A First Approach to Universal Second-Order Acceleration for Convex Minimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

This work proposes a universal and adaptive second-order method for minimizing second-order smooth, convex functions. Our algorithm achieves $O(\sigma/\sqrt{T})$ convergence when the oracle feedback is stochastic with variance σ^2, and improves its convergence to $O(1/T^3)$ with deterministic oracles, where T is the number of iterations. Our method also interpolates these rates without knowing the nature of the oracle apriori, which is enabled by a parameter-free adaptive step-size that is oblivious to the knowledge of smoothness modulus, variance bounds and the diameter of the constrained set. To our knowledge, this is the first universal algorithm with such global guarantees within the second-order optimization literature.

1 Introduction

Over the last few decades, first-order (convex) minimization methods have gained popularity for modern machine learning and optimization problems due to their efficient per-iteration cost and *global convergence* properties. The literature on first-order methods is rather dense and extensive with a concrete, thorough understanding of the optimal *global* convergence behavior. Focusing on the more relevant settings of smooth, convex minimization, the lower bounds have been well-established; $O(\sigma/\sqrt{T})$ when the gradient feedback is noisy with variance σ^2, and $O(1/T^2)$ under deterministic first-order oracles [39, 44]. Under slight variations of the aforementioned problem setting, there exists an extensive amount of work that enjoys the latter, "accelerated" rate [42, 43, 40, 50, 52, 29, 2, 31, 51, 13, 12, 26, 22].

On the contrary to its first-order analogue, the literature on *global convergence* of second-order, smooth methods is notably sparse with many open questions standing even in the simplest problem formulations. Following the pioneering works of Bennett [6], Kantorovich [24], Newton's method and its variations [30, 33] are considered as the staple of second-order methods in optimization. Although its powerful local convergence behavior has been repeatedly demonstrated [11, 28], studies on its global behavior are relatively limited. Prior attempts at tackling global convergence mostly make additional structural assumptions on the objective function [47, 34, 28] or assume extra regularity conditions on the Hessian [25] beyond the simplest smooth and convex setting. Over the last decade, we have witnessed important progress towards a more complete theory of globally-convergent second-order methods (more on this shortly), and yet there remains many important questions unanswered, which we will delve into in this paper.

To motivate the perspective in our technical endeavour, we take a small detour to introduce the idea of *universality*, which we particularly characterize as *adaptation to the level of noise in oracle feedback*. Enabled by the recent advances in online optimization, universal first-order algorithms essentially

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

attain the $O(\sigma/\sqrt{T} + 1/T^2)$ convergence, interpolating between stochastic and deterministic rates. There exist a plethora of algorithms that enjoy this rate under different sets of assumptions [29, 26, 16, 22] and we have a quite solid understanding of this optimal behavior. However, we observe that such universal results do not exist in second-order literature, hence, it is only natural to ask,

Can we design a simple second-order method that will achieve accelerated universal rates beyond $O(\sigma/\sqrt{T} + 1/T^2)$?

More recently, global sub-linear convergence rates for second-order methods have been characterized by [45] for second-order smooth and convex setting. Essentially, the so-called Cubic Regularized Newton's Method combines the quadratic Taylor approximation in the typical Newton update with a cubic regularization term. At the expense of solving a cubic problem, this method achieves $O(1/T^2)$ convergence rate. Shortly after, Nesterov [41] proposes an accelerated version of the cubic regularization idea with $O(1/T^3)$ value convergence, pioneering a new direction of research in the study of globally-convergent second-order methods [36]. This idea has been studied further for different settings in convex optimization [19, 20] with the same accelerated $O(1/T^3)$ rate and extended to non-convex realm [8, 9], obtaining the analogous rates of $O(1/T^{2/3})$ and $O(1/T^{1/3})$ for finding first-order and second-order stationary points, respectively, leading the way for further investigations [5, 15, 10].

Notice that accelerated cubic regularization is *sub*-optimal such that recent studies prove a respective lower-bound for second-order smooth, convex problems as $O(1/T^{7/2})$ [1, 4]. A different line of research, which we will call "bisection-type" methods, shrinks the gap between the upper and lower bounds for achieving an *almost*-optimal (more on this shortly) convergence [46]. Pioneered by Monteiro and Svaiter [37], these class of algorithms propose a conceptual method where the step-size of the algorithm *implicitly* depends on the next iterate. To resolve, the authors propose a bisection procedure that simultaneously finds a step-size/next iterate pair that satisfies the conditions of the iterative update, which enables the convergence rate of $O(1/T^{7/2})$, modulo the complexity of bisection procedure. This idea was very recently generalized for higher-order tensor methods [17]. Not so surprisingly, the same construction finds application in variational inequality (VI) and min-max optimization literature [7, 21].

Although there are promising results with an increasing interest into second-order –and also higher-order– methods, we identify three main shortcomings in the literature, which we will systematically address in the sequel. First, bisection-type methods achieve the optimal convergence rate however, the search procedure is computationally very prohibitive [46, 32] and the resulting algorithms are complicated with many interconnected components. Indeed, finding a computationally-efficient, optimal second-order method is still an open problem according to [46]. On the other hand, cubic regularization-based ideas propose a simple construction that achieves acceleration beyond $O(1/T^2)$ however, similar to bisection methods, they either require the knowledge of smoothness constant or need to execute a standard line-search procedure to estimate it locally. A common drawback for both approaches is that the algorithmic constructions are designed for handling *only* deterministic oracle information and it is an open question whether such frameworks could accommodate stochastic first and second-order information.

Our contributions: To address the aforementioned issues, we developed the first universal and adaptive second-order algorithm, EXTRANEWTON, for convex minimization. We summarize our contributions as follows:

1. We prove the global convergence rate of EXTRANEWTON that interpolates simultaneously between $\mathcal{O}(LD^2/T^3)$ for deterministic and $\mathcal{O}(\sigma/\sqrt{T})$ for stochastic oracles, achieving the first universal convergence result in the literature.
2. Our method is completely oblivious to any problem-dependent parameters including smoothness modulus, variance bounds on stochastic oracles, diameter of the constraint set and any possible bounds on the gradient and Hessian.

Table 1: A survey on first and second-order algorithms with key properties

	AGD [42]	**UniXGrad** [26]	**Reg. Newton** [36]	**Accelerated Cubic Reg.** [41]	**ANPE**[1] [37]	**ExtraNewton [ours]**
Convergence Rate	$\frac{1}{T^2}$	$\frac{\sigma}{\sqrt{T}} + \frac{1}{T^2}$	$\frac{1}{T^2}$	$\frac{1}{T^3}$	$\frac{1}{T^{7/2}}$	$\frac{\sigma}{\sqrt{T}} + \frac{1}{T^3}$
Bisection	✗	✗	✗	✗	✓	✗
Adapts to L	✗	✓	✗	Partial	✗	✓
Adapts to Oracle	✗	✓	✗	✗	✗	✓

3. We design the first adaptive step-size, in the sense of [14, 49], that successfully incorporates second-order information "on-the-fly". While doing so, we bypass any bisection or linesearch procedure, and propose a simple, intuitive algorithmic framework.

From a technical point of view, what will allow us to achieve these results is the combination of three principal ingredients: *(i)* proposing appropriate adjustments to Extra-Gradient [27] that was originally designed for solving variational inequalities and min/max problems; *(ii)* an "optimistic" weighted iterate averaging scheme accompanied by an appropriate gradient rescaling strategy in the spirit of [51, 13, 26] which allows us to obtain an accelerated rate of convergence by means of a generalized online-to-batch conversion (Theorem 3.3), and *(iii)* the glue that holds these elements together is an adaptive learning rate inspired by [49, 26, 3] which automatically rescales aggregated gradients and second order information. In what follows, we shall explicate these arguments.

2 Problem setup

Throughout the sequel, we will be focusing on solving (constrained) convex minimization problems of the general form:

$$\begin{aligned} \text{minimize} \quad & f(x), \\ \text{subject to} \quad & x \in \mathcal{X}. \end{aligned} \tag{Opt}$$

Formally, in the above $\mathcal{X}$ is a convex and compact subset of a d- dimensional normed space $\mathcal{V} \cong \mathbb{R}^d$, and $f : \mathcal{V} \to \mathbb{R} \cup \{+\infty\}$ is a proper, lower semi-continuous, convex function with $\mathrm{dom} f = \{x \in \mathbb{R}^d : f(x) < +\infty\} \subset \mathcal{X}$. To that end, we make the following blanket assumptions for (Opt). First in order to avoid trivialities we assume that the said problem admits at least a solution, i.e.

Assumption 2.1. *The solution set* $\mathcal{X}^* = \arg\min_{x \in \mathcal{X}} f(x)$ *of* (Opt) *is non-empty.*

Furthermore, we assume that there exists a Lipschitz continuous selection $x \mapsto \nabla^2 f(x) \in \mathbb{R}^{d \times d}$, i.e.,

$$\|\nabla^2 f(x) - \nabla^2 f(x')\| \leq L\|x - x'\| \quad \forall x, x' \in \mathcal{X} \tag{H-smooth}$$

and in addition it satisfies the second order approximation:

$$f(x) = f(x') + \langle \nabla f(x'), x - x' \rangle + \frac{1}{2}\langle \nabla^2 f(x')(x - x'), x - x' \rangle + O\left(\|x - x'\|^3\right) \tag{Taylor}$$

To that end, combining (H-smooth) and (Taylor) we readily get the following inequality:

$$\|\nabla f(x) - \nabla f(x') - \nabla^2 f(x')(x - x')\| \leq \frac{L}{2}\|x - x'\|^2 \tag{1}$$

The above equivalences are well-established and hence we omit their proofs (we defer for a panoramic view to [53])

Oracle feedback structure From an algorithmic point of view, we aim to solve (Opt) by using methods that require access to a (stochastic) first and second order-oracle. Before we move forward with the methodology,we shall introduce the definitions and notations for this oracle model which

[1]Note that the bisection procedure is computationally prohibitive, we defer the reader to [46], p.304-305.

we will use in algorithm definitions and technical discussions. For notational simplicity we denote stochastic gradients by $g(\cdot)$ and stochastic Hessian by $H(\cdot)$ such that

$$\begin{aligned} \mathbb{E}\left[\, g(x) \mid x \,\right] &= \nabla f(x), & \mathbb{E}\left[\, \|g(x) - \nabla f(x)\|^2 \mid x \,\right] &\leq \sigma_g^2 \\ \mathbb{E}\left[\, H(x) \mid x \,\right] &= \nabla^2 f(x), & \mathbb{E}\left[\, \|H(x) - \nabla^2 f(x)\|^2 \mid x \,\right] &\leq \sigma_H^2 \end{aligned} \tag{2}$$

Due to space constraints, we will also define an operator that accommodates second-order information and its respective stochastic counterpart.

$$\begin{aligned} \mathbf{F}(x; x') &= \nabla f(x') + \nabla^2 f(x')(x - x') \\ \tilde{\mathbf{F}}(x; x') &= g(x') + H(x')(x - x') \end{aligned} \tag{3}$$

where $\mathbf{F}$ is essentially the gradient (with respect to x) of the second-order Taylor polynomial. By definition, the operator $\mathbf{F}$ satisfies the second-order smoothness property in Eq. (1)

3 Method

In this section, we shall establish our universal second-order framework. Our presentation evolves around three key components: choosing the appropriate algorithmic template with the key motivations behind it, solving implementability issues that commonly arise in higher-order methods and finally designing a universal algorithm that can handle deterministic and noisy oracle feedback simultaneously without having prior knowledge. Our point of departure is the popular Extra-Gradient (EG) template; originally introduced by Korpelevich [27] and further developed in Nemirovski [38],

$$\begin{aligned} X_{t+1/2} &= \Pi_{\mathcal{X}}\left(X_t - \gamma_t \nabla f(x_t)\right) \\ X_{t+1} &= \Pi_{\mathcal{X}}\left(X_t - \gamma_t \nabla f(x_{t+1/2})\right), \end{aligned} \tag{EG}$$

where $\Pi_{\mathcal{X}}(x) = \arg\min_{z \in \mathcal{X}} \|x - z\|^2$ is the standard Euclidean projection onto the set $\mathcal{X}$. We assume throughout as a blanket assumption that $\mathcal{X}$ possesses favorable geometry which facilitates a tractable projection computation. In terms of output, the candidate solution returned by (EG) after T iterations is the so-called "ergodic average"

$$\bar{X}_T = \frac{\sum_{t=1}^{T} b_t X_{t+1/2}}{\sum_{t=1}^{T} b_t} \tag{4}$$

Then, taking $b_t = \gamma_t$ and assuming the method's step-size γ_t is chosen appropriately, $\bar{X}_T$ enjoys the following universal guarantee [23, 48]:

$$\mathbb{E}[f(\overline{X}_t) - f(x^*)] = \mathcal{O}\left(\frac{1}{T} + \frac{\sigma}{\sqrt{T}}\right) \tag{5}$$

where σ signifies the effect of the noisy feedback. However, as it becomes apparent, the vanilla (EG) template is not capable of matching the iconic $1/T^2$ for the smooth deterministic case. It is well-established in the literature of smooth, convex minimization that iterate averaging (or momentum in the sense of Nesterov [42]) is essential for matching the $O(1/T^2)$ lower bounds. In fact, plain uniform averaging is not sufficient; one needs to introduce new iterates with *increasing* weights. Precisely, this is equivalent to computing an average by taking $b_t = O(t)$. However, we cannot fully characterize the acceleration machinery without what we like to call "gradient weighting". On top of (weighted) iterate averaging, gradients must be multiplied by the *same order of weights* to achieve acceleration, which is a recurring theme in the literature of accelerated and universal optimization [50, 52, 29, 2, 31, 51, 12, 26, 22].

Going back to discussion on (EG), Wang and Abernethy [51] and Kavis et al. [26] provide useful insights into acceleration within the context of (EG). Wang and Abernethy [51] identifies a 2-player game with a particular structure called FENCHELGAME framework, which essentially reduces to minimizing a smooth, convex function when the players cooperate. By introducing an "optimistic" weighted iterate averaging along with a complementary gradient weighting strategy, the framework recovers different acceleration schemes of Nesterov [42, 43, 40]. On a related front, Diakonikolas

and Orecchia [13] proposes the first acceleration of (EG) by appropriately integrating the optimistic averaging idea [51] into the (EG) template as follows:

$$\tilde{X}_t = \frac{b_t X_t + \sum_{s=1}^{t-1} b_s X_{s+1/2}}{\sum_{s=1}^{t} b_s}, \qquad \bar{X}_{t+1/2} = \frac{\sum_{s=1}^{t} b_s X_{s+1/2}}{\sum_{s=1}^{t} b_s} \tag{6}$$

where $b_t = O(t)$ is the "iterate averaging" parameter. Later on, Kavis et al. [26] designs an adaptive, universal variant of accelerated Mirror-Prox following the same optimistic averaging idea as in Eq. (6). All in all, it is a recurring theme among accelerated algorithms to adopt weighted iterate averaging ($b_t = O(t)$) with proportionate gradient weighting, and not so surprisingly, prior work establishes clear connections between the degree of weighting and convergence rate. Cutkosky [12] designs a black-box reduction that accelerated a class of online algorithms and proves that the rate of convergence of the reduction is $O(1/\sum_{t=1}^{T} b_t)$ for $b_t \in [1, t]$. In retrospect, we aim at answering the following question; *What algorithmic construction would enable acceleration beyond $O(1/T^2)$?*

3.1 Implicit algorithm

We give a first affirmative answer to the above question by presenting our implicit accelerated algorithm which is constructed upon (EG), and establish its convergence properties. Note that the implicitness of the scheme serves as a gentle introduction to the actual explicit second order acceleration, which shall follow. Formally, our scheme is given via the following recursion:

$$\begin{aligned} X_{t+1/2} &= \Pi_{\mathcal{X}}\left(X_t - \gamma_t a_t \mathbf{F}(\bar{X}_{t+1/2}; \tilde{X}_t)\right) \\ X_{t+1} &= \Pi_{\mathcal{X}}\left(X_t - \gamma_t a_t \nabla f(\bar{X}_{t+1/2})\right) \end{aligned} \tag{Implicit Second-Order}$$

with $\tilde{X}_t$ and $\bar{X}_{t+1/2}$ being defined as in (6) and the adaptive step-size policy γ_t defined as:

$$\gamma_t = \frac{\gamma}{\sqrt{1 + \sum_{s=1}^{t-1} a_s^2 \|\nabla f(\bar{X}_{s+1/2}) - \mathbf{F}(\bar{X}_{s+1/2}; \tilde{X}_s)\|^2}}. \tag{7}$$

The implicit nature of (Implicit Second-Order) originates from $X_{t+1/2}$ update (which we shall refer to as (corrected) extrapolation step at times) since $\bar{X}_{t+1/2}$ depends upon $X_{t+1/2}$ itself. However, this scheme exhibits several key differences from the vanilla (EG), which constitute the fundamental parts of our second-order acceleration machinery. In particular, we have:

(i) integration of second-order updates for sharper extrapolation steps - first step of acceleration.
(ii) interplay between averaging (b_t) and gradient weighting (a_t) which allows more aggressive averaging - second step of acceleration.
(iii) adaptive step-size in the sense of Rakhlin and Sridharan [49] - key to adaptivity and universality.

Second-order updates: In particular, (EG) tries to approximate X_{t+1} by generating the extrapolated point $X_{t+1/2}$, and make use of the gradient at $X_{t+1/2}$ to take a step from X_t to X_{t+1}. Therefore, if the algorithm is able to compute a sharper estimate in the extrapolation step, it should be able live up to the fame of proximal point and display faster convergence. To this end, we augment the extrapolation step by introducing second-order term. Essentially, our algorithm makes use of *second-order Taylor approximation*, as opposed to first-order expansion, only for the extrapolation step, trading-off sharper approximation with second-order information.

Iterate averaging and gradient weighting: Now, we turn our attention to the second component in our acceleration machinery; averaging and weighting. Recall that the acceleration framework of Cutkosky [12] guarantees a value convergence rate of $O(1/t^{p+1})$ when weighting factor satisfies $b_t = O(t^p)$ with $p \in [0, 1]$. We take this result one step beyond in two fronts; our algorithm exploits higher-order smoothness in order to extend this bound for $p \in [0, 2]$, implying the accelerated rate of $O(1/T^3)$. Second, we observe that previous work restricts the choice of gradient weights and averaging weights by taking $a_t \approx b_t$. We decouple those weights by allowing the sequences a_t and b_t to be *different*, which in turn equips us with more aggressive iterate averaging when necessary.

Adaptive step-size: As the final component, we study the adaptive step-size (7) from the parameter adaptation perspective (i.e., adaptation to the Lipschitz modulus) and expand on its universal properties in the next section. The vast literature on adaptive methods predominantly rely on constructions of AdaGrad-like decreasing step-size policies by accumulating the observed gradient norms in its denominator. The intuition behind this choice is that whenever the method approaches a solution, the vanishing gradients bring about stabilization, ensuring progress around the solution's neighborhood. However, this idea fails for (compactly) constrained problems; when the solution lies on the boundary. So inspired by [49], we overcome this issue by accumulating $\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2}; \tilde{X}_t)\|^2$ which converges to 0 as $\bar{X}_{t+1/2} - \tilde{X}_t \to 0$; which in turn implies convergence of the algorithm. To our knowledge, this is the first adaptive step-size that accommodates second order information.

Having established the core components of our design, we are in position to present the first accelerated convergence rate guarantee for (Implicit Second-Order). Formally, this is given by the following.

Theorem 3.1. *Let $\{X_{t+1/2}\}_{t=1}^T$ be generated by* (Implicit Second-Order) *run with the adaptive step-size policy* (7) *where $a_t = t^2$, $b_t = t^3$. Assume that f satisfies* (H-smooth)*. Then, the following guarantee holds:*

$$f(\bar{X}_{T+1/2}) - f(x^*) \leq O\left(\frac{L\left(D^4/\gamma^2 + D\gamma\right)}{T^3}\right).$$

When $\gamma = D$, we obtain the converge rate $O\left(\frac{LD^2}{T^3}\right)$.

Remark 3.1. We emphasize that the above rate *does not* require any prior knowledge of problem paramaters such as L, D, time-horizon T and any bounds on gradient/Hessian norms. In order to have better dependence on D one could set $\gamma = D$, and our rate of $O(1/T^3)$ holds irrespective of γ.

3.2 Explicit algorithm

Despite the fact that (Implicit Second-Order) improves upon the accelerated rate of $O(1/T^2)$, one may easily observe that it exhibits the following drawbacks:

1. (Implicit Second-Order) is a conceptual algorithm and therefore, *not* implementable in practice.
2. A fortiori, it cannot provide rate interpolation guarantees as it does not have the machinery to simultaneously cope with deterministic and stochastic feedback.

As discussed earlier, a common strategy for overcoming this implicit construction is using a bisection/line-search procedure [21, 37, 7]. Depending on the context, this procedure serves two *distinct* purposes. Primarily, it tackles the implicit nature of the update rule by simultaneously finding a pair of $(\gamma_t, X_{t+1/2})$ and secondly, it enables adaptation to the second-order smoothness modulus. However, one may identify major setbacks with these approaches; first, it is not clear how to handle stochastic oracles for executing the search procedure, so it is not capable of satisfying any universal guarantees. Moreover, it yields a rather complicated procedure as a byproduct that has many moving parts. To that end, we propose an alternative approach which not only yields a simple scheme, but also provides a universal algorithm that is able to handle noisy feedback on-the-fly. Without further ado, we display our explicit algorithm, EXTRANEWTON, with appropriate modifications.

Algorithm 1: EXTRANEWTON

Input: $X_1 \in \mathcal{X}$, gradient weight $\{a_t\}_{t=1}^T$, averaging weight $\{b_t\}_{t=1}^T$, $\gamma > 0$

for $t = 1$ to T **do**

$\quad X_{t+1/2} = \arg\min_{x\in\mathcal{X}} \langle a_t g(\tilde{X}_t), x\rangle + \frac{a_t B_t}{2b_t}\langle H(\tilde{X}_t)(x - X_t), x - X_t\rangle + \frac{1}{2\gamma_t}\|x - X_t\|^2$

$\quad X_{t+1} \quad = \arg\min_{x\in\mathcal{X}} \langle a_t g(\bar{X}_{t+1/2}), x\rangle + \frac{1}{2\gamma_t}\|x - X_t\|^2$

end for

Having established our main algorithmic template, Algorithm 1, we would like to provide a more detailed description of its core components.

Universal step-size We modify our step-size (see Eq. (8)) in order to operate in the stochastic regime while making it noise-adaptive for achieving rate interpolation. Using the same optimistic weighted averaging scheme in Eq. (6), we define the universal counterpart of the adaptive step-size,

$$\gamma_t = \frac{\gamma}{\sqrt{1 + \sum_{s=1}^{t-1} a_s^2 \|g(\bar{X}_{s+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{s+1/2}; \tilde{X}_s)\|^2}}. \tag{8}$$

Note that γ_t is independent of any variable/randomness generated at iteration t; it accumulates $a_t^2 \|g(\bar{X}_{s+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{s+1/2}; \tilde{X}_s)\|^2$ up to $t-1$. Therefore, the step-size is decoupled from the explicit update, *a priori*. Then, what remains is a new algorithmic design that will retain the accelerated convergence properties demonstrated by (Implicit Second-Order) while having an explicit construction that is capable of automatically adjusting to noise level in the oracle feedback.

Before expanding upon the technical details of our proof strategy, let us take our time to explain the consequences of our explicit design in comparison to (Implicit Second-Order).

From implicit to explicit To obtain the explicit algorithm, (*i*) we write the projection sub-problem in the $\arg\min$ form; (*ii*) replace $X_{t+1/2}$ with the free variable x; (*iii*) introduce *stochastic* evaluations into the quadratic, second-order term; then, (*iv*) simplify:

$$\frac{a_t B_t}{2b_t} \langle H(\tilde{X}_t)(\frac{b_t x + \sum_{s=1}^{t-1} b_s X_{s+1/2}}{B_t} - \tilde{X}_t), \frac{b_t x + \sum_{s=1}^{t-1} b_s X_{s+1/2}}{B_t} - \tilde{X}_t \rangle$$
$$\Downarrow$$
$$\frac{a_t b_t}{2B_t} \langle H(\tilde{X}_t)(x - X_t), x - X_t \rangle$$

Given the bisection-type conceptual methods [37, 21, 7], it is surprising how smoothly we could transition from implicit to explicit *once* we decouple the step-size from the current iteration *apriori*. Moreover, the resulting update rule for the extrapolation step retains the quadratic structure as the X_{t+1} update rule.

Having analyzed the components of the explicit scheme, we will first present the universal convergence rates then provide a concise explanation of the proof strategy with particular emphasis on the principal components of the analysis.

Theorem 3.2. *Let $\{X_{t+1/2}\}_{t=1}^T$ be generated by Algorithm 1, run with the adaptive step-size policy (8) and $a_t = t^2, b_t = t^3$. Assume that f satisfies (H-smooth), and that Assumptions (2) hold with $\sigma = \max\{\sigma_g, \sigma_H\}$. Then, the following universal guarantee holds:*

$$\mathbb{E}\left[f(\bar{X}_{T+1/2}) - f(x^*) \right] \leq O\left(\frac{\sigma}{\sqrt{T}} + \frac{L\left(D^4/\gamma^2 + D\gamma\right)}{T^3} \right).$$

When $\gamma = D$, we obtain the target rate $O\left(\frac{\sigma}{\sqrt{T}} + \frac{LD^2}{T^3}\right)$.

Remark 3.2. Similar to Theorem 3.1, EXTRANEWTON achieves the preceding convergence rate independent of the knowledge of problem parameters.

Compatible with the (EG)-based algorithmic design, our proof has the following main steps

1. We perform an *offline* regret analysis of Alg. 1 and show adaptive regret bounds - see Prop. 3.1.
2. We prove an anytime online-to-batch conversion framework, which generalizes that of Cutkosky [12], through decoupling iterate averaging from gradient weighting - see Theorem 3.3.
3. Combining the adaptive regret bound with the conversion theorem immediately implies *universal, accelerated* value convergence of $O(\frac{\sigma}{\sqrt{T}} + \frac{LD^2}{T^3})$ - see Theorem 3.2.

Let us begin with clarifying what *offline regret* means for Algorithm 1. We define the (linear) regret considering the convention in both online learning [49, 12] and first-order acceleration literature [51, 26, 22]. We measure the performance of our decisions for the extrapolation sequence such that after playing $X_{t+1/2}$, our algorithm observes and suffers the linear (weighted) loss with respect to $a_t \nabla f(\bar{X}_{t+1/2})$. Hence, we define the regret as

$$\textsc{Reg}_{\mathrm{T}}(x) = \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - x \rangle \tag{Reg}$$

where we run the algorithm for T rounds. Next up, we provide our generalized conversion result.

Theorem 3.3. *Let $\textsc{Reg}_{\mathrm{T}}(x^*)$ denote the anytime regret for the decision sequence $\{X_{t+1/2}\}_{t=1}^{T}$ as in (Reg), and define two sequences of non-decreasing weights a_t and b_t such that $a_t, b_t \geq 1$. As long as a_t/b_t is ensured to be non-increasing,*

$$f(\bar{X}_T) - f(x^*) \leq \frac{\textsc{Reg}_{\mathrm{T}}(x^*)}{a_T \frac{B_T}{b_T}}$$

Remark 3.3. This conversion result holds independent of the order of smoothness of the objective as long as f is convex. Moreover, it allows averaging parameter b_t to be asymptotically larger than gradient weights a_t, enabling a more aggressive averaging strategy when necessary.

To complement the lower bound to the regret $\textsc{Reg}_{\mathrm{T}}(x^*)$, we will provide a template upper bound that will help us explain how we exploit second-order smoothness for a more aggressive weighting, hence the rate $O(1/T^3)$.

Proposition 3.1. *Let $\{X_{t+1/2}\}_{t=1}^{T}$ be generated by Algorithm 1, run with a non-increasing step-size sequence γ_t and non-decreasing sequences of weights $a_t, b_t \geq 1$ such that a_t/b_t is also non-increasing. Then, the following guarantee holds:*

$$\mathbb{E}\left[\,\textsc{Reg}_{\mathrm{T}}(x^*)\,\right] \leq \frac{1}{2}\mathbb{E}\left[\frac{3D^2}{\gamma_{T+1}} + \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|g(\bar{X}_{t+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{t+1/2}; \tilde{X}_t)\|^2 - \frac{\|X_{t+1/2} - X_t\|^2}{\gamma_{t+1}}\right]$$

It is important to note that the template inequality in Proposition 3.1 is agnostic to the design of our adaptive step-size as in Eq. (8) as well as the selection of the weights as we described in Theorem 3.2; it holds for Algorithm 1 in the most general parameter selection. It essentially applies to any non-increasing sequence of step-sizes and non-decreasing gradient weight sequence $a_t \geq 1$. To obtain it, we neither used convexity nor the smoothness of the objective f. In fact, the structure of the objective function, i.e., its convexity, will not be needed for upper-bounding the regret expression, and we use it only for the conversion in Theorem 3.3.

Now, let us explain how we make use of second-order smoothness for enjoying faster rates, and give a brief discussion of how the regret bound will look in its final form. First, we decompose the stochastic term $\|g(\bar{X}_{t+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{t+1/2}; \tilde{X}_t)\|^2$ into deterministic feedback and noise. Then, we argue that *the noisy component* grows as $O(\sigma T^{5/2})$ where we define $\sigma = \max\{\sigma_g, \sigma_H\}$ for brevity. On the other hand, achieving the accelerated $O(1/T^3)$ component of the universal rate amounts to showing that the regret has a constant, $O(1)$, component. In the worst-case sense, however, *the deterministic component itself* grows as $O(T^{5/2})$, too. Fortunately, we identify that the negative term is "large enough" in magnitude to control the growth of the deterministic term, which permits a constant component of order $O(LD^2)$ for the regret bound.

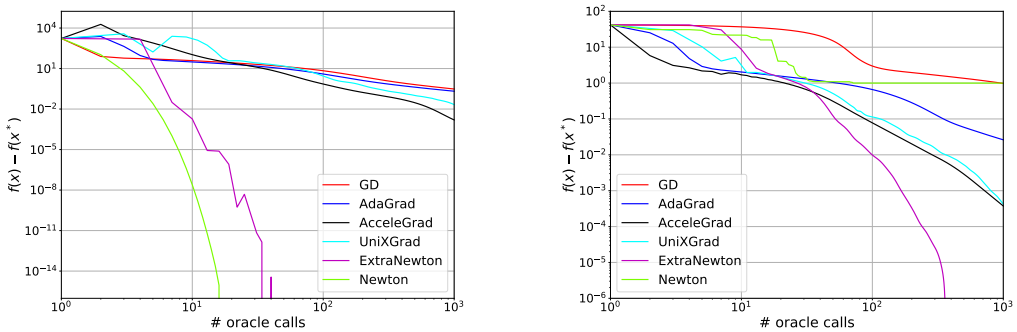
Although the regret bound of $O(LD^2 + \sigma T^{5/2})$ seems counter-intuitive from an online-learning perspective, it will make perfect sense when we discuss how second-order smoothness leads to "faster" conversion through more aggressive averaging. As a matter of fact, we will conclude our discussion with how second-order smoothness helps us accelerate. It turns out that using (H-smooth), iterate averaging as in Eq.(6) and compactness of $\mathcal{X}$, we can bound the negative term as,

$$-\frac{1}{\gamma_{t+1}}\|X_{t+1/2} - X_t\|^2 \leq -\frac{1}{L^2 D^2 \gamma_{t+1}} t^4 \|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2}; \tilde{X}_t)\|^2$$

Observe that to seamlessly combine the positive and negative terms, our analysis enforces that $a_t = O(t^2)$ and $b_t = \Omega(t^2)$. Then, the conversion implies a convergence rate of $\text{REG}_\text{T}(x^*)/T^3$, hence the recipe for acceleration. Therefore, the constant component of the regret amounts to $O(1/T^3)$ convergence rate, while the stochastic component of the regret implies $O(\sigma/\sqrt{T})$ rate, giving us the first universal acceleration beyond first-order smoothness.

4 Experiments

In this section, we will present practical performance of EXTRANEWTON against a set of algorithms, e.g., GD, SGD, ADAGRAD [14], ACCELEGRAD [31], UNIXGRAD [26], NEWTON'S, for least squares and logistic regression problems over a LIBSVM dataset, `a1a`. Our main objective is to convey two motivational ideas. First, when the objective has a favorable structure as in least squares, second-order method has cheap oracle costs (Hessian will be constant across the whole space) and display superior convergence behavior. Second, we want to demonstrate the improved rates of our algorithm against accelerated and non-accelerated first-order methods through the ℓ_2-regularized logistic regression problem. In the plots, the x-axis denotes *# of oracle calls* which counts any gradient or Hessian computation as one oracle call.



(a) Least-squares regression on `a1a`

(b) Logistic regression on `a1a`

Figure 1: Comparison of value convergence for regression problems with deterministic oracle access

When the problem is suitable, second-order methods show promising performance with truly superior run time. In Figure 1a, we display the result for least squares setting. Second-order methods are known to be suitable for quadratic problems, and our method exploits its hybrid construction to converge significantly faster than first-order methods, matching the behavior of NEWTON'S.

For the logistic regression problem, we regularize it with $g(x) = 1/2\|x\|^2$, but use a very small regularization constant to render the problem ill-conditioned, making things slightly more difficult for the algorithms [34, 36]. Although we implement NEWTON'S with line-search, we actually observed a sporadic convergence behavior; when the initial point is close to the solution it converges similarly to EXTRANEWTON, however when we initialize further away it doesn't converge. This non-convergent behavior has been known for NEWTON'S, even with line-search present [18]. On the contrary, EXTRANEWTON consistently converges; even if we perturb the initial step-size and make it adversarially large, it manages to recover due to its adaptive step-size.

5 Conclusion

In this work, we present the *first* universal, second-order algorithm, EXTRANEWTON, which enjoys the value convergence rate of $O(\sigma/\sqrt{T} + 1/T^3)$. To that end, an important open question is whether we could design a method that achieves an improved rate interpolation guarantee $O(\sigma/\sqrt{T} + 1/T^{7/2})$ without depending on any line-search/bisection mechanism. We defer this to a future work.

References

[1] Naman Agarwal and Elad Hazan. Lower bounds for higher-order convex optimization. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 774–792. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/agarwal18a.html.

[2] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent, 2016.

[3] Kimon Antonakopoulos, Elena Veronica Belmega, and Panayotis Mertikopoulos. Adaptive extra-gradient methods for min-max optimization and games. *CoRR*, abs/2010.12100, 2020. URL https://arxiv.org/abs/2010.12100.

[4] Yossi Arjevani, Ohad Shamir, and Ron Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, pages 1–34, 2019.

[5] Stefania Bellavia, Gianmarco Gurioli, and Benedetta Morini. Adaptive cubic regularization methods with dynamic inexact Hessian information and applications to finite-sum minimization. *IMA Journal of Numerical Analysis*, 41(1):764–799, 04 2020. ISSN 0272-4979. doi: 10.1093/imanum/drz076. URL https://doi.org/10.1093/imanum/drz076.

[6] Albert A. Bennett. Newton's method in general analysis. *Proceedings of the National Academy of Sciences*, 2(10):592–598, 1916. doi: 10.1073/pnas.2.10.592. URL https://www.pnas.org/doi/abs/10.1073/pnas.2.10.592.

[7] Brian Bullins and Kevin A. Lai. Higher-order methods for convex-concave min-max optimization and monotone variational inequalities, 2020. URL https://arxiv.org/abs/2007.04528.

[8] Coralia Cartis, Nicholas I. Gould, and Philippe L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. part i: Motivation, convergence and numerical results. *Math. Program.*, 127 (2):245–295, apr 2011. ISSN 0025-5610.

[9] Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. part ii: Worst-case function- and derivative-evaluation complexity. *Math. Program.*, 130(2):295–319, dec 2011. ISSN 0025-5610. doi: 10.1007/s10107-009-0337-y. URL https://doi.org/10.1007/s10107-009-0337-y.

[10] Xi Chen, Bo Jiang, Tianyi Lin, and Shuzhong Zhang. Accelerating adaptive cubic regularization of newton's method via random sampling. *Journal of Machine Learning Research*, 23(90):1–38, 2022. URL http://jmlr.org/papers/v23/20-910.html.

[11] Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint. *Trust Region Methods*. Society for Industrial and Applied Mathematics, 2000. doi: 10.1137/1.9780898719857. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898719857.

[12] Ashok Cutkosky. Anytime online-to-batch conversions, optimism, and acceleration. *the International Conference on Machine Learning (ICML)*, June 2019.

[13] Jelena Diakonikolas and Lorenzo Orecchia. Accelerated extra-gradient descent: A novel accelerated first-order method. In *ITCS*, 2018.

[14] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[15] Jean-Pierre Dussault and Dominique Orban. Scalable adaptive cubic regularization methods. 2021. doi: 10.13140/RG.2.2.18142.15680. URL http://rgdoi.net/10.13140/RG.2.2.18142.15680.

[16] Alina Ene, Huy L. Nguyen, and Adrian Vladu. Adaptive gradient methods for constrained convex optimization and variational inequalities, 2021.

[17] Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Evgeniya Vorontsova, Daniil Selikhanovych, and César A. Uribe. Optimal tensor methods in smooth convex and uniformly convexoptimization. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1374–1391. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/gasnikov19a.html.

[18] Florian Jarre and Philippe L. Toint. Simple examples for the failure of newton's method with line search for strictly convex minimization. *Math. Program.*, 158(1–2):23–34, jul 2016. ISSN 0025-5610. doi: 10.1007/s10107-015-0913-2. URL https://doi.org/10.1007/s10107-015-0913-2.

[19] Bo Jiang, Tianyi Lin, and Shuzhong Zhang. A unified scheme to accelerate adaptive cubic regularization and gradient methods for convex optimization, 2017. URL https://arxiv.org/abs/1710.04788.

[20] Bo Jiang, Tianyi Lin, and Shuzhong Zhang. A unified adaptive tensor approximation scheme to accelerate composite convex optimization. *SIAM Journal on Optimization*, 30(4):2897–2926, 2020. doi: 10.1137/19M1286025. URL https://doi.org/10.1137/19M1286025.

[21] Ruichen Jiang and Aryan Mokhtari. Generalized optimistic methods for convex-concave saddle point problems, 2022. URL https://arxiv.org/abs/2202.09674.

[22] Pooria Joulani, Anant Raj, Andras Gyorgy, and Csaba Szepesvari. A simpler approach to accelerated optimization: iterative averaging meets optimism. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4984–4993. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/joulani20a.html.

[23] Anatoli Juditsky, Arkadi Semen Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[24] L. V. Kantorovich. Functional analysis and applied mathematics. *Uspekhi Mat. Nauk*, 3:89–185, 1948.

[25] Sai Praneeth Karimireddy, Sebastian U. Stich, and Martin Jaggi. Global linear convergence of newton's method without strong-convexity or lipschitz gradients, 2018. URL https://arxiv.org/abs/1806.00413.

[26] Ali Kavis, Kfir Y. Levy, Francis Bach, and Volkan Cevher. Unixgrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6260–6269. Curran Associates, Inc., 2019.

[27] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12: 747–756, 1976.

[28] Dmitry Kovalev, Konstantin Mishchenko, and Peter Richtárik. Stochastic newton and cubic newton methods with simple local linear-quadratic rates, 2019. URL https://arxiv.org/abs/1912.01597.

[29] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

[30] Kenneth Levenberg. A method for the solution of certain non – linear problems in least squares. *Quarterly of Applied Mathematics*, 2:164–168, 1944.

[31] Kfir Y Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. In *Neural and Information Processing Systems (NeurIPS)*, December 2018.

[32] Tianyi Lin and Michael. I. Jordan. Perseus: A simple high-order regularization method for variational inequalities, 2022. URL https://arxiv.org/abs/2205.03202.

[33] Donald W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963. doi: 10.1137/0111030. URL https://doi.org/10.1137/0111030.

[34] Ulysse Marteau-Ferey, Francis R. Bach, and Alessandro Rudi. Globally convergent newton methods for ill-conditioned generalized self-concordant losses. In *NeurIPS*, 2019.

[35] H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *COLT '10: Proceedings of the 23rd Annual Conference on Learning Theory*, 2010.

[36] Konstantin Mishchenko. Regularized newton method with global $o(1/k^2)$ convergence, 2021. URL https://arxiv.org/abs/2112.02089.

[37] Renato D. C. Monteiro and B. F. Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2): 1092–1125, 2013. doi: 10.1137/110833786. URL https://doi.org/10.1137/110833786.

[38] Arkadi Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[39] Arkadii Nemirovskii, David Borisovich Yudin, and ER Dawson. Problem complexity and method efficiency in optimization. 1983.

[40] Yu Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, may 2005. ISSN 0025-5610. doi: 10.1007/s10107-004-0552-5. URL https://doi.org/10.1007/s10107-004-0552-5.

[41] Yu. Nesterov. Accelerating the cubic regularization of newton's method on convex problems. *Mathematical Programming*, 2008.

[42] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269:543–547, 1983.

[43] Yurii Nesterov. On an approach to the construction of optimal methods of minimization of smooth convex functions. *Ekonomika i Mateaticheskie Metody*, 24(3):509–517, 1988.

[44] Yurii Nesterov. Introductory lectures on convex optimization. 2004, 2003.

[45] Yurii Nesterov and Boris Polyak. Cubic regularization of newton method and its global performance. *Math. Program.*, 108:177–205, 08 2006. doi: 10.1007/s10107-006-0706-8.

[46] Yurii E. Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, 2018.

[47] Boris Polyak. Newton-kantorovich method and its global convergence. *Journal of Mathematical Sciences*, 133:1513–1523, 2006.

[48] Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *NIPS '13: Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2013.

[49] Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013.

[50] P Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 01 2008.

[51] Jun-Kun Wang and Jacob D Abernethy. Acceleration through optimistic no-regret dynamics. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3824–3834. Curran Associates, Inc., 2018.

[52] Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.

[53] NESTEROV Yurii. Implementable tensor methods in unconstrained convex optimization. LIDAM Discussion Papers CORE 2018005, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), March 2018. URL https://ideas.repec.org/p/cor/louvco/2018005.html.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2 and Theorem 3.1, 3.3, 3.2
   (b) Did you include complete proofs of all theoretical results? [Yes] See the Appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We provide mean and standard deviation on the curves over multiple runs
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

Table of Contents

A Preface 15

B Notation and Definitions 15

C Further Experimental Evaluation 16

D Generalized Online-to-batch Conversion (Theorem 3.3) 17

E Template Regret Bound (Proposition 3.1) 19

F Technical Lemma for the Main Proofs 22

G Implicit Accelerated Second-order Algorithm (Theorem 3.1) 22

H EXTRANEWTON: The First Universal Second-order Accelerated Method (Theorem 3.2) 27

A Preface

In Appendix B, we provide a complete list of notation and definitions that we have used throughout the manuscript.

In Appendix C, we showcase additional numerical evidence for the comparison we provided in the main text. Due to space constraints, we moved most of our plots to the appendix. We investigate the practical behavior in both deterministic and stochastic setting.

In Appendix D, we begin with the proof of the generalized online-to-batch conversion in Theorem 3.3 to form the connection between the offline regret $\textsc{Reg}_{\mathrm{T}}(x^*)$ and value convergence $f(\bar{X}_{T+1/2}) - f(x^*)$.

Then in Appendix E, we present the analysis for obtaining the template regret bound in Proposition 3.1. This template inequality is indeed the point where the analysis in the deterministic, implicit setting and universal, explicit setting part ways.

In Appendix F, we take a small detour to introduce a crucial numerical inequality that is commonly used in the analysis of adaptive methods.

In Appendix G, we share the analysis of our first main result: convergence of the implicit algorithm (Implicit Second-Order) in deterministic setting (Theorem 3.1). We also prove a corollary of Proposition 3.1 for the case of deterministic oracles in this section.

We conclude the appendix by the universal convergence analysis of EXTRANEWTON (Theorem 3.2) in Appendix H.

B Notation and Definitions

To complement the notation in Section 2, we will present a complete list of definitions and parameter descriptions to make it easier for the reader to follow the technical arguments in the whole paper.

Table 2: A complete list of parameters and expressions, their definitions and descriptions

	Formal Definition	**Description**
f	$f : \mathbb{R}^d \to \mathbb{R} + \{+\infty\}$	objective function
$\mathcal{X}$	$\mathcal{X} \subset \mathbb{R}^d$	convex and compact constraint set
x^*	$= \arg\min_{x \in \mathcal{X}} f(x)$	solution of the constrained problem (Opt)
D	$= \sup_{x,y \in \mathcal{X}} \lVert x - y \rVert$	diameter of the constraint set $\mathcal{X}$
L	$\lVert \nabla^2 f(x) - \nabla^2 f(x') \rVert \le L \lVert x - x' \rVert$	second-order smoothness constant of f
$g(\cdot)$	$\mathbb{E}\left[\, g(x) \mid x \,\right] = \nabla f(x)$	unbiased gradient estimate
$H(\cdot)$	$\mathbb{E}\left[\, H(x) \mid x \,\right] = \nabla^2 f(x)$	unbiased Hessian estimate
σ_g	$\mathbb{E}\left[\, \lVert g(x) - \nabla f(x) \rVert^2 \mid x \,\right] \le \sigma_g^2$	variance bound for gradient estimate
σ_H	$\mathbb{E}\left[\, \lVert H(x) - \nabla^2 f(x) \rVert^2 \mid x \,\right] \le \sigma_H^2$	variance bound for Hessian estimate
σ	$= \max\{\sigma_g, \sigma_H\}$	maximum variance of oracles
γ_t	Eq. (7) and Eq. (8)	adaptive step-size
a_t	$= t^2$	gradient weights
A_t	$= \sum_{s=1}^t a_s$	normalization factor for gradient weights a_t
b_t	$= t^p$, where $p \ge 2$	averaging weights
B_t	$= \sum_{s=1}^t b_s$	normalization factor for averaging weights b_t

C Further Experimental Evaluation

In this section we will present additional numerical experiments in two fronts;

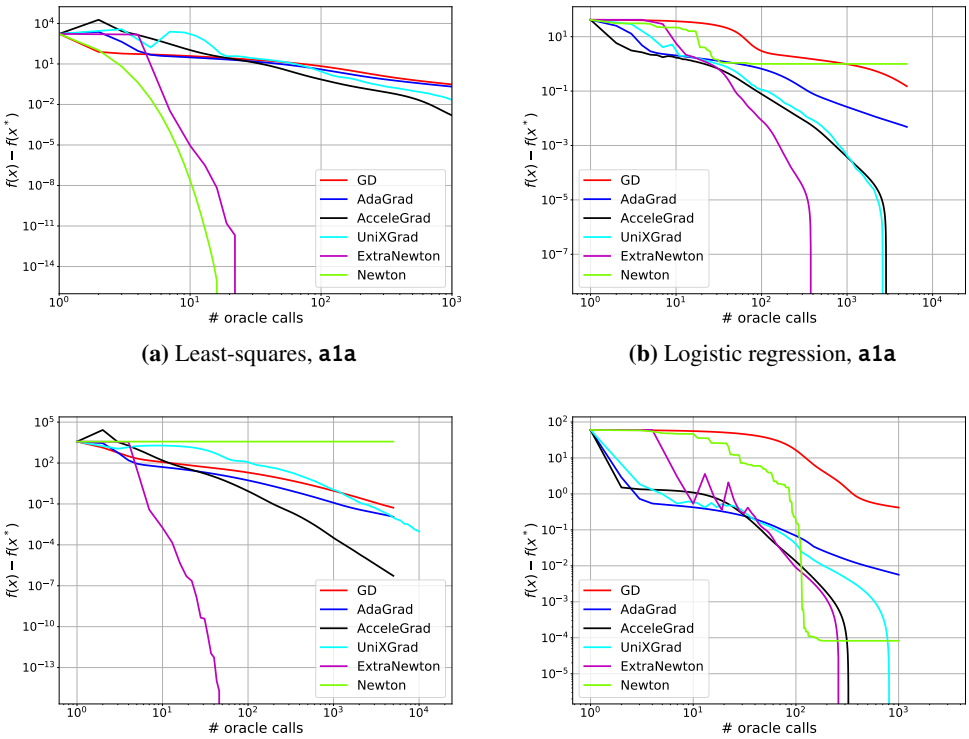
- we run logistic regression and least-squares regression under deterministic gradients with another LIBSVM datasets, `w1a`,

- and subsequently display results in the stochastic setting for the same datasets `a1a, w1a`.

Figure 2 shows the results for the deterministic experiments while Figure 3 focuses on the results of the stochastic setting. In both figures, we present results for the least-squares in the first column and the logistic regression in the second column.

In Figure 2, the x-axis represent the number of calls made to the *deterministic* oracle, and in Figure 3, x-axis corresponds to *number of full data passes (epochs)* to compute the stochastic gradient estimates. The deterministic setup is the same as we described in the main text. In the case of stochastic gradients, we compute mini-batch gradient estimates with a batch-size of 50 samples. We plot the mean of 5 trials for all the methods under mini-batch gradients and also display the variance as the shaded region around the mean curve.

For the case of logistic regression under deterministic gradients, our method performs better than the rest of the pack with `a1a` dataset but has almost matching performance with a smaller performance gap compared to accelerated first-order methods with `w1a` dataset. For both datasets, we tried to tune Newton's method for a randomly-chosen initialization but it was very difficult to find a parameter setting where Newton shows any reasonable behavior. One could notice that Newton's method doesn't converge to the solution for logistic regression problem for this random initial point.



(a) Least-squares, **`a1a`**

(b) Logistic regression, **`a1a`**

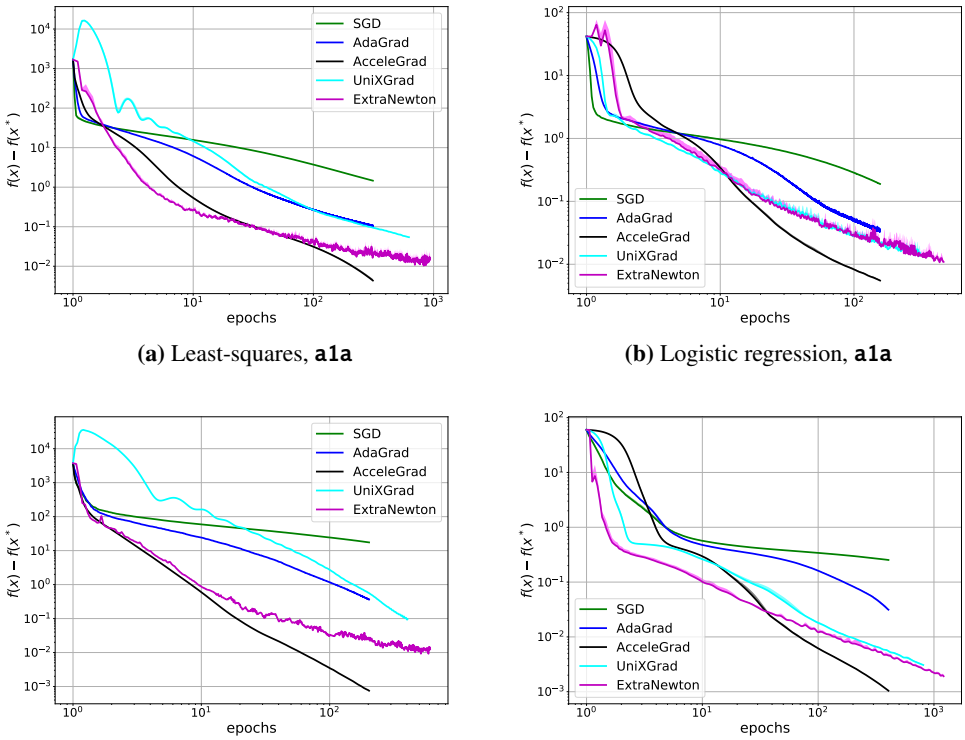
(c) Least-squares, **`w1a`**

(d) Logistic regression, **`w1a`**

Figure 2: Comparison of value convergence for regression problems with **deterministic** oracle access

We observe that the main advantage of our approach, and in general that of second-order methods, becomes apparent when the problem at hand has a compatible structure such as least-squares. Intuitively, second-order methods should benefit when the cost of computing the Hessian is comparable to gradient computation. In fact, quadratic problems like least-squares yield a constant Hessian for any point in the domain, granting a significant advantage to second-order methods. We exemplify this behavior for least-squares problem with deterministic oracles. With `w1a` dataset, we couldn't get Newton's method to converge once again. On the contrary, our method shows significant performance upgrade compared to first-order methods while converging consistently in all our trials.

Finally, we have the experiments under stochastic oracles. We essentially present these results for two main reasons; to show that our method works seamlessly with stochastic gradients without any modifications, and to demonstrate that EXTRANEWTON achieves the $O(1/\sqrt{T})$ rate (same as other methods we compare against) when the gradient information is noisy. We showcase both of these perspectives in Figure 3.



(a) Least-squares, **`a1a`**

(b) Logistic regression, **`a1a`**

(c) Least-squares, **`w1a`**

(d) Logistic regression, **`w1a`**

Figure 3: Comparison of value convergence for regression problems with **stochastic** oracle access

D Generalized Online-to-batch Conversion (Theorem 3.3)

In this section we present the online-to-batch conversion scheme which connects the optimality gap $f(\bar{X}_{T+1/2}) - f(x^*)$ with the "weighted" regret $\text{REG}_\text{T}(x^*) = \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - x^* \rangle$.

Theorem 3.3. *Let* $\text{REG}_\text{T}(x^*)$ *denote the anytime regret for the decision sequence* $\{X_{t+1/2}\}_{t=1}^{T}$ *as in* (Reg)*, and define two sequences of non-decreasing weights* a_t *and* b_t *such that* $a_t, b_t \geq 1$*. As long as* a_t/b_t *is ensured to be non-increasing,*

$$f(\bar{X}_{T+1/2}) - f(x^*) \leq \frac{\text{REG}_\text{T}(x^*)}{a_T \frac{B_T}{b_T}}$$

Proof. First, recall the definition of the offline regret:

$$\textsc{Reg}_{\mathrm{T}}(x^*) = \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+1/2}, X_{t+1/2} - x^* \rangle$$

Devising our analysis in the spirit of [12, 26], we need to relate $X_{t+1/2}$ to the average iterate $\bar{X}_{t+1/2}$ in order to exploit the convexity of the objective function. Notice that we could write the iterate $X_{t+1/2}$ as the difference of consecutive *average* iterates,

$$a_t X_{t+1/2} = a_t \frac{B_t}{b_t} \bar{X}_{t+1/2} - a_t \frac{B_{t-1}}{b_t} \bar{X}_{t-1/2}. \tag{9}$$

Also, we could subsequently express $a_t x^* = a_t \frac{B_t}{b_t} x^* - a_t \frac{B_{t-1}}{b_t} x^*$. Combining them together,

$$\begin{aligned}
\textsc{Reg}_{\mathrm{T}}(x^*) &= \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+1/2}, X_{t+1/2} - x^* \rangle \\
&= \sum_{t=1}^{T} a_t \frac{B_t}{b_t} \langle \nabla f(\bar{X}_{t+1/2}), \bar{X}_{t+1/2} - x^* \rangle - a_t \frac{B_{t-1}}{b_t} \langle \nabla f(\bar{X}_{t+1/2}, \bar{X}_{t-1/2} - x^* \rangle \\
&= \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+1/2}), \bar{X}_{t+1/2} - x^* \rangle + a_t \frac{B_{t-1}}{b_t} \langle \nabla f(\bar{X}_{t+1/2}), \bar{X}_{t+1/2} - \bar{X}_{t-1/2} \rangle
\end{aligned}$$

where we added and subtracted $a_t \frac{B_{t-1}}{b_t} \langle \nabla f(\bar{X}_{t+1/2}), \bar{X}_{t+1/2} \rangle$ to obtain the second equality. Having expressed both inner products in the form we want, we could apply convexity and telescope.

$$\begin{aligned}
&\sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+1/2}, X_{t+1/2} - x^* \rangle \\
&\quad \geq \sum_{t=1}^{T} a_t \left(f(\bar{X}_{t+1/2}) - f(x^*) \right) + a_t \frac{B_{t-1}}{b_t} \left(f(\bar{X}_{t+1/2}) - f(\bar{X}_{t-1/2}) \right) \\
&\quad = \sum_{t=1}^{T} a_t \left(f(\bar{X}_{t+1/2}) - f(x^*) \right) + a_t \frac{B_{t-1}}{b_t} \left(f(\bar{X}_{t+1/2}) - f(x^*) \right) - a_t \frac{B_{t-1}}{b_t} \left(f(\bar{X}_{t-1/2}) - f(x^*) \right) \\
&\quad = \sum_{t=1}^{T} a_t \frac{B_t}{b_t} \left(f(\bar{X}_{t+1/2}) - f(x^*) \right) - a_t \frac{B_{t-1}}{b_t} \left(f(\bar{X}_{t-1/2}) - f(x^*) \right) \\
&\quad = a_T \frac{B_T}{b_T} \left(f(\bar{X}_{T+1/2}) - f(x^*) \right) - a_1 \frac{B_0}{b_1} \left(f(\bar{X}_{-1/2}) - f(x^*) \right) + \sum_{t=1}^{T-1} B_t \left(\frac{a_t}{b_t} - \frac{a_{t+1}}{b_{t+1}} \right) \left(f(\bar{X}_{t+1/2}) - f(x^*) \right)
\end{aligned}$$

Setting $B_0 = 0$ eliminates the second term. To conclude the proof, we need to show that the summation term in the above expression is always non-negative. This is ensured when the sequence $\frac{a_t}{b_t}$ is monotonically non-increasing, which is specified in the theorem statement (and subsequently satisfied by the algorithms). Hence,

$$\begin{aligned}
&\sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+1/2}, X_{t+1/2} - x^* \rangle \\
&\quad = a_T \frac{B_T}{b_T} \left(f(\bar{X}_{T+1/2}) - f(x^*) \right) + \sum_{t=1}^{T-1} B_t \left(\frac{a_t}{b_t} - \frac{a_{t+1}}{b_{t+1}} \right) \left(f(\bar{X}_{t+1/2}) - f(x^*) \right) \\
&\quad \geq a_T \frac{B_T}{b_T} \left(f(\bar{X}_{T+1/2}) - f(x^*) \right).
\end{aligned}$$

Rearranging the terms gives us the final result

$$f(\bar{X}_{T+1/2}) - f(x^*) \leq \frac{\sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+1/2}, X_{t+1/2} - x^* \rangle}{a_T \frac{B_T}{b_T}} = \frac{\textsc{Reg}_{\mathrm{T}}(x^*)}{a_T \frac{B_T}{b_T}}.$$

■

E Template Regret Bound (Proposition 3.1)

In this section, we will prove the template inequality in Proposition 3.1 in the case of stochastic oracles. This inequality will give us the main departure point for both Theorem 3.1 and Theorem 3.2. We will prove a corollary of the following result later on, specifically for the deterministic setup, which will follow the same steps as Proposition 3.1.

For ease of navigation, we present EXTRANEWTON once more.

EXTRANEWTON

Input: $X_1 \in \mathcal{X}$, gradient weight $\{a_t\}_{t=1}^T$, averaging weight $\{b_t\}_{t=1}^T$, $\gamma > 0$

1: **for** $t = 1$ to T **do**

2: $\quad X_{t+1/2} = \arg\min_{x\in\mathcal{X}} \langle a_t g(\tilde{X}_t), x\rangle + \frac{a_t B_t}{2b_t}\langle H(\tilde{X}_t)(x - X_t), x - X_t\rangle + \frac{1}{2\gamma_t}\|x - X_t\|^2$

3: $\quad X_{t+1} \quad = \arg\min_{x\in\mathcal{X}} \langle a_t g(\bar{X}_{t+1/2}), x\rangle + \frac{1}{2\gamma_t}\|x - X_t\|^2$

4: **end for**

Proposition 3.1. *Let $\{X_{t+1/2}\}_{t=1}^T$ be generated by Algorithm 1, run with a non-increasing step-size sequence γ_t and non-decreasing sequences of weights $a_t, b_t \geq 1$ such that a_t/b_t is also non-increasing. Then, the following guarantee holds:*

$$\mathbb{E}\left[\,\textsc{Reg}_{\mathrm{T}}(x^*)\,\right] \leq \frac{1}{2}\mathbb{E}\left[\frac{3D^2}{\gamma_{T+1}} + \sum_{t=1}^{T}\gamma_{t+1}a_t^2\|g(\bar{X}_{t+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 - \frac{\|X_{t+1/2} - X_t\|^2}{\gamma_{t+1}}\right]$$

Proof. We take off from the optimality conditions associated with each update sequence for our explicit algorithm EXTRANEWTON (Algorithm 1). Optimality condition for $X_{t+1/2}$ implies for any $z_0 \in \mathcal{X}$,

$$\begin{aligned}
&\langle a_t g(\tilde{X}_t) + a_t\frac{b_t}{B_t}H(\tilde{X}_t)(X_{t+1/2} - X_t), X_{t+1/2} - z_0\rangle \\
&= \langle a_t g(\tilde{X}_t) + a_t H(\tilde{X}_t)(\bar{X}_{t+1/2} - \tilde{X}_t), X_{t+1/2} - z_0\rangle \\
&= \langle a_t\tilde{\mathbf{F}}(\bar{X}_{t+1/2};\tilde{X}_t), X_{t+1/2} - z_0\rangle \\
&\leq \frac{1}{\gamma_t}\langle X_{t+1/2} - X_t, z_0 - X_{t+1/2}\rangle \\
&= \frac{1}{2\gamma_t}\left(\|X_t - z_0\|^2 - \|X_{t+1/2} - z_0\|^2 - \|X_{t+1/2} - X_t\|^2\right)
\end{aligned} \tag{10}$$

Similarly, optimality of X_{t+1} update yields for any $z_1 \in \mathcal{X}$,

$$\begin{aligned}
\langle a_t g(\bar{X}_{t+1/2}), X_{t+1} - z_1\rangle &\leq \frac{1}{2\gamma_t}\langle X_{t+1} - X_t, z_1 - X_{t+1}\rangle \\
&= \frac{1}{2\gamma_t}\left(\|X_t - z_1\|^2 - \|X_{t+1} - z_1\|^2 - \|X_{t+1} - X_t\|^2\right)
\end{aligned} \tag{11}$$

First, we will set $z_1 = x^*$ to establish the telescoping summation over $\|X_t - x^*\|^2 - \|X_{t+1} - x^*\|^2$. Then, we will simply align the above expression with the regret as follows,

$$\begin{aligned}
&\langle a_t g(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star\rangle \\
&\qquad = \langle a_t g(\bar{X}_{t+1/2}), X_{t+1/2} - X_{t+1}\rangle + \langle a_t g(\bar{X}_{t+1/2}), X_{t+1} - x^\star\rangle \\
&\qquad \leq \langle a_t g(\bar{X}_{t+1/2}), X_{t+1/2} - X_{t+1}\rangle \\
&\qquad\quad + \frac{1}{2\gamma_t}\left(\|X_t - x^\star\|^2 - \|X_{t+1} - x^\star\|^2 - \|X_{t+1} - X_t\|^2\right)
\end{aligned} \tag{12}$$

Now, observe that setting $z_0 = X_{t+1}$ in Eq. (10) and rearranging we have

$$\begin{aligned}
&-\frac{1}{2\gamma_t}\|X_{t+1} - X_t\|^2 \\
&\leq -\langle a_t\tilde{\mathbf{F}}(\bar{X}_{t+1/2};\tilde{X}_t), X_{t+1/2} - X_{t+1}\rangle - \frac{1}{2\gamma_t}\left(\|X_{t+1/2} - X_{t+1}\|^2 + \|X_{t+1/2} - X_t\|^2\right)
\end{aligned}$$

Plugging the above expression into Eq. (12) and summing over $t = 1, ..., T$, we will obtain,

$$\begin{aligned}
&\sum_{t=1}^{T} \langle a_t g(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star \rangle \\
&\quad \leq \sum_{t=1}^{T} a_t \langle g(\bar{X}_{t+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{t+1/2}; \tilde{X}_t), X_{t+1/2} - X_{t+1} \rangle \\
&\qquad + \sum_{t=1}^{T} \frac{1}{2\gamma_t} \left(\|X_t - x^\star\|^2 - \|X_{t+1} - x^\star\|^2 - \|X_{t+1/2} - X_{t+1}\|^2 - \|X_{t+1/2} - X_t\|^2 \right)
\end{aligned}$$

First off, we bound the inner product term using Cauchy-Schwarz and a slight generalization of Young's inequality [49]

$$\begin{aligned}
&\sum_{t=1}^{T} a_t \langle g(\bar{X}_{t+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{t+1/2}; \tilde{X}_t), X_{t+1/2} - X_{t+1} \rangle \\
&\quad \leq \sum_{t=1}^{T} a_t \|g(\bar{X}_{t+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{t+1/2}; \tilde{X}_t)\| \|X_{t+1/2} - X_{t+1}\| \\
&\quad \leq \frac{1}{2} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|g(\bar{X}_{t+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{t+1/2}; \tilde{X}_t)\|^2 + \frac{1}{\gamma_{t+1}} \|X_{t+1/2} - X_{t+1}\|^2.
\end{aligned}$$

We merge the expressions together,

$$\begin{aligned}
&\sum_{t=1}^{T} \langle a_t g(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star \rangle \\
&\quad \leq \frac{1}{2} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|g(\bar{X}_{t+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{t+1/2}; \tilde{X}_t)\|^2 + \frac{1}{\gamma_{t+1}} \|X_{t+1/2} - X_{t+1}\|^2 \\
&\qquad + \sum_{t=1}^{T} \frac{1}{2\gamma_t} \left(\|X_t - x^\star\|^2 - \|X_{t+1} - x^\star\|^2 - \|X_{t+1/2} - X_{t+1}\|^2 - \|X_{t+1/2} - X_t\|^2 \right)
\end{aligned}$$

It is important that we invoke generalized Young's inequality with step-size at time $t+1$. Since the step-size lags one iteration behind, γ_t does not include $\|g(\bar{X}_{t+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{t+1/2}; \tilde{X}_t)\|^2$ and this would pose some problems in the later stages of the proof. Hence, we add/subtract $\frac{1}{\gamma_{t+1}} \|X_{t+1/2} - X_t\|^2$

and regroup the terms,

$$
\begin{aligned}
&\sum_{t=1}^{T}\langle a_t g(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star\rangle \\
&\quad\le \frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|g(\bar{X}_{t+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 - \frac{1}{\gamma_{t+1}}\|X_{t+1/2} - X_t\|^2 \\
&\qquad + \frac{1}{2}\sum_{t=1}^{T}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\left(\|X_{t+1/2} - X_{t+1}\|^2 + \|X_{t+1/2} - X_t\|^2\right) \\
&\qquad + \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\gamma_t}\left(\|X_t - x^\star\|^2 - \|X_{t+1} - x^\star\|^2\right) \\
&\quad\le \frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|g(\bar{X}_{t+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 - \frac{1}{\gamma_{t+1}}\|X_{t+1/2} - X_t\|^2 \\
&\qquad + \frac{\|X_1 - x^*\|^2}{2\gamma_1} + \frac{1}{2}\sum_{t=1}^{T-1}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\|X_{t+1} - x^\star\|^2 + D^2\sum_{t=1}^{T}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right) \\
&\quad\le \frac{3D^2}{2\gamma_{T+1}} + \frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|g(\bar{X}_{t+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 - \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\gamma_{t+1}}\|X_{t+1/2} - X_t\|^2
\end{aligned}
$$

where we have rewritten the telescoping summation for $\|X_t - x^\star\|^2 - \|X_{t+1} - x^\star\|^2$ and used that $D^2 = \sup_{x,y\in\mathcal{X}}\|x - y\|^2$ (diameter of the constraint set) to obtain the second inequality. The final line follows from telescoping the summations, plugging in the diameter D and rearranging the resulting terms.

Now, what remains is to obtain the (expected) regret from $\sum_{t=1}^{T}\langle a_t g(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star\rangle$. Taking expectation over all randomness,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\langle a_t g(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} a_t\langle g(\bar{X}_{t+1/2}) - \nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star\rangle + a_t\langle\nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[\, a_t\langle g(\bar{X}_{t+1/2}) - \nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star\rangle \mid X_{t+1/2}\,\right]\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} a_t\langle\nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} a_t\langle\mathbb{E}\left[\, g(\bar{X}_{t+1/2}) \mid X_{t+1/2}\,\right] - \nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star\rangle\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} a_t\langle\nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} a_t\langle\nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star\rangle\right]
\end{aligned}
$$

We used towering property of expectation (equivalently total law of expectation) to have the second inequality, and the last line from the unbiasedness assumption of gradient oracles in Eq. (2) such that

$\mathbb{E}\left[\, g(\bar{X}_{t+1/2}) \mid X_{t+1/2} \,\right] = \nabla f(\bar{X}_{t+1/2})$. Hence, we obtain that

$$\begin{aligned}\mathbb{E}\left[\,\textsc{Reg}_{\mathrm{T}}(x^*)\,\right] &= \mathbb{E}\left[\sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star\rangle\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \langle a_t g(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star\rangle\right],\end{aligned}$$

which concludes the target result,

$$\mathbb{E}\left[\,\textsc{Reg}_{\mathrm{T}}(x^*)\,\right] \leq \frac{1}{2}\mathbb{E}\left[\frac{3D^2}{\gamma_{T+1}} + \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|g(\bar{X}_{t+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{t+1/2}; \tilde{X}_t)\|^2 - \frac{\|X_{t+1/2} - X_t\|^2}{\gamma_{t+1}}\right]$$

■

F Technical Lemma for the Main Proofs

Before proceeding with the proofs of our main results, we need to establish the following technical result, due to [35] and [31], which has been commonly used in the analysis of adaptive methods. We make use of it for the proof of Theorem 3.1 and Theorem 3.2.

Lemma F.1 (35, 31)**.** *For all non-negative numbers $\alpha_1, \dots \alpha_t$, the following inequality holds:*

$$\sqrt{\sum_{t=1}^{T} \alpha_t} \leq \sum_{t=1}^{T} \frac{\alpha_t}{\sqrt{\sum_{i=1}^{t} \alpha_i}} \leq 2\sqrt{\sum_{t=1}^{T} \alpha_t} \tag{13}$$

G Implicit Accelerated Second-order Algorithm (Theorem 3.1)

In this section, we will provide the analysis of the implicit algorithm (Implicit Second-Order) under deterministic oracles. To do so, we will first start with a corollary result based on Proposition 3.1 that essentially proves the same template inequality under deterministic oracle model. In fact, one could easily show that Proposition 3.1 holds exactly up to replacing stochastic evaluations $g(\cdot)$ and $\tilde{\mathbf{F}}(\cdot;\cdot)$ with $\nabla f(\cdot)$ and $\mathbf{F}(\cdot;\cdot)$. For completeness, we will provide the proof of the aforementioned result in Proposition G.1 which follows the same steps as the proof of Proposition 3.1.

Proposition G.1. *Let $\{X_{t+1/2}\}_{t=1}^{T}$ be generated by* (Implicit Second-Order)*, run with a non-increasing step-size sequence γ_t and non-decreasing sequences of weights $a_t, b_t \geq 1$ such that a_t/b_t is also non-increasing. Then, the following guarantee holds:*

$$\textsc{Reg}_{\mathrm{T}}(x^*) \leq \frac{1}{2}\left(\frac{3D^2}{\gamma_{T+1}} + \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2}; \tilde{X}_t)\|^2 - \frac{\|X_{t+1/2} - X_t\|^2}{\gamma_{t+1}}\right).$$

Proof. We take off from the optimality conditions associated with each update sequence for the implicit algorithm (Implicit Second-Order). Optimality condition for $X_{t+1/2}$ implies for any $z_0 \in \mathcal{X}$,

$$\begin{aligned}&\langle a_t \mathbf{F}(\bar{X}_{t+1/2}; \tilde{X}_t), X_{t+1/2} - z_0\rangle \\ &\leq \frac{1}{\gamma_t}\langle X_{t+1/2} - X_t, z_0 - X_{t+1/2}\rangle \\ &= \frac{1}{2\gamma_t}\left(\|X_t - z_0\|^2 - \|X_{t+1/2} - z_0\|^2 - \|X_{t+1/2} - X_t\|^2\right)\end{aligned} \tag{14}$$

Similarly, optimality of X_{t+1} update yields for any $z_1 \in \mathcal{X}$,

$$\begin{aligned}\langle a_t \nabla f(\bar{X}_{t+1/2}), X_{t+1} - z_1\rangle &\leq \frac{1}{2\gamma_t}\langle X_{t+1} - X_t, z_1 - X_{t+1}\rangle \\ &= \frac{1}{2\gamma_t}\left(\|X_t - z_1\|^2 - \|X_{t+1} - z_1\|^2 - \|X_{t+1} - X_t\|^2\right)\end{aligned} \tag{15}$$

First, we will set $z_1 = x^*$ to establish the telescoping summation over $\|X_t - x^*\|^2 - \|X_{t+1} - x^*\|^2$. Then, we will simply align the above expression with the regret as follows,

$$\begin{aligned}
&\langle a_t \nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star \rangle \\
&\qquad = \langle a_t \nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - X_{t+1} \rangle + \langle a_t g(\bar{X}_{t+1/2}), X_{t+1} - x^\star \rangle \\
&\qquad \leq \langle a_t \nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - X_{t+1} \rangle \\
&\qquad\quad + \frac{1}{2\gamma_t} \left(\|X_t - x^\star\|^2 - \|X_{t+1} - x^\star\|^2 - \|X_{t+1} - X_t\|^2 \right)
\end{aligned} \tag{16}$$

Now, observe that setting $z_0 = X_{t+1}$ in Eq. (14) and rearranging we have

$$\begin{aligned}
&- \frac{1}{2\gamma_t} \|X_{t+1} - X_t\|^2 \\
&\leq -\langle a_t \mathbf{F}(\bar{X}_{t+1/2}; \tilde{X}_t), X_{t+1/2} - X_{t+1} \rangle - \frac{1}{2\gamma_t} \left(\|X_{t+1/2} - X_{t+1}\|^2 + \|X_{t+1/2} - X_t\|^2 \right)
\end{aligned}$$

Plugging the above expression into Eq. (16) and summing over $t = 1, ..., T$, we will obtain the offline weighted regret,

$$\begin{aligned}
&\textsc{Reg}_{\mathrm{T}}(x^*) \\
&\quad = \sum_{t=1}^{T} \langle a_t \nabla f(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star \rangle \\
&\quad \leq \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2}; \tilde{X}_t), X_{t+1/2} - X_{t+1} \rangle \\
&\qquad + \sum_{t=1}^{T} \frac{1}{2\gamma_t} \left(\|X_t - x^\star\|^2 - \|X_{t+1} - x^\star\|^2 - \|X_{t+1/2} - X_{t+1}\|^2 - \|X_{t+1/2} - X_t\|^2 \right)
\end{aligned}$$

First off, we bound the inner product term using Cauchy-Schwarz and a slight generalization of Young's inequality [49]

$$\begin{aligned}
&\sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2}; \tilde{X}_t), X_{t+1/2} - X_{t+1} \rangle \\
&\quad \leq \sum_{t=1}^{T} a_t \|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2}; \tilde{X}_t)\| \|X_{t+1/2} - X_{t+1}\| \\
&\quad \leq \frac{1}{2} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2}; \tilde{X}_t)\|^2 + \frac{1}{\gamma_{t+1}} \|X_{t+1/2} - X_{t+1}\|^2.
\end{aligned}$$

We merge the expressions together,

$$\begin{aligned}
&\textsc{Reg}_{\mathrm{T}}(x^*) \\
&\quad \leq \frac{1}{2} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2}; \tilde{X}_t)\|^2 + \frac{1}{\gamma_{t+1}} \|X_{t+1/2} - X_{t+1}\|^2 \\
&\qquad + \sum_{t=1}^{T} \frac{1}{2\gamma_t} \left(\|X_t - x^\star\|^2 - \|X_{t+1} - x^\star\|^2 - \|X_{t+1/2} - X_{t+1}\|^2 - \|X_{t+1/2} - X_t\|^2 \right)
\end{aligned}$$

It is important that we invoke generalized Young's inequality with step-size at time $t+1$. Since the step-size lags one iteration behind, γ_t does not include $\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2}; \tilde{X}_t)\|^2$ and this would pose some problems in the later stages of the proof. Hence, we add/subtract $-\frac{1}{\gamma_{t+1}} \|X_{t+1/2} - X_t\|^2$

and regroup the terms,

$$
\begin{aligned}
&\textsc{Reg}_{\mathrm{T}}(x^*) \\
&\quad \leq \frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 - \frac{1}{\gamma_{t+1}}\|X_{t+1/2} - X_t\|^2 \\
&\qquad + \frac{1}{2}\sum_{t=1}^{T}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\left(\|X_{t+1/2} - X_{t+1}\|^2 + \|X_{t+1/2} - X_t\|^2\right) \\
&\qquad + \frac{1}{2}\sum_{t=1}^{T}\frac{1}{\gamma_t}\left(\|X_t - x^\star\|^2 - \|X_{t+1} - x^\star\|^2\right) \\
&\quad \leq \frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 - \frac{1}{\gamma_{t+1}}\|X_{t+1/2} - X_t\|^2 \\
&\qquad + \frac{\|X_1 - x^*\|^2}{2\gamma_1} + \frac{1}{2}\sum_{t=1}^{T-1}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right)\|X_{t+1} - x^\star\|^2 + D^2\sum_{t=1}^{T}\left(\frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t}\right) \\
&\quad \leq \frac{1}{2}\left(\frac{3D^2}{\gamma_{T+1}} + \sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 - \frac{1}{\gamma_{t+1}}\|X_{t+1/2} - X_t\|^2\right),
\end{aligned}
$$

where we have rewritten the telescoping summation for $\|X_t - x^\star\|^2 - \|X_{t+1} - x^\star\|^2$ and used that $D^2 = \sup_{x,y\in\mathcal{X}}\|x - y\|^2$ (diameter of the constraint set) to obtain the second inequality. The final line follows from telescoping the summations, plugging in the diameter D and rearranging the resulting terms, which gives us the target inequality. ■

Theorem 3.1. *Let $\{X_{t+1/2}\}_{t=1}^T$ be generated by* (Implicit Second-Order) *run with the adaptive step-size policy* (7) *where $a_t = t^2$, $b_t = t^3$. Assume that f satisfies* (H-smooth) *and denote the diameter of the set as D with $\sigma = \max\{\sigma_g, \sigma_H\}$. Then, the following guarantee holds:*

$$
f(\bar{X}_{T+1/2}) - f(x^*) \leq O\left(\frac{L\left(D^4/\gamma^2 + D\gamma\right)}{T^3}\right).
$$

When $\gamma = D$, we obtain the converge rate $O\left(\frac{LD^3}{T^3}\right)$.

Proof. We will initiate our proof at template regret inequality as we proved in Proposition G.1. Our overall strategy is straightforward; we first prove a constant upper bound for the offline weighted regret, then make use of the conversion result in Theorem 3.3 to obtain a convergence rate of order $O(1/T^3)$.

Due to Proposition G.1 we have,

$$
\textsc{Reg}_{\mathrm{T}}(x^*) \leq \frac{1}{2}\left(\frac{3D^2}{\gamma_{T+1}} + \sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 - \frac{\|X_{t+1/2} - X_t\|^2}{\gamma_{t+1}}\right)
$$

We will merge the first two terms and express the first term in the form of the second one using Lemma F.1.

$$
\begin{aligned}
&\textsc{Reg}_{\mathrm{T}}(x^*) \\
&\leq \frac{3D^2}{2\gamma}\sqrt{1+\sum_{t=1}^{T} a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2} \\
&\quad + \frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 - \frac{\|X_{t+1/2}-X_t\|^2}{\gamma_{t+1}} \\
&\leq \frac{3D^2}{2\gamma} + \frac{3D^2}{2\gamma}\sum_{t=1}^{T}\frac{a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2}{\sqrt{1+\sum_{s=1}^{t} a_s^2\|\nabla f(\bar{X}_{s+1/2}) - \mathbf{F}(\bar{X}_{s+1/2};\tilde{X}_s)\|^2}} \\
&\quad + \frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 - \frac{\|X_{t+1/2}-X_t\|^2}{\gamma_{t+1}} \\
&= \frac{3D^2}{2\gamma} + \frac{3D^2}{2\gamma^2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 \\
&\quad + \frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 - \frac{\|X_{t+1/2}-X_t\|^2}{\gamma_{t+1}} \\
&= \frac{3D^2}{2\gamma} + \frac{1}{2}\sum_{t=1}^{T}\frac{3D^2+\gamma^2}{\gamma^2}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 - \frac{\|X_{t+1/2}-X_t\|^2}{\gamma_{t+1}},
\end{aligned}
$$

where we obtain the second inequality due to Lemma F.1 and the last two lines follow from the definition of the step-size in Eq. (7) and appropriate regrouping. As we have discussed in the main text, the key for faster rates beyond $O(1/T^2)$ is understanding how to manipulate the negative term in the above expression. A crucial part of our analysis is understanding the implications of second-order smoothness and how to unlock its potential. This next derivation will demonstrate how (H-smooth) allows for a more aggressive gradient weighting and in turn faster convergence rate implied by our generalized conversion technique. Next, we will relate the negative term to the positive terms using smoothness and primal averaging, similar to the approaches in [51, 26].

$$
\begin{aligned}
-\frac{\|X_{t+1/2}-X_t\|^2}{\gamma_{t+1}} &= -\frac{D^2}{D^2\gamma_{t+1}}\|X_{t+1/2}-X_t\|^2 \\
&\leq -\frac{1}{D^2\gamma_{t+1}}\|X_{t+1/2}-X_t\|^4 \\
&= -\frac{1}{D^2\gamma_{t+1}}\frac{B_t^4}{b_t^4}\|\frac{b_t}{B_t}X_{t+1/2} - \frac{b_t}{B_t}X_t\|^4 \\
&= -\frac{1}{D^2\gamma_{t+1}}\frac{B_t^4}{b_t^4}\|\frac{b_tX_{t+1/2}+\sum_{s=1}^{t-1}b_sX_{s+1/2}}{B_t} - \frac{b_tX_t+\sum_{s=1}^{t-1}a_sX_{s+1/2}}{B_t}\|^4 \\
&= -\frac{1}{D^2\gamma_{t+1}}c^4t^4\|\bar{X}_{t+1/2}-\tilde{X}_t\|^4 \\
&\leq -\frac{4c^4}{L^2D^2\gamma_{t+1}}t^4\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2
\end{aligned}
$$

First, notice that for any sequence $b_t = O(t^p)$ with $p \geq 0$, we have $B_t = \sum_{s=1}^{t} b_s = O(t^{p+1})$, which implies $\frac{B_t}{b_t} \leq ct$, where $c > 0$ is an absolute constant depending on how b_t is defined. Then, we use the definitions of average sequences $\bar{X}_{t+1/2}$ and $\tilde{X}_t$ to go from $\|X_{t+1/2}-X_t\|^4$ to $\|\bar{X}_{t+1/2}-\tilde{X}_t\|^4$ to obtain equalities 3-5, and apply smoothness to obtain the last line. On a related note, we want to highlight the importance of optimistic weighted averaging that is central for obtaining the above

expression. Since the averaged pairs $\bar{X}_{t+1/2}$ and $\tilde{X}_t$ differ by only the last element, we can seamlessly relate $\|X_{t+1/2} - X_t\|$ to $\|\bar{X}_{t+1/2} - \tilde{X}_t\|$.

Now, we are at a position to explain how we will go beyond $O(1/T^2)$ convergence rate, which fundamentally depends on the gradient weights a_t and jointly relies on our generalized online-to-batch conversion in Theorem 3.3. The negative term above is monotonically decreasing (increases in magnitude) which is essential to (partially) control the growth of remaining positive term. More specifically, one can notice that in order to align the summands of the positive and negative term, the algebra dictates that we need to select $a_t = O(t^2)$, which implies $b_t = \Omega(t^2)$. Notice that our averaging and weighting parameters grow at least $O(t)$ faster than the existing accelerated schemes for first-order smoothness, which grants the improved $O(1/T^3)$ rate. On the contrary, first-order smoothness would only allow t^2 factor in front of the norm, leading to the slower rate.

Setting $a_t = t^2$, plugging the bound on the negative term into the original expression we have,

$$\leq \frac{3D^2}{2\gamma} + \frac{1}{2}\sum_{t=1}^{T}\left(\frac{3D^2+\gamma^2}{\gamma^2} - \frac{4c^4}{L^2D^2\gamma_{t+1}^2}\right)\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 \quad (17)$$

Our main objective is to show that the above summation is summable so we could show the constant upper bound for the offline regret, hence the acceleration. First off, notice that $\left(\frac{3D^2+\gamma^2}{\gamma^2} - \frac{4c^4}{L^2D^2\gamma_{t+1}^2}\right)$ is a non-increasing quantity and we are interested in the time point at which this quantity become negative. For that reason, we define the following time point,

$$T_0 = \max\left\{t \in \mathbb{N} \mid \left(\frac{3D^2+\gamma^2}{\gamma^2} - \frac{4c^4}{L^2D^2\gamma_{t+1}^2}\right) \geq 0\right\}.$$

This immediately implies that for any $t \leq T_0$,

$$\frac{1}{\gamma_{t+1}} \leq \frac{LD\sqrt{3D^2+\gamma^2}}{2\gamma c^2}. \quad (18)$$

To paint a complete picture, we would like to have a brief discussion on the possible values for T_0.

1. $T_0 = 0$ implies that the step-size is small enough from the very beginning and that the summation term in Eq. (17) is always bounded by a constant, which immediately implies constant regret and $O(1/T^3)$ rate.
2. $T_0 = \infty$ implies that the step-size is always lower bounded by the *inverse of the constant on the right-hand side* of Eq.(18). This is equivalent to saying $\sum_{t=1}^{\infty} a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 \leq C$ for some constant C, which in turn ensures that the summation in Eq. (17) is summable. Once again, we will have the constant regret and $O(1/T^3)$ rate.
3. When T_0 is a finite positive integer, we can upper bound the summation in Eq. (17) with the same summation up to iteration T_0. Note that it is not important whether T is larger or smaller than T_0, as the summands change sign and become negative after T_0.

Therefore, we can safely unify all the 3 cases above and simply upper bound the expression in Eq. (17) by rewriting the summation up to T_0. Therefore,

$$
\begin{aligned}
&\leq \frac{3D^2}{2\gamma} + \frac{1}{2}\sum_{t=1}^{T}\left(\frac{3D^2+\gamma^2}{\gamma^2} - \frac{4c^4}{L^2D^2\gamma_{t+1}^2}\right)\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 \\
&\leq \frac{3D^2}{2\gamma} + \frac{3D^2+\gamma^2}{2\gamma^2}\sum_{t=1}^{T_0}\gamma_{t+1}a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 \\
&= \frac{3D^2}{2\gamma} + \frac{3D^2+\gamma^2}{2\gamma}\sum_{t=1}^{T_0}\frac{a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2}{\sqrt{1+\sum_{s=1}^{t}a_s^2\|\nabla f(\bar{X}_{s+1/2}) - \mathbf{F}(\bar{X}_{s+1/2};\tilde{X}_s)\|^2}} \\
&\leq \frac{3D^2+\gamma^2}{\gamma}\sqrt{1+\sum_{t=1}^{T_0}a_t^2\|\nabla f(\bar{X}_{t+1/2}) - \mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2} \\
&= \left(3D^2+\gamma^2\right)\frac{1}{\gamma_{T_0+1}} \\
&\leq \frac{LD\left(3D^2+\gamma^2\right)^{3/2}}{2\gamma c^2}
\end{aligned}
$$

Hence, we established the constant regret bound

$$
\textsc{Reg}_{\mathrm{T}}(x^*) \leq \frac{LD\left(3D^2+\gamma^2\right)^{3/2}}{2\gamma c^2}
$$

Before concluding the convergence proof, we would like to have a quick detour on the value of c. The value of c is roughly between $[1/p, 1]$, where p is the exponent of the averaging weight, $b_t = t^p$. For instance, when we pick $b_t = t^2$, we have $t^3/3 \leq B_t \leq t^3$; and when $b_t = t^3$, $t^4/4 \leq B_t \leq t^4$. Hence, we can avoid its effect in the final bound. Therefore, the online-to-batch conversion implies a conversion of order $O(\textsc{Reg}_{\mathrm{T}}(x^*)/T^3)$. Plugging this result in its place we obtain the convergence rate,

$$
f(\bar{X}_{T+1/2}) - f(x^*) \leq \frac{L\left(D^4/\gamma + D\gamma^2\right)}{T^3}
$$

■

H EXTRANEWTON: The First Universal Second-order Accelerated Method (Theorem 3.2)

Theorem 3.2. *Let $\{X_{t+1/2}\}_{t=1}^{T}$ be generated by Algorithm 1, run with the adaptive step-size policy (8) and $a_t = t^2, b_t = t^3$. Assume that f satisfies (H-smooth), and that Assumptions (2) hold with $\sigma = \max\{\sigma_g, \sigma_H\}$. Then, the following universal guarantee holds:*

$$
\mathbb{E}\left[f(\bar{X}_{T+1/2}) - f(x^*)\right] \leq O\left(\frac{\max\{D,1\}\left(D^2/\gamma+\gamma\right)\sigma}{\sqrt{T}} + \frac{L\left(D^4/\gamma + D\gamma^2\right)}{T^3}\right).
$$

When $\gamma = D$, we obtain the target rate $O\left(\frac{\max\{D,1\}D\sigma}{\sqrt{T}} + \frac{LD^3}{T^3}\right)$.

Proof. We take Proposition 3.1 as our departure point for the analysis. After proving an offline regret bound, we will use Theorem 3.3 to obtain the optimality gap from the regret bound. Recall the template regret bound,

$$
\mathbb{E}\left[\textsc{Reg}_{\mathrm{T}}(x^*)\right] \leq \frac{1}{2}\mathbb{E}\left[\frac{3D^2}{\gamma_{T+1}} + \sum_{t=1}^{T}\gamma_{t+1}a_t^2\|g(\bar{X}_{t+1/2}) - \tilde{\mathbf{F}}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 - \frac{\|X_{t+1/2}-X_t\|^2}{\gamma_{t+1}}\right]
$$

Now, we want to unify the first two terms through numerical inequalities. Different from the analysis of the implicit algorithm with deterministic oracles, we will write the *second term in terms of the first term*, instead. Due to Lemma F.1, we can upper the bound second term as,

$$\begin{aligned}
&\frac{1}{2}\sum_{t=1}^{T}\gamma_{t+1}a_t^2\|g(\bar{X}_{t+1/2})-\tilde{\mathbf{F}}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2 \\
&\qquad = \frac{\gamma}{2}\sum_{t=1}^{T}\frac{a_t^2\|g(\bar{X}_{t+1/2})-\tilde{\mathbf{F}}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2}{\sqrt{1+\sum_{s=1}^{t}a_s^2\|g(\bar{X}_{s+1/2})-\tilde{\mathbf{F}}(\bar{X}_{s+1/2};\tilde{X}_s)\|^2}} \\
&\qquad \le \gamma\sqrt{1+\sum_{t=1}^{T}a_t^2\|g(\bar{X}_{t+1/2})-\tilde{\mathbf{F}}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2}-\frac{\gamma}{2}
\end{aligned}$$

Plugging this back into the original expression gives us

$$\begin{aligned}
&\mathbb{E}\left[\,\textsc{Reg}_{\mathrm{T}}(x^*)\,\right] \\
&\le \left(\frac{3D^2}{2\gamma}+\gamma\right)\sqrt{1+\sum_{t=1}^{T}a_t^2\|g(\bar{X}_{t+1/2})-\tilde{\mathbf{F}}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2}-\frac{1}{2}\sum_{t=1}^{T}\frac{1}{\gamma_{t+1}}\|X_{t+1/2}-X_t\|^2
\end{aligned}$$

Next up, we will handle the negative term in the above expression. We will do so in exactly the same way as we did for the analysis of Theorem 3.1 and make use of the second-order smoothness to express it in a compatible form with the other terms.

$$\begin{aligned}
-\frac{\|X_{t+1/2}-X_t\|^2}{\gamma_{t+1}} &= -\frac{D^2}{D^2\gamma_{t+1}}\|X_{t+1/2}-X_t\|^2 \\
&\le -\frac{1}{D^2\gamma_{t+1}}\|X_{t+1/2}-X_t\|^4 \\
&= -\frac{1}{D^2\gamma_{t+1}}\frac{B_t^4}{b_t^4}\|\frac{b_t}{B_t}X_{t+1/2}-\frac{b_t}{B_t}X_t\|^4 \\
&= -\frac{1}{D^2\gamma_{t+1}}\frac{B_t^4}{b_t^4}\|\frac{b_tX_{t+1/2}+\sum_{s=1}^{t-1}b_sX_{s+1/2}}{B_t}-\frac{b_tX_t+\sum_{s=1}^{t-1}b_sX_{s+1/2}}{B_t}\|^4 \\
&= -\frac{1}{D^2\gamma_{t+1}}c^4t^4\|\bar{X}_{t+1/2}-\tilde{X}_t\|^4 \\
&\le -\frac{4c^4t^4}{L^2D^2\gamma_{t+1}}\|\nabla f(\bar{X}_{t+1/2})-\mathbf{F}(\bar{X}_{t+1/2};\tilde{X}_t)\|^2
\end{aligned}$$

We obtain the last line the same way as we discussed in the proof of Theorem 3.1 in the previous section. Please refer to the two paragraphs above Eq.(17) for a detailed discussion on the choice of a_t and b_t, as well as the key to achieving acceleration beyond $O(1/T^2)$. Due to (margin-wise) space constraints, we will use a slightly more compact notation for certain expressions. Let us first define a shorthand notation for noise in gradient and Hessian evaluations, respectively.

$$\begin{aligned}
\epsilon_t &= [g(\bar{X}_{t+1/2})-g(\tilde{X}_t)]-[\nabla f(\bar{X}_{t+1/2})-\nabla f(\tilde{X}_t)] \\
\delta_t &= H(\tilde{X}_t)-\nabla^2 f(\tilde{X}_t)
\end{aligned} \tag{19}$$

Then, we define following deterministic/stochastic placeholders:

$$\begin{aligned}
\nabla_t &= \nabla f(\bar{X}_{t+1/2})-\mathbf{F}(\bar{X}_{t+1/2},\tilde{X}_t) \\
\widetilde{\nabla}_t &= g(\bar{X}_{t+1/2})-\tilde{\mathbf{F}}(\bar{X}_{t+1/2},\tilde{X}_t)=\nabla_t+\epsilon_t+\delta_t(\bar{X}_{t+1/2}-\tilde{X}_t)
\end{aligned} \tag{20}$$

Setting $a_t = t^2$, combining all the terms and introducing the compact notation,

$$\sum_{t=1}^{T} \langle a_t g(\bar{X}_{t+1/2}), X_{t+1/2} - x^\star \rangle$$
$$\leq \left(\frac{3D^2}{2\gamma} + \gamma \right) \sqrt{1 + \sum_{t=1}^{T} a_t^2 \|\widetilde{\nabla}_t\|^2} - \sum_{t=1}^{T} \frac{2c^4}{L^2 D^2 \gamma_{t+1}} a_t^2 \|\nabla_t\|^2$$

At this point, we need to understand how to relate $\|\nabla_t\|^2$ and $\|\widetilde{\nabla}_t\|^2$ while treating the step-size γ_{t+1} accordingly. The issue is that the step-size is agnostic to deterministic oracle information since we accumulate $\|\widetilde{\nabla}_t\|^2$. From the perspective of step-size, we need to find a relevant, if not matching, lower bound for $\|\nabla_t\|^2$ and $\|\widetilde{\nabla}_t\|^2$. Indeed, we follow the ideas presented in [26], and begin by (trivially) lower bounding both terms with the same expression,

$$\begin{aligned} \|\widetilde{\nabla}_t\|^2 &\geq \min \left\{ \|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2 \right\} \\ \|\nabla_t\|^2 &\geq \min \left\{ \|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2 \right\} \end{aligned} \tag{21}$$

Now, we will decompose $\|\widetilde{\nabla}_t\|^2$ into $\|\nabla_t\|^2$ and the noise terms. Using the definitions in Eq. (19) and (20) and applying triangular inequality with quadratic expansion,

$$\|\widetilde{\nabla}_t\|^2 \leq 2\|\nabla_t\|^2 + 4\|\delta_t(\bar{X}_{t+1/2} - \tilde{X}_t)\|^2 + 4\|\epsilon_t\|^2 \tag{22}$$

We can also have the following trivial upper bound,

$$\begin{aligned} \|\widetilde{\nabla}_t\|^2 &\leq 2\|\widetilde{\nabla}_t\|^2 \\ &\leq 2\|\widetilde{\nabla}_t\|^2 + 4\|\delta_t(\bar{X}_{t+1/2} - \tilde{X}_t)\|^2 + 4\|\epsilon_t\|^2 \end{aligned} \tag{23}$$

Let us simplify the relationship between the bounds in Eq. (22) and Eq. (23); if $\|\nabla_t\|^2 \leq \|\widetilde{\nabla}_t\|^2$, then Eq. (22) is tighter, otherwise Eq. (23) is tighter. Hence, we could select the minimum of $\|\nabla_t\|^2$ and $\|\widetilde{\nabla}_t\|$:

$$\|\widetilde{\nabla}_t\|^2 \leq 2 \min \left\{ \|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2 \right\} + 4\|\delta_t(\bar{X}_{t+1/2} - \tilde{X}_t)\|^2 + 4\|\epsilon_t\|^2 \tag{24}$$

Using this intuition, we can construct a variable λ_t that always upper bounds the step-size.

$$\lambda_t = \frac{\gamma}{\sqrt{1 + \sum_{s=1}^{t-1} a_s^2 \min \left\{ \|\widetilde{\nabla}_s\|^2, \|\nabla_s\|^2 \right\}}} \tag{25}$$

It is immediate that $\gamma_t \leq \lambda_t$. Essentially, we will replace the terms $\|\nabla_t\|^2$ and $\|\widetilde{\nabla}_t\|$ with $\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}$, $\|\delta_t(\bar{X}_{t+1/2} - \tilde{X}_t)\|^2$ and $\|\epsilon_t\|^2$.

$$
\begin{aligned}
&\mathbb{E}\left[\,\textsc{Reg}_{\mathrm{T}}(x^*)\,\right] \\
&\leq \mathbb{E}\left[\frac{3D^2+2\gamma^2}{2\gamma}\sqrt{1+\sum_{t=1}^{T} a_t^2\|\widetilde{\nabla}_t\|^2} - \sum_{t=1}^{T}\frac{2c^4}{L^2D^2\gamma_{t+1}}a_t^2\|\nabla_t\|^2\right] \\
&\leq \mathbb{E}\left[\frac{3D^2+2\gamma^2}{2\gamma}\sqrt{1+\sum_{t=1}^{T} 2a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\} + 4a_t^2\|\delta_t(\bar{X}_{t+1/2} - \tilde{X}_t)\|^2 + 4a_t^2\|\epsilon_t\|^2}\right. \\
&\quad \left. - \sum_{t=1}^{T}\frac{2c^4}{L^2D^2\lambda_{t+1}}a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}\right] \\
&\leq \mathbb{E}\left[\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma}\sqrt{1+\sum_{t=1}^{T} a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}} - \sum_{t=1}^{T}\frac{2c^4a_t^2}{L^2D^2\lambda_{t+1}}\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}\right. \\
&\quad \left. + 2\left(\frac{3D^2}{2\gamma}+\gamma\right)\sqrt{\sum_{t=1}^{T} a_t^2\|\delta_t(\bar{X}_{t+1/2} - \tilde{X}_t)\|^2} + 2\left(\frac{3D^2}{2\gamma}+\gamma\right)\sqrt{\sum_{t=1}^{T} a_t^2\|\epsilon_t\|^2}\ \right] \\
&\leq \mathbb{E}\left[\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma^2}\sum_{t=1}^{T}\lambda_{t+1}a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\} - \sum_{t=1}^{T}\frac{2c^4a_t^2}{L^2D^2\lambda_{t+1}}\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}\right. \\
&\quad \left. + 2\left(\frac{3D^2}{2\gamma}+\gamma\right)\sqrt{\sum_{t=1}^{T} a_t^2\|\delta_t(\bar{X}_{t+1/2} - \tilde{X}_t)\|^2} + 2\left(\frac{3D^2}{2\gamma}+\gamma\right)\sqrt{\sum_{t=1}^{T} a_t^2\|\epsilon_t\|^2}\ \right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma^2} - \frac{2c^4}{L^2D^2\lambda_{t+1}^2}\right)\lambda_{t+1}a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2, \|\nabla_t\|^2\right\}\right. \\
&\quad \left. + 2\left(\frac{3D^2}{2\gamma}+\gamma\right)\sqrt{\sum_{t=1}^{T} a_t^2\|\delta_t(\bar{X}_{t+1/2} - \tilde{X}_t)\|^2} + 2\left(\frac{3D^2}{2\gamma}+\gamma\right)\sqrt{\sum_{t=1}^{T} a_t^2\|\epsilon_t\|^2}\ \right]
\end{aligned}
$$

Next, we will simplify the first summation and eventually show that it has a finite, constant upper bound. First off, notice that $\left(\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma^2} - \frac{2c^4}{L^2D^2\lambda_{t+1}^2}\right)$ is a decreasing quantity and we are interested in the time point at which it changes signs. Let us define,

$$
T_0 = \max\left\{t \in \mathbb{N} \mid \left(\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma^2} - \frac{2c^4}{L^2D^2\lambda_{t+1}^2}\right) \geq 0\right\}.
$$

This immediately implies that for any $t \leq T_0$,

$$
\frac{1}{\lambda_{t+1}} \leq \frac{LD\sqrt{3D^2+2\gamma^2}}{2^{3/4}\gamma c^2}. \tag{26}
$$

Observe that the way we define T_0 enables us to *upper bound* the summation up to T, with the summation up to T_0. Hence,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma^2}-\frac{2c^4}{L^2D^2\lambda_{t+1}^2}\right)\lambda_{t+1}a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2,\|\nabla_t\|^2\right\}\right]\\
&\quad\le\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma}+\sum_{t=1}^{T_0}\left(\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma^2}-\frac{2c^4}{L^2D^2\lambda_{t+1}^2}\right)\lambda_{t+1}a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2,\|\nabla_t\|^2\right\}\\
&\quad\le\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma}+\frac{3D^2+2\gamma^2}{\sqrt{2}\gamma}\sum_{t=1}^{T_0}\frac{a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2,\|\nabla_t\|^2\right\}}{\sqrt{1+\sum_{s=1}^{t-1}a_s^2\min\left\{\|\widetilde{\nabla}_s\|^2,\|\nabla_s\|^2\right\}}}\\
&\quad\le\frac{3\sqrt{2}D^2+2\sqrt{2}\gamma^2}{\gamma}\sqrt{\sum_{t=1}^{T_0}a_t^2\min\left\{\|\widetilde{\nabla}_t\|^2,\|\nabla_t\|^2\right\}}\\
&\quad=\le\left(3\sqrt{2}D^2+2\sqrt{2}\gamma^2\right)\frac{1}{\lambda_{T_0+1}}\\
&\quad\le\frac{LD\left(3D^2+2\gamma^2\right)^{3/2}}{2^{1/4}\gamma c^2}
\end{aligned}
$$

This gives us the constant part of the regret, which will lead to the $O(1/T^3)$ part of the convergence rate. Now, what remains is to handle the "stochasticity". We will bound the remaining stochastic terms with respect to the stochastic gradient and the stochastic Hessian. Plugging the expected regret in to the bound and combining all the expressions together,

$$
\begin{aligned}
&\mathbb{E}\left[\,\textsc{Reg}_{\mathrm{T}}(x^*)\,\right]\\
&\le\frac{3D^2+2\gamma^2}{\gamma}\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}a_t^2\|\delta_t(\bar{X}_{t+1/2}-\tilde{X}_t)\|^2}+\sqrt{\sum_{t=1}^{T}a_t^2\|\epsilon_t\|^2}\right]+\frac{LD\left(3D^2+2\gamma^2\right)^{3/2}}{2^{1/4}\gamma c^2}\\
&\le\frac{3D^2+2\gamma^2}{\gamma}\left(\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\,a_t^2\|\delta_t(\bar{X}_{t+1/2}-\tilde{X}_t)\|^2\,\right]}+\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\,a_t^2\|\epsilon_t\|^2\,\right]}\right)\\
&\quad+\frac{LD\left(3D^2+2\gamma^2\right)^{3/2}}{2^{1/4}\gamma c^2}\\
&=\frac{3D^2+2\gamma^2}{\gamma}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\,a_t^2\mathbb{E}\left[\,\|\delta_t(\bar{X}_{t+1/2}-\tilde{X}_t)\|^2\mid\bar{X}_{t+1/2}\,\right]\right]}\\
&\quad+\frac{3D^2+2\gamma^2}{\gamma}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\,a_t^2\mathbb{E}\left[\,\|\epsilon_t\|^2\mid\bar{X}_{t+1/2}\,\right]\right]}+\frac{LD\left(3D^2+2\gamma^2\right)^{3/2}}{2^{1/4}\gamma c^2}\\
&\le\frac{3D^2+2\gamma^2}{\gamma}\left(\sqrt{\sigma_H^2\sum_{t=1}^{T}a_t^2\mathbb{E}\left[\,\|\bar{X}_{t+1/2}-\tilde{X}_t\|^2\,\right]}+\sqrt{4\sigma_g^2\sum_{t=1}^{T}a_t^2}\right)+\frac{LD\left(3D^2+2\gamma^2\right)^{3/2}}{2^{1/4}\gamma c^2}\\
&\le\frac{3D^2+2\gamma^2}{\gamma}\left(\sqrt{\sigma_H^2D^2\sum_{t=1}^{T}a_t^2\frac{b_t^2}{B_t^2}}+2\sigma_gT^{5/2}\right)+\frac{LD\left(3D^2+2\gamma^2\right)^{3/2}}{2^{1/4}\gamma c^2}\\
&\le\frac{LD\left(3D^2+2\gamma^2\right)^{3/2}}{2^{1/4}\gamma c^2}+\frac{3D^3+2D\gamma^2}{c\gamma}\sigma_HT^{3/2}+\frac{6D^2+4\gamma^2}{\gamma}\sigma_gT^{5/2}
\end{aligned}
$$

In order not to repeat the same arguments on the value of c, which we discussed in the last paragraph of the proof of Theorem 3.1, we will use the fact that it is an absolute constant for any fixed definition

of b_t. Running the above expression through Theorem 3.3 we obtain,

$$f(\bar{X}_{T+1/2}) - f(x^*) \leq O\left(\frac{\left(\frac{D^2}{\gamma} + \gamma\right)\sigma_g}{\sqrt{T}} + \frac{\left(\frac{D^3}{\gamma} + D\gamma\right)\sigma_H}{T^{3/2}} + \frac{L\left(\frac{D^4}{\gamma} + D\gamma^2\right)}{T^3}\right)$$

To write this result in a more compact manner, we define $\sigma = \max\{\sigma_g, \sigma_H\}$ and rearrange the terms to get

$$f(\bar{X}_{T+1/2}) - f(x^*) \leq O\left(\frac{\max\{D, 1\}\left(\frac{D^2}{\gamma} + \gamma\right)\sigma}{\sqrt{T}} + \frac{L\left(\frac{D^4}{\gamma} + D\gamma^2\right)}{T^3}\right)$$

■