EXTRA-NEWTON: A First Approach to Noise-Adaptive Accelerated Second-Order Methods

Kimon Antonakopoulos*  
EPFL (LIONS)  
kimon.antonakopoulos@epfl.ch

Ali Kavis*  
EPFL (LIONS)  
ali.kavis@epfl.ch

Volkan Cevher  
EPFL (LIONS)  
volkan.cevher@epfl.ch

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 $\sigma^2$, 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 $\sigma^2$, and $O(1/T^2)$ under deterministic first-order oracles [52, 58]. Under slight variations of the aforementioned problem setting, there exists an extensive amount of work that enjoys the latter, “accelerated” rate [56, 57, 53, 65, 69, 39, 2, 41, 67, 19, 18, 35, 31, 6, 44].

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 [11], Kantorovich [33], Newton’s method and its variations [40, 46] are considered as the staple of second-order methods in optimization. Although its powerful local convergence behavior has been repeatedly demonstrated [17, 38], studies on its global behavior are relatively limited. Prior attempts at tackling global convergence mostly make additional structural assumptions on the objective function [61, 47, 38] or assume extra regularity conditions on the Hessian [34] 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 attain the $O(\sigma/\sqrt{T} + 1/T^2)$ convergence for convex minimization problems, interpolating between stochastic and deterministic rates. There exist a plethora of algorithms that enjoy this rate under

*Alphabetical order, equal contribution

different sets of assumptions for both minimization scenarios (for convex and non-convex settings, we refer the reader to [39, 35, 22, 31, 6] and [68, 42, 36, 45], respectively), and the more general framework of variational inequalities [8, 5, 66, 25, 3, 26, 4]. 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 [59] 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 [55] 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 [49]. This idea has been studied further for different settings in convex optimization [28, 29] with the same accelerated \(O(1/T^3)\) rate and extended to non-convex realm [14, 15], 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 [10, 21, 16].

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, 7]. The first line of research that shrinks the gap between the upper and lower bounds for achieving an almost-optimal (more on this shortly) convergence [60] is the so-called “bisection-type” methods. Pioneered by Monteiro and Svaiter [50], 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 [23]. Not so surprisingly, the same construction finds application in variational inequality (VI) and min-max optimization literature [12, 30]. Very recently and concurrently to our work, [13] propose the first bisection free acceleration for second-order methods, that achieves the optimal \(O(1/T^{7/2})\). The authors define an explicit, deterministic procedure called MS oracle and compute the step-size using a standard line-search procedure enabling them to achieve optimal rates while adaptively computing the step-size without needing to know the smoothness constant.

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 [60, 43] and the resulting algorithms are complicated with many interconnected components. On the other hand, cubic regularization-based ideas propose a simple construction that achieves acceleration beyond \(O(1/T^2)\) however, similar to previous 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 oracles and it is an open question whether such frameworks could immediately accommodate stochastic first and second-order information.

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

1. We prove EXTRA-NEWTON achieves the global convergence rate of \(O(\frac{\sigma_g}{\sqrt{T}} + \frac{\sigma_H}{T} + \frac{LD^2}{T})\) that adapts simultaneously to the variance in the gradient oracle (\(\sigma_g\)) and Hessian oracle (\(\sigma_H\)) 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

<table>
<thead>
<tr>
<th></th>
<th>AGD</th>
<th>UniXGrad</th>
<th>Reg. Newton</th>
<th>Accel. Cubic Reg.</th>
<th>ANPE^2</th>
<th>OptMS</th>
<th>Extra Newton (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate</td>
<td>$\frac{1}{T}$</td>
<td>$\frac{1}{\sqrt{T}} + \frac{1}{T}$</td>
<td>$\frac{1}{T}$</td>
<td>$\frac{1}{T}$</td>
<td>$\frac{1}{T^2}$</td>
<td>$\frac{1}{T^2}$</td>
<td>$\frac{1}{T^2} + \frac{\sigma g}{T^2} + \frac{1}{3}$</td>
</tr>
<tr>
<td>Bisection-free</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Adapts to L</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Noise-adaptive</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>

3. We design the first adaptive step-size, in the sense of [20, 63], 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 [37] 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 [67, 19, 35] 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 [63, 35, 4] 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:

$$\text{minimize} \quad f(x)$$

subject to $x \in \mathcal{X}$. (Opt)

Formally, in the above $\mathcal{X}$ is a convex and compact subset of a $d$-dimensional normed space $\mathcal{V} \cong \mathbb{R}^d$ with diameter $D = \max_{x,y \in \mathcal{X}} \|x - y\|$, and $f: \mathcal{V} \to \mathbb{R} \cup \{+\infty\}$ is a proper, lower semi-continuous, convex function with $\text{dom} f = \{x \in \mathbb{R}^d : f(x) < +\infty\} \subset \mathcal{X}$. To that end, we make a set of blanket assumptions for (Opt). Following the vast literature of constrained convex minimization [54, 9], we consider “simple” constraint sets, i.e.,

**Assumption 2.1.** The constraint set $\mathcal{X}$ of (Opt) possesses favorable geometry which facilitates a tractable projection operator.

In order to avoid trivialities, we also assume that the said problem admits at least a solution, i.e.

**Assumption 2.2.** 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.,

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

(H-smooth)

and in addition it satisfies the second order approximation:

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

(Taylor)

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

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

(1)

The above equivalences are well-established and hence we omit their proofs (we defer for a panoramic view to [70]).

^2Note that the bisection procedure is computationally prohibitive, we defer the reader to [60], p.304-305.
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 we will use in algorithm definitions and technical discussions. Let \( g(x, \xi) \) denote the stochastic gradient evaluated at \( x \) with randomness defined by \( \xi \) and \( H(x, \xi) \) be the stochastic Hessian at \( x \) with \( \xi \) describing the randomness of the oracle, such that

\[
\mathbb{E} \left[ g(x, \xi) \mid x \right] = \nabla f(x), \quad \mathbb{E} \left[ \| g(x, \xi) - \nabla f(x) \|_2^2 \mid x \right] \leq \sigma_g^2 \\
\mathbb{E} \left[ H(x, \xi) \mid x \right] = \nabla^2 f(x), \quad \mathbb{E} \left[ \| H(x, \xi) - \nabla^2 f(x) \|_2^2 \mid x \right] \leq \sigma_H^2
\]  

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

\[
F(x; x') = \nabla f(x') + \frac{1}{2} \nabla^2 f(x')(x - x') \\
\tilde{F}(x; x', \xi) = g(x', \xi) + \frac{1}{2} H(x', \xi)(x - x')
\]

where \( F \) is essentially the gradient (with respect to \( x \)) of the second-order Taylor polynomial. By definition, the operator \( F \) 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 [37] and further developed in Nemirovski [51],

\[
X_{t+\frac{1}{2}} = \Pi_X (X_t - \gamma_t \nabla f(x_t)) \\
X_{t+1} = \Pi_X (X_t - \gamma_t \nabla f(x_{t+1/2}))
\]

(EG)

where \( \Pi_X (x) = \arg \min_{z \in X} \| x - z \|^2 \) is the standard Euclidean projection onto the set \( X \). In terms of output, the candidate solution returned by (EG) after \( T \) iterations is the so-called “ergodic average”

\[
\bar{X}_T = \frac{\sum_{t=1}^{T} b_t X_{t+\frac{1}{2}}}{\sum_{t=1}^{T} b_t}
\]

Then, taking \( b_t = \gamma_t \) and assuming the method’s step-size \( \gamma_t \) is chosen appropriately, \( \bar{X}_T \) enjoys the following universal guarantee [32, 62]:

\[
\mathbb{E} \left[ f(\bar{X}_T) - f(x^*) \right] = O \left( \frac{1}{T} + \frac{\sigma}{\sqrt{T}} \right)
\]

where \( \sigma \) 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 [56]) 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 [65, 69, 39, 2, 41, 67, 18, 35, 31].

Going back to discussion on (EG), Wang and Abernethy [67] and Kavis et al. [35] provide useful insights into acceleration within the context of (EG). Wang and Abernethy [67] 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
recover different acceleration schemes of Nesterov [56, 57, 53]. On a related front, Diakonikolas and Orecchia [19] proposes the first acceleration of (EG) by appropriately integrating the optimistic averaging idea [67] into the (EG) template as follows:

\[
\tilde{X}_t = \frac{b_t X_t + \sum_{s=1}^{t-1} b_s X_{s+\frac{1}{2}}}{\sum_{s=1}^{t} b_s}, \quad \tilde{X}_{t+\frac{1}{2}} = \frac{\sum_{s=1}^{t} b_s X_{s+\frac{1}{2}}}{\sum_{s=1}^{t} b_s} \tag{6}
\]

where \(b_t = O(t)\) is the “iterate averaging” parameter. Later on, Kavis et al. [35] 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 - second step of acceleration. Formally, our scheme is given via the following recursion:

\[
X_{t+\frac{1}{2}} = \Pi_X \left( X_t - \gamma_t a_t F(X_{t+\frac{1}{2}}; \tilde{X}_t) \right)
= \arg \min_{x \in \mathcal{X}} a_t (\nabla f(X_t)) + \frac{1}{2} \nabla^2 f(\tilde{X}_t)(\tilde{X}_{t+\frac{1}{2}} - X_t) + \frac{\|x - X_t\|^2}{2\gamma_t}
\]

\[
X_{t+1} = \Pi_X \left( X_t - \gamma_t a_t \nabla f(X_{t+\frac{1}{2}}) \right)
= \arg \min_{x \in \mathcal{X}} a_t (\nabla f(X_{t+\frac{1}{2}})), x - X_t + \frac{\|x - X_t\|^2}{2\gamma_t}
\]

with \(\Pi_X (x)\) denoting the Euclidean projection of \(x\) onto \(\mathcal{X}\), average sequences \(\tilde{X}_t\) and \(\tilde{X}_{t+\frac{1}{2}}\), defined as in (6) and the adaptive step-size \(\gamma_t\) defined as (for some \(\gamma, \beta_0 > 0\)):

\[
\gamma_t = \frac{\gamma}{\sqrt{\beta_0 + \sum_{s=1}^{t-1} a_s^2 \|\nabla f(X_{s+\frac{1}{2}}) - F(X_{s+\frac{1}{2}}; \tilde{X}_s)\|^2}}. \tag{7}
\]

The implicit nature of \((\text{Implicit})\) originates from \(X_{t+\frac{1}{2}}\) update (which we shall refer to as (corrected) extrapolation step at times) since \(\tilde{X}_{t+\frac{1}{2}}\) depends upon \(X_{t+\frac{1}{2}}\) 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 [63] - key to adaptivity and universality.

**Second-order updates:** First, we will consider the particular interpretation of (EG) as an approximation to the Proximal Point method [64] which serves as motivation for the accommodation of second-order information in our scheme.

\[
X_{t+1} = X_t - \gamma_t \nabla f(X_t). \tag{PP}
\]

In particular, (EG) tries to approximate \(X_{t+1}\) by generating the extrapolated point \(X_{t+\frac{1}{2}}\), and make use of the gradient at \(X_{t+\frac{1}{2}}\) 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 (PP) 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 [18] 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 [63], we design a constraint-aware step-size by accumulating $\|\nabla f(\bar{X}_{t+\frac{1}{2}}) - F(\bar{X}_{t+\frac{1}{2}}; X_t)\|^2$ which converges to 0 as $\bar{X}_{t+\frac{1}{2}} - \bar{X}_t \rightarrow 0$; 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). Formally, this is given by the following.

**Theorem 3.1.** Let $\{X_{t+\frac{1}{2}}\}_{t=1}^T$ be generated by (Implicit) run with the adaptive step-size policy (7) where $a_t = t^2$, $b_t = t^p$ with $p \geq 2$. Assume that $f$ satisfies (H-smooth) then, it is ensured that:

$$f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq O\left(\max\left\{\frac{\sqrt{n_0}D^2}{T^3}, L\frac{D^3 + Dn^3}{\gamma} \right\}\right)$$

When $\gamma = D$, we obtain the converge rate $O\left(\frac{\max\{LD^3, \sqrt{n_0}D\}}{T^3}\right)$.

**Remark 3.1.** We emphasize that the above rate does not require any prior knowledge of problem parameters 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 $\gamma$.

### 3.2 Explicit algorithm

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

1. (Implicit) 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 [30, 50, 12]. 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 $(\gamma_t, X_{t+\frac{1}{2}})$ and secondly, it enables adaptation to the second-order smoothness. 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
Algorithm 1: EXTRA-NEWTON

Input: \(X_1 \in X\), \(a_t = t^2\) and \(A_t = \sum_{s=1}^t a_s\), \(b_t = t^p\) \((p \geq 2)\) and \(B_t = \sum_{s=1}^t b_s\), \(\gamma > 0\), \(\xi_t \sim \text{i.i.d.}\)

1: for \(t = 1\) to \(T\) do
2: \(\gamma_t = \frac{\gamma}{\sqrt{\beta_0 + \sum_{s=1}^{t-1} a_s^2 \|g(X_{s+\frac{1}{2}}, \xi_{s+\frac{1}{2}}) - \bar{F}(X_{s+\frac{1}{2}}, \xi_s)\|^2}}\)
3: \(X_{t+\frac{1}{2}} = \arg \min_{x \in X} \langle a_t g(\bar{X}_t, \xi_t), x \rangle + \frac{a_t b_t}{2B_t} (H(\bar{X}_t, \xi_t)(x - X_t), x - X_t) + \frac{1}{2\gamma_t} \|x - X_t\|^2\)
4: \(X_{t+1} = \arg \min_{x \in X} \langle a_t g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), x \rangle + \frac{1}{2\gamma_t} \|x - X_t\|^2\)
5: end for

Universal step-size We modify our step-size (see Eq. (2)) in order to operate in the stochastic regime while making it noise-adaptive for rate interpolation. Using the same weighted averaging scheme in Eq. (6), we define the universal counterpart of the adaptive step-size, Note that \(\gamma_t\) is independent of any variable/randomness generated at iteration \(t\); it accumulates \(a_t^2 \|g(X_{s+\frac{1}{2}}, \xi_{s+\frac{1}{2}}) - \bar{F}(X_{s+\frac{1}{2}}, \xi_s)\|^2\) up to \(t - 1\). Therefore, the step-size is decoupled from the explicit update, \(a \text{ priori}\).

Now, what remains is a new algorithmic design that will retain the accelerated convergence properties demonstrated by (Implicit) 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 strategy, let us take our time to explain the consequences of our explicit design compared to (Implicit).

From implicit to explicit To obtain the explicit algorithm, (i) we write the projection sub-problem in the \(\arg \min\) form; (ii) introduce stochastic oracle feedback; (iii) for the second-order term, replace \(X_{t+\frac{1}{2}}\) in \(\bar{X}_{t+\frac{1}{2}}\) with the free variable \(\gamma\); then, (iv) simplify as follows:

\[
\frac{a_t}{2} \langle H(\bar{X}_t, \xi_t)(\bar{X}_{t+\frac{1}{2}} - \bar{X}_t), x - X_t \rangle
\]

\[
\frac{a_t}{2} \langle H(\bar{X}_t, \xi_t) \left(\frac{b_t X_{t+\frac{1}{2}} + \sum_{s=1}^{t-1} b_s X_{s+\frac{1}{2}}}{B_t} - \frac{b_t X_t + \sum_{s=1}^{t-1} b_s X_{s+\frac{1}{2}}}{B_t}\right), x - X_t \rangle
\]

\[
\frac{a_t}{2} \langle H(\bar{X}_t, \xi_t) \left(\frac{b_t x + \sum_{s=1}^{t-1} b_s X_{s+\frac{1}{2}}}{B_t} - \frac{b_t X_t + \sum_{s=1}^{t-1} b_s X_{s+\frac{1}{2}}}{B_t}\right), x - X_t \rangle
\]

\[
\frac{a_t b_t}{2B_t} \langle H(\bar{X}_t, \xi_t)(x - X_t), x - X_t \rangle
\]

Given the bisection-type conceptual methods \([50, 30, 12]\), it is surprising how smoothly we could transition from implicit to explicit once we decouple the step-size from the current iteration \(a \text{ priori}\). 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 \(\{X_{t+\frac{1}{2}}\}_{t=1}^T\) be a sequence generated by Algorithm 1, run with the adaptive step-size policy (2) and \(a_t = t^2\), \(b_t = t^p\) with \(p \geq 2\). Assume that \(f\) satisfies (H-smooth), and that Assumptions (2) hold. Then, the following universal guarantee holds:

\[
f(\bar{X}_{T+\frac{1}{2}}) - f(x^\star) \leq O \left( \frac{D^2 + \gamma^2 \sigma_g}{\sqrt{T}} + \frac{D^3 + D^2 \gamma^2 \sigma_H}{T^{3/2}} + \max \left\{ \frac{L D^4 + D^2 \gamma^3}{T^3}, \sqrt{\frac{D^2 + \gamma^2}{T}} \right\} \right)
\]
When $\gamma = D$, we obtain the target rate $O\left(\frac{D\sigma_X}{\sqrt{T}} + \frac{D^2\sigma_H}{T^{3/2}} + \frac{\max\{LD^3, \sqrt{D}\}}{T^3}\right)$.

Remark 3.2. Similar to Theorem 3.1, EXTRA-NEWTON 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 [18], 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 $O\left(\frac{D\sigma_X}{\sqrt{T}} + \frac{D^2\sigma_H}{T^{3/2}} + \frac{\max\{LD^3, \sqrt{D}\}}{T^3}\right)$ - 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 [63, 18] and first-order acceleration literature [67, 35, 31]. We measure the performance of our decisions for the extrapolation sequence such that after playing $X_{t+\frac{1}{2}}$, our algorithm observes and suffers the linear (weighted) loss with respect to $a_t \nabla f(\bar{X}_{t+\frac{1}{2}})$. Hence, we define the regret as

$$R_T(x) = \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x \rangle$$

(Reg)

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

Theorem 3.3. Let $R_T(x^*)$ denote the anytime regret for the decision sequence $\{X_{t+\frac{1}{2}}\}_{t=1}^{T}$ 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,

$$f(\bar{X}_T) - f(x^*) \leq \frac{R_T(x^*)}{a_T b_T}$$

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 $R_T(x^*)$, we present an upper bound that helps us explain how we exploit second-order smoothness for a more aggressive weighting, hence the rate $O(1/T^3)$.

Proposition 3.1. Let $\{X_{t+\frac{1}{2}}\}_{t=1}^{T}$ be generated by Algorithm 1, run with a non-increasing step-size sequence $\gamma_t$ 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:

$$\mathbb{E}R_T(x^*) \leq \frac{1}{2} \mathbb{E} \left[ \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|g(\bar{X}_{t+\frac{1}{2}}; \tau_{t+\frac{1}{2}}) - \bar{F}(\bar{X}_{t+\frac{1}{2}}; \bar{X}_t, \tau_t)\|_{\gamma_{t+1}}^2 \right]$$

Observe that the inequality in Proposition 3.1 is agnostic to the design of our step-size in Eq. (2) as well as the selection of the weights as described in Theorem 3.2. 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. In fact, the structure of the objective function, i.e., its convexity, will not be needed for upper-bounding the regret expression, and required 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 $\|g(\bar{X}_{t+\frac{1}{2}}; \tau_{t+\frac{1}{2}}) - \bar{F}(\bar{X}_{t+\frac{1}{2}}; \bar{X}_t, \tau_t)\|_{\gamma_{t+1}}^2$ into deterministic feedback and noise. Then, we argue that the noisy component grows as $O(\sigma_H^2 T^{3/2} + \sigma_H^5 T^5/2)$. 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 $O(T^5/2)$. Fortunately, we identify that the negative term is “large enough” in magnitude to control the growth of the deterministic term, permitting a constant component $O(LD^2)$ for the regret.
Although the regret bound of $O(LD^3 + D^2\sigma_H T^{3/2} + D\sigma_g T^{5/2})$ seems counterintuitive 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 continue 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,

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

Observe that to seamlessly combine the positive and negative terms, our analysis enforces that $a_t = O(t^2)$ and $b_t = O(t^2).$ Then, the conversion implies a convergence rate of $R_T(x^*)/T^3,$ 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 $O(\sigma_H/T^{3/2} + \sigma_g/\sqrt{T})$ rate, giving us the first universal acceleration beyond first-order smoothness.

Let us conclude by discussing the intricate relationship between the universal step-size and the regret bounds. Simply put, growth of the summation in the denominator of $\gamma_t$ is of the same order as the regret bound. Under stochastic gradient and Hessian oracles, the regret bound is of order $O(T^{5/2})$, and we can trivially show using variance bounds that the step-size is lower bounded by $O(T^{-5/2}).$ On the other extreme, the regret bound described in Proposition 3.1 is bounded by a constant under deterministic oracles, which implies that the summation in the denominator of the step-size is in turn summable, i.e., the step-size has a positive, constant lower bound. This adaptive behavior of our step-size enables automatic adaptation to noise levels and thus the universal rates.

4 Experiments

In this section, we will present practical performance of EXTRA-NEWTON against a set of first-order algorithms, e.g., GD, SGD, ADAGRAD [20], ACCELERAGRAD [41], UNICODEGRAD [35]; and second-order methods, e.g., NEWTON’S, Optimal Monteiro-Svaiter (OPTMS) [13], Cubic Regularization of Newton’s method (CRN) [59] and Accelerated CRN (ACRN) [55] for least squares and logistic regression problems over a LIBSVM datasets, a1a and a9a. Our main objective is three-folds. First, when the objective has a favorable structure as in least squares, second-order method has cheap oracle costs 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 $\ell_2$-regularized logistic regression problem. Finally, we compare our methods with respect to other second-order methods that achieve (almost) optimal rates. In the plots, the statement # of oracle calls on the x-axis counts any gradient or Hessian computation as one oracle call. Also note that we consider the black-box oracle model in which the algorithms only have access to gradient and Hessians without knowing the actual objective function.

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 [47, 49]. 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 EXTRA-NEWTON, 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 [27]. On the contrary, EXTRA-NEWTON 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. We complement our numerical tests by comparing EXTRA-NEWTON with a set of second-order methods. To that end, we implemented our method within the framework presented in [13]. Using the implementation and the experimental setup provided in their GitHub repository [24], we implemented our method in their code and compared against NEWTON’S, CRN, ACRN and OPTMS algorithms. Figure 2 shows that EXTRA-NEWTON has comparable performance to OPTMS, which has the theoretically faster rate $O(1/T^{7/2}),$ and marginally outperforms with respect to number of linear system solutions since the linesearch procedure of OPTMS might require multiple system solutions per iteration. While CRN and ACRN has worse convergence than EXTRA-NEWTON, NEWTON’S seems to have the fastest.

9
Note that the initialization favors Newton’s as it lies in a close neighborhood of the solution, and Newton’s performance sporadically deteriorates when initialized arbitrarily.

5 Conclusion

In this work, we present the first universal, second-order algorithm, EXTRA-NEWTON, which enjoys the value convergence rate of $O(\sigma_g/\sqrt{T} + \sigma_H/T^{3/2} + 1/T^3)$. By extending the notion of bounded variance on stochastic gradients to stochastic Hessian, we prove adaptation to the noise in first and second-order oracles, simultaneously, while showing accelerated rates matching that of Nesterov [55] under the fully deterministic oracle model. To that end, an important open question is whether we could design a method that achieves an improved rate interpolation guarantee $O(\sigma_g/\sqrt{T} + \sigma_H/T^{3/2} + 1/T^{7/2})$ without depending on any line-search/bisection mechanism. We defer this to a future work.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement n° 725594 - time-data) and the Swiss National Science Foundation (SNSF) under grant number 200021_205011.
References


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

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Preface</td>
<td>16</td>
</tr>
<tr>
<td>B</td>
<td>Notation and Definitions</td>
<td>16</td>
</tr>
<tr>
<td>C</td>
<td>Further Experimental Evaluation</td>
<td>17</td>
</tr>
<tr>
<td>D</td>
<td>Generalized Online-to-batch Conversion (Theorem 3.3)</td>
<td>18</td>
</tr>
<tr>
<td>E</td>
<td>Template Regret Bound (Proposition 3.1)</td>
<td>20</td>
</tr>
<tr>
<td>F</td>
<td>Technical Lemma for the Main Proofs</td>
<td>23</td>
</tr>
<tr>
<td>G</td>
<td>EXTRA-NEWTON: The First Universal Second-order Accelerated Method (Theorem 3.2)</td>
<td>23</td>
</tr>
<tr>
<td>H</td>
<td>Implicit Accelerated Second-order Algorithm (Theorem 3.1)</td>
<td>28</td>
</tr>
</tbody>
</table>
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 $R_T(x^*)$ and value convergence $f(X_{T+\frac{1}{2}}) - f(x^*)$.

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.

We present the universal convergence analysis of EXTRA-NEWTON (Theorem 3.2) in Appendix G.

In Appendix H, we share the analysis of our conceptual framework: convergence of the implicit algorithm (Implicit) in deterministic setting (Theorem 3.1), with the appropriate corollary of Proposition 3.1 for the case of deterministic oracles in this section.

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 |
| $x^*$ | $\text{arg min}_{x \in \mathcal{X}} f(x)$ solution of the constrained problem (Opt) |
| $D$ | $\sup_{x,y \in \mathcal{X}} \|x - y\|$ diameter of the constraint set $\mathcal{X}$ |
| $L$ | $\|\nabla^2 f(x) - \nabla^2 f(x')\| \leq L \|x - x'\|$ second-order smoothness constant of $f$ |
| $g(\cdot, \xi)$ | $\mathbb{E}[g(x, \xi) \mid x] = \nabla f(x)$, $x \perp \perp \xi$ unbiased gradient estimate |
| $H(\cdot, \xi)$ | $\mathbb{E}[H(x, \xi) \mid x] = \nabla^2 f(x)$, $x \perp \perp \xi$ unbiased Hessian estimate |
| $\mathcal{F}_t$ | $\sigma(\xi_1, \xi_{t+\frac{1}{2}}, \ldots, \xi_t)$ $\sigma$-algebra generated by random variables up to $\xi_t$ |
| $\mathcal{F}_{t+\frac{1}{2}}$ | $\sigma(\xi_1, \xi_{t+\frac{1}{2}}, \ldots, \xi_t, \xi_{t+\frac{1}{2}})$ $\sigma$-algebra generated by random variables up to $\xi_{t+\frac{1}{2}}$ |
| $\sigma_g$ | $\mathbb{E}\left[\|g(x) - \nabla f(x)\|^2 \mid x\right] \leq \sigma_g^2$ variance bound for gradient estimate |
| $\sigma_H$ | $\mathbb{E}\left[\|H(x) - \nabla^2 f(x)\|^2 \mid x\right] \leq \sigma_H^2$ variance bound for Hessian estimate |
| $\sigma$ | $\max\{\sigma_g, \sigma_H\}$ maximum variance of oracles |
| $\gamma_t$ | Eq. (7) and Eq. (2) 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 $p \geq 2$ 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, \( \text{w1a} \),
- and subsequently display results in the stochastic setting for the same datasets \( \text{a1a, w1a} \).

Figure 3 shows the results for the deterministic experiments while Figure 4 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 3, the x-axis represent the number of calls made to the deterministic oracle, and in Figure 4, 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 \( \text{a1a} \) dataset but has almost matching performance with a smaller performance gap compared to accelerated first-order methods with \( \text{w1a} \) 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.

![Figure 3: Comparison of value convergence for regression problems with deterministic oracle access](image)

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 EXTRA-NEWTON 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 4.

![Figure 4: Comparison of value convergence for regression problems with stochastic oracle access](image)

**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 \( f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \) with the "weighted" regret \( R_T(x^*) = \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle \).

**Theorem 3.3.** Let \( R_T(x^*) \) denote the anytime regret for the decision sequence \( \{X_{t+\frac{1}{2}}\}_{t=1}^{T} \) 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 \( \frac{a_t}{b_t} \) is ensured to be non-increasing,

\[
f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq \frac{R_T(x^*)}{\frac{1}{a_T} \frac{1}{b_T}}
\]

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

\[
R_T(x^*) = \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle
\]
Devising our analysis in the spirit of [18, 35], we need to relate $X_{t+\frac{1}{2}}$ to the average iterate $\bar{X}_{t+\frac{1}{2}}$ in order to exploit the convexity of the objective function. Notice that we could write the iterate $X_{t+\frac{1}{2}}$ as the difference of consecutive average iterates,

$$a_t X_{t+\frac{1}{2}} = a_t \frac{B_t}{b_t} \bar{X}_{t+\frac{1}{2}} - a_t \frac{B_{t-1}}{b_t} \bar{X}_{t-\frac{1}{2}}. \tag{8}$$

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,

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

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

$$\sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^*) \rangle \geq \sum_{t=1}^{T} a_t \left( f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right) + a_t \frac{B_{t-1}}{b_t} \left( f(\bar{X}_{t+\frac{1}{2}}) - f(\bar{X}_{t-\frac{1}{2}}) \right)$$

$$= \sum_{t=1}^{T} a_t \left( f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right) + a_t \frac{B_{t-1}}{b_t} \left( f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right) - a_t \frac{B_{t-1}}{b_t} \left( f(\bar{X}_{t-\frac{1}{2}}) - f(x^*) \right)$$

$$= \sum_{t=1}^{T} a_t \frac{B_t}{b_t} \left( f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right) + a_t \frac{B_{t-1}}{b_t} \left( f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right) - a_t \frac{B_{t-1}}{b_t} \left( f(\bar{X}_{t-\frac{1}{2}}) - f(x^*) \right)$$

$$= \sum_{t=1}^{T} a_t \frac{B_T}{b_T} \left( f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \right) - a_1 \frac{B_0}{b_1} \left( f(\bar{X}_{1/2}) - f(x^*) \right) + \sum_{t=1}^{T-1} B_t \left( a_t \frac{a_{t+1}}{b_t} - a_1 \frac{a_{t+1}}{b_{t+1}} \right) \left( f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right)$$

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,

$$\sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^*) \rangle = \frac{B_T}{b_T} \left( f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \right) + \sum_{t=1}^{T-1} B_t \left( a_t \frac{a_{t+1}}{b_t} - a_1 \frac{a_{t+1}}{b_{t+1}} \right) \left( f(\bar{X}_{t+\frac{1}{2}}) - f(x^*) \right)$$

$$\geq \frac{B_T}{b_T} \left( f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \right).$$

Rearranging the terms gives us the final result

$$f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq \frac{\sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}, X_{t+\frac{1}{2}} - x^*) \rangle}{aT \frac{B_T}{b_T}} = \frac{R_T(x^*)}{aT \frac{B_T}{b_T}}.$$
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 EXTRA-NEWTON once more.

**Proposition 3.1.** Let \( \{X_{t+\frac{1}{2}}\}_{t=1}^{T} \) be generated by Algorithm 1, run with a non-increasing step-size sequence \( \gamma_t \) 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:

\[
\mathbb{E} \left[ T \right] \leq \frac{1}{2} \mathbb{E} \left[ 3D^2 + \sum_{t=1}^{T} \frac{\gamma_t}{\gamma_{t+1}} \left( g(X_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{F}(X_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) \right)^2 - \frac{\|X_{t+\frac{1}{2}} - X_t\|^2}{\gamma_{t+1}} \right]
\]

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

\[
\langle a_t g(\tilde{X}_t, \xi_t), (X_{t+\frac{1}{2}} - z_0) \rangle = \langle a_t g(\tilde{X}_t, \xi_t), (X_{t+\frac{1}{2}} - X_t) \rangle - \frac{1}{\gamma_t} \langle X_{t+\frac{1}{2}} - X_t, z_0 - X_{t+\frac{1}{2}} \rangle
\]

(9)

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

\[
\langle a_t g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), (X_{t+1} - z_1) \rangle \leq \frac{1}{2\gamma_t} \langle X_{t+1} - X_t, z_1 - X_{t+1} \rangle
\]

(10)

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

\[
\langle a_t g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), (X_{t+\frac{1}{2}} - x^\ast) \rangle = \langle a_t g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), (X_{t+\frac{1}{2}} - X_{t+1}) \rangle + \langle a_t g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), (X_{t+1} - x^\ast) \rangle
\]

(11)
Now, observe that setting $z_0 = X_{t+1}$ in Eq. (9) and rearranging we have

$$-\frac{1}{2\gamma_t} \|X_{t+1} - X_t\|^2$$

$$\leq -(a_t \tilde{F}(\tilde{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t), X_{t+\frac{1}{2}} - X_{t+1}) - \frac{1}{2\gamma_t} \left(\|X_{t+\frac{1}{2}} - X_{t+1}\|^2 + \|X_{t+\frac{1}{2}} - X_t\|^2\right)$$

Plugging the above expression into Eq. (11) and summing over $t = 1, ..., T$, we will obtain,

$$\sum_{t=1}^{T} (a_t g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^*)$$

$$\leq \sum_{t=1}^{T} a_t (g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{F}(\tilde{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t), X_{t+\frac{1}{2}} - X_{t+1})$$

$$+ \sum_{t=1}^{T} \frac{1}{2\gamma_t} \left(\|X_t - x^*\|^2 - \|X_{t+1} - x^*\|^2 - \|X_{t+\frac{1}{2}} - X_{t+1}\|^2 - \|X_{t+\frac{1}{2}} - X_t\|^2\right)$$

First off, we bound the inner product term using Cauchy-Schwarz and a slight generalization of Young’s inequality [63]

$$\sum_{t=1}^{T} a_t (g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{F}(\tilde{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t), X_{t+\frac{1}{2}} - X_{t+1})$$

$$\leq \sum_{t=1}^{T} a_t \|g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{F}(\tilde{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t)\| \|X_{t+\frac{1}{2}} - X_{t+1}\|$$

$$\leq \frac{1}{2} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{F}(\tilde{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t)\|^2 + \frac{1}{\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_{t+1}\|^2.$$

We merge the expressions together,

$$\sum_{t=1}^{T} (a_t g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^*)$$

$$\leq \frac{1}{2} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{F}(\tilde{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t)\|^2 + \frac{1}{\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_{t+1}\|^2$$

$$+ \sum_{t=1}^{T} \frac{1}{2\gamma_t} \left(\|X_t - x^*\|^2 - \|X_{t+1} - x^*\|^2 - \|X_{t+\frac{1}{2}} - X_{t+1}\|^2 - \|X_{t+\frac{1}{2}} - X_t\|^2\right)$$

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, $\gamma_t$ does not include $\|g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{F}(\tilde{X}_{t+\frac{1}{2}}; \tilde{X}_t, \xi_t)\|^2$ and this would pose some problems in the later stages of the proof. Hence, we add/subtract
$\frac{1}{\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_t\|^2$ and regroup the terms,

$$\sum_{t=1}^{T} \langle a_t g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle$$

$$\leq \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\gamma_{t+1}} a_t^2 \|g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{F}(\bar{X}_{t+\frac{1}{2}}, \xi_t)\|^2 - \frac{1}{2\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_t\|^2$$

$$+ \frac{1}{2} \sum_{t=1}^{T} \left( \frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} \right) \left( \|X_{t+\frac{1}{2}} - X_{t+1}\|^2 + \|X_{t+\frac{1}{2}} - X_t\|^2 \right)$$

$$+ \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\gamma_t} \left( \|X_t - x^*\|^2 - \|X_{t+1} - x^*\|^2 \right)$$

$$\leq \frac{1}{2} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{F}(\bar{X}_{t+\frac{1}{2}}, \xi_t)\|^2 - \frac{1}{2\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_t\|^2$$

$$+ \frac{\|X_t - x^*\|^2}{2\gamma_t} + \frac{1}{2} \sum_{t=1}^{T-1} \left( \frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} \right) \|X_{t+1} - x^*\|^2 + D^2 \sum_{t=1}^{T} \left( \frac{1}{\gamma_{t+1}} - \frac{1}{\gamma_t} \right)$$

$$\leq \frac{3D^2}{2\gamma_{T+1}} + \frac{1}{2} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \|g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{F}(\bar{X}_{t+\frac{1}{2}}, \xi_t)\|^2 - \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\gamma_{t+1}} \|X_{t+\frac{1}{2}} - X_t\|^2$$

where we have rewritten the telescoping summation for $\|X_t - x^*\|^2 - \|X_{t+1} - x^*\|^2$ and used $D^2 = \sup_{x, y \in \mathcal{X}} \|x - y\|^2$ (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 $\sum_{t=1}^{T} \langle a_t g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle$. Recall the definitions of $\mathcal{F}_t = \sigma(\xi_1, \xi_{t+\frac{1}{2}}, \cdots, \xi_t)$ and $\mathcal{F}_{t+\frac{1}{2}} = \sigma(\xi_1, \xi_{t+\frac{1}{2}}, \cdots, \xi_t, \xi_{t+\frac{1}{2}})$ from Table 2. Taking expectation over all randomness,

$$\mathbb{E} \left[ \sum_{t=1}^{T} \langle a_t g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle \right]$$

$$= \mathbb{E} \left[ \sum_{t=1}^{T} a_t \langle g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle + a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle \right]$$

$$= \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{E} \left[ a_t \langle g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle \mid \mathcal{F}_t \right] \right]$$

$$+ \mathbb{E} \left[ \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle \right]$$

$$= \mathbb{E} \left[ \sum_{t=1}^{T} a_t \mathbb{E} \left[ g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) \mid \mathcal{F}_t \right] - \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \right]$$

$$+ \mathbb{E} \left[ \sum_{t=1}^{T} a_t \langle \nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle \right]$$

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
\( \mathbb{E} \left[ g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) \mid \mathcal{F}_t \right] = \nabla f(\bar{X}_{t+\frac{1}{2}}) \). Hence, we obtain that
\[
\mathbb{E} \left[ R_T(x^*) \right] = \mathbb{E} \left[ \sum_{t=1}^T a_t (\nabla f(\bar{X}_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^*) \right] = \mathbb{E} \left[ \sum_{t=1}^T (a_t g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^*) \right],
\]
which concludes the target result,
\[
\mathbb{E} \left[ R_T(x^*) \right] \leq \frac{1}{2} \mathbb{E} \left[ \frac{3D^2}{\gamma_{T+1}} + \sum_{t=1}^T (\gamma_{t+1} a_t^2 \|g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{F}(\bar{X}_{t+\frac{1}{2}}; \bar{X}_t, \xi_t)\|^2 - \frac{\|X_{t+\frac{1}{2}} - X_t\|^2}{\gamma_{t+1}}) \right].
\]

**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 [48] and [41], 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** (48, 41). For all non-negative numbers \( \alpha_1, \ldots, \alpha_t \), the following inequality holds:
\[
\sqrt{\sum_{t=1}^T \alpha_t^2} \leq \frac{\sum_{t=1}^T \alpha_t}{\sqrt{\sum_{t=1}^t \alpha_t}} \leq 2 \sqrt{\sum_{t=1}^T \alpha_t}. \tag{12}
\]

**G EXTRA-NEWTON: The First Universal Second-order Accelerated Method (Theorem 3.2)**

**Theorem 3.2.** Let \( \{X_{t+\frac{1}{2}}\}_{t=1}^T \) be a sequence generated by Algorithm 1, run with the adaptive step-size policy (2) and \( a_t = t^2, b_t = t^p \) for \( p \geq 2 \). Assume that \( f \) satisfies (H-smooth), and that Assumptions (2) hold. Then, the following universal guarantee holds:
\[
f(\bar{X}_{T+\frac{1}{2}}) - f(x^*) \leq O \left( \frac{D^2 + \gamma^2 \sigma_s}{\sqrt{T}} + \frac{D^3 + \gamma^2 \sigma_H}{T^{3/2}} + \max \left\{ \frac{L D^2 + 3 \gamma^2}{\gamma T}, \sqrt{\frac{D}{\gamma}} \right\} \right).
\]

When \( \gamma = D \), we obtain the target rate \( O \left( \frac{D \sigma_s}{\sqrt{T}} + \frac{D^2 \sigma_H}{T^{3/2}} + \frac{L D^3 \sqrt{T}}{T^3} \right) \).

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

Now, we want to unify the first two terms through numerical inequalities. We will write the second term in terms of the first term. Due to Lemma F.1, we can upper the bound second term as,
\[
\frac{1}{2} \sum_{t=1}^T \gamma_{t+1} a_t^2 \|g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{F}(\bar{X}_{t+\frac{1}{2}}; \bar{X}_t, \xi_t)\|^2
\]
\[
= \frac{\gamma}{2} \sum_{t=1}^T \frac{a_t^2 \|g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{F}(\bar{X}_{t+\frac{1}{2}}; \bar{X}_t, \xi_t)\|^2}{\sqrt{\beta_0 + \sum_{s=1}^t a_s^2 \|g(\bar{X}_{s+\frac{1}{2}}, \xi_{s+\frac{1}{2}}) - \tilde{F}(\bar{X}_{s+\frac{1}{2}}; \bar{X}_s, \xi_s)\|^2}}
\]
\[
\leq \frac{\gamma}{2 \sqrt{\beta_0}} \sum_{t=1}^T a_t^2 \|g(\bar{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \tilde{F}(\bar{X}_{t+\frac{1}{2}}; \bar{X}_t, \xi_t)\|^2.
\]
Plugging this back into the original expression gives us

\[
\mathbb{E} [ R_T(x^*) ] \leq \left( \frac{3D^2}{2\gamma} + \gamma \right) \sqrt{\beta_0 + \sum_{t=1}^{T} a_t^2 ||g(\hat{X}_{t+\frac{1}{2}}; \xi_t) - \tilde{F}(\hat{X}_{t+\frac{1}{2}}; \hat{X}_t, \xi_t)||^2} - \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\gamma_{t+1}} ||X_{t+\frac{1}{2}} - X_t||^2
\]

Next up, we will handle the negative term in the above expression. 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 [67, 35].

\[
- \frac{||X_{t+\frac{1}{2}} - X_t||^2}{\gamma_{t+1}} = \frac{D^2}{D^2\gamma_{t+1}} ||X_{t+\frac{1}{2}} - X_t||^2 
\]

\[
\leq - \frac{1}{D^2\gamma_{t+1}} ||X_{t+\frac{1}{2}} - X_t||^4 
\]

\[
= - \frac{1}{D^2\gamma_{t+1}} \frac{B_t^4}{b_t^4} ||b_t X_{t+\frac{1}{2}} - b_t X_t||^4 
\]

\[
= - \frac{1}{D^2\gamma_{t+1}} \frac{B_t^4}{b_t^4} ||b_t X_{t+\frac{1}{2}} + \sum_{s=1}^{t-1} b_s X_{s+\frac{1}{2}} - b_t X_t + \sum_{s=1}^{t-1} b_s X_{s+\frac{1}{2}}||^4 
\]

\[
= - \frac{1}{D^2\gamma_{t+1}} c^4 t^4 ||\tilde{X}_{t+\frac{1}{2}} - X_t||^4 
\]

\[
\leq - \frac{4c^4 t^4}{L^2 D^2\gamma_{t+1}} ||\nabla f(\tilde{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_{t+\frac{1}{2}}; \tilde{X}_t)||^2 
\]

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 \(\frac{b_t}{b_s} \leq ct\), where \(c > 0\) is an absolute constant depending on how \(b_t\) is defined. Then, we use the definitions of average sequences \(\tilde{X}_{t+\frac{1}{2}}\) and \(\tilde{X}_t\) to go from \(||X_{t+\frac{1}{2}} - X_t||^4\) to \(||\tilde{X}_{t+\frac{1}{2}} - \tilde{X}_t||^4\) 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 \(\tilde{X}_{t+\frac{1}{2}}\) and \(\tilde{X}_t\) differ by only the last element, we can seamlessly relate \(||X_{t+\frac{1}{2}} - X_t||\) to \(||\tilde{X}_{t+\frac{1}{2}} - \tilde{X}_t||\).

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.

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.

\[
\epsilon_t = [g(\tilde{X}_{t+\frac{1}{2}}; \xi_{t+\frac{1}{2}}) - g(\tilde{X}_t; \xi_t)] - [\nabla f(\tilde{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_t)] 
\]

\[
\delta_t = H(\tilde{X}_t; \xi_t) - \nabla^2 f(\tilde{X}_t) 
\]

Then, we define following deterministic/stochastic placeholders:

\[
\nabla_t = \nabla f(\tilde{X}_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_{t+\frac{1}{2}}; \tilde{X}_t) 
\]

\[
\nabla_t = g(\tilde{X}_{t+\frac{1}{2}}; \xi_{t+\frac{1}{2}}) - g(\tilde{X}_t; \xi_t) = \nabla_t + \epsilon_t - \delta_t(\tilde{X}_{t+\frac{1}{2}} - \tilde{X}_t) 
\]

(14)
Setting \( a_t = t^2 \), combining all the terms and introducing the compact notation,

\[
\sum_{t=1}^{T} \langle a_t g(X_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^* \rangle \leq \left( \frac{3D^2}{2\gamma} + \gamma \right) \sqrt{\beta_0 + \sum_{t=1}^{T} a_t^2 \| \tilde{\nabla}_t \|^2 - \sum_{t=1}^{T} \frac{2c^4}{L^2 D^2 \gamma t+1} a_t^2 \| \nabla \|^2}.
\]

At this point, we need to understand how to relate \( \| \nabla \|^2 \) and \( \| \tilde{\nabla}_t \|^2 \) while treating the step-size \( \gamma_{t+1} \) accordingly. The issue is that the step-size is agnostic to deterministic oracle information since we accumulate \( \| \tilde{\nabla}_t \|^2 \). From the perspective of step-size, we need to find a relevant, if not matching, lower bound for \( \| \nabla \|^2 \) and \( \| \tilde{\nabla}_t \|^2 \). Indeed, we follow the ideas presented in [35], and begin by (trivially) lower bounding both terms with the same expression,

\[
\| \tilde{\nabla}_t \|^2 \geq \min \{ \| \nabla \|^2, \| \tilde{\nabla}_t \|^2 \}
\]

(15)

Now, we will decompose \( \| \tilde{\nabla}_t \|^2 \) into \( \| \nabla \|^2 \) and the noise terms. Using the definitions in Eq. (13) and (14) and applying triangular inequality with quadratic expansion,

\[
\| \tilde{\nabla}_t \|^2 \leq 2 \| \nabla \|^2 + 4 \| \delta_t (\hat{X}_{t+\frac{1}{2}} - \hat{X}_t) \|^2 + 4 \| \epsilon_t \|^2
\]

(16)

We can also have the following trivial upper bound,

\[
\| \tilde{\nabla}_t \|^2 \leq 2 \| \tilde{\nabla}_t \|^2
\]

\[
\leq 2 \| \tilde{\nabla}_t \|^2 + 4 \| \delta_t (\hat{X}_{t+\frac{1}{2}} - \hat{X}_t) \|^2 + 4 \| \epsilon_t \|^2
\]

(17)

Let us simplify the relationship between the bounds in Eq. (16) and Eq. (17); if \( \| \nabla \|^2 \leq \| \tilde{\nabla}_t \|^2 \), then Eq. (16) is tighter, otherwise Eq. (17) is tighter. Hence, we could select the minimum of \( \| \nabla \|^2 \) and \( \| \tilde{\nabla}_t \|^2 \):

\[
\| \tilde{\nabla}_t \|^2 \leq 2 \min \{ \| \nabla \|^2, \| \tilde{\nabla}_t \|^2 \} + 4 \| \delta_t (\hat{X}_{t+\frac{1}{2}} - \hat{X}_t) \|^2 + 4 \| \epsilon_t \|^2
\]

(18)

Using this intuition, we can construct a variable \( \lambda_t \) that always upper bounds the step-size.

\[
\lambda_t = \frac{\gamma}{\sqrt{\beta_0 + \sum_{s=1}^{t-1} a_s^2 \min \{ \| \nabla_s \|^2, \| \nabla_s \|^2 \}}}
\]

(19)
It is immediate that $\gamma_t \leq \lambda_t$. Essentially, we will replace the terms $\|\nabla_t\|^2$ and $\|\nabla_t\|^2$ with

\[
\min \left\{ \|\nabla_t\|^2, \|\nabla_t\|^2 \right\}, \|\delta_t(X_t + \frac{1}{2} - \tilde{X}_t)\|^2 \text{ and } \|\epsilon_t\|^2.
\]

\[
\mathbb{E} \left[ R_T(x^*) \right] = \mathbb{E} \left[ \frac{3D^2 + 2\gamma^2}{2\gamma} \left( \min_{t=1}^{T} \sum_{t=1}^{T} a_t \|\nabla_t\|^2 \right) \right.
\]

\[
\leq \mathbb{E} \left[ \frac{3D^2 + 2\gamma^2}{2\gamma} \left( \min_{t=1}^{T} \sum_{t=1}^{T} a_t \|\nabla_t\|^2 \right) \right.
\]

\[
- \sum_{t=1}^{T} \frac{2c^4}{L^2D^2\lambda_{t+1}} a_t \left\{ \|\nabla_t\|^2, \|\nabla_t\|^2 \right\}
\]

\[
+ 2 \left( \frac{3D^2}{2\gamma} + \gamma \right) \sqrt{\frac{\sum_{t=1}^{T} a_t^2 \|\delta_t(X_t + \frac{1}{2} - \tilde{X}_t)\|^2}{\sum_{t=1}^{T} a_t^2 \|\epsilon_t\|^2}} + 2 \left( \frac{3D^2}{2\gamma} + \gamma \right) \sqrt{\frac{\sum_{t=1}^{T} a_t^2 \|\epsilon_t\|^2}{\sum_{t=1}^{T} a_t^2 \|\epsilon_t\|^2}} \right]
\]

\[
\leq \mathbb{E} \left[ \frac{3D^2 + 2\gamma^2}{2\gamma} \left( \gamma \min_{t=1}^{T} \sum_{t=1}^{T} \lambda_{t+1} a_t^2 \left\{ \|\nabla_t\|^2, \|\nabla_t\|^2 \right\} \right) - \sum_{t=1}^{T} \frac{2c^4a_t^2}{L^2D^2\lambda_{t+1}} \min \left\{ \|\nabla_t\|^2, \|\nabla_t\|^2 \right\}
\]

\[
+ 2 \left( \frac{3D^2}{2\gamma} + \gamma \right) \sqrt{\frac{\sum_{t=1}^{T} a_t^2 \|\delta_t(X_t + \frac{1}{2} - \tilde{X}_t)\|^2}{\sum_{t=1}^{T} a_t^2 \|\epsilon_t\|^2}} + 2 \left( \frac{3D^2}{2\gamma} + \gamma \right) \sqrt{\frac{\sum_{t=1}^{T} a_t^2 \|\epsilon_t\|^2}{\sum_{t=1}^{T} a_t^2 \|\epsilon_t\|^2}} \right]
\]

\[
\leq \frac{3D^2 + 2\gamma^2}{2\gamma} \int_{\beta_0}^{\beta_0} \mathbb{E} \left[ \frac{\sum_{t=1}^{T} \left( \frac{3D^2 + 2\gamma^2}{\sqrt{2}\gamma^2} - \frac{2c^4}{L^2D^2\lambda_{t+1}} \right) \lambda_{t+1} a_t^2 \min \left\{ \|\nabla_t\|^2, \|\nabla_t\|^2 \right\}}{\sum_{t=1}^{T} a_t^2 \|\epsilon_t\|^2} \right] \right]
\]

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

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

\[\text{This immediately implies that for any } t \leq T_0,
\]

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

\[\text{(20)}
\]

\[\text{There is a critical cut-off point for the possible values of } T_0 \text{ depending on the value of } \beta_0. \text{ When the initial step-size is small enough, i.e., } \beta_0 \text{ is too large, then } T_0 < 0. \text{ This occurs when } \beta_0 \geq \frac{L^2D^2(3D^2 + 2\gamma^2)}{2^3/4\gamma c^2}, \text{ which implies,}
\]

\[\mathbb{E} \left[ \sum_{t=1}^{T} \left( \frac{3D^2 + 2\gamma^2}{\sqrt{2}\gamma^2} - \frac{2c^4}{L^2D^2\lambda_{t+1}} \right) \lambda_{t+1} a_t^2 \min \left\{ \|\nabla_t\|^2, \|\nabla_t\|^2 \right\} \right] \leq 0
\]
We get the same bound when $T_0 = 0$. For any other value of $T_0$, i.e., $T_0 > 0$, 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,

$$
\frac{3D^2 + 2\gamma^2}{\gamma} \sqrt{\beta_0} + \mathbb{E} \left[ \sum_{t=1}^{T} \left( \frac{3D^2 + 2\gamma^2}{\gamma} - \frac{2c^4}{L^2D^2\lambda_{t+1}^2} \right) \lambda_{t+1} a_t^2 \min \left\{ \|v_t\|^2, \|v_t\|^2 \right\} \right]
\leq \frac{3D^2 + 2\gamma^2}{\sqrt{\gamma}} \sqrt{\beta_0} + \mathbb{E} \left[ \sum_{t=1}^{T_0} \left( \frac{3D^2 + 2\gamma^2}{\gamma} - \frac{2c^4}{L^2D^2\lambda_{t+1}^2} \right) \lambda_{t+1} a_t^2 \min \left\{ \|v_t\|^2, \|v_t\|^2 \right\} \right]
\leq \frac{3D^2 + 2\gamma^2}{\gamma} \sqrt{\beta_0} + \frac{3D^2 + 2\gamma^2}{\gamma} \sum_{t=1}^{T_0} a_t^2 \min \left\{ \|v_t\|^2, \|v_t\|^2 \right\}
\leq \frac{3\sqrt{2}D^2 + 2\sqrt{2}\gamma^2}{\gamma} \sqrt{\beta_0} + \sum_{t=1}^{T_0} a_t^2 \min \left\{ \|v_t\|^2, \|v_t\|^2 \right\}
= \left( \frac{3\sqrt{2}D^2 + 2\sqrt{2}\gamma^2}{\gamma} \right) \frac{1}{\lambda_{T_0 + 1}}
\leq \frac{LD \left( 3D^2 + 2\gamma^2 \right)^{3/2}}{2^{1/4} \gamma^2}
$$

To make sure we incorporate the effect of the initial step-size, we combine the bounds to get

$$
\frac{3D^2 + 2\gamma^2}{\gamma} \sqrt{\beta_0} + \mathbb{E} \left[ \sum_{t=1}^{T} \left( \frac{3D^2 + 2\gamma^2}{\gamma} - \frac{2c^4}{L^2D^2\lambda_{t+1}^2} \right) \lambda_{t+1} a_t^2 \min \left\{ \|v_t\|^2, \|v_t\|^2 \right\} \right]
\leq \frac{3D^2 + 2\gamma^2}{\gamma} \sqrt{\beta_0} + \frac{3D^2 + 2\gamma^2}{2^{1/4} \gamma^2} \max \left\{ \sqrt{\beta_0}, \frac{LD \sqrt{3D^2 + 2\gamma^2}}{c^2} \right\}
$$

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,

$$
\mathbb{E} \left[ R_T(x^*) \right]
\leq \frac{3D^2 + 2\gamma^2}{\gamma} \sqrt{\beta_0} \left[ \sum_{t=1}^{T} a_t^2 \|\delta_t(\tilde{X}_{t+\frac{1}{2}} - \tilde{X}_t)\|^2 + \sum_{t=1}^{T} a_t^2 \|\epsilon_t\|^2 \right] + \frac{3D^2 + 2\gamma^2}{2^{1/4} \gamma^2} \max \left\{ \sqrt{\beta_0}, \frac{LD \sqrt{3D^2 + 2\gamma^2}}{c^2} \right\}
\leq \frac{3D^2 + 2\gamma^2}{2^{1/4} \gamma^2} \max \left\{ \sqrt{\beta_0}, \frac{LD \sqrt{3D^2 + 2\gamma^2}}{c^2} \right\} + \frac{3D^2 + 2\gamma^2}{\gamma} \sqrt{\beta_0} \left[ \sum_{t=1}^{T} a_t^2 \|\delta_t\|^2 \|\tilde{X}_{t+\frac{1}{2}} - \tilde{X}_t\|^2 \right]
\leq \frac{3D^2 + 2\gamma^2}{2^{1/4} \gamma^2} \max \left\{ \sqrt{\beta_0}, \frac{LD \sqrt{3D^2 + 2\gamma^2}}{c^2} \right\} + \frac{3D^2 + 2\gamma^2}{\gamma} \sqrt{D^2 \sum_{t=1}^{T} a_t^2 \beta_t^2 \mathbb{E} \left[ \|\delta_t\|^2 \right]} + \frac{3D^2 + 2\gamma^2}{\gamma} \sum_{t=1}^{T} a_t^2 \mathbb{E} \left[ \|g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_{t+\frac{1}{2}})\|^2 \right] + \mathbb{E} \left[ \|g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}} - \nabla f(\tilde{X}_{t+\frac{1}{2}})\|^2 \right] + \mathbb{E} \left[ \|g(\tilde{X}_{t+\frac{1}{2}}, \xi_{t+\frac{1}{2}}) - \nabla f(\tilde{X}_{t+\frac{1}{2}})\|^2 \right]
$$

27
When we obtain,

\[ \text{Theorem 3.1.} \]

Let the value of \( c \) be such that \( \text{regret} \), then make use of the conversion result in Theorem 3.3 to obtain a convergence rate of order \( \frac{\sqrt{c}}{2^{1/4}} \) when \( c \leq \frac{2D^3 + 2 \gamma^3}{\gamma^4} \).

Overall strategy is straightforward; we first prove a constant upper bound for the offline weighted policy. The proof of this theorem is analogous to that of Proposition 3.1 in Section E, up to replacing stochastic evaluations \( g(\cdot) \) and \( \mathbb{F}(\cdot; \cdot) \) with \( \nabla f(\cdot) \) and \( \mathbb{F}(\cdot; \cdot) \). For completeness, we will formalize the aforementioned result in Proposition H.1 which follows the same steps as the proof of Proposition 3.1.

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

\[
R_T(x^*) \leq \frac{1}{2} \left( \frac{3D^2}{\gamma T^3} + \sum_{t=1}^{T} \gamma_t \|\nabla f(X_{t+\frac{1}{2}}) - \mathbb{F}(X_{t-\frac{1}{2}}; X_t)\|^2 - \frac{\|X_{t+\frac{1}{2}} - X_t\|^2}{\gamma_t} \right).
\]

**Proof.** The proof of this theorem is analogous to that of Proposition 3.1 in Section E, up to replacing the stochastic feedback with the deterministic oracle calls.

**Theorem 3.1.** Let \( \{X_{t+\frac{1}{2}}\}_{t=1}^T \) be a sequence generated by (Implicit), 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 \( D \). Then, the following guarantee holds:

\[
f(X_{T+\frac{1}{2}}) - f(x^*) \leq O \left( \max \left\{ \frac{\sqrt{D^2}}{\gamma}, \frac{L D^3 + D^3}{\gamma} \right\} \right)
\]

When \( \gamma = D \), we obtain the converge rate \( O(1/T^3) \).

**Proof.** We will initiate our proof at template regret inequality as we proved in Proposition H.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 H.1 we have,

\[ R_T(x^*) \leq \frac{1}{2} \left( \frac{3D^2}{\gamma T + 1} + \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \| \nabla f(\bar{x}_{t+\frac{1}{2}}) - F(\bar{x}_{t+\frac{1}{2}}; \hat{x}_t) \| ^2 \right) \]

We will merge the first two terms and express the first term in the form of the second one using Lemma F.1. Observe that for the proof of Theorem 3.2, we did the opposite and converted the summation into the form of the first term, \( \frac{3D^2}{2\gamma} \).

\[
R_T(x^*) \leq \frac{3D^2}{2\gamma} \sqrt{\beta_0 + \sum_{t=1}^{T} a_t^2 \| \nabla f(\bar{x}_{t+\frac{1}{2}}) - F(\bar{x}_{t+\frac{1}{2}}; \hat{x}_t) \| ^2} + \frac{1}{2} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \| \nabla f(\bar{x}_{t+\frac{1}{2}}) - F(\bar{x}_{t+\frac{1}{2}}; \hat{x}_t) \| ^2 - \frac{\| X_{t+\frac{1}{2}} - X_t \| ^2}{\gamma_{t+1}}
\]

\[
\leq \frac{3D^2 \sqrt{\beta_0}}{2\gamma} + \frac{3D^2}{2\gamma} \sum_{t=1}^{T} \frac{a_t^2 \| \nabla f(\bar{x}_{t+\frac{1}{2}}) - F(\bar{x}_{t+\frac{1}{2}}; \hat{x}_t) \| ^2}{\beta_0 + \sum_{s=1}^{t} a_s^2 \| \nabla f(\bar{x}_{s+\frac{1}{2}}) - F(\bar{x}_{s+\frac{1}{2}}; X_s) \| ^2} + \frac{1}{2} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \| \nabla f(\bar{x}_{t+\frac{1}{2}}) - F(\bar{x}_{t+\frac{1}{2}}; \hat{x}_t) \| ^2 - \frac{\| X_{t+\frac{1}{2}} - X_t \| ^2}{\gamma_{t+1}}
\]

\[
= \frac{3D^2 \sqrt{\beta_0}}{2\gamma} + \frac{3D^2}{2\gamma} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \| \nabla f(\bar{x}_{t+\frac{1}{2}}) - F(\bar{x}_{t+\frac{1}{2}}; \hat{x}_t) \| ^2 + \frac{1}{2} \sum_{t=1}^{T} \gamma_{t+1} a_t^2 \| \nabla f(\bar{x}_{t+\frac{1}{2}}) - F(\bar{x}_{t+\frac{1}{2}}; \hat{x}_t) \| ^2 - \frac{\| X_{t+\frac{1}{2}} - X_t \| ^2}{\gamma_{t+1}}
\]

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. Similar to the proof in the explicit algorithm, we upper bound the negative term using appropriate averaging constants and smoothness.

\[
- \frac{\| X_{t+\frac{1}{2}} - X_t \| ^2}{\gamma_{t+1}} \leq - \frac{4c^4}{L^2 D^2 \gamma_{t+1}} t^4 \| \nabla f(\bar{x}_{t+\frac{1}{2}}) - F(\bar{x}_{t+\frac{1}{2}}; \hat{x}_t) \| ^2
\]

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

\[
\leq \frac{3D^2 \sqrt{\beta_0}}{2\gamma} + \frac{1}{2} \sum_{t=1}^{T} \left( \frac{3D^2 + \gamma^2}{\gamma^2} - \frac{4c^4}{L^2 D^2 \gamma_{t+1}^2} \right) \gamma_{t+1} a_t^2 \| \nabla f(\bar{x}_{t+\frac{1}{2}}) - F(\bar{x}_{t+\frac{1}{2}}; \hat{x}_t) \| ^2
\]

(21)

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 \( \left( \frac{3D^2 + \gamma^2}{\gamma^2} - \frac{4c^4}{L^2 D^2 \gamma_{t+1}^2} \right) \) is a non-increasing quantity and we are interested in the time point at which this quantity becomes negative. For that reason, we define the following time point,

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

This immediately implies that for any \( t \leq T_0 \),

\[
\frac{1}{\gamma_{t+1}} \leq \frac{LD\sqrt{3D^2 + \gamma^2}}{2\gamma c^2}.
\]

(22)

To paint a complete picture, we would like to have a brief discussion on the possible values for \( T_0 \).
1. $T_0 \leq 0$ implies that the step-size is small enough from the very beginning and that the summation term in Eq. (21) 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. (22). This is equivalent to saying $\sum_{t=1}^{\infty} a_t^2 \|\nabla f(\bar{x}_{t+\frac{1}{2}}) - F(\bar{x}_{t+\frac{1}{2}}; \bar{x}_t)\|^2 \leq C$ for some constant $C$, which in turn ensures that the summation in Eq. (21) 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. (21) 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 summants change sign and become negative after $T_0$.

Same as in the proof of EXTRA-NEWTON, we need to understand the effect of the initial step-size choice due to $\beta_0$. Imagine the case $\sqrt{\beta_0} \geq \frac{LD\sqrt{3D^2 + \gamma^2}}{2\gamma^2}$. This implies that $T_0 < 0$ and that the step-size is already small enough to make the summation negative from the first step onwards. In that scenario, the condition in Eq. (22) doesn’t hold so we should consider the effect of this initial setup for the final bound. For the case when $T_0 > 0$, we can safely unify all the 3 cases above and simply upper bound the expression in Eq. (21) by rewriting the summation up to $T_0$. Therefore,

We combine the case for $T_0 < 0$ with the one above to establish the constant regret bound

$$R_T(x^*) \leq O\left(\max\left\{\sqrt{\beta_0} D^2 \gamma, \frac{LD^4 + D\gamma^3}{\gamma}\right\}\right)$$

Plugging this result in its place we obtain the convergence rate,

$$f(\bar{x}_{T+\frac{1}{2}}) - f(x^*) \leq O\left(\frac{\max\left\{\sqrt{\beta_0} D^2, \frac{LD^4 + D\gamma^3}{T^3}\right\}}{T^3}\right)$$