the same argument as in (13) (and using the fact that $[\cdot]_+$ is nondecreasing), we can estimate the right-hand side of our balance equation (16) and conclude that $(H_{k+1} - H_k) D^2 \leq \frac{(1-\nu) L_\nu^{2/(1-\nu)}}{(1+\nu) H_k^{(1+\nu)/(1-\nu)}}$. (Assume, for simplicity, that $\nu < 1$; to rigorously handle the case $\nu = 1$ we need a more careful argument.) This is a certain recurrent inequality that we can use to estimate the rate of growth of H_k. This would be especially simple if we had, say, $2H_{k+1}^{(1+\nu)/(1-\nu)}$ instead of $H_k^{(1+\nu)/(1-\nu)}$:

$$(H_{k+1} - H_k) D^2 \leq \frac{(1-\nu) L_\nu^{2/(1-\nu)}}{2(1+\nu) H_{k+1}^{(1+\nu)/(1-\nu)}}. \tag{19}$$

Then, a simple integration argument (see Lemma E.4 with $p = \frac{1+\nu}{1-\nu}$ for which $p + 1 = \frac{2}{1-\nu}$) would show that

$$H_k \leq \frac{L_\nu}{D^{1-\nu}} k^{(1-\nu)/2}, \tag{20}$$

provided that the initial step-size coefficient was chosen appropriately: $H_0 = 0$. (Note that this would not cause any

Algorithm 1 Universal Line-Search-Free Gradient Method

```
1: Initialize: $x_0 \in \operatorname{dom}\psi$, diameter $D > 0$, $H_0 := 0$.
2: for $k = 0, 1, \dots$ do
3:     Compute $g_k \in \partial f(x_k)$.
4:     $x_{k+1} = \operatorname*{argmin}_{x \in \operatorname{dom}\psi} \{\langle g_k, x\rangle + \psi(x) + \frac{H_k}{2}\|x - x_k\|^2\}$.
5:     $H_{k+1} := H_k + \frac{[\beta_{k+1} - \frac{1}{2} H_k r_{k+1}^2]_+}{D^2 + \frac{1}{2} r_{k+1}^2}$, where
       $r_{k+1} := \|x_k - x_{k+1}\|$, $\beta_{k+1} := \beta_f^{g_k}(x_k, x_{k+1})$.
6: end for
```

Algorithm 2 Universal Stochastic Gradient Method

```
1: Initialize: $x_0 \in \operatorname{dom}\psi$, $D > 0$, $H_0 := 0$, $g_0 \sim \hat{g}(x_0)$.
2: for $k = 0, 1, \dots$ do
3:     $x_{k+1} = \operatorname*{argmin}_{x \in \operatorname{dom}\psi} \{\langle g_k, x\rangle + \psi(x) + \frac{H_k}{2}\|x - x_k\|^2\}$.
4:     $g_{k+1} \sim \hat{g}(x_{k+1})$.
5:     $H_{k+1} := H_k + \frac{[\hat{\beta}_{k+1} - \frac{1}{2} H_k r_{k+1}^2]_+}{D^2 + \frac{1}{2} r_{k+1}^2}$, where
       $r_{k+1} = \|x_{k+1} - x_k\|$, $\hat{\beta}_{k+1} = \langle g_{k+1} - g_k, x_{k+1} - x_k\rangle$.
6: end for
```

problems for the iteration (8) being well-defined since we assume that $\operatorname{dom}\psi$ is a bounded set.)

Of course, we cannot argue that that our "real" version of (19) (the one with $H_k^{(1+\nu)/(1-\nu)}$ instead of $2H_{k+1}^{(1+\nu)/(1-\nu)}$) implies the desired (19) (in fact, the relationship is exactly the opposite since $H_k \le H_{k+1}$). However, we can slightly modify the reasoning we used to pass from (15) to (18) and the corresponding recurrent inequality for H_k. Specifically, we can rewrite the $-\frac{H_k}{2} r_{k+1}^2$ term in the right-hand side of (15) as $-\frac{H_{k+1}}{2} r_{k+1}^2 + \frac{1}{2}(H_{k+1} - H_k) r_{k+1}^2$, and then upper bound $r_{k+1} \le D$. As a result, we get $\beta_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2 + (H_{k+1} - H_k) D^2$ in the right-hand of (15), and can now choose the coefficient H_{k+1} using the following balance equation instead of (16):

$$\boxed{(H_{k+1} - H_k) D^2 = \Big[\beta_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2\Big]_+.} \tag{21}$$

Although this is no longer a linear equation in H_{k+1}, it always has a unique solution $H_{k+1} \ge H_k$, which can be easily computed: if $\beta_{k+1} \le \frac{H_k}{2} r_{k+1}^2$, then $H_{k+1} = H_k$; otherwise, H_{k+1} is the solution of the linear equation $(H_{k+1} - H_k) D^2 = \beta_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2$ (see Lemma E.1). Proceeding exactly is the same way as before, we get (18) but with $2D^2$ instead of D^2, and, most importantly, the desired (19). As a result, (20) indeed holds and we get

$$F(x_k^*) - F^* \le \frac{2 H_k D^2}{k} \le \inf_{\nu \in [0,1]} \frac{2 L_\nu D^{1+\nu}}{k^{(1+\nu)/2}},$$

where the infimum is due to the fact that $\nu \in [0,1]$ was allowed to be arbitrary in our analysis.

3.2. The Method

Summarizing the outlined above considerations into a formal algorithmic scheme, we arrive at Algorithm 1. This is essentially the classical (Composite) Gradient Method (8) but equipped with our novel step-size adjusting rule (21) (the formula for H_{k+1} at Line 5 is the explicitly written solution of the balance equation (21)).

Theorem 3.1. *Let Algorithm 1 be applied to problem (5) under Assumptions 2.1 and 2.2, and let x_k^* be the "best" iterate as defined in (12). Then, for all $k \ge 1$, we have*

$$F(x_k^*) - F^* \le \inf_{\nu \in [0,1]} \frac{2 L_\nu D^{1+\nu}}{k^{(1+\nu)/2}}.$$

To reach $F(x_k^*) - F^* \le \epsilon$ for any $\epsilon > 0$, it thus suffices to make $\inf_{\nu \in [0,1]} [\frac{2 L_\nu}{\epsilon}]^{2/(1+\nu)} D^2$ iterations.

Comparing the efficiency bound from Theorem 3.1 with the corresponding bound (14) for the line-search method, we see that they are almost identical. The only difference is that our method has the diameter of the feasible set D instead of the initial distance D_0. However, as we show next, our method can be easily extended to stochastic problems.

4. Universal Gradient Method for Stochastic Optimization

Now we assume that f in problem (5) is accessible only via the *stochastic gradient oracle* $\hat{g}$. Formally, this is a pair (g, ξ) consisting of a random variable ξ and a mapping $g\colon \operatorname{dom} f \times \operatorname{Im}\xi \to \mathbb{R}^n$ (with $\operatorname{Im}\xi$ being the image of ξ). When queried at a point $x \in \operatorname{dom}\psi$, the oracle automatically generates an independent copy ξ of its randomness, and then returns $s = g(x, \xi)$ (notation: $s \sim \hat{g}(x)$)—a random estimate of a subgradient of f at x.

We make the following standard assumption on the oracle:

Assumption 4.1. *The function f in problem (5) is accessible only via an unbiased stochastic gradient oracle $\hat{g} = (g, \xi)$ with bounded variance:*

$$f'(x) := \mathbb{E}_\xi[g(x, \xi)] \in \partial f(x), \tag{22}$$

$$\sigma^2 := \sup_{x \in \operatorname{dom}\psi} \mathbb{E}_\xi[\|g(x, \xi) - f'(x)\|_*^2] < +\infty. \tag{23}$$

4.1. The (Stochastic) Method

Our Universal Gradient Method for problem (5) with stochastic gradient oracle is shown in Algorithm 2. As we can see, this method is very similar to its deterministic counterpart (Algorithm 1) with two fundamental differences to accommodate stochastic feedback. First, instead of the exact subgradients $f'(x_k)$, we now use their stochastic estimates g_k. Second, instead of the exact Bregman distance $\beta_f(x_k, x_{k+1})$, we now use $\hat{\beta}_{k+1} = \langle g_{k+1} - g_k, x_k - x_{k+1} \rangle$, which can be seen as the stochastic approximation of the *symmetrized Bregman distance* $\hat{\beta}_f(x_k, x_{k+1}) := \langle f'(x_{k+1}) - f'(x_k), x_{k+1} - x_k \rangle$. (Note that, for any $x, y \in \operatorname{dom} f$, we have $\hat{\beta}_f(x, y) = \beta_f(x, y) + \beta_f(y, x)$.)

Next, we present the main result on the convergence of Algorithm 2 (see Appendix B for the proof).

Theorem 4.2. *Let Algorithm 2 be applied to problem (5) under Assumptions 2.1, 2.2 and 4.1, and let $\bar{x}_k := \frac{1}{k} \sum_{i=1}^k x_i$ be the average iterate. Then, for all $k \geq 1$,*

$$\mathbb{E}[F(\bar{x}_k)] - F^* \leq \inf_{\nu \in [0,1]} \frac{8 L_\nu D^{1+\nu}}{k^{(1+\nu)/2}} + \frac{4\sigma D}{\sqrt{k}}.$$

To reach $\mathbb{E}[F(\bar{x}_k)] - F^* \leq \epsilon$ for any $\epsilon > 0$, it then suffices to make $O\big(\inf_{\nu \in [0,1]} [\frac{L_\nu}{\epsilon}]^{2/(1+\nu)} D^2 + \frac{\sigma^2 D^2}{\epsilon^2}\big)$ oracle calls.

4.2. Main Idea and Outline of Analysis

Let us briefly explain the motivation behind the specific formula for $\hat{\beta}_{k+1}$ in Algorithm 2 and sketch the corresponding convergence analysis. The formal proof with all the details can be found in Appendix B.

From the definition of x_{k+1}, it follows that (Lemma E.2)

$$\begin{aligned} f(x_k) + \langle g_k, x_{k+1} - x_k \rangle + \psi(x_{k+1}) + \frac{H_k}{2} r_{k+1}^2 + \frac{H_k}{2} d_{k+1}^2 \\ \leq f(x_k) + \langle g_k, x^* - x_k \rangle + \psi(x^*) + \frac{H_k}{2} d_k^2, \end{aligned}$$

where $r_{k+1} := \|x_{k+1} - x_k\|$ and $d_k := \|x_k - x^*\|$.

Observe that $\mathbb{E}_{\xi_k}[f(x_k) + \langle g_k, x^* - x_k \rangle] = f(x_k) + \langle f'(x_k), x^* - x_k \rangle \leq f(x^*)$, where ξ_k is the oracle's randomness defining $g_k \equiv g(x_k, \xi_k)$. However, if we attempted to follow the line-search idea from Section 3.1 by first dividing both sides in the previous display by H_k, then we would not be able to use the oracle's unbiasedness as H_k and x_k would depend on each other.

Nevertheless, our line-search-free idea still works. Specifically, passing to expectations in the above display and using the lower bound on $f(x^*)$ from the previous paragraph, and then rearranging, we obtain

$$\begin{aligned} \mathbb{E}\Big[F(x_{k+1}) - F^* + \frac{H_{k+1}}{2} d_{k+1}^2 - \frac{H_k}{2} d_k^2\Big] \\ \leq \mathbb{E}\Big[\beta_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2 + (H_{k+1} - H_k) D^2\Big], \end{aligned} \tag{24}$$

where $\beta_{k+1} := f(x_{k+1}) - f(x_k) - \langle g_k, x_{k+1} - x_k \rangle$, and the $(H_{k+1} - H_k) D^2$ term corresponds to the upper bound on $\frac{1}{2}(H_{k+1} - H_k)(d_{k+1}^2 + r_{k+1}^2)$.

The problem is that we cannot compute β_{k+1} since it involves the exact function values $f(x_{k+1})$ and $f(x_k)$. However, we may replace it with an appropriate stochastic approximation. Indeed, for our goals it suffices to know some $\hat{\beta}_{k+1}$ which is an upper estimate for β_{k+1} in expectation: $\mathbb{E}\, \beta_{k+1} \leq \mathbb{E}\, \hat{\beta}_{k+1}$. To get an appropriate $\hat{\beta}_{k+1}$, we could, in principle, ask the oracle to provide not only stochastic gradients but also stochastic function values. However, this would require imposing extra requirements for the oracle.

Instead, we take another, simpler, approach. By the convexity of f, we can estimate

$$\beta_{k+1} \leq \langle f'(x_{k+1}) - g_k, x_{k+1} - x_k \rangle = \mathbb{E}_{\xi_{k+1}}[\hat{\beta}_{k+1}], \tag{25}$$

where $\hat{\beta}_{k+1} := \langle g_{k+1} - g_k, x_{k+1} - x_k \rangle$ can be calculated in the algorithm and ξ_{k+1} is the oracle's randomness defining $g_{k+1} \equiv g(x_{k+1}, \xi_{k+1})$. It is important for the final identity that ξ_{k+1} is generated after x_k and x_{k+1}.

This leads us to the balance equation

$$\boxed{(H_{k+1} - H_k) D^2 = [\hat{\beta}_{k+1} - \tfrac{1}{2} H_{k+1} r_{k+1}^2]_+,} \tag{26}$$

whose solution is given at Line 5 in Algorithm 2.

Taking into account our balance equation and (25), we obtain exactly the same simple-to-telescope inequality as (17) (valid in expectation) which then leads to

$$\mathbb{E}[F(\bar{x}_k)] - F^* \leq \frac{2\, \mathbb{E}[H_k] D^2}{k}. \tag{27}$$

The rest of the analysis focuses on estimating the (expected) rate of growth of H_k. The key idea is that we can estimate

$$\begin{aligned} \hat{\beta}_{k+1} &= \langle f'(x_{k+1}) - f'(x_k) + \Delta_{k+1}, x_{k+1} - x_k \rangle \\ &\leq L_\nu r_{k+1}^{1+\nu} + \sigma_{k+1} r_{k+1}, \end{aligned}$$

where $f'(x_k) := \mathbb{E}_{\xi_k}[g_k] \in \partial f(x_k)$ and $\Delta_{k+1} := \delta_{k+1} - \delta_k$ with $\delta_k := g_k - f'(x_k)$ being the error of the stochastic gradient (such that $\mathbb{E}\|\delta_k\|^2 \leq \sigma^2$), and $\sigma_{k+1} := \|\Delta_{k+1}\|$. This gives

$$(H_{k+1} - H_k) D^2 = [L_\nu r_{k+1}^{1+\nu} + \sigma_{k+1} r_{k+1} - \tfrac{1}{2} H_{k+1} r_{k+1}^2]_+.$$

Eliminating r_{k+1} from this inequality (by maximizing the right-hand side in this variable), we get a certain recurrence for H_{k+1}, which is similar to (19) but with an additional $\frac{\sigma_{k+1}^2}{H_{k+1}}$ term in the right-hand side; carefully analyzing the resulting recurrence (see Lemma E.7), we get $H_k \leq O\big(\frac{L_\nu}{D^{1-\nu}} k^{(1-\nu)/2} + \frac{1}{D} (\sum_{i=1}^k \sigma_i^2)^{1/2}\big)$. This gives us

$$\mathbb{E}[H_k] \leq O\Big(\frac{L_\nu}{D^{1-\nu}} k^{(1-\nu)/2} + \frac{\sigma}{D} \sqrt{k}\Big)$$

after taking expectations. Substituting this bound into (27), we get the convergence result from Theorem 4.2.

4.3. Comparison with AdaGrad-type Methods

Let us compare the step-size coefficient H_k from Algorithm 2 with that of AdaGrad-type methods. Denote $\gamma_{k+1} := \|g_{k+1} - g_k\|_*$. From the definitions of $\hat{\beta}_{k+1}$ and r_{k+1}, it follows that $\hat{\beta}_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2 \leq \gamma_{k+1} r_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2 \leq \frac{\gamma_{k+1}^2}{2 H_{k+1}}$. Substituting this into the balance equation (26) (using the monotonicity of $[\cdot]_+$), we get

$$(H_{k+1} - H_k) D^2 \leq \frac{\gamma_{k+1}^2}{2 H_{k+1}}. \tag{28}$$

From this and $H_0 = 0$, it follows that (see Lemma E.4)

$$H_k \leq H_k' := \frac{1}{D} \Bigl(\sum_{i=1}^{k} \gamma_i^2 \Bigr)^{1/2}. \tag{29}$$

Note that H_k' is the step-size coefficient used by a variety of other AdaGrad-type algorithms[3] (see, e.g., (Kavis et al., 2019; Ene & Lê Nguyen, 2022)). Thus, the “step size” $\frac{1}{H_k}$ in our algorithm is at least as large as $\frac{1}{H_k'}$ used by AdaGrad.

In fact, the theoretical reasoning we used in Section 4.2 to arrive at our formula for the step-size coefficient, can be seen as a more precise theoretical analysis of the Stochastic Gradient Method with adaptive step sizes. Specifically, coming back to our preliminary recurrence (24), we see that AdaGrad first estimates $\hat{\beta}_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2 \leq \frac{\gamma_{k+1}^2}{2 H_{k+1}}$ and only then attempts to balance the terms. This corresponds to the idea of choosing the coefficient H_{k+1} in such a way that (28) becomes an identity (and then we not only have (29) but also the similar lower bound $H_k \geq \frac{1}{\sqrt{2}} H_k'$ (see Lemma E.5), which means that $H_k = \Theta(H_k')$). In contrast, our reasoning suggests that this extra estimation step is unnecessary.

5. Universal Fast Gradient Method for Stochastic Optimization

We now present, in Algorithm 3, the accelerated version of our universal stochastic method for solving problem (5). This algorithm is essentially one of the standard variants of the Fast Gradient Method known as the Method of Similar Triangles (see, e.g., Section 6.1.3 in (Nesterov, 2018)), which uses stochastic gradients instead of the exact ones and is equipped with our novel rule for adjusting the step-size coefficient H_k. The algorithm enjoys the following efficiency estimate (see Appendix C for the proof):

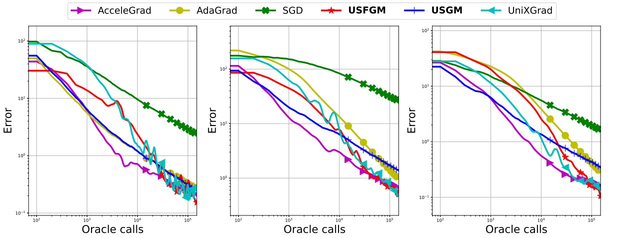
Theorem 5.1. *Let Algorithm 3 be applied to problem (5) under Assumptions 2.1, 2.2 and 4.1. Then, for all $k \geq 1$,*

$$\mathbb{E}[F(x_k)] - F^* \leq \inf_{\nu \in [0,1]} \frac{32 L_\nu D^{1+\nu}}{k^{(1+3\nu)/2}} + \frac{8 \sigma D}{\sqrt{3k}}.$$

To reach $\mathbb{E}[F(x_k)] - F^* \leq \epsilon$ for any $\epsilon > 0$, it then suffices to make $O\bigl(\inf_{\nu \in [0,1]} [\frac{L_\nu D^{1+\nu}}{\epsilon}]^{2/(1+3\nu)} + \frac{\sigma^2 D^2}{\epsilon^2}\bigr)$ oracle calls.

Algorithm 3 Universal Stochastic Fast Gradient Method

```
Initialize: $x_0 = v_0 \in \operatorname{dom} \psi$, $D > 0$, $H_0 := A_0 := 0$.
for $k = 0, 1, \ldots$ do
    $a_{k+1} := k + 1$,   $A_{k+1} := A_k + a_{k+1}$ $(> 0)$.
    $y_k := \frac{A_k}{A_{k+1}} x_k + \frac{a_{k+1}}{A_{k+1}} v_k$,   $g_k^y \sim \hat{g}(y_k)$.
    $v_{k+1} = \operatorname{argmin}_{x \in \operatorname{dom} \psi} \{ a_{k+1} [\langle g_k^y, x \rangle + \psi(x)] + \frac{H_k}{2} \|x - v_k\|^2 \}$
    $x_{k+1} := \frac{A_k}{A_{k+1}} x_k + \frac{a_{k+1}}{A_{k+1}} v_{k+1}$.
    $H_{k+1} := H_k + \frac{[A_{k+1} \hat{\beta}_{k+1} - \frac{1}{2} H_k r_{k+1}^2]_+}{D^2 + \frac{1}{2} r_{k+1}^2}$, where
    $r_{k+1} = \|v_{k+1} - v_k\|$, $\hat{\beta}_{k+1} = \langle g_{k+1}^x - g_k^y, x_{k+1} - y_k \rangle$
    with $g_{k+1}^x \sim \hat{g}(x_{k+1})$.
end for
```



(a) Least-square, diabetes. (b) Logistic regression, a1a. (c) Logistic regression, ionosphere.

Figure 1. Comparison of different stochastic algorithms on convex optimization problems.

In the deterministic case ($\sigma = 0$), the efficiency bound from Theorem 5.1 coincides with that of the Universal Fast Gradient Method from (Nesterov, 2015). It is worth mentioning that, in this case, instead of the symmetrized Bregman distance, we can use the standard one, $\hat{\beta}_{k+1} = \beta_f^{g_k^y}(y_k, x_{k+1})$, in Algorithm 3, and get similar convergence estimates but with slightly better absolute constants (see Theorem D.1).

6. Experiments

In this section, we present some preliminary computational experiments for the proposed methods.

6.1. Convex optimization

Least-Squares. Let us consider the following problem:

$$\min_{x \in \mathbb{R}^n} \Bigl\{ F(x) := \frac{1}{2} \|Ax - b\|^2 : \|x\| \leq 1 \Bigr\}, \tag{30}$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. We run the experiment on real-world diabetes dataset from LIBSVM[4]. In the stochastic

[3] The classical AdaGrad uses $\gamma_i = \|g_i\|_*$ but such a choice does not work properly for smooth constrained optimization when $\nabla f(x^*) \neq 0$.

[4] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

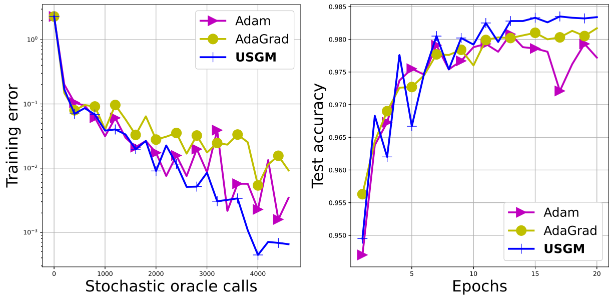


Figure 2. Comparison between the proposed universal stochastic gradient method, Adam, and AdaGrad in neural network training.

setting, we run the proposed USGM (Algorithm 2) and USFGM (Algorithm 3), and compare them against SGD, AdaGrad, UnixGrad (Kavis et al., 2019), and AcceleGrad (Levy et al., 2018). The result in Figure 1.(a) shows that the proposed method attains a convergence rate comparable to AdaGrad and AcceleGrad.

Logistic regression. We also focus on the logistic loss:

$$\min_{x\in\mathbb{R}^n}\Big\{F(x) := \sum_{i=1}^{m}\log\big(1+\exp(-b_i\langle a_i,x\rangle)\big) : \|x\|\le 1\Big\},$$

where $a_i \in \mathbb{R}^n$ is the feature vector and $b_i \in \{0,1\}$ is the label. We run the experiment on the a1a and ionosphere datasets from LIBSVM. The remaining setup is the same as in the case of Least-Squares. We present the result in Figure 1.(b-c), where ASUGM and SUGM are slightly faster than AdaGrad while performing similarly to UniXGrad.

6.2. Non-convex neural networks training

We now show that the proposed method can also be applied to non-convex neural network training. Specifically, we focus on classification tasks with the cross-entropy loss on MNIST dataset. A three-layer fully connected networks with layer dimensions $[28\times 28, 256, 256, 10]$ and ReLU activation function are selected. We compare the proposed method against AdaGrad and Adam. We select the mini-batch size as 256. Step-size of each method is tuned by a parameter sweep over $\{10, 1, 0.1, 0.01, 0.001, 0.0001\}$. The diameter of the proposed method is tuned by sweeping over $\{50, 35, 20, 10, 5\}$. We present the result in Figure 2, where we can see the proposed stochastic universal gradient method can solve non-convex problems as well.

7. Conclusion

We have proposed first universal gradient methods that are provably adaptive simultaneously to the noise level in the gradient feedback, the Hölder exponent and the associated Hölder constant of the objective function, while achieving optimal efficiency bounds. Unlike the majority of the works on adaptive methods relying on AdaGrad step-size constructions, our algorithm design is inspired by the line search approach. We have proposed a nonlinear balance equation for updating the step-size coefficient, which results in a tighter analysis of adaptive stochastic gradient algorithms compared to existing AdaGrad methods.

Note that our analysis exploits the fact that the feasible set has the bounded diameter D, knowledge of which is available to our algorithms. While this assumption may seem rather restrictive, it is nevertheless quite similar to the classical assumption on the knowledge of an upper bound R_0 on the distance $\|x_0 - x^*\|$ from the initial point to the minimizer. Indeed, if we know R_0, we can easily convert our original problem (5) into an equivalent one, $\min_{x\in \operatorname{dom}\psi_D}[f(x)+\psi_D(x)]$, where ψ_D is the restriction of ψ onto the ball $B_0 := \{x\in\mathbb{R}^n : \|x-x_0\|\le R_0\}$ or, in other words, the sum of ψ and the indicator function of B_0. For this new problem, we can run our methods with diameter $D = 2R_0$. The only detail that one needs to address is how to compute the proximal-point step for the function ψ_D via the corresponding operation for ψ. But this is usually not difficult and requires solving a certain one-dimensional nonlinear equation, which can be done very efficiently by Newton's method (at no extra queries to the stochastic gradient oracle). In some special cases, this equation can even be solved analytically, e.g., when the original problem is unconstrained, one simply needs to perform the projection on the B_0 ball. Nonetheless, it is still an interesting open question whether the same type of results can be obtained without this (somewhat artificial) replacement of the original problem. More importantly, it would be interesting to obtain parameter-free versions of our algorithms similar to (Carmon & Hinder, 2022; Ivgi et al., 2023), which could work with a sufficiently loose approximation of R_0.

Having that said, we would like to make a few remarks regarding the technical challenges involved in the design of optimal universal methods for stochastic optimization. Note that the existing accelerated adaptive methods for minimizing smooth convex functions (Levy et al., 2018; Kavis et al., 2019; Joulani et al., 2020), which assume bounded domains and use AdaGrad-inspired step-sizes, do not trivially extend to the Hölder class of functions. Essentially, they rely on the knowledge that the objective function is either Lipschitz smooth ($\nu = 1$) or Lipschitz continuous ($\nu = 0$), and the analysis is not directly compatible with intermediate values of $\nu \in (0,1)$. Our approach is based on different, more suitable, techniques and yields a new adaptive step-size schedule that enables fast universal rates in the noisy setting.

Acknowledgement

This work was supported by Hasler Foundation Program: Hasler Responsible AI (project number 21043). Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-24-1-0048. This work was supported by the Swiss National Science Foundation (SNSF) under grant number 200021_205011. Research of Ali Kavis is funded by the Swiss National Science Foundation (SNSF) under grant number P500PT_217942. The authors are also thankful to the four anonymous reviewers for their valuable feedback.

Impact Statement

This paper presents a theory work whose goal is to advance the field of Machine Learning and Optimization. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Armijo, L. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16:1–3, 1966.

Beck, A. and Teboulle, M. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Carmon, Y. and Hinder, O. Making SGD Parameter-Free. In Loh, P.-L. and Raginsky, M. (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pp. 2360–2389. PMLR, 2022.

Cauchy, A. Méthode générale pour la résolution des systèmes d'équations simultanées. *C.R. Acad. Sci. Paris*, 25:536–538, 1847.

Cutkosky, A. and Boahen, K. Online Learning Without Prior Information. In *Conference on Learning Theory*, pp. 643–677. PMLR, 2017.

Cutkosky, A. and Orabona, F. Black-Box Reductions for Parameter-free Online Learning in Banach Spaces. In *Conference On Learning Theory*, pp. 1493–1529. PMLR, 2018.

Defazio, A. and Mishchenko, K. Learning-Rate-Free Learning by D-Adaptation. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 7449–7479. PMLR, 2023.

Diakonikolas, J. and Orecchia, L. Accelerated Extra-Gradient Descent: A Novel Accelerated First-Order Method. In Karlin, A. R. (ed.), *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94, pp. 23:1–23:19. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018.

Duchi, J., Hazan, E., and Singer, Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12(7), 2011.

Ene, A. and Lê Nguyen, H. Adaptive and Universal Algorithms for Variational Inequalities with Optimal Convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6559–6567, 2022.

Ene, A., Nguyen, H. L., and Vladu, A. Adaptive Gradient Methods for Constrained Convex Optimization and Variational Inequalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7314–7321, 2021.

Gasnikov, A. V. and Nesterov, Y. E. Universal Method for Stochastic Composite Optimization Problems. *Computational Mathematics and Mathematical Physics*, 58(1): 48–64, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Ivgi, M., Hinder, O., and Carmon, Y. DoG is SGD's Best Friend: A Parameter-Free Dynamic Step Size Schedule. In *International Conference on Machine Learning*, pp. 14465–14499. PMLR, 2023.

Jacobsen, A. and Cutkosky, A. Unconstrained Online Learning with Unbounded Losses. In *International Conference on Machine Learning*, pp. 14590–14630. PMLR, 2023.

Joulani, P., Raj, A., Gyorgy, A., and Szepesvári, C. A simpler approach to accelerated optimization: iterative averaging meets optimism. In *International Conference on Machine Learning*, pp. 4984–4993. PMLR, 2020.

Kavis, A., Levy, K. Y., Bach, F., and Cevher, V. UniXGrad: A Universal, Adaptive Algorithm with Optimal Guarantees for Constrained Optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Khaled, A., Mishchenko, K., and Jin, C. DoWG Unleashed: An Efficient Universal Parameter-Free Gradient Descent Method. *Advances in Neural Information Processing Systems*, 36:6748–6769, 2023.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations*, 2015.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. 2009.

Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133:365–397, 2012.

Levy, K. Y., Yurtsever, A., and Cevher, V. Online adaptive methods, universality and acceleration. *Advances in Neural Information Processing Systems*, 31, 2018.

Li, T. and Lan, G. A simple uniformly optimal method without line search for convex optimization. *arXiv preprint arXiv:2310.10082*, 2023.

Malitsky, Y. and Mishchenko, K. Adaptive Gradient Descent without Descent. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6702–6712. PMLR, 2020.

Malitsky, Y. and Mishchenko, K. Adaptive proximal gradient method for convex optimization. *arXiv preprint arXiv:2308.02261*, 2023.

McMahan, B. and Streeter, M. No-Regret Algorithms for Unconstrained Online Convex Optimization. *Advances in Neural Information Processing Systems*, 25, 2012.

McMahan, H. B. and Streeter, M. Adaptive Bound Optimization for Online Convex Optimization. *COLT 2010*, pp. 244, 2010.

Mhammedi, Z. and Koolen, W. M. Lipschitz and Comparator-Norm Adaptivity in Online Learning. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pp. 2858–2887. PMLR, 2020.

Mishchenko, K. and Defazio, A. Prodigy: An Expeditiously Adaptive Parameter-Free Learner. *arXiv preprint arXiv:2306.06101*, 2023.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 19(4): 1574–1609, 2009. doi: 10.1137/070704277.

Nemirovsky, A. and Yudin, D. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

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

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103:127–152, 2005.

Nesterov, Y. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112:159–181, 2008.

Nesterov, Y. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152 (1-2):381–404, 2015.

Nesterov, Y. *Lectures on Convex Optimization*, volume 137. Springer, 2nd edition, 2018.

Nocedal, J. and Wright, S. *Numerical Optimization*. Springer Science & Business Media, 2nd edition, 2006.

Orabona, F. Simultaneous Model Selection and Optimization through Parameter-free Stochastic Learning. *Advances in Neural Information Processing Systems*, 27, 2014.

Orabona, F. Normalized Gradients for All. *arXiv preprint arXiv:2308.05621*, 2023.

Paquette, C. and Scheinberg, K. A Stochastic Line Search Method with Expected Complexity Analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020. doi: 10.1137/18M1216250.

Rakhlin, S. and Sridharan, K. Optimization, Learning, and Games with Predictable Sequences. *Advances in Neural Information Processing Systems*, 26, 2013.

Reddi, S., Kale, S., and Kumar, S. On the Convergence of Adam and Beyond. In *International Conference on Learning Representations*, 2018.

Robbins, H. and Monro, S. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3): 400 – 407, 1951. doi: 10.1214/aoms/1177729586.

Tieleman, T. and Hinton, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4 (2), 2012.

Tseng, P. On Accelerated Proximal Gradient Methods for Convex-Concave Optimization. *submitted to SIAM Journal on Optimization*, 2008.

Van Erven, T. and Koolen, W. M. MetaGrad: Multiple Learning Rates in Online Learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Wang, G., Lu, S., and Zhang, L. Adaptivity and Optimality: A Universal Algorithm for Online Convex Optimization. In Adams, R. P. and Gogate, V. (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115, pp. 659–668. PMLR, 2020.

Wang, J.-K. and Abernethy, J. D. Acceleration through Optimistic No-Regret Dynamics. *Advances in Neural Information Processing Systems*, 31, 2018.

Wolfe, P. Convergence Conditions for Ascent Methods. *SIAM Review*, 11(2):226–235, 1969. doi: 10.1137/1011036.

Xiao, L. Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization. *Journal of Machine Learning Research*, 11(88):2543–2596, 2010.

Yan, Y., Zhao, P., and Zhou, Z. Universal Online Learning with Gradient Variations: A Multi-layer Online Ensemble Approach. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, 2023.

Zhang, L., Wang, G., Yi, J., and Yang, T. A Simple yet Universal Strategy for Online Convex Optimization. In *International Conference on Machine Learning*, pp. 26605–26623. PMLR, 2022.

A. Proof of Theorem 3.1

Theorem 3.1. *Let Algorithm 1 be applied to problem (5) under Assumptions 2.1 and 2.2, and let x_k^* be the "best" iterate as defined in (12). Then, for all $k \geq 1$, we have*

$$F(x_k^*) - F^* \leq \inf_{\nu \in [0,1]} \frac{2L_\nu D^{1+\nu}}{k^{(1+\nu)/2}}.$$

To reach $F(x_k^*) - F^* \leq \epsilon$ for any $\epsilon > 0$, it thus suffices to make $\inf_{\nu \in [0,1]} [\frac{2L_\nu}{\epsilon}]^{2/(1+\nu)} D^2$ iterations.

Proof. i. We are going to prove that, for any $k \geq 1$,

$$F(x_k^*) - F^* \leq F(x_k^*) - \Phi_k^* \leq \frac{1}{k} \sum_{i=1}^{k} F(x_i) - \Phi_k^* \leq \frac{2H_k D^2}{k} \leq \inf_{\nu \in [0,1]} \frac{2L_\nu D^{1+\nu}}{k^{(1+\nu)/2}}. \tag{31}$$

where

$$\Phi_k^* \coloneqq \min_{x \in \operatorname{dom} \psi} \Bigl\{ \Phi_k(x) \coloneqq \frac{1}{k} \sum_{i=0}^{k-1} [f(x_i) + \langle g_i, x - x_i \rangle] + \psi(x) \Bigr\} \quad (\leq F^*). \tag{32}$$

(The inequality follows from the fact that $g_i \in \partial f(x_i)$ for all $i \geq 0$ and (5).)

Note that the function Φ_k, defined in (32), is the sum of an affine function and ψ. Since ψ is simple by our assumptions, we can easily compute its minimal value Φ_k^*. This value allows us to compute the quantities $\epsilon_k^* \coloneqq F(x_k^*) - \Phi_k^*$ and $\bar{\epsilon}_k \coloneqq \frac{1}{k} \sum_{i=1}^{k} F(x_i) - \Phi_k^*$, appearing in (31), and thus equip Algorithm 1 with a reliable stopping criterion $\epsilon_k^* \leq \epsilon$ (or $\bar{\epsilon}_k \leq \epsilon$) which guarantees that $F(x_k^*) - F^* \leq \epsilon$ for some given $\epsilon > 0$.

The first inequality in (31) follows from (32). The second one follows from the definition of x_k^* in (12).

ii. Let us prove the third inequality in (31).

For each $k \geq 0$, let $\zeta_k \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be the function

$$\zeta_k(x) \coloneqq f(x_k) + \langle g_k, x - x_k \rangle + \psi(x). \tag{33}$$

Let $k \geq 0$ and $x \in \operatorname{dom} \psi$ be arbitrary. By the definition of x_{k+1} at Line 4 and Lemma E.2, we have

$$\zeta_k(x) + \tfrac{1}{2} H_k \|x - x_k\|^2 \geq \zeta_k(x_{k+1}) + \tfrac{1}{2} H_k \|x_{k+1} - x_k\|^2 + \tfrac{1}{2} H_k \|x - x_k\|^2. \tag{34}$$

According to (33), (4) and (5),

$$\zeta_k(x_{k+1}) = f(x_k) + \langle g_k, x_{k+1} - x_k \rangle + \psi(x_{k+1}) = F(x_{k+1}) - \beta_f^{g_k}(x_k, x_{k+1}).$$

Substituting this into (34) and taking into account the definitions of r_{k+1} and β_{k+1}, we get

$$\begin{aligned} \zeta_k(x) + \tfrac{1}{2} H_k \|x - x_k\|^2 &\geq F(x_{k+1}) - \beta_f^{g_k}(x_k, x_{k+1}) + \tfrac{1}{2} H_k \|x_{k+1} - x_k\|^2 + \tfrac{1}{2} H_k \|x - x_k\|^2 \\ &= F(x_{k+1}) - \beta_{k+1} + \tfrac{1}{2} H_k r_{k+1}^2 + \tfrac{1}{2} H_k \|x - x_k\|^2. \end{aligned}$$

Consequently,

$$\begin{aligned} F(x_{k+1}) + \tfrac{1}{2} H_{k+1} \|x - x_{k+1}\|^2 \leq \zeta_k(x) + \tfrac{1}{2} H_k \|x - x_k\|^2 + [\beta_{k+1} - \tfrac{1}{2} H_{k+1} r_{k+1}^2] \\ + \tfrac{1}{2} (H_{k+1} - H_k)(\|x - x_{k+1}\|^2 + r_{k+1}^2). \end{aligned} \tag{35}$$

Note that, by construction, $H_k \leq H_{k+1}$. Also, in view of Assumption 2.1 (and the fact that $x_i \in \operatorname{dom} \psi$ for any $i \geq 0$), we have $r_{k+1} \leq D$ and $\|x - x_{k+1}\| \leq D$. Therefore,

$$\tfrac{1}{2} (H_{k+1} - H_k)(\|x - x_{k+1}\|^2 + r_{k+1}^2) \leq (H_{k+1} - H_k) D^2. \tag{36}$$

At the same time, by the definition of H_{k+1} at Line 5 and Lemma E.1, it satisfies the following equation:

$$(H_{k+1} - H_k) D^2 = [\beta_{k+1} - \tfrac{1}{2} H_{k+1} r_{k+1}^2]_+. \tag{37}$$

Therefore,

$$\beta_{k+1} - \tfrac{1}{2} H_{k+1} r_{k+1}^2 \leq [\beta_{k+1} - \tfrac{1}{2} H_{k+1} r_{k+1}^2]_+ = (H_{k+1} - H_k) D^2. \tag{38}$$

Substituting (36) and (38) into (35), we get

$$F(x_{k+1}) + \tfrac{1}{2} H_{k+1} \|x - x_{k+1}\|^2 \leq \zeta_k(x) + \tfrac{1}{2} H_k \|x - x_k\|^2 + 2(H_{k+1} - H_k) D^2. \tag{39}$$

Let $k \geq 1$ be arbitrary. Summing up (39) for all indices $0 \leq k' \leq k-1$ and using the fact that $H_0 = 0$, we obtain

$$\sum_{i=1}^{k} F(x_i) \leq \sum_{i=0}^{k-1} \zeta_i(x) + 2 H_k D^2 = k \Phi_k(x) + 2 H_k D^2,$$

where the final identity follows from (33) and (32) (and we have dropped the nonnegative term $\frac{1}{2} H_k \|x - x_k\|^2$ from the left-hand side). Since $x \in \operatorname{dom} \psi$ was arbitrary, this proves the second inequality in (31).

iii. It remains to estimate the rate of growth of the coefficients H_k.

Let $\nu \in [0, 1]$ be arbitrary such that $H_\nu < +\infty$. From (7) and the definitions of β_{k+1} and r_{k+1}, we obtain $\beta_{k+1} \leq \frac{L_\nu}{1+\nu} r_{k+1}^{1+\nu}$ for any $k \geq 0$. Hence, according to (37), for all $k \geq 0$, we have the following bound:

$$(H_{k+1} - H_k) D^2 \leq \Bigl[\frac{L_\nu}{1+\nu} r_{k+1}^{1+\nu} - \frac{1}{2} H_{k+1} r_{k+1}^2 \Bigr]_+.$$

Applying Lemma E.6 (with $\Omega := D^2$, $M := L_\nu$, $\gamma_k := 1$), we get, for all $k \geq 1$,

$$H_k \leq \Bigl[\frac{1}{(1+\nu) D^2} k \Bigr]^{(1-\nu)/2} L_\nu \leq \frac{L_\nu}{D^{1-\nu}} k^{(1-\nu)/2}.$$

Substituting this estimate into (31) and using the fact that $\nu \in [0, 1]$ was arbitrary, we obtain the final inequality in (31). □

B. Proof of Theorem 4.2

Theorem 4.2. *Let Algorithm 2 be applied to problem (5) under Assumptions 2.1, 2.2 and 4.1, and let $\bar{x}_k := \frac{1}{k} \sum_{i=1}^{k} x_i$ be the average iterate. Then, for all $k \geq 1$,*

$$\mathbb{E}[F(\bar{x}_k)] - F^* \leq \inf_{\nu \in [0,1]} \frac{8 L_\nu D^{1+\nu}}{k^{(1+\nu)/2}} + \frac{4 \sigma D}{\sqrt{k}}.$$

To reach $\mathbb{E}[F(\bar{x}_k)] - F^* \leq \epsilon$ for any $\epsilon > 0$, it then suffices to make $O\bigl(\inf_{\nu \in [0,1]} [\frac{L_\nu}{\epsilon}]^{2/(1+\nu)} D^2 + \frac{\sigma^2 D^2}{\epsilon^2}\bigr)$ oracle calls.

Proof. i. We will show that

$$\mathbb{E}[F(\bar{x}_k)] - F^* \leq \frac{2\, \mathbb{E}[H_k] D^2}{k} \leq \inf_{\nu \in [0,1]} \frac{8 L_\nu D^{1+\nu}}{k^{(1+\nu)/2}} + \frac{4 \sigma D}{\sqrt{k}}. \tag{40}$$

ii. Let $k \geq 0$ be arbitrary. Let $\zeta_k \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be the function

$$\zeta_k(x) := f(x_k) + \langle g_k, x - x_k \rangle + \psi(x). \tag{41}$$

Note that, by definition, $g_k = g(x_k, \xi_k)$, where ξ_k is independent of x_k. Therefore, in view of (22) and (5), in expectation, ζ_k is a global lower bound on the objective function F: for all $x \in \operatorname{dom} \psi$, we have

$$\mathbb{E}_{\xi_k}[\zeta_k(x)] = \mathbb{E}_{\xi_k}[f(x_k) + \langle f'(x_k), x - x_k \rangle] + \psi(x) \leq f(x) + \psi(x) = F(x). \tag{42}$$

Let $x \in \operatorname{dom} \psi$ be arbitrary. By the definition of x_{k+1} (Line 3) and Lemma E.2,

$$\zeta_k(x) + \tfrac{1}{2} H_k \|x - x_k\|^2 \geq \zeta_k(x_{k+1}) + \tfrac{1}{2} H_k \|x_{k+1} - x_k\|^2 + \tfrac{1}{2} H_k \|x - x_{k+1}\|^2. \tag{43}$$

According to (41) and (5), we have

$$\zeta_k(x_{k+1}) = f(x_k) + \langle g_k, x_{k+1} - x_k \rangle + \psi(x_{k+1}) = F(x_{k+1}) - \beta_{k+1}, \tag{44}$$

where $\beta_{k+1} \coloneqq f(x_{k+1}) - f(x_k) - \langle g_k, x_{k+1} - x_k \rangle$. Using (22), we can estimate

$$\beta_{k+1} \leq \langle f'(x_{k+1}) - g_k, x_{k+1} - x_k \rangle = \langle g_{k+1} - g_k, x_{k+1} - x_k \rangle + \Delta_{k+1}, \tag{45}$$

where $\Delta_{k+1} \coloneqq \langle f'(x_{k+1}) - g_{k+1}, x_{k+1} - x_k \rangle$. Recall that $g_{k+1} = g(x_{k+1}, \xi_{k+1})$ with ξ_{k+1} being independent of x_k and x_{k+1}. Therefore, according to (22),

$$\mathbb{E}_{\xi_{k+1}}[\Delta_{k+1}] = 0. \tag{46}$$

Substituting (44) and (45) into (43), we obtain

$$\begin{aligned} \zeta_k(x) + \tfrac{1}{2} H_k \|x - x_k\|^2 &\geq F(x_{k+1}) - \langle g_{k+1} - g_k, x_{k+1} - x_k \rangle - \Delta_{k+1} \\ &\qquad + \tfrac{1}{2} H_k \|x_{k+1} - x_k\|^2 + \tfrac{1}{2} H_k \|x - x_{k+1}\|^2 \\ &= F(x_{k+1}) - \hat{\beta}_{k+1} + \tfrac{1}{2} H_k r_{k+1}^2 + \tfrac{1}{2} H_k \|x - x_{k+1}\|^2 - \Delta_{k+1}, \end{aligned}$$

where the last identity is due to the definitions of r_{k+1} and β_{k+1}. Rearranging, we get

$$\begin{aligned} F(x_{k+1}) + \tfrac{1}{2} H_{k+1} \|x - x_{k+1}\|^2 &\leq \zeta_k(x) + \tfrac{1}{2} H_k \|x - x_k\|^2 + [\hat{\beta}_{k+1} - \tfrac{1}{2} H_{k+1} r_{k+1}^2] \\ &\quad + \tfrac{1}{2}(H_{k+1} - H_k)(\|x - x_{k+1}\|^2 + r_{k+1}^2) + \Delta_{k+1}. \end{aligned} \tag{47}$$

By construction, $H_k \leq H_{k+1}$ (Line 5). Also, in view of Assumption 2.1 (and the fact that $x_i \in \operatorname{dom} \psi$ for all $i \geq 0$), $\|x - x_{k+1}\| \leq D$ and $r_{k+1} \leq D$. Therefore,

$$\tfrac{1}{2}(H_{k+1} - H_k)(\|x - x_{k+1}\|^2 + r_{k+1}^2) \leq (H_{k+1} - H_k) D^2. \tag{48}$$

Further, by the definition of H_{k+1} at Line 5, it satisfies the following equation (see Lemma E.1):

$$(H_{k+1} - H_k) D^2 = [\hat{\beta}_{k+1} - \tfrac{1}{2} H_{k+1} r_{k+1}^2]_+. \tag{49}$$

Substituting (48) and (49) into (47), we obtain

$$F(x_{k+1}) + \tfrac{1}{2} H_{k+1} \|x - x_{k+1}\|^2 \leq \zeta_k(x) + \tfrac{1}{2} H_k \|x - x_k\|^2 + 2(H_{k+1} - H_k) D^2 + \Delta_{k+1}. \tag{50}$$

Let $k \geq 1$ be arbitrary. Summing up (50) for all indices $0 \leq k' \leq k - 1$ and using the fact that $H_0 = 0$, we get

$$\sum_{i=1}^{k} F(x_i) \leq \sum_{i=0}^{k-1} \zeta_i(x) + 2 H_k D^2 + \sum_{i=1}^{k} \Delta_i.$$

Hence, by the convexity of F and the definition of $\bar{x}_k$,

$$F(\bar{x}_k) \leq \frac{1}{k} \sum_{i=1}^{k} F(x_i) \leq \frac{1}{k} \sum_{i=0}^{k-1} \zeta_i(x) + \frac{2 H_k D^2}{k} + \frac{1}{k} \sum_{i=1}^{k} \Delta_i.$$

Passing to expectations and taking into account (42) and (46), we conclude that

$$\mathbb{E}[F(\bar{x}_k)] \leq F(x) + \frac{2\, \mathbb{E}[H_k] D^2}{k}.$$

This proves the first inequality in (40) since $x \in \operatorname{dom} \psi$ was arbitrary.

iii. Let us estimate the rate of growth of the coefficients H_k.

For each $k \geq 0$, denote

$$\delta_k := g_k - f'(x_k) = g(x_k, \xi_k) - f'(x_k). \tag{51}$$

Note that, according to (22) and (23), we have

$$\mathbb{E}_{\xi_k}[\delta_k] = 0, \qquad \mathbb{E}_{\xi_k}[\|\delta_k\|_*^2] \leq \sigma^2. \tag{52}$$

Let $\nu \in [0, 1]$ be arbitrary such that $L_\nu < +\infty$. Let $k \geq 0$ be arbitrary. By the definitions of $\hat{\beta}_{k+1}$ and r_{k+1}, and by (51), we have

$$\begin{aligned} \hat{\beta}_{k+1} &= \langle g_{k+1} - g_k, x_{k+1} - x_k \rangle = \langle f'(x_{k+1}) - f'(x_k), x_{k+1} - x_k \rangle + \langle \delta_{k+1} - \delta_k, x_{k+1} - x_k \rangle \\ &\leq \|f'(x_{k+1}) - f'(x_k)\|_* r_{k+1} + \|\delta_{k+1} - \delta_k\|_* r_{k+1} \leq L_\nu r_{k+1}^{1+\nu} + \sigma_{k+1} r_{k+1}, \end{aligned} \tag{53}$$

where $\sigma_{k+1} := \|\delta_{k+1} - \delta_k\|_*$, the first inequality is the Cauchy–Schwartz, and the final one is due to (6) and (22). Recall that ξ_{k+1} is independent of x_k and ξ_k. Hence, it is also independent of δ_k (see (51)). Therefore, according to (3) and (52), we have

$$\begin{aligned} \mathbb{E}_{\xi_k, \xi_{k+1}}[\sigma_{k+1}^2] &= \mathbb{E}_{\xi_k, \xi_{k+1}}[\|\delta_{k+1}\|_*^2 + \|\delta_k\|_*^2 + \langle \delta_{k+1}, B^{-1} \delta_k \rangle] \\ &= \mathbb{E}_{\xi_k}\big[\mathbb{E}_{\xi_{k+1}}[\|\delta_{k+1}\|_*^2] + \|\delta_k\|_*^2\big] \leq \mathbb{E}_{\xi_k}[\sigma^2 + \|\delta_k\|_*^2] \leq 2\sigma^2. \end{aligned} \tag{54}$$

In particular, the same inequality holds for the full expectation.

Substituting (53) into (49) (using the monotonicity of $[\cdot]_+$), we get, for any $k \geq 0$,

$$(H_{k+1} - H_k) D^2 \leq [L_\nu r_{k+1}^{1+\nu} + \sigma_{k+1} r_{k+1} - \tfrac{1}{2} H_{k+1} r_{k+1}^2]_+.$$

Applying Lemma E.7 (with $\Omega := D^2$, $L := L_\nu$, $\alpha_k := 1$, $\gamma_k := \sigma_k$), we obtain, for all $k \geq 1$,

$$H_k \leq [2(1+\nu)]^{(1+\nu)/2} \frac{L_\nu}{D^{1-\nu}} k^{(1-\nu)/2} + \Big(\frac{2}{D^2} \sum_{i=1}^k \sigma_i^2 \Big)^{1/2}.$$

By Jensen's inequality $\mathbb{E}[X^{1/2}] \leq (\mathbb{E}[X])^{1/2}$ and (54), for all $k \geq 1$, it holds

$$\mathbb{E}\Big[\Big(\frac{2}{D^2} \sum_{i=1}^k \sigma_i^2 \Big)^{1/2} \Big] \leq \Big(\frac{2}{D^2} \sum_{i=1}^k \mathbb{E}[\sigma_i^2] \Big)^{1/2} \leq \sqrt{\frac{2}{D^2} (2\sigma^2) k} = \frac{2\sigma}{D} \sqrt{k}.$$

Thus, for all $k \geq 1$,

$$\mathbb{E}[H_k] \leq [2(1+\nu)]^{(1+\nu)/2} \frac{L_\nu}{D^{1-\nu}} k^{(1-\nu)/2} + \frac{2\sigma}{D} \sqrt{k} \leq \frac{4 L_\nu}{D^{1-\nu}} k^{(1-\nu)/2} + \frac{2\sigma}{D} \sqrt{k}.$$

Substituting this estimate into (40) and taking into account the fact that $\nu \in [0, 1]$ was arbitrary, we obtain the final inequality in (40). □

C. Proof of Theorem 5.1

Theorem 5.1. *Let Algorithm 3 be applied to problem* (5) *under Assumptions 2.1, 2.2 and 4.1. Then, for all* $k \geq 1$,

$$\mathbb{E}[F(x_k)] - F^* \leq \inf_{\nu \in [0,1]} \frac{32 L_\nu D^{1+\nu}}{k^{(1+3\nu)/2}} + \frac{8\sigma D}{\sqrt{3k}}.$$

To reach $\mathbb{E}[F(x_k)] - F^* \leq \epsilon$ *for any* $\epsilon > 0$, *it then suffices to make* $O\big(\inf_{\nu \in [0,1]} [\frac{L_\nu D^{1+\nu}}{\epsilon}]^{2/(1+3\nu)} + \frac{\sigma^2 D^2}{\epsilon^2}\big)$ *oracle calls.*

Proof. i. We will show that

$$\mathbb{E}[F(x_k)] - F^* \leq \frac{2\,\mathbb{E}[H_k] D^2}{A_k} \leq \frac{4\,\mathbb{E}[H_k] D^2}{k(k+1)} \leq \inf_{\nu \in [0,1]} \frac{32 L_\nu D^{1+\nu}}{k^{(1+3\nu)/2}} + \frac{8\sigma D}{\sqrt{3k}}. \tag{55}$$

ii. For each $k \geq 0$, denote

$$\delta_k^y \coloneqq g_k^y - f'(y_k) \equiv g(y_k, \xi_k^y) - f'(y_k), \tag{56}$$

$$\delta_{k+1}^x \coloneqq g_{k+1}^x - f'(x_{k+1}) \equiv g(x_{k+1}, \xi_{k+1}^x) - f'(x_{k+1}), \tag{57}$$

where ξ_k^y, ξ_{k+1}^x, $k = 0, 1, \ldots$ are independent copies of the oracle's randomness.

Note that ξ_k^y (resp., ξ_{k+1}^y) is generated after (and independently) of y_k (resp., x_{k+1}). Therefore, according to (22) and (23),

$$\mathbb{E}_{\xi_k^y}[\delta_k^y] = 0, \qquad \mathbb{E}_{\xi_k^y}[\|\delta_k^y\|_*^2] \leq \sigma^2, \tag{58}$$

$$\mathbb{E}_{\xi_{k+1}^x}[\delta_{k+1}^x] = 0, \qquad \mathbb{E}_{\xi_{k+1}^x}[\|\delta_{k+1}^x\|_*^2] \leq \sigma^2. \tag{59}$$

iii. Let $k \geq 0$ be arbitrary. Let $\zeta_k \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be the global lower bound on the objective function F obtained by linearizing f at y_k:

$$\zeta_k(x) \coloneqq f(y_k) + \langle f'(y_k), x - y_k \rangle + \psi(x) \quad (\leq F(x)), \tag{60}$$

and let $\hat{\zeta}_k \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be its stochastic approximation:

$$\hat{\zeta}_k(x) \coloneqq f(y_k) + \langle g_k^y, x - y_k \rangle + \psi(x) = \zeta_k(x) + \Delta_k^y(x), \tag{61}$$

where

$$\Delta_k^y(x) \coloneqq \langle \delta_k^y, x - y_k \rangle. \tag{62}$$

Recall that ξ_k^y is independent of y_k. Therefore, for any (possibly random variable) $x \in \mathbb{R}^n$ that is also independent of ξ_k^y, we have

$$\mathbb{E}_{\xi_k^y}[\Delta_k^y(x)] = 0. \tag{63}$$

Let $x \in \operatorname{dom} \psi$ be an arbitrary (deterministic) point. Applying Lemma E.2 to the definition of v_{k+1} at Line 5 and taking into account (61), we obtain

$$\begin{aligned} a_{k+1}\hat{\zeta}_k(x) + \tfrac{1}{2}H_k\|x - v_k\|^2 &\geq a_{k+1}\hat{\zeta}_k(v_{k+1}) + \tfrac{1}{2}H_k\|v_{k+1} - v_k\|^2 + \tfrac{1}{2}H_k\|x - v_{k+1}\|^2 \\ &= a_{k+1}\hat{\zeta}_k(v_{k+1}) + \tfrac{1}{2}H_k\|x - v_{k+1}\|^2 + \tfrac{1}{2}H_k r_{k+1}^2, \end{aligned} \tag{64}$$

where the last identity follows from the definition of r_{k+1}.

In view of (60) and (61), we have

$$\begin{aligned} A_kF(x_k) + a_{k+1}\hat{\zeta}_k(v_{k+1}) &\geq A_k\zeta_k(x_k) + a_{k+1}\hat{\zeta}_k(v_{k+1}) \\ &= A_k\hat{\zeta}_k(x_k) + a_{k+1}\hat{\zeta}_k(v_{k+1}) - A_k\Delta_k^y(x_k) \geq A_{k+1}\hat{\zeta}_k(x_{k+1}) - A_k\Delta_k^y(x_k), \end{aligned} \tag{65}$$

where the last inequality follows from the convexity of $\hat{\zeta}_k$ and the definitions of x_{k+1} and A_{k+1} at Lines 6 and 3, respectively. According to (61) and (5),

$$\hat{\zeta}_k(x_{k+1}) = f(y_k) + \langle g_k^y, x_{k+1} - y_k \rangle + \psi(x_{k+1}) = F(x_{k+1}) - \beta_{k+1}, \tag{66}$$

where $\beta_{k+1} \coloneqq f(x_{k+1}) - f(y_k) - \langle g_k^y, x_{k+1} - y_k \rangle$. Using (22) and the definition of $\hat{\beta}_{k+1}$, we can estimate

$$\beta_{k+1} \leq \langle f'(x_{k+1}) - g_k^y, x_{k+1} - y_k \rangle = \hat{\beta}_{k+1} + \Delta_{k+1}^x, \tag{67}$$

where $\Delta_{k+1}^x \coloneqq \langle \delta_{k+1}^x, y_k - x_{k+1} \rangle$ (see (57)). Note that ξ_{k+1}^x is generated after (and independently) of x_{k+1} and y_k. Hence, in view of (59),

$$\mathbb{E}_{\xi_{k+1}^x}[\Delta_{k+1}^x] = 0. \tag{68}$$

Putting together (65)–(67) we obtain

$$\begin{aligned} A_kF(x_k) + a_{k+1}\hat{\zeta}_k(v_{k+1}) &\geq A_{k+1}[F(x_{k+1}) - \beta_{k+1}] - A_k\Delta_k^y(x_k) \\ &\geq A_{k+1}[F(x_{k+1}) - \hat{\beta}_{k+1}] - A_k\Delta_k^y(x_k) - A_{k+1}\Delta_{k+1}^x. \end{aligned}$$

Combining the above inequality with (64), we get

$$\begin{aligned}&A_k F(x_k) + a_{k+1}\hat{\zeta}_k(x) + \tfrac{1}{2}H_k\|x - v_k\|^2 \\ &\qquad \geq A_{k+1}[F(x_{k+1}) - \hat{\beta}_{k+1}] + \tfrac{1}{2}H_k r_{k+1}^2 + \tfrac{1}{2}H_k\|x - v_{k+1}\|^2 - A_k\Delta_k^y(x_k) - A_{k+1}\Delta_{k+1}^x.\end{aligned}$$

After rearranging, we can write

$$\begin{aligned}&A_{k+1}F(x_{k+1}) + \tfrac{1}{2}H_{k+1}\|x - v_{k+1}\|^2 \\ &\qquad \leq A_k F(x_k) + \tfrac{1}{2}H_k\|x - v_k\|^2 + a_{k+1}\hat{\zeta}_k(x) + [A_{k+1}\hat{\beta}_{k+1} - \tfrac{1}{2}H_{k+1}r_{k+1}^2] \\ &\qquad\quad + \tfrac{1}{2}(H_{k+1} - H_k)[\|x - v_{k+1}\|^2 + r_{k+1}^2] + A_k\Delta_k^y(x_k) + A_{k+1}\Delta_{k+1}^x.\end{aligned}$$

Note that, by construction, $H_k \leq H_{k+1}$ (Line 7). Further, in view of Assumption 2.1 (and the fact that $v_i \in \operatorname{dom}\psi$ for all $i \geq 0$), we have $\|x - v_{k+1}\| \leq D$ and $r_{k+1} \leq D$. Hence,

$$\tfrac{1}{2}(H_{k+1} - H_k)[\|x - v_{k+1}\|^2 + r_{k+1}^2] \leq (H_{k+1} - H_k)D^2.$$

On the other hand, from the definition of H_{k+1} at Line 7 and Lemma E.1 (with $\beta := A_{k+1}\hat{\beta}_{k+1}$, $\rho := \frac{1}{2}r_{k+1}^2$, $\Omega := D^2$), it follows that

$$(H_{k+1} - H_k)D^2 = [A_{k+1}\hat{\beta}_{k+1} - \tfrac{1}{2}H_{k+1}r_{k+1}^2]_+. \tag{69}$$

Combining the above three displays, we obtain

$$\begin{aligned}&A_{k+1}F(x_{k+1}) + \tfrac{1}{2}H_{k+1}\|x - v_{k+1}\|^2 \\ &\qquad \leq A_k F(x_k) + \tfrac{1}{2}H_k\|x - v_k\|^2 + a_{k+1}\hat{\zeta}_k(x) + 2(H_{k+1} - H_k)D^2 + A_k\Delta_k^y(x_k) + A_{k+1}\Delta_{k+1}^x.\end{aligned} \tag{70}$$

Note that this inequality is valid for any $k \geq 0$.

Let $k \geq 1$ be arbitrary. Summing up (70) for all indices $0 \leq k' \leq k - 1$ and taking into account that $H_0 = A_0 = 0$ (by definition), we get

$$A_k F(x_k) \leq \sum_{i=0}^{k-1} a_{i+1}\hat{\zeta}_i(x) + 2H_k D^2 + \sum_{i=0}^{k-1}(A_i\Delta_i^y(x_i) + A_{i+1}\Delta_{i+1}^x),$$

where we have additionally dropped the nonnegative term $\frac{1}{2}H_k\|x - v_k\|^2$ from the left-hand side. Combining this with (61) and (60), we obtain

$$\begin{aligned}A_k F(x_k) &\leq \sum_{i=0}^{k-1} a_{i+1}[\zeta_i(x) + \Delta_i^y(x)] + 2H_k D^2 + \sum_{i=0}^{k-1}(A_i\Delta_i^y(x_i) + A_{i+1}\Delta_{i+1}^x) \\ &\leq A_k F(x) + 2H_k D^2 + \sum_{i=0}^{k-1}(a_{i+1}\Delta_i^y(x) + A_i\Delta_i^y(x_i) + A_{i+1}\Delta_{i+1}^x),\end{aligned}$$

where we have used the fact that $A_k = \sum_{i=1}^k a_i$ (see Line 3).

Observe that, by definitions at Line 3, the coefficients a_i and A_i are deterministic for each $i \geq 0$. Also, recall that x is assumed to be deterministic as well. Therefore, passing to expectations in the above inequality, we get

$$A_k\,\mathbb{E}[F(x_k)] \leq A_k F(x) + 2\,\mathbb{E}[H_k]D^2 + \sum_{i=0}^{k-1}(a_{i+1}\,\mathbb{E}[\Delta_i^y(x)] + A_i\,\mathbb{E}[\Delta_i^y(x_i)] + A_{i+1}\,\mathbb{E}[\Delta_{i+1}^x]).$$

Note that, for any $i \geq 0$, the random variable ξ_i^y is generated after x_i, and hence they are independent. Therefore, according to (62) and (68), for each $i \geq 0$, we have $\mathbb{E}[\Delta_i^y(x)] = \mathbb{E}[\Delta_i^y(x_i)] = \mathbb{E}[\Delta_{i+1}^x] = 0$. Thus, the above display reads

$$A_k\,\mathbb{E}[F(x_k)] \leq A_k F(x) + 2\,\mathbb{E}[H_k]D^2.$$

This proves the first inequality in (55) since $x \in \operatorname{dom}\psi$ was arbitrary.

iv. From the definitions at Line 3 and the fact that $A_0 = 0$, it follows that

$$A_k = \sum_{i=1}^{k} a_i = \sum_{i=1}^{k} i = \tfrac{1}{2} k(k+1) \quad (\geq \tfrac{1}{2} k^2) \tag{71}$$

for any $k \geq 1$. This proves the second inequality in (55).

v. To prove the final inequality in (55), it remains to estimate the expected growth rate of regularization parameters H_k.

Let $\nu \in [0,1]$ be arbitrary such that $L_\nu < +\infty$. Let $k \geq 0$ be arbitrary. According to the definition of $\hat{\beta}_{k+1}$ and (56) and (57), we have

$$\begin{aligned} \hat{\beta}_{k+1} &= \langle g^x_{k+1} - g^y_k, x_{k+1} - y_k \rangle = \langle f'(x_{k+1}) - f'(y_k), x_{k+1} - y_k \rangle + \langle \delta^x_{k+1} - \delta^y_k, x_{k+1} - y_k \rangle \\ &\leq \| f'(x_{k+1}) - f'(y_k) \|_* \| x_{k+1} - y_k \| + \sigma_{k+1} \| x_{k+1} - y_k \| \leq L_\nu \| x_{k+1} - y_k \|^{1+\nu} + \sigma_{k+1} \| x_{k+1} - y_k \|, \end{aligned} \tag{72}$$

where $\sigma_{k+1} \coloneqq \| \delta^x_{k+1} - \delta^y_k \|_*$, the first inequality is the Cauchy–Schwarz inequality, and the final one is due to (6) and (22). Recall that δ^y_k is a function of y_k and ξ^y_k (see (56)), and ξ^x_{k+1} is generated after y_k and ξ^y_k. Therefore, ξ^x_{k+1} is independent of δ^y_k. Hence, according to (3), (58) and (59),

$$\begin{aligned} \mathbb{E}_{\xi^y_k, \xi^x_{k+1}}[\sigma^2_{k+1}] &= \mathbb{E}_{\xi^y_k, \xi^x_{k+1}}[\| \delta^x_{k+1} \|^2_* + \| \delta^y_k \|^2_* + 2 \langle \delta^x_{k+1}, B^{-1} \delta^y_k \rangle] \\ &= \mathbb{E}_{\xi^y_k} \big[\mathbb{E}_{\xi^x_{k+1}}[\| \delta^x_{k+1} \|^2_*] + \| \delta^y_k \|^2_* \big] \leq \mathbb{E}_{\xi^y_k}[\sigma^2 + \| \delta^y_k \|^2_*] \leq 2\sigma^2. \end{aligned} \tag{73}$$

Further, by the definitions of y_k and x_{k+1} at Lines 4 and 6, $x_{k+1} - y_k = \frac{a_{k+1}}{A_{k+1}}(v_{k+1} - v_k)$, which means that $\| x_{k+1} - y_k \| = \frac{a_{k+1}}{A_{k+1}} r_{k+1}$. Substituting this into (72) and using the definition of a_{k+1} at Line 3 together with (71), we obtain

$$\begin{aligned} A_{k+1} \hat{\beta}_{k+1} &\leq A_{k+1} \Big[L_\nu \Big(\frac{a_{k+1}}{A_{k+1}} r_{k+1} \Big)^{1+\nu} + \sigma_{k+1} \frac{a_{k+1}}{A_{k+1}} r_{k+1} \Big] = L_\nu \frac{a^{1+\nu}_{k+1}}{A^\nu_{k+1}} r^{1+\nu}_{k+1} + a_{k+1} \sigma_{k+1} r_{k+1} \\ &\leq L_\nu \frac{(k+1)^{1+\nu}}{[(\frac{1}{2}(k+1)^2]^\nu} r^{1+\nu}_{k+1} + (k+1) \sigma_{k+1} r_{k+1} = 2^\nu L_\nu (k+1)^{1-\nu} r^{1+\nu}_{k+1} + (k+1) \sigma_{k+1} r_{k+1}. \end{aligned}$$

Combining the above inequality with (69) (using the monotonicity of $[\cdot]_+$), we come to the following recurrence relation:

$$(H_{k+1} - H_k) D^2 \leq [2^\nu L_\nu (k+1)^{1-\nu} r^{1+\nu}_{k+1} + (k+1) \sigma_{k+1} r_{k+1} - \tfrac{1}{2} H_{k+1} r^2_{k+1}]_+,$$

which is valid for any $k \geq 0$.

Let $k \geq 1$ be arbitrary. Applying Lemma E.7 (with $\Omega \coloneqq D^2$, $L \coloneqq 2^\nu L_\nu$, $\alpha_k \coloneqq k$ and $\gamma_k \coloneqq k \sigma_k$), we conclude that

$$\begin{aligned} H_k &\leq [2(1+\nu)]^{(1+\nu)/2} 2^\nu L_\nu \Big(\frac{1}{D^2} \sum_{i=1}^{k} i^2 \Big)^{(1-\nu)/2} + \Big(\frac{2}{D^2} \sum_{i=1}^{k} (i \sigma_i)^2 \Big)^{1/2} \\ &= 2^{(1+3\nu)/2} (1+\nu)^{(1+\nu)/2} \frac{L_\nu}{D^{1-\nu}} \Big(\sum_{i=1}^{k} i^2 \Big)^{(1-\nu)/2} + \Big(\frac{2}{D^2} \sum_{i=1}^{k} i^2 \sigma_i^2 \Big)^{1/2}. \end{aligned}$$

Note that, by Jensen's inequality $\mathbb{E}[X^{1/2}] \leq (\mathbb{E}[X])^{1/2}$ and (73),

$$\mathbb{E}\Big[\Big(\frac{2}{D^2} \sum_{i=1}^{k} i^2 \sigma_i^2 \Big)^{1/2} \Big] \leq \Big(\frac{2}{D^2} \sum_{i=1}^{k} i^2 \, \mathbb{E}[\sigma_i^2] \Big)^{1/2} \leq \frac{2\sigma}{D} \Big(\sum_{i=1}^{k} i^2 \Big)^{1/2}.$$

Thus,

$$\begin{aligned} \mathbb{E}[H_k] &\leq 2^{(1+3\nu)/2} (1+\nu)^{(1+\nu)/2} \frac{L_\nu}{D^{1-\nu}} \Big(\sum_{i=1}^{k} i^2 \Big)^{(1-\nu)/2} + \frac{2\sigma}{D} \Big(\sum_{i=1}^{k} i^2 \Big)^{1/2} \\ &\leq 2^{(1+3\nu)/2} (1+\nu)^{(1+\nu)/2} \frac{L_\nu}{D^{1-\nu}} \Big(\frac{1}{3} k(k+1)^2 \Big)^{(1-\nu)/2} + \frac{2\sigma}{D} \Big(\frac{1}{3} k(k+1)^2 \Big)^{1/2} \\ &= \frac{2^{(1+3\nu)/2} (1+\nu)^{(1+\nu)/2}}{3^{(1-\nu)/2}} \frac{L_\nu}{D^{1-\nu}} k^{(1-\nu)/2} (k+1)^{1-\nu} + \frac{2}{\sqrt{3}} \frac{\sigma}{D} \sqrt{k}\, (k+1) \\ &\leq \frac{8 L_\nu}{D^{1-\nu}} k^{(1-\nu)/2} (k+1)^{1-\nu} + \frac{2}{\sqrt{3}} \frac{\sigma}{D} \sqrt{k}\, (k+1), \end{aligned}$$

where we have used the fact that $\sum_{i=1}^k i^2 = \frac{1}{6}k(k+1)(2k+1) \le \frac{1}{3}k(k+1)^2$. Consequently,

$$\frac{4\,\mathbb{E}[H_k]D^2}{k(k+1)} \le 32L_\nu D^{1+\nu}\frac{k^{(1-\nu)/2}(k+1)^{1-\nu}}{k(k+1)} + \frac{8\sigma D}{\sqrt{3}}\frac{\sqrt{k}\,(k+1)}{k(k+1)}$$
$$= \frac{32L_\nu D^{1+\nu}}{k^{(1+\nu)/2}(k+1)^\nu} + \frac{8\sigma D}{\sqrt{3k}} \le \frac{32L_\nu D^{1+\nu}}{k^{(1+3\nu)/2}} + \frac{8\sigma D}{\sqrt{3k}}.$$

This proves the final inequality in (55) since $\nu \in [0,1]$ was arbitrary. □

D. Universal Line-Search-Free Fast Gradient Method

Theorem D.1. *Let Algorithm 3 be applied for solving problem* (5) *under Assumptions 2.1 and 2.2 with the deterministic gradient oracle $g(x,\xi) \equiv \nabla f(x)$ and with $\hat{\beta}_{k+1} := \beta_f^{g_k^y}(y_k, x_{k+1})$ (Line 7) at each iteration $k \ge 0$. Then, for all $k \ge 1$, it holds that*

$$F(x_k) - F^* \le \frac{4H_k D^2}{k(k+1)} \le \inf_{\nu\in[0,1]} \frac{8L_\nu D^{1+\nu}}{k^{(1+3\nu)/2}}. \tag{74}$$

Proof. We proceed exactly in the same way as in the proof of Theorem 5.1 (Appendix C) but do not upper bound $\beta_{k+1} = \beta_f(y_k, x_{k+1})$ with $\hat{\beta}_{k+1}$ in (67). We then arrive, exactly as before, at the following inequality that holds for any $k \ge 1$:

$$F(x_k) - F^* \le \frac{4H_k D^2}{k(k+1)}. \tag{75}$$

To upper bound H_k, we use, as before, the following equation:

$$(H_{k+1} - H_k)D^2 = [A_{k+1}\beta_{k+1} - \tfrac{1}{2}H_{k+1}r_{k+1}^2]_+,$$

that holds for any $k \ge 0$, but now we can upper bound

$$\beta_{k+1} \le \frac{L_\nu}{1+\nu}A_{k+1}\|x_{k+1} - y_k\|^{1+\nu}.$$

This is essentially the same bound that we had in (72) with the formal change of L_ν to $L'_\nu := \frac{L_\nu}{1+\nu}$. Proceeding exactly as before, we then obtain

$$A_{k+1}\beta_{k+1} \le \frac{2^\nu L_\nu}{1+\nu}(k+1)^{1-\nu}r_{k+1}^{1+\nu}$$

for any $k \ge 0$ and arbitrary $\nu \in [0,1]$, which gives us

$$(H_{k+1} - H_k)D^2 \le \Big[\frac{2^\nu L_\nu}{1+\nu}(k+1)^{1-\nu}r_{k+1}^{1+\nu} - \frac{H_{k+1}}{2}r_{k+1}^2\Big]_+.$$

Instead of Lemma E.7, we can now apply a slightly more precise result (in terms of absolute constants)—Lemma E.6 (with $\Omega := D^2$, $M := 2^\nu L_\nu$, $\gamma_k := k$)—to conclude that, for all $k \ge 1$,

$$H_k \le \Big[\frac{1}{(1+\nu)D^2}\sum_{i=1}^k i^2\Big]^{(1-\nu)/2} 2^\nu L_\nu \le 2^\nu\Big[\frac{1}{3(1+\nu)}k(k+1)^2\Big]^{(1-\nu)/2}\frac{L_\nu}{D^{1-\nu}} \le \frac{2L_\nu}{D^{1-\nu}}k^{(1-\nu)/2}(k+1)^{1-\nu},$$

where the second inequality is due to $\sum_{i=1}^k i^2 = \frac{1}{6}k(k+1)(2k+1) \le \frac{1}{3}k(k+1)^2$, and the final inequality follows from the fact that $2^\nu/[3(1+\nu)]^{(1-\nu)/2}$ monotonically increases in $\nu \in [0,1]$. Substituting the above bound into (75), we get

$$F(x_k) - F^* \le 2L_\nu D^{1+\nu}\frac{k^{(1-\nu)/2}(k+1)^{1-\nu}}{k(k+1)} = \frac{2L_\nu D^{1+\nu}}{k^{(1+\nu)/2}(k+1)^\nu} \le \frac{2L_\nu D^{1+\nu}}{k^{(1+3\nu)/2}}.$$ □

E. Auxiliary Results

Lemma E.1. *Let $H, \beta, \rho \geq 0$ and $\Omega > 0$. Then, the equation*

$$(H_+ - H)\Omega = [\beta - H_+\rho]_+ \tag{76}$$

has a unique solution given by

$$H_+ := H + \frac{[\beta - H\rho]_+}{\Omega + \rho}. \tag{77}$$

Proof. Denote the left- and right-hand sides in (76) (as functions of H_+) by $\zeta_1(H_+)$ and $\zeta_2(H_+)$, respectively, and let $\zeta(H_+) := \zeta_1(H_+) - \zeta_2(H_+)$. Note that both ζ_1 and ζ_2 are continuous functions, ζ_1 is strictly increasing, while ζ_2 is decreasing, hence ζ is a continuous strictly increasing function. When $H_+ = H$, we have $\zeta_1(H) = 0$, while $\zeta_2(H) \geq 0$, hence $\zeta(H) \leq 0$. When $H_+ \to +\infty$, $\zeta_1(H_+)$ tends to $+\infty$, while $\zeta_2(H_+)$ tends to a finite number (either 0 if $\rho > 0$, or β if $\rho = 0$), hence $\zeta(H_+)$ tends to $+\infty$. Thus, there exists a unique point $H_+ \geq H$ such that $\zeta(H_+) = 0$. This point is exactly the unique solution of equation (76).

It remains to show that (77) is indeed a solution to (76). But this is simple. Indeed, if $\beta \leq H\rho$, then, according to (77), $H_+ = H + (\beta - H\rho)/(\Omega + \rho)$, and hence

$$\beta - H_+\rho = \beta - H\rho - \frac{\beta - H\rho}{\Omega + \rho}\rho = \frac{\Omega}{\Omega + \rho}(\beta - H\rho) = (H_+ - H)\Omega \quad (\geq 0),$$

which means that H_+ satisfies (76). If $\beta > H\rho$, then, by (77), $H_+ = H$, and hence

$$[\beta - H_+\rho]_+ = [\beta - H\rho]_+ = 0 = (H_+ - H)\Omega,$$

which also means that H_+ satisfies (76). □

Lemma E.2. *Let $\zeta\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function, $\bar{x} \in \operatorname{dom} \zeta$, $H \geq 0$. Then, for any $x^* \in \operatorname{Argmin}_{x \in \operatorname{dom} \zeta}\{\zeta(x) + \frac{1}{2}H\|x - \bar{x}\|^2\}$ and any $x \in \operatorname{dom} \zeta$, we have*

$$\zeta(x) + \tfrac{1}{2}H\|x - \bar{x}\|^2 \geq \zeta(x^*) + \tfrac{1}{2}H\|x^* - \bar{x}\|^2 + \tfrac{1}{2}H\|x - x^*\|^2.$$

Proof. This is a standard result that can be seen as a consequence of the fact that $\zeta_H(x) := \zeta(x) + \frac{H}{2}\|x - \bar{x}\|^2$ is a strongly convex function with constant H, and hence $\zeta_H(x) \geq \zeta_H(x^*) + \frac{H}{2}\|x - x^*\|^2$ for any $x \in \operatorname{dom} \zeta$. □

Lemma E.3. *Let $\nu \in [0, 1)$, $M \geq 0$ and $H > 0$. Then,*

$$\max_{r \geq 0}\Bigl\{\frac{M}{1 + \nu}r^{1+\nu} - \frac{H}{2}r^2\Bigr\} = \frac{1 - \nu}{2(1 + \nu)}\frac{M^{2/(1-\nu)}}{H^{(1+\nu)/(1-\nu)}}. \tag{78}$$

Proof. After the change of variables $t = r^{1+\nu}$, the objective function inside the $\max$ becomes concave in t (since $r^2 = t^{2/(1+\nu)}$ with $\frac{2}{1+\nu} \geq 1$). Computing its derivative and setting to zero, we see that the maximum is attained at the point $r_* := (M/H)^{1/(1-\nu)}$. Thus,

$$\begin{aligned}\max_{r \geq 0}\Bigl\{\frac{M}{1 + \nu}r^{1+\nu} - \frac{H}{2}r^2\Bigr\} &= \frac{M}{1 + \nu}\Bigl(\frac{M}{H}\Bigr)^{(1+\nu)/(1-\nu)} - \frac{H}{2}\Bigl(\frac{M}{H}\Bigr)^{2/(1-\nu)} \\ &= \frac{1}{1 + \nu}\frac{M^{2/(1-\nu)}}{H^{(1+\nu)/(1-\nu)}}\Bigl(1 - \frac{1}{2}(1 + \nu)\Bigr) = \frac{1 - \nu}{2(1 + \nu)}\frac{M^{2/(1-\nu)}}{H^{(1+\nu)/(1-\nu)}}.\end{aligned}$$

□

Lemma E.4. *Let $(H_k)_{k=0}^{\infty}$ be a nonnegative nondecreasing sequence of reals such that, for any $k \geq 0$,*

$$(p + 1)H_{k+1}^p(H_{k+1} - H_k) \leq \alpha_{k+1},$$

where $p \geq 0$ is real and $(\alpha_k)_{k=1}^{\infty}$ is a nonnegative sequence of reals. Then, for any $k \geq 1$, it holds that

$$H_k \leq \Bigl(H_0^{p+1} + \sum_{i=1}^{k}\alpha_i\Bigr)^{1/(p+1)}.$$

Proof. Since $H_k \le H_{k+1}$ for any $k \ge 0$ and $p \ge 0$, we can estimate

$$\alpha_{k+1} \ge (p+1) H_{k+1}^p (H_{k+1} - H_k) \ge (p+1) \int_{H_k}^{H_{k+1}} t^p dt = H_{k+1}^{p+1} - H_k^{p+1}.$$

Telescoping these inequalities, we obtain, for any $k \ge 1$,

$$H_k^{p+1} - H_0^{p+1} \le \sum_{i=1}^{k} \alpha_i,$$

and the claim follows. □

Lemma E.5. *Let $(H_k)_{k=0}^{\infty}$ be a nonnegative nondecreasing sequence of reals such that, for any $k \ge 0$,*

$$H_{k+1}(H_{k+1} - H_k) \ge \alpha_{k+1},$$

where $(\alpha)_{k=1}^{\infty}$ is a nonnegative sequence of reals. Then, for any $k \ge 0$, it holds that

$$H_k \ge \Bigl(H_0^2 + \sum_{i=1}^{k} \alpha_i \Bigr)^{1/2}.$$

Proof. Indeed, for any $k \ge 0$, we can estimate

$$\alpha_{k+1} \le H_{k+1}(H_{k+1} - H_k) \le (H_{k+1} + H_k)(H_{k+1} - H_k) = H_{k+1}^2 - H_k^2.$$

Summing up these inequalities and rearranging, we obtain the claim. □

Lemma E.6. *Let $(H_k)_{k=0}^{\infty}$ be a nondecreasing sequence such that $H_0 = 0$ and, for all $k \ge 0$, it holds*

$$(H_{k+1} - H_k)\Omega \le \Bigl[\frac{1}{1+\nu} M \gamma_{k+1}^{1-\nu} r_{k+1}^{1+\nu} - \frac{1}{2} H_{k+1} r_{k+1}^2 \Bigr]_+, \tag{79}$$

where $\Omega > 0$, $M \ge 0$, $\nu \in [0,1]$ are certain constants, and $(\gamma_k)_{k=1}^{\infty}$ and $(r_k)_{k=1}^{\infty}$ are certain positive and nonnegative sequences, respectively. Then, for all $k \ge 1$, we have

$$H_k \le \Bigl[\frac{1}{(1+\nu)\Omega} \sum_{i=1}^{k} \gamma_i^2 \Bigr]^{(1-\nu)/2} M. \tag{80}$$

Proof. Suppose $\nu = 1$. Then, according to (79), for all $k \ge 0$, we have

$$(H_{k+1} - H_k)\Omega \le [\tfrac{1}{2} M r_{k+1}^2 - \tfrac{1}{2} H_{k+1} r_{k+1}^2]_+ = [M - H_{k+1}]_+ \tfrac{1}{2} r_{k+1}^2. \tag{81}$$

Since $H_0 = 0 \le M$, this implies that $H_k \le M$ for all $k \ge 0$ (which is exactly (80) for $\nu = 1$). Indeed, if $H_k \le M < H_{k+1}$ for some $k \ge 0$, then the left-hand side in (81) is strictly positive, while the right-hand side is zero, which is a contradiction.

From now on, suppose $\nu < 1$. Without loss of generality, we can assume that $H_{k+1} > 0$ for all $k \ge 0$. Let $k \ge 0$ be arbitrary. Applying Lemma E.3 to bound the right-hand side in (79) (and using the monotonicity of $[\cdot]_+$), we obtain

$$(H_{k+1} - H_k)\Omega \le \frac{1-\nu}{2(1+\nu)} \frac{(M\gamma_{k+1}^{1-\nu})^{2/(1-\nu)}}{H_{k+1}^{(1+\nu)/(1-\nu)}} = \frac{1-\nu}{2(1+\nu)} \frac{M^{2/(1-\nu)}}{H_{k+1}^{(1+\nu)/(1-\nu)}} \gamma_{k+1}^2.$$

Applying Lemma E.4 (with $p = \frac{1+\nu}{1-\nu}$ for which $p + 1 = \frac{2}{1-\nu}$) and using the fact that $H_0 = 0$, we conclude that

$$H_k \le \Bigl[\frac{M^{2/(1-\nu)}}{2(1+\nu)\Omega} \sum_{i=1}^{k} \gamma_i^2 \Bigr]^{(1-\nu)/2} = \Bigl[\frac{1}{(1+\nu)\Omega} \sum_{i=1}^{k} \gamma_i^2 \Bigr]^{(1-\nu)/2} M. \qquad □$$

Lemma E.7. *Let $(H_k)_{k=0}^{\infty}$ be a nondecreasing sequence such that $H_0 = 0$ and, for all $k \geq 0$, it holds*

$$(H_{k+1} - H_k)\Omega \leq [L\alpha_{k+1}^{1-\nu} r_{k+1}^{1+\nu} + \gamma_{k+1} r_{k+1} - \tfrac{1}{2} H_{k+1} r_{k+1}^2]_+, \tag{82}$$

where $\Omega > 0$, $M \geq 0$, $\nu \in [0, 1]$ are certain constants, $(\alpha_k)_{k=1}^{\infty}$ is a certain positive sequence, and $(r_k)_{k=1}^{\infty}$ and $(\gamma_k)_{k=1}^{\infty}$ are certain nonnegative sequences. Then, for all $k \geq 1$,

$$H_k \leq [2(1+\nu)]^{(1+\nu)/2} L \Big(\frac{1}{\Omega} \sum_{i=1}^{k} \alpha_i^2 \Big)^{(1-\nu)/2} + \Big(\frac{2}{\Omega} \sum_{i=1}^{k} \gamma_i^2 \Big)^{1/2}. \tag{83}$$

Remark E.8. Setting $\gamma_k \equiv 0$ in Lemma E.7, we recover Lemma E.6.

Proof. i. Without loss of generality, we can assume that $H_{k+1} > 0$ for all $k \geq 0$. Indeed, otherwise, either $H_k = 0$ for all $k \geq 0$, and (83) is trivial, or we can work with the subsequence $(H_k)_{k=k_0}^{\infty}$, where $k_0 \geq 0$ is the first integer such that $H_{k_0+1} > 0$.

ii. Suppose[5] $\nu = 1$. In this case, (82) reads

$$(H_{k+1} - H_k)\Omega \leq [(L - \tfrac{1}{2} H_{k+1}) r_{k+1}^2 + \gamma_{k+1} r_{k+1}]_+ \tag{84}$$

for all $k \geq 0$, and we need to prove that, for all $k \geq 1$,

$$H_k \leq 4L + \Big(\frac{2}{\Omega} \sum_{i=1}^{k} \gamma_i^2 \Big)^{1/2}. \tag{85}$$

(This is exactly (83) for $\nu = 1$.)

Since $H_0 = 0$, we can assume that there exists an index $k_0 \geq 0$ such that

$$H_{k_0} \leq 4L < H_{k_0+1}. \tag{86}$$

(Otherwise, $H_k \leq 4L$ for all $k \geq 0$, and (85) is trivial.) As $(H_k)_{k=0}^{\infty}$ is nondecreasing, (85) is clearly valid for all indices $0 \leq k \leq k_0$. Let us prove that it is also valid for all $k \geq k_0 + 1$.

Let $k \geq k_0$ be arbitrary. By monotonicity of $(H_i)_{i=0}^{\infty}$, from (86), it follows that $H_{k+1} \geq H_{k_0+1} > 4L$. Therefore,

$$\Big(L - \frac{1}{2} H_{k+1} \Big) r_{k+1}^2 + \gamma_{k+1} r_{k+1} \leq \gamma_{k+1} r_{k+1} - \frac{1}{4} H_{k+1} r_{k+1}^2 \leq \frac{\gamma_{k+1}^2}{H_{k+1}},$$

where the final inequality follows from Lemma E.3 (with $\nu := 0$ and $H := \frac{1}{2} H_{k+1}$). Combining this with (84) (using the monotonicity of $[\cdot]_+$), we get

$$(H_{k+1} - H_k)\Omega \leq \frac{\gamma_{k+1}^2}{H_{k+1}}.$$

Thus, for all $k \geq k_0$, we have

$$(H_{k+1}^2 - H_k^2)\Omega \leq 2H_{k+1}(H_{k+1} - H_k)\Omega \leq 2\gamma_{k+1}^2. \tag{87}$$

(Recall that $H_k \leq H_{k+1}$.)

Let $k \geq k_0 + 1$ be arbitrary. Summing up (87) for all indices $k_0 \leq k' \leq k - 1$ and rearranging, we get

$$H_k^2 \leq H_{k_0}^2 + \frac{2}{\Omega} \sum_{i=k_0+1}^{k} \gamma_i^2 \leq (4L)^2 + \frac{2}{\Omega} \sum_{i=1}^{k} \gamma_i^2,$$

where the last inequality is due to (86) (and the fact that $k_0 \geq 0$). Using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$, we obtain (85).

[5] In principle, we can cover the case $\nu = 1$ by only considering the values of $\nu \in [0, 1)$ and then passing to the limit as $\nu \to 1$. However, we prefer to present a more explicit proof without using the limiting argument.

iii. Now suppose $\nu < 1$. Let $k \geq 0$ be arbitrary. Applying Lemma E.3 twice, we obtain

$$\begin{aligned} L\alpha_{k+1}^{1-\nu} r_{k+1}^{1+\nu} + \gamma_{k+1} r_{k+1} - \tfrac{1}{2} H_{k+1} r_{k+1}^2 &= [L\alpha_{k+1}^{1-\nu} r_{k+1}^{1+\nu} - \tfrac{1}{4} H_{k+1} r_{k+1}^2] + [\gamma_{k+1} r_{k+1} - \tfrac{1}{4} H_{k+1} r_{k+1}^2] \\ &\leq \frac{1-\nu}{2(1+\nu)} \frac{[(1+\nu) L \alpha_{k+1}^{1-\nu}]^{2/(1-\nu)}}{(\frac{1}{2} H_{k+1})^{(1+\nu)/(1-\nu)}} + \frac{1}{2} \frac{\gamma_{k+1}^2}{\frac{1}{2} H_{k+1}} = (1-\nu) \frac{M^{2/(1-\nu)}}{H_{k+1}^{(1+\nu)/(1-\nu)}} \alpha_{k+1}^2 + \frac{\gamma_{k+1}^2}{H_{k+1}}, \end{aligned}$$

where

$$M \coloneqq \frac{2^{(1+\nu)/2}(1+\nu)}{[2(1+\nu)]^{(1-\nu)/2}} L = 2^\nu (1+\nu)^{(1+\nu)/2} L. \tag{88}$$

Combining this with (82) (using the monotonicity of $[\cdot]_+$), we get

$$(H_{k+1} - H_k)\Omega \leq (1-\nu) \frac{M^{2/(1-\nu)}}{H_{k+1}^{(1+\nu)/(1-\nu)}} \alpha_{k+1}^2 + \frac{\gamma_{k+1}^2}{H_{k+1}}.$$

Since $H_k \leq H_{k+1}$, it follows that

$$\frac{1}{2}(H_{k+1}^2 - H_k^2)\Omega \leq H_{k+1}(H_{k+1} - H_k)\Omega \leq (1-\nu) \frac{M^{2/(1-\nu)}}{H_{k+1}^{2\nu/(1-\nu)}} \alpha_{k+1}^2 + \gamma_{k+1}^2.$$

Note that this inequality is valid for all $k \geq 0$.

Applying Lemma E.9 (with $C_k \coloneqq H_k^2$, $\alpha'_k \coloneqq \frac{2}{\Omega}\alpha_k^2$ and $\gamma'_k \coloneqq \frac{2}{\Omega}\gamma_k^2$), we conclude that, for all $k \geq 1$,

$$\begin{aligned} H_k^2 &\leq M^2 \Bigl(\sum_{i=1}^k \frac{2}{\Omega} \alpha_i^2 \Bigr)^{1-\nu} + \sum_{i=1}^k \frac{2}{\Omega} \gamma_i^2 = 2^{2\nu}(1+\nu)^{1+\nu} L^2 \Bigl(\sum_{i=1}^k \frac{2}{\Omega} \alpha_i^2 \Bigr)^{1-\nu} + \frac{2}{\Omega} \sum_{i=1}^k \gamma_i^2 \\ &= [2(1+\nu)]^{1+\nu} L^2 \Bigl(\frac{1}{\Omega} \sum_{i=1}^k \alpha_i^2 \Bigr)^{1-\nu} + \frac{2}{\Omega} \sum_{i=1}^k \gamma_i^2, \end{aligned}$$

where the second identity follows from (88). Using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$, we obtain (83). □

Lemma E.9. *Let $(C_k)_{k=1}^\infty$ be a positive sequence satisfying, for all $k \geq 0$,*

$$C_{k+1} - C_k \leq (1-\nu) \frac{M^{2/(1-\nu)}}{C_{k+1}^{\nu/(1-\nu)}} \alpha_{k+1} + \gamma_{k+1}, \tag{89}$$

where $C_0 \coloneqq 0$, and $M \geq 0$, $\nu \in [0, 1)$ are certain constants, and $(\alpha_k)_{k=1}^\infty$ and $(\gamma_k)_{k=1}^\infty$ are certain nonnegative sequences. Then, for all $k \geq 1$, we have

$$C_k \leq M^2 \Bigl(\sum_{i=1}^k \alpha_i \Bigr)^{1-\nu} + \sum_{i=1}^k \gamma_i. \tag{90}$$

Proof. For each $k \geq 0$, let $\hat{C}_k$ be the right-hand side of (90):

$$\hat{C}_k \coloneqq M^2 A_k^{1-\nu} + \sum_{i=1}^k \gamma_i, \qquad A_k \coloneqq \sum_{i=1}^k \alpha_i, \tag{91}$$

with the convention that $\hat{C}_0 = A_0 = 0$. Note that $\hat{C}_k > 0$ for all $k \geq 1$. Indeed, if $\hat{C}_k = 0$ for some $k \geq 1$, then $\hat{C}_1 = 0$ (by the monotonicity of $(\hat{C}_k)_{k=1}^\infty$), which means that $\gamma_1 = 0$ and either $M = 0$ or $\alpha_1 = 0$; but then, according to (89), $C_1 - C_0 \leq 0$; since $C_0 = 0$, this implies $C_1 \leq 0$, which contradicts our assumption about the positivity of $(C_k)_{k=1}^\infty$.

Let us prove by induction that $C_k \leq \hat{C}_k$ for all $k \geq 0$. Clearly, this inequality is satisfied for $k = 0$ since $C_0 = \hat{C}_0 = 0$. Now suppose that $C_k \leq \hat{C}_k$ for some $k \geq 0$, and let us prove that $C_{k+1} \leq \hat{C}_{k+1}$.

Let $\chi_{k+1} \colon (0, +\infty) \to \mathbb{R}$ be the function

$$\chi_{k+1}(C) \coloneqq C - (1-\nu) \frac{M^{2/(1-\nu)}}{C^{\nu/(1-\nu)}} \alpha_{k+1}. \tag{92}$$

According to (89) and the inductive hypothesis, we have

$$\chi_{k+1}(C_{k+1}) \leq C_k + \gamma_{k+1} \leq \hat{C}_k + \gamma_{k+1}. \tag{93}$$

Since the function χ_{k+1} is strictly increasing, to prove that $C_{k+1} \leq \hat{C}_{k+1}$, it suffices to show that $\chi_{k+1}(C_{k+1}) \leq \chi_{k+1}(\hat{C}_{k+1})$. According to (93), for this, it suffices to show that

$$\hat{C}_k + \gamma_{k+1} \leq \chi_{k+1}(\hat{C}_{k+1}).$$

Substituting (92) and rearranging, we see that we need to prove that $(\hat{C}_i)_{i=0}^{\infty}$ satisfies (90) with the reversed sign:

$$\hat{C}_{k+1} - \hat{C}_k \geq (1-\nu) \frac{M^{2/(1-\nu)}}{\hat{C}_{k+1}^{\nu/(1-\nu)}} \alpha_{k+1} + \gamma_{k+1}.$$

In view of (91), we have

$$\hat{C}_{k+1} - \hat{C}_k = M^2 [A_{k+1}^{1-\nu} - A_k^{1-\nu}] + \gamma_{k+1}.$$

Thus, we need to check if

$$M^2 [A_{k+1}^{1-\nu} - A_k^{1-\nu}] \geq (1-\nu) \frac{M^{2/(1-\nu)}}{\hat{C}_{k+1}^{\nu/(1-\nu)}} \alpha_{k+1},$$

or, equivalently, if

$$\hat{C}_{k+1}^{\nu/(1-\nu)} [A_{k+1}^{1-\nu} - A_k^{1-\nu}] \geq (1-\nu) M^{2\nu/(1-\nu)} \alpha_{k+1}.$$

From (91), it follows that $\hat{C}_{k+1} \geq M^2 A_{k+1}^{1-\nu}$. Hence,

$$\hat{C}_{k+1}^{\nu/(1-\nu)} [A_{k+1}^{1-\nu} - A_k^{1-\nu}] \geq M^{2\nu/(1-\nu)} A_{k+1}^{\nu} [A_{k+1}^{1-\nu} - A_k^{1-\nu}].$$

Thus, it suffices to show that

$$A_{k+1}^{\nu} [A_{k+1}^{1-\nu} - A_k^{1-\nu}] \geq (1-\nu) \alpha_{k+1}.$$

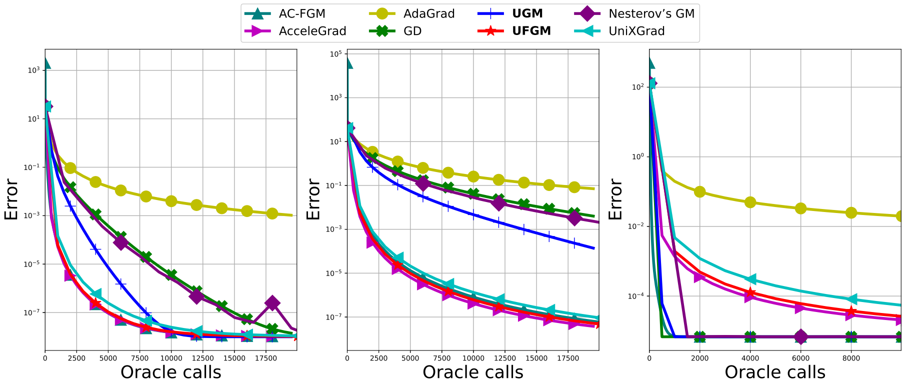
But this is indeed true, as for any $0 \leq t_1 \leq t_2$, by the concavity of $t \mapsto t^{1-\nu}$, we have $t_2^{\nu}(t_2^{1-\nu} - t_1^{1-\nu}) \geq (1-\nu)(t_2 - t_1)$, while $A_{k+1} - A_k = \alpha_{k+1}$ according to (91). □

F. Additional Related Work

Within the context of Problem (5), we most commonly consider that f and ψ are both convex and ψ is a *simple*, non-smooth function, such that we could solve (5) efficiently by means of a *proximity* function.

Classical methods: Focusing on the setting where ψ is the indicator function of a compact set Q, we can solve the problem at a rate of $O(1/\sqrt{k})$ when f is non-smooth while we can accelerate the convergence to $O(1/k^2)$ when f has Lipschitz continous gradients, i.e., f is smooth. These rates are shown to be tight when the gradient feedback is noiseless (Nemirovsky & Yudin, 1983). When the first-order oracle is stochastic with variance σ^2, the lower bounds imply a convergence rate of $O(\sigma/\sqrt{k})$ (Nemirovski et al., 2009; Lan, 2012).

The simple (sub-)gradient descent (GD) (Cauchy, 1847), i.e., $x_{k+1} = x_k - \gamma_k g(x_k)$, with a sufficiently small γ_0 that decays as $O(1/\sqrt{k})$ achieves $O(1/\sqrt{k})$ rate for non-smooth minimization. Although this rate matches the information theoretic lower bounds, the same method converges at an $O(1/k)$ rate under smoothness, which is suboptimal. Nesterov (1983) introduced the idea of "momentum" and proposed the first order-optimal algorithm, accelerated gradient descent (AGD), which manages to decrease the error at a rate of $O(1/k^2)$. Since then, various different interpretations of Nesterov's acceleration has been proposed. For a broad review of acceleration mechanisms, we refer the reader to Nesterov (2005); Xiao (2010); Tseng (2008); Beck & Teboulle (2009); Diakonikolas & Orecchia (2018); Wang & Abernethy (2018) and references therein.



(a) Synthetic, n=100, m=200. (b) Synthetic, n=200, m=400. (c) Diabetes.

Figure 3. Comparison of different deterministic algorithms on convex optimization problems.

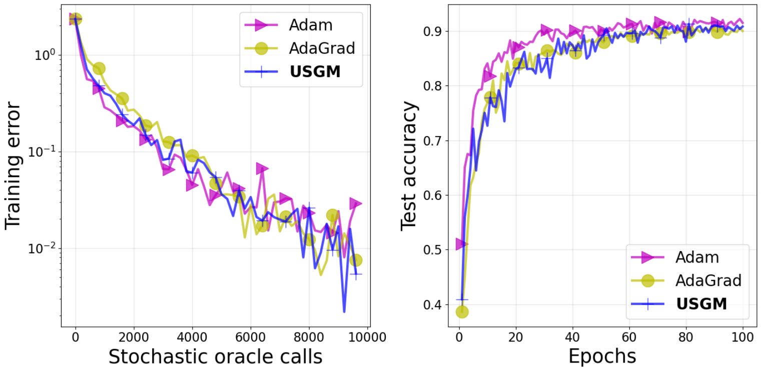


Figure 4. Comparison of different stochastic algorithms on non-convex optimization problems.

An integral components of the classical methods, such as GD, AGD and its variants, is the dependence on the knowledge of problem parameters, specifically the Lipschitz constant of the problem. When the step-size is not selected sufficiently small with respect to the Lipschitz constant, then these methods are destined to diverge. Similar arguments hold for stochastic methods such that the initial step-size needs to be sufficiently small to guarantee convergence for smooth problems (Nemirovsky & Yudin, 1983). Additionally, the step-size must decay optimally at the rate of $O(1/\sqrt{k})$, irrespective of the smoothness of the problem, to control the effect of noise and ensure convergence to the set of solutions (Robbins & Monro, 1951).

Line-search methods: A fundamental technique to overcome the dependence on problem parameters is the line-search machinery (Armijo, 1966; Wolfe, 1969; Nocedal & Wright, 2006), which dynamically selects step-size every iteration by using local information. There are several strategies such as exact line-search and backtracking line-research, which could be implemented with appropriate "sufficient decrease" and curvature conditions. Essentially, line-search helps estimate a *locally-valid* step-size, enabling larger step-sizes than using the globally worst-case Lipschitz constant. When equipped with an appropriate line-search mechanism, GD and AGD could achieve the same convergence rates without the need to know the Lipschitz parameter. However, this comes at the expense of an iterative search procedure which demands function value evaluations per iteration of the line-search subroutine. Similarly, stochastic variants of line-search are available, nonetheless, they enforce extra assumptions on the objective and gradient information (Paquette & Scheinberg, 2020).

G. Additional Experiments

In this section, we first elaborate on our experiments in the deterministic setting. We focus on the least-square problem in (30). We first run the experiment on real-world diabetes dataset from LIBSVM. Next, we consider a synthetic dataset, where we randomly generate an optimal solution x^* in the surface of a unit ball. Next, we sample each element of A from a uniform distribution over $[0, 1]$ and set $b = Ax^*$. We test the proposed Algorithm 1, denoted by UGM, and its accelerated version (AUGM). The baselines include GD, Nesterov's GM (Nesterov, 2015), AdaGrad, UnixGrad (Kavis et al., 2019), AcceleGrad (Levy et al., 2018) and AC-FGM (Li & Lan, 2023). In GD, we set the step size as $1/L$ while other methods are tuned via grid search. The result is presented in Figure 3, where we observe that the proposed UGM shows better performance than UniXGrad and AcceleGrad.

Next, we include additional experiments on the stochastic setting. To be specific, we train a ResNet18 (He et al., 2016) on CIFAR-10 (Krizhevsky & Hinton, 2009). We select the mini-batch size as 512. The step size of each method is tuned by a parameter sweep over $\{10, 1, 0.1, 0.01, 0.001, 0.0001\}$. The diameter of the proposed method is tuned by sweeping over $\{50, 35, 20, 10, 5\}$. We show the result in Figure 4, where we can observe that the proposed stochastic universal gradient method can be applied on non-convex problems as well.