Adaptive Accelerated (Extra-)Gradient Methods with Variance Reduction

Zijian Liu [*1] **Ta Duy Nguyen** [*1] **Alina Ene** [1] **Huy L. Nguyen** [2]

Abstract

In this paper, we study the finite-sum convex optimization problem focusing on the general convex case. Recently, the study of variance reduced (VR) methods and their accelerated variants has made exciting progress. However, the step size used in the existing VR algorithms typically depends on the smoothness parameter, which is often unknown and requires tuning in practice. To address this problem, we propose two novel adaptive VR algorithms: *Adaptive Variance Reduced Accelerated Extra-Gradient* (AdaVRAE) and *Adaptive Variance Reduced Accelerated Gradient* (AdaVRAG). Our algorithms do not require knowledge of the smoothness parameter. AdaVRAE uses $\mathcal{O}\left(n\log\log n+\sqrt{\frac{n\beta}{\epsilon}}\right)$ and AdaVRAG uses $\mathcal{O}\left(n\log\log n+\sqrt{\frac{n\beta\log\beta}{\epsilon}}\right)$ gradient evaluations to attain an $\mathcal{O}(\epsilon)$-suboptimal solution, where n is the number of functions in the finite sum and β is the smoothness parameter. This result matches the best-known convergence rate of non-adaptive VR methods and it improves upon the convergence of the state of the art adaptive VR method, AdaSVRG. We demonstrate the superior performance of our algorithms compared with previous methods in experiments on real-world datasets.

1. Introduction

In this paper, we consider the finite-sum optimization problem in the form of

$$\min_{x\in\mathcal{X}}\left\{\frac{1}{n}\sum_{i=1}^{n}f_i(x)+h(x)\right\} \tag{1}$$

where each function f_i is convex and β-smooth, h is convex and potentially nonsmooth but admitting an efficient proximal operator, and $\mathcal{X}\subseteq\mathbb{R}^d$ is a closed convex set. Additionally, we further assume that $\mathcal{X}$ is compact when β is unknown. Problem (1) has found a wide range of applications in machine learning, typically in empirical risk minimization problems, and has been extensively studied in the past few years.

Among existing approaches to solve this problem, variance reduced (VR) methods (Johnson & Zhang, 2013; Defazio et al., 2014; Schmidt et al., 2017; Roux et al., 2012) have recently shown significant improvement over the classic stochastic gradient methods such as stochastic gradient descent (SGD) and its variants. For example, in strongly convex problems, VR methods such as (Allen-Zhu, 2017; Lan et al., 2019; Lin et al., 2015) can achieve the optimal number of gradient evaluations of $\mathcal{O}\left((n+\sqrt{n\kappa})\log\frac{1}{\epsilon}\right)$ to attain an $\mathcal{O}(\epsilon)$-suboptimal solution, where κ is the condition number, which improves over full-batch gradient descent ($\mathcal{O}\left(n\kappa\log\frac{1}{\epsilon}\right)$) and Nesterov's accelerated gradient descent (Nesterov, 1983; 2003) ($\mathcal{O}\left(n\sqrt{\kappa}\log\frac{1}{\epsilon}\right)$). For general convex problems, the current state-of-the-art VR methods, namely VRADA (Song et al., 2020) can find an $\mathcal{O}(\epsilon)$-suboptimal solution using $\mathcal{O}\left(n\log\log n+\sqrt{\frac{n\beta}{\epsilon}}\right)$ gradient evaluations, which nearly-matches the lower bound of $\Omega\left(n+\sqrt{\frac{n\beta}{\epsilon}}\right)$ (Woodworth & Srebro, 2016).

However, most of existing VR gradient methods have the same limitation as classic gradient methods; that is, they require the prior knowledge of the smoothness parameter in order to set the step size. Lacking this information, one may have to carefully perform hyper-parameter tuning to avoid the situation that the algorithm divergences or converges too slowly due to too large or too small step size. This limitation of gradient methods motivates the development of methods that aim to adapt to unknown problem structures. A notable line of work starting with the influential AdaGrad algorithm has designed a family of gradient descent based methods that set the step size based on the gradients or iterates observed in previous iterations (McMahan & Streeter, 2010; Duchi et al., 2011; Kingma & Ba, 2014; Levy, 2017; Levy et al., 2018; Bach & Levy, 2019; Cutkosky, 2019; Kavis et al., 2019; Joulani et al., 2020; Ene et al., 2021; Antonakopoulos et al.,

[*]Equal contribution [1]Department of Computer Science, Boston University [2]Khoury College of Computer and Information Science, Northeastern University. Correspondence to: Ta Duy Nguyen <taduy@bu.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Table 1. Our results and comparison with prior works.

Algorithm	General convex	Adaptive
SVRG (Johnson & Zhang, 2013)	-	No
SVRG^{++} (Allen-Zhu & Yuan, 2016)	$\mathcal{O}\left(n\log\frac{\beta}{\epsilon}+\frac{\beta}{\epsilon}\right)$	No
Katyusha (Allen-Zhu, 2017)	$\mathcal{O}\left(n\log\frac{\beta}{\epsilon}+\sqrt{\frac{n\beta}{\epsilon}}\right)$	No
VARAG (Lan et al., 2019)	$\mathcal{O}\left(n\min\left\{\log\frac{\beta}{\epsilon},\log n\right\}+\sqrt{\frac{n\beta}{\epsilon}}\right)$	No
VRADA (Song et al., 2020)	$\mathcal{O}\left(n\min\left\{\log\log\frac{\beta}{\epsilon},\log\log n\right\}+\sqrt{\frac{n\beta}{\epsilon}}\right)$	No
AdaSVRG (Dubois-Taine et al., 2021)	$\mathcal{O}\left(\frac{n\beta}{\epsilon}\right)$ (fixed sized inner loop, only if $\epsilon=\Omega(\frac{\beta}{n})$) $\mathcal{O}\left(n\log\frac{\beta}{\epsilon}+\frac{\beta}{\epsilon}\right)$ (multi-stage)	Yes
AdaVRAE (unknown β) **(This Paper)**	$\mathcal{O}\left(n\min\left\{\log\log\frac{\beta}{\epsilon},\log\log n\right\}+\sqrt{\frac{n\beta}{\epsilon}}\right)$	Yes
VRAE (known β) **(This Paper)**	$\mathcal{O}\left(n\min\left\{\log\log\frac{\beta}{\epsilon},\log\log n\right\}+\sqrt{\frac{n\beta}{\epsilon}}\right)$	No
AdaVRAG (unknown β) **(This Paper)**	$\mathcal{O}\left(n\min\left\{\log\log\frac{\beta\log\beta}{\epsilon},\log\log n\right\}+\sqrt{\frac{n\beta\log\beta}{\epsilon}}\right)$	Yes
VRAG (known β) **(This Paper)**	$\mathcal{O}\left(n\min\left\{\log\log\frac{\beta}{\epsilon},\log\log n\right\}+\sqrt{\frac{n\beta}{\epsilon}}\right)$	No
Lower Bound (Woodworth & Srebro, 2016)	$\Omega\left(n+\sqrt{\frac{n\beta}{\epsilon}}\right)$	-

2021; Ene & Nguyen, 2021). Remarkably, these works have shown that, in the setting where we have access to the exact full gradient in each iteration, it is possible to match the convergence rates of both unaccelerated and accelerated gradient descent methods without any prior knowledge of the smoothness parameter. These methods have also been analyzed in the stochastic setting under a bounded variance assumption, and they achieve a convergence rate that is comparable to that of SGD.

Given the theoretical and practical success of adaptive methods, it is natural to ask whether one can design VR methods that achieve state of the art convergence guarantees without any prior knowledge of the smoothness parameter. The recent work of (Dubois-Taine et al., 2021) gives the first adaptive VR method — AdaSVRG — with the gradient complexity of $\mathcal{O}\left(n\log\frac{\beta}{\epsilon}+\frac{\beta}{\epsilon}\right)$. AdaSVRG builds on the AdaGrad (Duchi et al., 2011) and SVRG algorithms (Johnson & Zhang, 2013), both of which are not accelerated.

Our contributions: In this work, we take this line of work further and design the first accelerated VR methods that do not require any prior knowledge of the smoothness parameter. Our algorithms, *Adaptive Variance Reduced Accelerated Extra-Gradient* (AdaVRAE) and *Adaptive Variance Reduced Accelerated Gradient* (AdaVRAG), only use $\mathcal{O}\left(n\log\log n+\sqrt{\frac{n\beta}{\epsilon}}\right)$ and $\mathcal{O}\left(n\log\log n+\sqrt{\frac{n\beta\log\beta}{\epsilon}}\right)$ gradient evaluations respectively to attain an $\mathcal{O}(\epsilon)$-suboptimal solution when β is unknown, both of which significantly improve the convergence rate of AdaSVRG. Table 1 compares our algorithms and prior VR methods and Section 2 discusses our algorithmic approaches and techniques. The convergence rate of AdaVRAE matches up to constant factors the best-known convergence rate of non-adaptive VR methods (Song et al., 2020; Joulani et al., 2020). Both of our algorithms follow a different approach from these methods and are based on extra-gradient and mirror descent, instead of dual averaging.

We demonstrate the efficiency of our algorithms in practice on multiple real-world datasets. We show that AdaVRAG and AdaVRAE are competitive with existing standard and adaptive VR methods while having the advantage of not requiring hyperparameter tuning, and in many cases AdaVRAG outperforms these benchmarks.

1.1. Related work

Variance reduced gradient methods: Variance reduction technique (Roux et al., 2012; Schmidt et al., 2017; Shalev-

Shwartz & Zhang, 2013; Mairal, 2013; Johnson & Zhang, 2013; Defazio et al., 2014) has been proposed to improve the convergence rate of stochastic gradient descent algorithms in the finite sum problem and has since become widely-used in many successful algorithms. Notable improvements can be seen in strongly convex optimization problems where earliest algorithms such as SVRG (Johnson & Zhang, 2013) or SAGA (Defazio et al., 2014) obtain $\mathcal{O}\left((n+\kappa)\log\frac{1}{\epsilon}\right)$ convergence rate compared with $\mathcal{O}\left(\frac{\sigma^2\kappa}{\beta\epsilon}\right)$ of plain SGD, with the latter requiring an additional assumption on the σ^2-boundedness of the variance term, i.e., $\mathbb{E}_i\left[\|\nabla f_i(x)-\nabla f(x)\|^2\right]\le\sigma^2$. However, these non-accelerated methods do not achieve the optimal convergence rate. Recent works such as (Lin et al., 2015; Allen-Zhu, 2017; Lan et al., 2019) focus on designing accelerated methods and successfully match the optimal lower bound for strongly convex optimization of $\Omega\left((n+\sqrt{n\kappa})\log\frac{1}{\epsilon}\right)$ given by (Lan & Zhou, 2018).

In non-strongly convex problems, however, existing works do not yet match the lower bound of $\Omega\left(n+\sqrt{\frac{\beta n}{\epsilon}}\right)$ shown in (Woodworth & Srebro, 2016). The best effort so far can be found in the line of accelerated methods started by (Allen-Zhu, 2017) and followed by (Allen-Zhu, 2018; Lan et al., 2019; Li, 2021) that rely on incorporating the checkpoint in each update. AdaVRAG follows the same idea but offers simpler update and more efficient choice of coefficients that results in a better convergence rate, equivalent to VRADA (Song et al., 2020). By comparison, while VRADA is a dual-averaging scheme, AdaVRAG is a mirror descent method and AdaVRAE is an extra-gradient algorithm.

In a different line of research (Allen-Zhu & Hazan, 2016; Fang et al., 2018; Zhou et al., 2018), variance reduction has been applied to non-convex optimization to find critical points with much better convergence rate.

Adaptive methods with variance reduction: There has been extensive research on adaptive methods (Duchi et al., 2011; Kingma & Ba, 2014; Reddi et al., 2018; Tieleman et al., 2012; Dozat, 2016) in the setting where we compute a full gradient in each iteration. However, there are only few works combining adaptive methods with VR techniques in the finite sum setup. Most relevant for our work is AdaSVRG (Dubois-Taine et al., 2021). This algorithm is built upon SVRG which as mentioned earlier is a non-accelerated method and has a slower convergence rate. AdaSVRG uses the gradient norm to update the step size, similar to (Duchi et al., 2011) and the step is reset in every epoch, which could lead to step sizes that are too large in later stages. In contrast, both AdaVRAG and AdaVRAE are accelerated VR methods and use a cumulative step size. AdaVRAG uses the iterate movement to update the step size, as in (Bach & Levy, 2019; Ene et al., 2021). AdaVRAE improves the convergence rate by a $\sqrt{\log\beta}$ factor by using the gradient difference similarly to (Mohri & Yang, 2016; Joulani et al., 2020; Ene & Nguyen, 2021). In a different direction,(Xu et al., 2017) propose an adaptive VR algorithm adaptive to the unknown growth parameter instead of the smoothness parameter.

A different line of work considers VR methods that set the step size using stochastic line search (Schmidt et al., 2017; Mairal, 2013) or Barzilai-Borwein step size (Tan et al., 2016; Li et al., 2020). The former methods do not have theoretical guarantees, and the latter methods require knowledge of the smoothness parameter in order to obtain theoretical bounds.

Recent works design variance-reduced methods for non-convex optimization. STORM (Cutkosky & Orabona, 2019) and STORM$^+$(Levy et al., 2021) design an adaptive step size, though the former still requires the smoothness parameter in the step size. Super-Adam (Huang et al., 2021) also requires their parameters to satisfy some inequality involving the smoothness parameter like STORM.

1.2. Notation and problem setup

Let $[n]$ denote the set $\{1,2,\cdots,n\}$. For simplicity, we only consider the Euclidean norm $\|\cdot\|:=\|\cdot\|_2$ (Our work can be extended to $\|x\|_A:=\sqrt{x^\top Ax}$ for any $A\succ 0$ with almost no change). x^+ represents $\max\{x,0\}$.

We are interested in solving the following problem

$$\min_{x\in\mathcal{X}}\{F(x)=f(x)+h(x)\}$$

where $f(x):=\frac{1}{n}\sum_{i=1}^n f_i(x)$ and for $i\in[n]$, $f_i:\mathbb{R}^d\to\mathbb{R}$ and $h:\mathcal{X}\to\mathbb{R}$ are convex functions with a closed convex set $\mathcal{X}\subseteq\mathbb{R}^d$. Let $x^*=\arg\min_{x\in\mathcal{X}}F(x)$. We say a function G is β-smooth if $\|\nabla G(x)-\nabla G(y)\|\le\beta\|x-y\|$ for all $x,y\in\mathbb{R}^d$. Equivalently, we have $G(y)\le G(x)+\langle\nabla G(x),y-x\rangle+\frac{\beta}{2}\|y-x\|^2$. In this paper we always assume that each f_i is β-smooth, which implies that f is also β-smooth. We assume that we can efficiently solve optimization problems of the form $\arg\min_{x\in\mathcal{X}}\left(\gamma h(x)+\frac{1}{2}\|x-v\|^2\right)$ where $\gamma\ge 0$ and $v\in\mathbb{R}^d$. When the smoothness parameter β is unknown, we additionally assume that $\mathcal{X}$ is compact with diameter D, i.e., $\sup_{x,y\in\mathcal{X}}\|x-y\|\le D$.

2. Our algorithms and convergence guarantees

In this section, we describe our algorithms and state their convergence guarantees. Our algorithm AdaVRAE shown in Algorithm 1 is a novel accelerated scheme that uses past extra-gradient update steps in the inner loop and novel averaging to achieve acceleration. In each inner iteration, the

Algorithm 1 AdaVRAE

Input: initial point $u^{(0)}$, domain diameter D.
Parameters: $\{a^{(s)}\}$,$\{T_s\}$, $A_{T_0}^{(0)} > 0$, $\eta > 0$.
$\overline{x}_0^{(1)} = z_0^{(1)} = u^{(0)}$, compute $\nabla f(u^{(0)})$
Initialize $\gamma_0^{(1)} = \gamma$, where γ is any small constant
for $s = 1$ **to** S**:**
  $A_0^{(s)} = A_{T_{s-1}}^{(s-1)} - T_s \left(a^{(s)}\right)^2$
  for $t = 1$ **to** T_s**:**
    $x_t^{(s)} = \arg\min_{x\in\mathcal{X}} \left\{ a^{(s)} \left\langle g_{t-1}^{(s)}, x \right\rangle + a^{(s)} h(x) + \frac{\gamma_{t-1}^{(s)}}{2} \left\| x - z_{t-1}^{(s)} \right\|^2 \right\}$
    Let $A_t^{(s)} = A_{t-1}^{(s)} + a^{(s)} + \left(a^{(s)}\right)^2$
    $\overline{x}_t^{(s)} = \frac{1}{A_t^{(s)}} \left(A_{t-1}^{(s)} \overline{x}_{t-1}^{(s)} + a^{(s)} x_t^{(s)} + \left(a^{(s)}\right)^2 u^{(s-1)} \right)$
    if $t \neq T_s$**:**
      Pick $i_t^{(s)} \sim$ Uniform $([n])$
      $g_t^{(s)} = \nabla f_{i_t^{(s)}}(\overline{x}_t^{(s)}) - \nabla f_{i_t^{(s)}}(u^{(s-1)}) + \nabla f(u^{(s-1)})$
    else:
      $g_t^{(s)} = \nabla f(\overline{x}_t^{(s)})$
    $\gamma_t^{(s)} = \frac{1}{\eta} \sqrt{\eta^2 \left(\gamma_{t-1}^{(s)}\right)^2 + \left(a^{(s)}\right)^2 \left\| g_t^{(s)} - g_{t-1}^{(s)} \right\|^2}$
    $z_t^{(s)} = \arg\min_{z\in\mathcal{X}} \left\{ a^{(s)} \left\langle g_t^{(s)}, z \right\rangle + a^{(s)} h(z) + \frac{\gamma_{t-1}^{(s)}}{2} \left\| z - z_{t-1}^{(s)} \right\|^2 + \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2} \left\| z - x_t^{(s)} \right\|^2 \right\}$
  $u^{(s)} = \overline{x}_0^{(s+1)} = \overline{x}_{T_s}^{(s)}$, $z_0^{(s+1)} = z_{T_s}^{(s)}$, $g_0^{(s+1)} = g_{T_s}^{(s)}$, $\gamma_0^{(s+1)} = \gamma_{T_s}^{(s)}$
return $u^{(S)}$

Algorithm 2 AdaVRAG

Input: initial point $u^{(0)}$, domain diameter D.
Parameters: $\{a^{(s)}\}$, $a^{(s)} \in (0,1)$, $\{q^{(s)}\}$, $\{T_s\}$, $\eta > 0$.
$x_0^{(1)} = u^{(0)}$
Initialize $\gamma_0^{(1)} = \gamma$, where γ is any small constant
for $s = 1$ **to** S**:**
  $\overline{x}_0^{(s)} = a^{(s)} x_0^{(s)} + (1 - a^{(s)}) u^{(s-1)}$, compute $\nabla f(u^{(s-1)})$
  for $t = 1$ **to** T_s**:**
    Pick $i_t^{(s)} \sim$ Uniform $([n])$
    $g_t^{(s)} = \nabla f_{i_t^{(s)}}(\overline{x}_{t-1}^{(s)}) - \nabla f_{i_t^{(s)}}(u^{(s-1)}) + \nabla f(u^{(s-1)})$
    $x_t^{(s)} = \arg\min_{x\in\mathcal{X}} \left\{ \left\langle g_t^{(s)}, x \right\rangle + h(x) + \frac{\gamma_{t-1}^{(s)} q^{(s)}}{2} \left\| x - x_{t-1}^{(s)} \right\|^2 \right\}$
    $\overline{x}_t^{(s)} = a^{(s)} x_t^{(s)} + (1 - a^{(s)}) u^{(s-1)}$
    Option I: $\gamma_t^{(s)} = \gamma_{t-1}^{(s)} \sqrt{1 + \frac{\left\| x_t^{(s)} - x_{t-1}^{(s)} \right\|^2}{\eta^2}}$
    Option II: $\gamma_t^{(s)} = \gamma_{t-1}^{(s)} + \frac{\left\| x_t^{(s)} - x_{t-1}^{(s)} \right\|^2}{\eta^2}$
  $u^{(s)} = \frac{1}{T_s} \sum_{t=1}^{T_s} \overline{x}_t^{(s)}$, $x_0^{(s+1)} = x_{T_s}^{(s)}$, $\gamma_0^{(s+1)} = \gamma_{T_s}^{(s)}$
return $u^{(S)}$

new average iterate $\overline{x}_t^{(s)}$ is obtained by combining the old average iterate $\overline{x}_{t-1}^{(s)}$, the new iterate $x_t^{(s)}$ and the checkpoint $u^{(s-1)}$ with coefficients $A_{t-1}^{(s)}$, $a^{(s)}$ and $\left(a^{(s)}\right)^2$ normalized by the sum of them, i.e by $A_t^{(s)} = A_{t-1}^{(s)} + a^{(s)} + \left(a^{(s)}\right)^2$. We will explain the intuition behind this choice of coefficients in the analysis outline. At the beginning of each epoch, we set $A_0^{(s)} = A_{T_{s-1}}^{(s-1)} - T_s \left(a^{(s)}\right)^2$ so that at the end, we only accumulate the coefficients of the new iterates $x_t^{(s)}$. AdaVRAE adaptively sets the step sizes based on the stochastic gradient difference. Our choice of step sizes is a novel adaptation to the VR setting of the step sizes used by the works (Mohri & Yang, 2016; Kavis et al., 2019; Joulani et al., 2020; Ene & Nguyen, 2021) in the batch/full-gradient setting. Our algorithm builds on the work (Ene & Nguyen, 2021), which provides an unaccelerated past extra-gradient algorithm in the batch/full-gradient setting.

Theorem 2.1 states the parameter choices and the convergence guarantee for AdaVRAE, and we give its proof in Section A in the appendix. The convergence rate of AdaVRAE matches up to constant factors the rate of the state of the art non-adaptive VR methods (Joulani et al., 2020; Song et al., 2020). The initial step size $\gamma_0^{(1)}$ can be set to any small constant γ, which in practice we choose $\gamma = 0.01$. Similarly to AdaGrad, setting $\eta = \Theta(D)$ gives us the optimal dependence of the convergence rate in the domain diameter. For simplicity, we state the convergence in Theorem 2.1 and 2.2 when $\eta = \Theta(D)$. We refer the reader to Theorems A.1 and B.1 in the appendix for the precise choice of parameters as well as the full dependence of the convergence rate on arbitrary choices of γ and η. In both Theorem 2.1 and 2.2, we measure convergence using the number of individual gradient evaluations ∇f_i, assuming that the exact computation of ∇f takes n gradient evaluations.

Theorem 2.1. *(Convergence of AdaVRAE) Define* $s_0 = \lceil \log_2 \log_2 4n \rceil$, $c = \frac{3}{2}$. *Suppose we set the parameters of Algorithm 1 as follows:*

$$a^{(s)} = \begin{cases} (4n)^{-0.5^s} & 1 \le s \le s_0 \\ \frac{s - s_0 - 1 + c}{2c} & s_0 < s \end{cases},$$
$$T_s = n,$$
$$A_{T_0}^{(0)} = \frac{5}{4}.$$

Suppose that $\mathcal{X}$ *is a compact convex set with diameter* D *and we set* $\eta = \Theta(D)$. *The number of individual gradient evaluations to achieve a solution* $u^{(S)}$ *such that* $\mathbb{E}\left[F(u^{(S)}) - F(x^*)\right] \le \epsilon$ *for Algorithm 1 is*

$$\#grads = \begin{cases} \mathcal{O}\left(n \log\log \frac{V_1}{\epsilon}\right) & \text{if } \epsilon \ge \frac{V_1}{n} \\ \mathcal{O}\left(n \log\log n + \sqrt{\frac{nV_1}{\epsilon}}\right) & \text{if } \epsilon < \frac{V_1}{n} \end{cases}$$

where $V_1 = \mathcal{O}\left(F(u^{(0)}) - F(x^*) + (\gamma + \beta) D^2\right)$.

Our algorithm AdaVRAG is shown in Algorithm 2. Compared with AdaVRAE, AdaVRAG has a worse dependence on the smoothness parameter β but it performs only one projection onto $\mathcal{X}$ in each inner iteration. Additionally, as we discuss in more detail below, it uses adaptive step sizes based on the iterate movement.

AdaVRAG follows a similar framework to existing VR methods such as VARAG (Lan et al., 2019) and VRADA (Song et al., 2020). Similarly to VRADA, the algorithm achieves acceleration at the epoch level, where an epoch is an iteration of the outer loop. The iterations in an epoch update the main iterates via mirror descent with novel choices of step sizes and coefficients. The stochastic gradient is computed at a point that is a convex combination between the current iterate and the checkpoint; the coefficients of this combination remain fixed throughout the epoch. The step sizes are adaptively set based on the iterate movement.

The structure of the inner iterations of our algorithm differs from both VARAG and VRADA in several notable aspects. VARAG also uses mirror descent to update the main iterates and it computes the stochastic gradient at suitable combinations of the iterates and the checkpoint. AdaVARAG uses a different averaging of the iterates to compute the snapshots. Moreover, it uses a very different and simpler choice for the coefficient used to combine the main iterates and the checkpoint in order to obtain the points at which the stochastic gradients are evaluated. In VARAG, this coefficient is set to a constant (namely, $1/2$) in the initial iterations, whereas in AdaVRAG, it starts from a small number and is increased gradually. This choice is critical for improving the first term in the convergence from $\mathcal{O}(n \log n)$ to $\mathcal{O}(n \log \log n)$. In a similar manner, VRADA attains the same convergence by a new choice of coefficient. However, this is achieved via a very different approach based on dual-averaging.

The step sizes used by AdaVRAG have two components: the step $\gamma_t^{(s)}$ that is updated based on the iterate movement and the per-epoch coefficient $q^{(s)}$ to achieve acceleration at the epoch level. Our analysis is flexible and allows the use of several approaches for updating the steps $\gamma_t^{(s)}$. One approach, shown as option I in Algorithm 2, is based on the multiplicative update rule of AdaGrad+ (Ene et al., 2021) which generalizes the AdaGrad update to the constrained setting. We also propose a different variant, shown as option II, that updates the steps in an additive manner. Our analysis shows a similar convergence guarantee for both options, with the main difference being in the dependence on the smoothness: option I incurs a dependence of $\sqrt{\beta \log \beta}$, whereas option II has a worse dependence of β. Option II achieved improved performance in our experiments.

Theorem 2.2 states the parameter choices and the convergence guarantee for AdaVRAG, and we give its proof in Section B in the appendix. Analogously to AdaVRAE, the initial step size γ can be set to any small constant.

Theorem 2.2. *(Convergence of AdaVRAG) Define* $s_0 = \lceil \log_2 \log_2 4n \rceil$, $c = \frac{3+\sqrt{33}}{4}$. *Suppose we set the parameters of Algorithm 2 as follows:*

$$a^{(s)} = \begin{cases} 1 - (4n)^{-0.5^s} & 1 \le s \le s_0 \\ \frac{c}{s - s_0 + 2c} & s_0 < s \end{cases},$$

$$q^{(s)} = \begin{cases} \frac{1}{\left(1-a^{(s)}\right)a^{(s)}} & 1 \le s \le s_0 \\ \frac{8\left(2-a^{(s)}\right)a^{(s)}}{3\left(1-a^{(s)}\right)} & s_0 < s \end{cases},$$

$$T_s = n.$$

Suppose that $\mathcal{X}$ is a compact convex set with diameter D and we set $\eta = \Theta(D)$. Additionally, we assume that $2\eta^2 > D^2$ if Option I is used for setting the step size. The number of individual gradient evaluations to achieve a solution $u^{(S)}$ such that $\mathbb{E}\left[F(u^{(S)}) - F(x^*)\right] \le \epsilon$ for Algorithm 2 is

$$\#grads = \begin{cases} \mathcal{O}\left(n \log \log \frac{V_2}{\epsilon}\right) & \epsilon \ge \frac{V_2}{n} \\ \mathcal{O}\left(n \log \log n + \sqrt{\frac{nV_2}{\epsilon}}\right) & \epsilon < \frac{V_2}{n} \end{cases},$$

where

$$V_2 = \begin{cases} \mathcal{O}\left(F(u^{(0)}) - F(x^*) + \left(\gamma + \beta \log\left(\frac{\beta}{\gamma}\right)\right) D^2\right) & \text{for Option I} \\ \mathcal{O}\left(F(u^{(0)}) - F(x^*) + \left(\gamma + \beta^2\right) D^2\right) & \text{for Option II} \end{cases}$$

Comparison to AdaSVRG: As noted in the introduction, the state of the art adaptive VR method is the AdaSVRG algorithm (Dubois-Taine et al., 2021), which is a non-accelerated method. Both of our algorithms achieve a faster convergence using different approaches and step sizes. AdaSVRG resets the step sizes in each epoch, whereas our algorithms use a cumulative update approach for the step sizes. In our experimental evaluation, the resetting of the step sizes led to slower convergence. AdaSVRG (multi-stage variant) uses varying epoch lengths similarly to SVRG^{++} (Allen-Zhu & Yuan, 2016), whereas our algorithms use epoch lengths that are set to n. Using an epoch of length n allows for implementing the random sampling via a random permutation of $[n]$ and is the preferred approach in practice.

Both our algorithms and AdaSVRG require that the domain $\mathcal{X}$ has bounded diameter. This is a restriction that is shared by almost all existing adaptive methods. Recent work (Antonakopoulos et al., 2021; Ene & Nguyen, 2021) in the batch/full-gradient setting have proposed unaccelerated methods that are suitable for unbounded domains, at a loss of additional factors in the convergence. All of the existing accelerated methods require that the domain is bounded,

even in the batch/full-gradient setting. We note that our analysis holds for arbitrary compact domains, whereas the analysis of AdaSVRG only applies to domains that contain the global optimum. Similarly to AdaGrad, both our algorithms and AdaSVRG can be used in the unconstrained setting under the promise that the iterates do not move too far from the optimum.

Non-adaptive variants of our algorithms: In the setting where the smoothness parameter is known, we can set the step sizes of our algorithms based on the smoothness, as shown in Algorithms 3 and 4 (Sections C and D in the appendix). Both algorithms match the convergence rates of the state of the art VR methods (Joulani et al., 2020; Song et al., 2020) using different algorithmic approaches based on mirror descent and extra-gradient instead of dual-averaging. We experimentally compare the non-adaptive algorithms to existing methods in Section E of the appendix.

2.1. Analysis outline

We outline some of the key steps in the analysis of AdaVRAE. For the purpose of simplicity, we assume $h = 0$ and $\eta = D$. Starting from the observation $\overline{x}_t^{(s)} = \frac{1}{A_t^{(s)}}\left(A_{t-1}^{(s)}\overline{x}_{t-1}^{(s)} + a^{(s)}x_t^{(s)} + \left(a^{(s)}\right)^2 u^{(s-1)}\right)$ and $A_t^{(s)} = A_{t-1}^{(s)} + a^{(s)} + \left(a^{(s)}\right)^2$, we have

$$\overline{x}_t^{(s)} - \overline{x}_{t-1}^{(s)} = \frac{a^{(s)}}{A_{t-1}^{(s)}}\left(x_t^{(s)} - \overline{x}_t^{(s)}\right) + \frac{\left(a^{(s)}\right)^2}{A_{t-1}^{(s)}}\left(u^{(s-1)} - \overline{x}_t^{(s)}\right)$$

which allows us to carry out the analysis for the function progress in one iteration, i.e, $f(\overline{x}_t^{(s)}) - f(\overline{x}_{t-1}^{(s)})$ and obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(A_t^{(s)} - \left(a^{(s)}\right)^2\right)\left(f(\overline{x}_t^{(s)}) - f(x^*)\right)\right. \\
&\quad \left. - A_{t-1}^{(s)}\left(f(\overline{x}_{t-1}^{(s)}) - f(x^*)\right)\right] \\
\leq &\mathbb{E}\left[a^{(s)}\underbrace{\left\langle g_t^{(s)}, x_t^{(s)} - x^*\right\rangle}_{\text{stochastic regret}}\right] \\
&+ \mathbb{E}\left[\left(a^{(s)}\right)^2\left\langle \nabla f(\overline{x}_t^{(s)}), u^{(s-1)} - \overline{x}_t^{(s)}\right\rangle\right] \\
&- \mathbb{E}\left[\frac{A_{t-1}^{(s)}}{2\beta}\left\|\nabla f(\overline{x}_t^{(s)}) - \nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right].
\end{aligned}$$

By building on the standard analysis of the stochastic regret for extra-gradient methods, we obtain the following result for the progress of one iteration:

$$\begin{aligned}
&\mathbb{E}\left[\left(A_t^{(s)} - \left(a^{(s)}\right)^2\right)\left(f(\overline{x}_t^{(s)}) - f(x^*)\right)\right. \\
&\quad \left. - A_{t-1}^{(s)}\left(f(\overline{x}_{t-1}^{(s)}) - f(x^*)\right)\right] \\
\leq &\mathbb{E}\left[\frac{\gamma_{t-1}^{(s)}}{2}\left\|z_{t-1}^{(s)} - x^*\right\|^2 - \frac{\gamma_t^{(s)}}{2}\left\|z_t^{(s)} - x^*\right\|^2\right] \\
&+ \mathbb{E}\left[\frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2\right] \\
&+ \mathbb{E}\left[\left(a^{(s)}\right)^2\left\langle \nabla f(\overline{x}_t^{(s)}), u^{(s-1)} - \overline{x}_t^{(s)}\right\rangle\right] \\
&+ \mathbb{E}\left[\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2\right] \\
&\underbrace{- \mathbb{E}\left[\frac{A_{t-1}^{(s)}}{2\beta}\left\|\nabla f(\overline{x}_t^{(s)}) - \nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right]}_{\text{gain}}. \qquad (2)
\end{aligned}$$

In comparison to the standard analysis, the coefficient for the checkpoint appears in the coefficient of $f(\overline{x}_t^{(s)}) - f(x^*)$, which becomes $\left(A_t^{(s)} - \left(a^{(s)}\right)^2\right)$ instead of the usual $A_t^{(s)}$, making the sum not telescope immediately. To resolve this, we first turn our attention to the analysis of the stochastic gradient difference $\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2$. The key idea is to split $\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2$ into $\left(\frac{1}{2\gamma_t^{(s)}} - \frac{1}{16\beta}\right)\left(a^{(s)}\right)^2\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2 + \frac{\left(a^{(s)}\right)^2}{16\beta}\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2$, and bound each term in turn. For the first term, we build on the techniques from prior work in the batch/full-gradient setting (Ene & Nguyen, 2021) when taking the sum over the iterations and epochs. For intuition, part of the gain $-\frac{1}{16\beta}\left(a^{(s)}\right)^2\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2$ will be used to cancel out the term $\mathbb{E}\left[\frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2\right]$. We then only need to consider the time before $\gamma_t^{(s)}$ goes above $\mathcal{O}(\beta)$, and notice that $\left(a^{(s)}\right)^2\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2 = \left(\gamma_t^{(s)}\right)^2 - \left(\gamma_{t-1}^{(s)}\right)^2$, thus we can upperbound the first term via the last $\gamma_t^{(s)}$ that is still small than $\mathcal{O}(\beta)$. We provide more details below.

For the second term, we use Young's inequality to write $\mathbb{E}\left[\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2\right] \leq \mathbb{E}\left[4\left\|\nabla f(\overline{x}_t^{(s)}) - g_t^{(s)}\right\|^2 + 4\left\|\nabla f(\overline{x}_{t-1}^{(s)}) - g_{t-1}^{(s)}\right\|^2\right] +$

$\mathbb{E}\left[2\left\|\nabla f(\overline{x}_t^{(s)}) - \nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right]$. The gradient difference loss term is cancelled by the gain term in (2), and thus we can focus on the first two variance terms. We apply the usual variance reduction technique put forward by (Lan et al., 2019) (see Lemma A.2) to bound the two variance terms, as follows:

$$\mathbb{E}\left[\left\|g_t^{(s)} - \nabla f(\overline{x}_t^{(s)})\right\|^2\right] \leq \mathbb{E}\left[2\beta\left(f(u^{(s-1)}) - f(\overline{x}_t^{(s)}) - \left\langle \nabla f(\overline{x}_t^{(s)}), u^{(s-1)} - \overline{x}_t^{(s)}\right\rangle\right)\right].$$

Thus we obtain an upper bound on $\frac{\left(a^{(s)}\right)^2}{16\beta}\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2$ in terms of $\left(a^{(s)}\right)^2\left(f(u^{(s-1)}) - f(\overline{x}_t^{(s)})\right)$. This is the reason for setting the coefficient for the checkpoint to $\left(a^{(s)}\right)^2$, so that the LHS of (2) can become the usual telescoping sum $A_t^{(s)}\left(f(\overline{x}_t^{(s)}) - f(x^*)\right) - A_{t-1}^{(s)}\left(f(\overline{x}_{t-1}^{(s)}) - f(x^*)\right)$. Using the convexity of f, we obtain the following key result for the progress of each epoch:

$$\begin{aligned}
&\mathbb{E}\left[A_{T_s}^{(s)}\left(f(\overline{x}_{T_s}^{(s)}) - f(x^*)\right) - A_0^{(s)}\left(f(\overline{x}_0^{(s)}) - f(x^*)\right)\right] \\
\leq&\mathbb{E}\left[\frac{\gamma_0^{(s)}}{2}\left\|z_0^{(s)} - x^*\right\|^2 - \frac{\gamma_{T_s}^{(s)}}{2}\left\|z_{T_s}^{(s)} - x^*\right\|^2\right] \\
&+ \mathbb{E}\left[\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2\right] \\
&+ \mathbb{E}\left[T_s\left(a^{(s)}\right)^2\left(f(u^{(s-1)}) - f(x^*)\right)\right] \\
&+ \mathbb{E}\left[\sum_{t=1}^{T_s}\left(\frac{1}{2\gamma_t^{(s)}} - \frac{1}{16\beta}\right)\left(a^{(s)}\right)^2\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2\right].
\end{aligned}$$

Intuitively, we want to have another telescoping sum when summing up the above inequality across all epochs s. To do so, we can set the starting points of the next epoch to be the ending points of the previous one, i.e., $\overline{x}_{T_s}^{(s)} = \overline{x}_0^{(s+1)} = u^{(s)}$, $\gamma_{T_s}^{(s)} = \gamma_0^{(s+1)}$, $z_{T_s}^{(s)} = z_0^{(s+1)}$. However, an extra term $T_s\left(a^{(s)}\right)^2\left(f(u^{(s-1)}) - f(x^*)\right)$ appears on the RHS. We need to reset the new starting coefficient in the new epoch $A_0^{(s)}$ to $A_{T_{s-1}}^{(s-1)} - T_s\left(a^{(s)}\right)^2$ so that we can telescope the LHS.

To bound the term $\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2 + \left(\frac{1}{2\gamma_t^{(s)}} - \frac{1}{16\beta}\right)\left(a^{(s)}\right)^2\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2$, since $\gamma_{T_s}^{(s)} = \gamma_0^{(s+1)}$ and, the sequence $\left(\gamma_t^{(s)}\right)$ is not decreasing, we can make the first part of the sum telescope: $\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2 \leq \sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}D^2 = \frac{D^2}{2}\left(\gamma_{T_S}^{(S)} - \gamma_0^{(1)}\right)$. Moreover, $g_{T_s}^{(s)} = g_0^{(s+1)}$, we can consider the doubly indexed sequences $\left(\gamma_t^{(s)}\right)$ and $\left(g_t^{(s)}\right)$ as two singly indexed sequences (γ_k) and (g_k) and the coefficient $a^{(s)}$ to be another sequence (a_k). Then we can employ the following two inequalities:

$$\begin{aligned}
&\frac{D^2}{2}\left(\gamma_K - \gamma_0\right) - \frac{1}{48\beta}\sum_{k=1}^{K}a_k^2\left\|g_k - g_{k-1}\right\|^2 \\
\leq&12\beta D^2 \\
&\sum_{k=1}^{K}\left(\frac{1}{2\gamma_k} - \frac{1}{24\beta}\right)a_k^2\left\|g_k - g_{k-1}\right\|^2 \\
\leq&12\beta D^2
\end{aligned}$$

Finally, we need to choose the parameters $a^{(s)}$ so that the conditions needed for our analysis are satisfied and $A_{T_s}^{(s)}$ is sufficiently large, so that we attain a fast convergence. We have to choose $a^{(s)}$ such that $\left(a^{(s)}\right)^2 \leq 4A_{t-1}^{(s)}$ for all $s, t \geq 1$ and that $A_0^{(s)} = A_{T_{s-1}}^{(s-1)} - T_s\left(a^{(s)}\right)^2 \geq 0$. The main idea is to divide the epochs into two phases: in the first phase, $A_{T_s}^{(s)}$ quickly rises to $\Omega(n)$ and in the second phase, to achieve the optimal $\sqrt{\frac{n\beta}{\epsilon}}$ rate, $A_{T_s}^{(s)} = \Omega(n^2)$. The nearly-optimal choice of $a^{(s)}$ in the first phase is $(4n)^{-0.5^s}$, stopping at $s = s_0 = \lceil \log_2 \log_2 4n \rceil$, while in the second phase, we have to be more conservative and choose $a^{(s)} = \frac{s - s_0 + \frac{1}{2}}{3}$. With this we can obtain the convergence rate of $\mathcal{O}\left(n\min\left\{\log\log\frac{\beta}{\epsilon}, \log\log n\right\} + \sqrt{\frac{n\beta}{\epsilon}}\right)$.

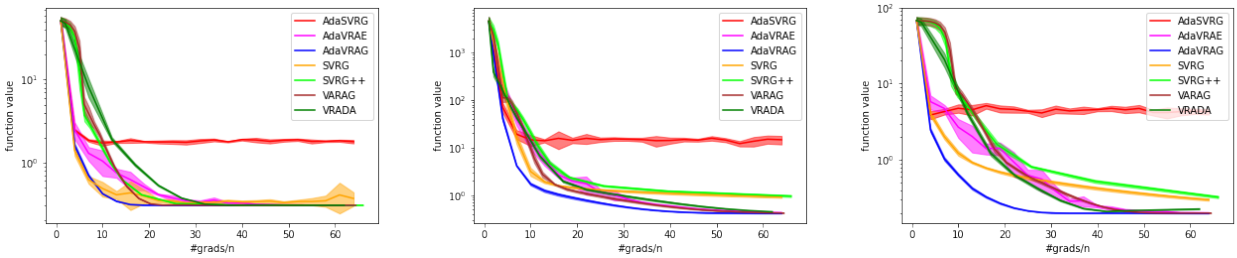
3. Experiments

In this section we demonstrate the performances of AdaVRAG and AdaVRAE in comparison with the existing standard and adaptive VR methods. We use the experimental setup and the code base of (Dubois-Taine et al., 2021)[1].

Datasets and loss functions: We experiment with binary classification on four standard LIBSVM datasets: a1a, mushrooms, w8a and phishing (Chang & Lin, 2011). For each dataset, we show the results for three different objective functions: logistic, squared and huber loss. Following the setting in (Dubois-Taine et al., 2021) we add a ℓ_2-regularization term to the loss function, with regularization set to $1/n$.

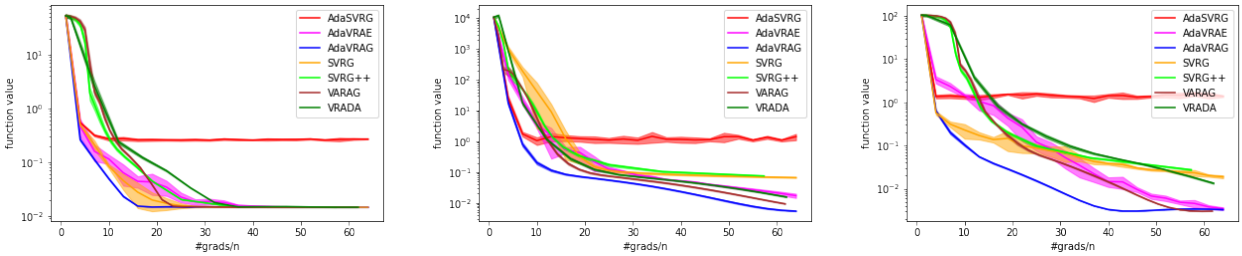
Constraint: In all experiments, we evaluate the algorithms under a ball constraint. That is, the domain of each problem in our experiment is a ball of radius $R = 100$ around the initial point, which means for every algorithm, in the update step, we need to do a projection onto this ball.

[1]Their code can be found at https://github.com/bpauld/AdaSVRG



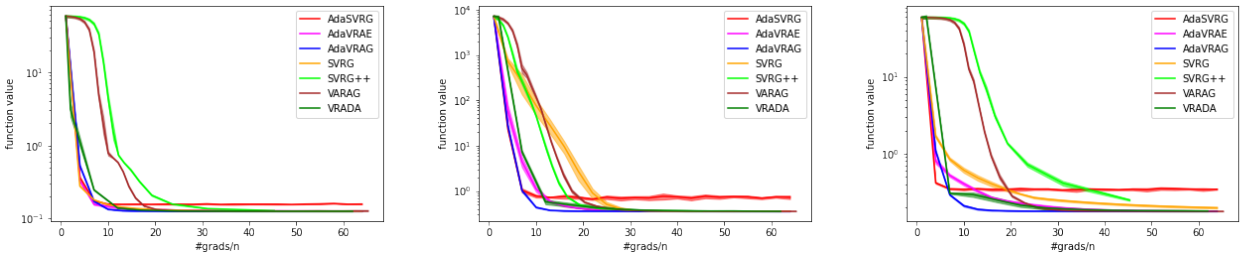
(a) Logistic loss (b) Squared loss (c) Huber loss

Figure 1. a1a



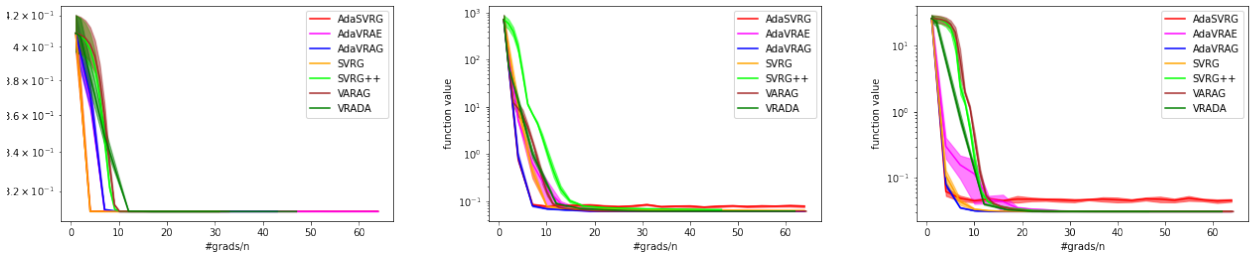
(a) Logistic loss (b) Squared loss (c) Huber loss

Figure 2. mushrooms



(a) Logistic loss (b) Squared loss (c) Huber loss

Figure 3. w8a



(a) Logistic loss (b) Squared loss (c) Huber loss

Figure 4. phishing

Algorithms and hyperparameter selection: We compare AdaVRAE and AdaVRAG with the common VR algorithms: SVRG (Johnson & Zhang, 2013), SVRG^{++}(Allen-Zhu & Yuan, 2016), VARAG (Lan et al., 2019), VRADA (Song et al., 2020), and AdaSVRG (Dubois-Taine et al., 2021) (in the experiment the multi-stage variant performs worse than the fixed-sized inner loop variant, and we omit it from the plots). Among these, only AdaSVRG is an adaptive VR method, which does not require parameter tuning. For the non-adaptive methods we chose the step size (or equivalently, the inverse of the smoothness parameter ($1/\beta$) for VRADA) via hyperparameter search over $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 100\}$. For each experiment, we used the choice that led to the best performance, and we report the parameters used in Table 2. The adaptive methods — AdaSVRG, AdaVRAE, AdaVRAG — do not require any hyperparameter tuning and we set their parameters as prescribed by the theoretical analysis. For AdaSVRG, we used $\eta = D/\sqrt{2} = \sqrt{2}R$ as recommended in the original paper. For AdaVRAE and AdaVRAG, we used $\gamma = 0.01$ and $\eta = D/2 = R$.

Implementation and initialization: For all algorithms, in the inner loop, we use a random permutation of the data points to select a function. We also fix the batch size to 1 in all cases to match the theoretical setting. We initialize $u^{(0)}$ to be a random point in $[0, 10]^d$ where each dimension is uniformly chosen in $[0, 10]$. Each experiment is repeated five times with different initial point, which is kept the same across all algorithms.

Results: The results are shown in Figures 1, 2, 3, 4. For each experiment, we plot the mean value and standard deviation of the training objective against the number of gradient evaluations normalized by the number of examples.

Discussion: We observe that, in all experiments, AdaVRAG consistently performs competitively with all methods and generally have the best performances. The non-accelerated methods in general converge more slowly compared with accelerated methods, especially in the later epochs. In some cases, VARAG suffers from a slow convergence rate in the first phase. This is possibly due to the fact that it sets to $1/2$ the coefficient for the checkpoint in the first phase. VRADA sometimes exhibits similar behavior but to a lesser extent. In AdaVRAG and AdaVRAE, the coefficient for the checkpoint is set to be small in the beginning and gradually increased over time when the quality of the checkpoint is improved. The other adaptive method, AdaSVRG, exhibits slow convergence in many cases. One reason might be that AdaSVRG resets the step size in every epoch and, in later epochs, the step size may be too large for the algorithm to converge. In contrast, AdaVRAG and AdaVRAE use cumulative step sizes.

4. Conclusion and future work

In this paper, we propose two accelerated variance reduced algorithms for the general finite-sum convex optimization problem with the step size set adaptively to the smoothness parameter. By a careful design of the coefficient choices, the first extra-gradient algorithm, AdaVRAE, which sets the step size via the gradient difference, uses $\mathcal{O}\left(n \log\log n + \sqrt{\frac{n\beta}{\epsilon}}\right)$ gradient evaluations to attain an $\mathcal{O}(\epsilon)$-suboptimal solution, matching the best-known convergence rate of non-adaptive VR methods, while removing the requirement of the knowledge about the smoothness parameter. The second algorithm, AdaVRAG, which uses the iterate moment in the step size, needs $\mathcal{O}\left(n \log\log n + \sqrt{\frac{n\beta \log \beta}{\epsilon}}\right)$ gradient evaluations, but having the advantage of using a single projection in each iteration and performing better in practice. For both algorithms, as well as the other state-of-the-art VR algorithms, there is still a gap to the lower bound convergence $\left(\Omega\left(n + \sqrt{\frac{n\beta}{\epsilon}}\right)\right)$ for the general finite-sum convex optimization problem. Finding an algorithm that can achieve this lower bound and making it adaptive remain an open question for the future work.

Acknowledgments

ZL, TN, and AE were supported in part by NSF CAREER grant CCF-1750333, NSF grant III-1908510, and an Alfred P. Sloan Research Fellowship. HN was supported in part by NSF CAREER grant CCF-1750716 and NSF grant CCF-1909314.

References

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Allen-Zhu, Z. Katyusha x: Practical momentum method for stochastic sum-of-nonconvex optimization. *arXiv preprint arXiv:1802.03866*, 2018.

Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pp. 699–707. PMLR, 2016.

Allen-Zhu, Z. and Yuan, Y. Improved svrg for non-strongly-convex or sum-of-non-convex objectives. In *International conference on machine learning*, pp. 1080–1089. PMLR, 2016.

Antonakopoulos, K., Belmega, V., and Mertikopoulos, P. Adaptive extra-gradient methods for min-max optimization and games. In *International Conference on Learning Representations (ICLR)*, 2021.

Bach, F. and Levy, K. Y. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Conference on Learning Theory*, pp. 164–194. PMLR, 2019.

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Cutkosky, A. Anytime online-to-batch, optimism and acceleration. In *International Conference of Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1446–1454. PMLR, 2019.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pp. 1646–1654, 2014.

Dozat, T. Incorporating nesterov momentum into adam. 2016.

Dubois-Taine, B., Vaswani, S., Babanezhad, R., Schmidt, M., and Lacoste-Julien, S. Svrg meets adagrad: Painless variance reduction. *arXiv preprint arXiv:2102.09645*, 2021.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Ene, A. and Nguyen, H. L. Adaptive and universal algorithms for variational inequalities with optimal convergence s. *arXiv preprint arXiv:2010.07799*, 2021.

Ene, A., Nguyen, H. L., and Vladu, A. Adaptive gradient methods for constrained convex optimization and variational inequalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7314–7321, 2021.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *arXiv preprint arXiv:1807.01695*, 2018.

Huang, F., Li, J., and Huang, H. Super-adam: Faster and universal framework of adaptive gradients. *arXiv preprint arXiv:2106.08208*, 2021.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323, 2013.

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

Kavis, A., Levy, K. Y., Bach, F., and Cevher, V. Unixgrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6257–6266, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lan, G. and Zhou, Y. Random gradient extrapolation for distributed and stochastic optimization. *SIAM Journal on Optimization*, 28(4):2753–2782, 2018.

Lan, G., Li, Z., and Zhou, Y. A unified variance-reduced accelerated gradient method for convex optimization. *arXiv preprint arXiv:1905.12412*, 2019.

Levy, K., Kavis, A., and Cevher, V. Storm+: Fully adaptive sgd with recursive momentum for nonconvex optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

Levy, K. Y. Online to offline conversions, universality and adaptive minibatch sizes. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1613–1622, 2017.

Levy, K. Y., Yurtsever, A., and Cevher, V. Online adaptive methods, universality and acceleration. In *Advances in*

Neural Information Processing Systems (NeurIPS), pp. 6500–6509, 2018.

Li, B., Wang, L., and Giannakis, G. B. Almost tune-free variance reduction. In *International Conference on Machine Learning*, pp. 5969–5978. PMLR, 2020.

Li, Z. Anita: An optimal loopless accelerated variance-reduced gradient method. *arXiv preprint arXiv:2103.11333*, 2021.

Lin, H., Mairal, J., and Harchaoui, Z. A universal catalyst for first-order optimization. *arXiv preprint arXiv:1506.02186*, 2015.

Mairal, J. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pp. 783–791. PMLR, 2013.

McMahan, H. B. and Streeter, M. J. Adaptive bound optimization for online convex optimization. In *Conference on Learning Theory (COLT)*, pp. 244–256. Omnipress, 2010.

Mohri, M. and Yang, S. Accelerating online convex optimization via adaptive prediction. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 848–856, 2016.

Nesterov, Y. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady an ussr*, volume 269, pp. 543–547, 1983.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

Roux, N. L., Schmidt, M., and Bach, F. A stochastic gradient method with an exponential convergence rate for finite training sets. *arXiv preprint arXiv:1202.6258*, 2012.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Shalev-Shwartz, S. and Zhang, T. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(2), 2013.

Song, C., Jiang, Y., and Ma, Y. Variance reduction via accelerated dual averaging for finite-sum optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Tan, C., Ma, S., Dai, Y.-H., and Qian, Y. Barzilai-borwein step size for stochastic gradient descent. *Advances in Neural Information Processing Systems*, 29:685–693, 2016.

Tieleman, T., Hinton, G., et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4 (2):26–31, 2012.

Woodworth, B. E. and Srebro, N. Tight complexity bounds for optimizing composite objectives. *Advances in neural information processing systems*, 29:3639–3647, 2016.

Xu, Y., Lin, Q., and Yang, T. Adaptive svrg methods under error bound conditions with unknown growth parameter. *Advances in Neural Information Processing Systems*, 30, 2017.

Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduction for nonconvex optimization. *arXiv preprint arXiv:1806.07811*, 2018.

A. Analysis of algorithm 1

In this section, we analyze Algorithm 1 and prove the following convergence guarantee:

Theorem A.1 (Convergence of AdaVRAE)**.** *Define* $s_0 = \lceil \log_2 \log_2 4n \rceil$, $c = \frac{3}{2}$. *If we choose parameters as follows*

$$a^{(s)} = \begin{cases} (4n)^{-0.5^s} & 1 \le s \le s_0 \\ \frac{s - s_0 - 1 + c}{2c} & s_0 < s \end{cases},$$
$$T_s = n,$$
$$A_{T_0}^{(0)} = \frac{5}{4}.$$

Assuming $\mathcal{X}$ is a compact convex set with diameter D, the number of individual gradient evaluations to achieve a solution $u^{(S)}$ such that $\mathbb{E}\left[F(u^{(S)}) - F(x^*)\right] \le \epsilon$ for Algorithm 1 is

$$\#grads = \begin{cases} \mathcal{O}\left(n \log\log \frac{V}{\epsilon}\right) & \text{if } \epsilon \ge \frac{V}{n} \\ \mathcal{O}\left(n \log\log n + \sqrt{\frac{Vn}{\epsilon}}\right) & \text{if } \epsilon < \frac{V}{n} \end{cases}$$

where $V = \frac{5}{2}\left(F(u^{(0)}) - F(x^*)\right) + \gamma \left\|u^{(0)} - x^*\right\|^2 + \frac{16\beta\left(D^4 + 2\eta^4\right)}{\eta^2}$.

To start with, we state and prove the following variance reduction lemma commonly used in accelerated methods:

Lemma A.2. *(Variance Reduction) Let $i \sim \mathrm{Uniform}([n])$ and $g = \nabla f_i(x) - \nabla f_i(u) + \nabla f(u)$ be an estimate of the gradient of f at x. We have*

$$\mathbb{E}_i\left[\|g - \nabla f(x)\|^2\right] \le 2\beta\left(f(u) - f(x) - \langle \nabla f(x), u - x\rangle\right).$$

Proof. By the definition of g,

$$\begin{aligned} \mathbb{E}_i\left[\|g - \nabla f(x)\|^2\right] &= \mathbb{E}_i\left[\|\nabla f_i(x) - \nabla f_i(u) + \nabla f(u) - \nabla f(x)\|^2\right] \\ &\overset{(a)}{\le} \mathbb{E}_i\left[\|\nabla f_i(u) - \nabla f_i(x)\|^2\right] \\ &\overset{(b)}{\le} \mathbb{E}_i\left[2\beta\left(f_i(u) - f_i(x) - \langle \nabla f_i(x), u - x\rangle\right)\right] \\ &\overset{(c)}{=} 2\beta\left(f(u) - f(x) - \langle \nabla f(x), u - x\rangle\right), \end{aligned}$$

where (a) is because $\mathbb{E}_i\left[\nabla f_i(u) - \nabla f_i(x)\right] = \nabla f(u) - \nabla f(x)$ and $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \le \mathbb{E}\left[\|X\|^2\right]$, (b) is by the convexity and β-smoothness of f_i, (c) is by $i \sim \mathrm{Uniform}([n])$ and the definition of f. □

A.1. Single iteration progress

We first analyze the progress in function value made in a single iteration of an epoch. The analysis follows the standard method as in (Ene & Nguyen, 2021); however, we need to pay attention to the extra term for the checkpoint that appears in the convex combination for $\overline{x}_t^{(s)}$. We start off by the following observation

Lemma A.3. *For any $s \ge 1$ and $t \in [T_s]$,*

$$\overline{x}_t^{(s)} - \overline{x}_{t-1}^{(s)} = \frac{a^{(s)}}{A_{t-1}^{(s)}}\left(x_t^{(s)} - \overline{x}_t^{(s)}\right) + \frac{\left(a^{(s)}\right)^2}{A_{t-1}^{(s)}}\left(u^{(s-1)} - \overline{x}_t^{(s)}\right).$$

Proof. We note that the definition $\overline{x}_t^{(s)} = \frac{1}{A_t^{(s)}}\left(A_{t-1}^{(s)}\overline{x}_{t-1}^{(s)} + a^{(s)}x_t^{(s)} + \left(a^{(s)}\right)^2 u^{(s-1)}\right)$ implies

$$
\begin{aligned}
A_t^{(s)}\overline{x}_t^{(s)} &= A_{t-1}^{(s)}\overline{x}_{t-1}^{(s)} + a^{(s)}x_t^{(s)} + \left(a^{(s)}\right)^2 u^{(s-1)} \\
\overset{(a)}{\Leftrightarrow} A_{t-1}^{(s)}(\overline{x}_t^{(s)} - \overline{x}_{t-1}^{(s)}) &= a^{(s)}(x_t^{(s)} - \overline{x}_t^{(s)}) + \left(a^{(s)}\right)^2 (u^{(s-1)} - \overline{x}_t^{(s)}) \\
\Leftrightarrow \overline{x}_t^{(s)} - \overline{x}_{t-1}^{(s)} &= \frac{a^{(s)}}{A_{t-1}^{(s)}}\left(x_t^{(s)} - \overline{x}_t^{(s)}\right) + \frac{\left(a^{(s)}\right)^2}{A_{t-1}^{(s)}}\left(u^{(s-1)} - \overline{x}_t^{(s)}\right),
\end{aligned}
$$

where (a) is by $A_t^{(s)} = A_{t-1}^{(s)} + a^{(s)} + \left(a^{(s)}\right)^2$. □

Next, we bound the function progress in a single epoch via the stochastic regret. Note that, this lemma is somewhat weaker than we would desire, due to the appearance the coefficient of the checkpoint, making the LHS not immediately telescope. We will account for this factor later in the analysis.

Lemma A.4. *For all epochs $s \geq 1$ and all iterations $t \in [T_s]$*

$$
\begin{aligned}
&\mathbb{E}\left[\left(A_t^{(s)} - \left(a^{(s)}\right)^2\right)\left(F(\overline{x}_t^{(s)}) - F(x^*)\right) - A_{t-1}^{(s)}\left(F(\overline{x}_{t-1}^{(s)}) - F(x^*)\right)\right] \\
\leq &\mathbb{E}\left[a^{(s)}\underbrace{\left\langle g_t^{(s)}, x_t^{(s)} - x^*\right\rangle}_{\textit{stochastic regret}} + \left(a^{(s)}\right)^2\left\langle \nabla f(\overline{x}_t^{(s)}), u^{(s-1)} - \overline{x}_t^{(s)}\right\rangle\right] \\
&- \mathbb{E}\left[\frac{A_{t-1}^{(s)}}{2\beta}\left\|\nabla f(\overline{x}_t^{(s)}) - \nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right] \\
&+ \mathbb{E}\left[\left(A_t^{(s)} - \left(a^{(s)}\right)^2\right)\left(h(\overline{x}_t^{(s)}) - h(x^*)\right) - A_{t-1}^{(s)}\left(h(\overline{x}_{t-1}^{(s)}) - h(x^*)\right)\right].
\end{aligned}
$$

Proof. Using the observation in Lemma A.3, we have

$$
\begin{aligned}
&F(\overline{x}_t^{(s)}) - F(\overline{x}_{t-1}^{(s)}) \\
=&f(\overline{x}_t^{(s)}) - f(\overline{x}_{t-1}^{(s)}) + h(\overline{x}_t^{(s)}) - h(\overline{x}_{t-1}^{(s)}) \\
\overset{(a)}{\leq} &\left\langle \nabla f(\overline{x}_t^{(s)}), \overline{x}_t^{(s)} - \overline{x}_{t-1}^{(s)}\right\rangle - \frac{1}{2\beta}\left\|\nabla f(\overline{x}_t^{(s)}) - \nabla f(\overline{x}_{t-1}^{(s)})\right\|^2 + h(\overline{x}_t^{(s)}) - h(\overline{x}_{t-1}^{(s)}) \\
\overset{(b)}{=} &\frac{a^{(s)}}{A_{t-1}}\left\langle \nabla f(\overline{x}_t^{(s)}), x_t^{(s)} - \overline{x}_t^{(s)}\right\rangle + \frac{\left(a^{(s)}\right)^2}{A_{t-1}^{(s)}}\left\langle \nabla f(\overline{x}_t^{(s)}), u^{(s-1)} - \overline{x}_t^{(s)}\right\rangle \\
&- \frac{1}{2\beta}\left\|\nabla f(\overline{x}_t^{(s)}) - \nabla f(\overline{x}_{t-1}^{(s)})\right\|^2 + h(\overline{x}_t^{(s)}) - h(\overline{x}_{t-1}^{(s)})
\end{aligned}
$$

where (a) is due to the smoothness of f and (b) comes from Lemma A.3. By the convexity of f, we also have

$$
\begin{aligned}
&F(\overline{x}_t^{(s)}) - F(x^*) \\
=&f(\overline{x}_t^{(s)}) - f(x^*) + h(\overline{x}_t^{(s)}) - h(x^*) \\
\leq &\left\langle \nabla f(\overline{x}_t^{(s)}), \overline{x}_t^{(s)} - x^*\right\rangle + h(\overline{x}_t^{(s)}) - h(x^*)
\end{aligned}
$$

We combine the two inequalities and obtain

$$\begin{aligned}
&A_{t-1}^{(s)}\left(F(\overline{x}_t^{(s)})-F(\overline{x}_{t-1}^{(s)})\right)+a^{(s)}\left(F(\overline{x}_t^{(s)})-F(x^*)\right)\\
\leq& a^{(s)}\left\langle\nabla f(\overline{x}_t^{(s)}),x_t^{(s)}-x^*\right\rangle+\left(a^{(s)}\right)^2\left\langle\nabla f(\overline{x}_t^{(s)}),u^{(s-1)}-\overline{x}_t^{(s)}\right\rangle\\
&-\frac{A_{t-1}}{2\beta}\left\|\nabla f(\overline{x}_t^{(s)})-\nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\\
&+A_{t-1}^{(s)}\left(h(\overline{x}_t^{(s)})-h(\overline{x}_{t-1}^{(s)})\right)+a^{(s)}\left(h(\overline{x}_t^{(s)})-h(x^*)\right)\\
=&a^{(s)}\left\langle g_t^{(s)},x_t^{(s)}-x^*\right\rangle+\left(a^{(s)}\right)^2\left\langle\nabla f(\overline{x}_t^{(s)}),u^{(s-1)}-\overline{x}_t^{(s)}\right\rangle\\
&+a^{(s)}\left\langle\nabla f(\overline{x}_t^{(s)})-g_t^{(s)},x_t^{(s)}-x^*\right\rangle-\frac{A_{t-1}^{(s)}}{2\beta}\left\|\nabla f(\overline{x}_t^{(s)})-\nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\\
&+A_{t-1}^{(s)}\left(h(\overline{x}_t^{(s)})-h(\overline{x}_{t-1}^{(s)})\right)+a^{(s)}\left(h(\overline{x}_t^{(s)})-h(x^*)\right).
\end{aligned}$$

Note that we can rearrange the terms

$$\begin{aligned}
&A_{t-1}^{(s)}\left(F(\overline{x}_t^{(s)})-F(\overline{x}_{t-1}^{(s)})\right)+a^{(s)}\left(F(\overline{x}_t^{(s)})-F(x^*)\right)\\
=&\left(A_t^{(s)}-\left(a^{(s)}\right)^2\right)\left(F(\overline{x}_t^{(s)})-F(x^*)\right)-A_{t-1}^{(s)}\left(F(\overline{x}_{t-1}^{(s)})-F(x^*)\right),\\
&A_{t-1}^{(s)}\left(h(\overline{x}_t^{(s)})-h(\overline{x}_{t-1}^{(s)})\right)+a^{(s)}\left(h(\overline{x}_t^{(s)})-h(x^*)\right)\\
=&\left(A_t^{(s)}-\left(a^{(s)}\right)^2\right)\left(h(\overline{x}_t^{(s)})-h(x^*)\right)-A_{t-1}^{(s)}\left(h(\overline{x}_{t-1}^{(s)})-h(x^*)\right).
\end{aligned}$$

Thus we obtain

$$\begin{aligned}
&\left(A_t^{(s)}-\left(a^{(s)}\right)^2\right)\left(F(\overline{x}_t^{(s)})-F(x^*)\right)-A_{t-1}^{(s)}\left(F(\overline{x}_{t-1}^{(s)})-F(x^*)\right)\\
\leq&a^{(s)}\left\langle g_t^{(s)},x_t^{(s)}-x^*\right\rangle+\left(a^{(s)}\right)^2\left\langle\nabla f(\overline{x}_t^{(s)}),u^{(s-1)}-\overline{x}_t^{(s)}\right\rangle\\
&+a^{(s)}\left\langle\nabla f(\overline{x}_t^{(s)})-g_t^{(s)},x_t^{(s)}-x^*\right\rangle-\frac{A_{t-1}^{(s)}}{2\beta}\left\|\nabla f(\overline{x}_t^{(s)})-\nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\\
&+\left(A_t^{(s)}-\left(a^{(s)}\right)^2\right)\left(h(\overline{x}_t^{(s)})-h(x^*)\right)-A_{t-1}^{(s)}\left(h(\overline{x}_{t-1}^{(s)})-h(x^*)\right).
\end{aligned}\tag{3}$$

Observe that for $t<T_s$

$$\begin{aligned}
\mathbb{E}\left[a^{(s)}\left\langle\nabla f(\overline{x}_t^{(s)})-g_t^{(s)},x_t^{(s)}-x^*\right\rangle\right]&=\mathbb{E}\left[\mathbb{E}_{i_t^{(s)}}\left[a^{(s)}\left\langle\nabla f(\overline{x}_t^{(s)})-g_t^{(s)},x_t^{(s)}-x^*\right\rangle\right]\right]\\
&=0.
\end{aligned}$$

and for $t=T_s$,we have $\nabla f(\overline{x}_t^{(s)})=g_t^{(s)}$ thus $\mathbb{E}\left[a^{(s)}\left\langle\nabla f(\overline{x}_t^{(s)})-g_t^{(s)},x_t^{(s)}-x^*\right\rangle\right]=0$. By taking expectations w.r.t. both sides of (3), we get

$$\begin{aligned}
&\mathbb{E}\left[\left(A_t^{(s)}-\left(a^{(s)}\right)^2\right)\left(F(\overline{x}_t^{(s)})-F(x^*)\right)-A_{t-1}^{(s)}\left(F(\overline{x}_{t-1}^{(s)})-F(x^*)\right)\right]\\
\leq&\mathbb{E}\left[a^{(s)}\left\langle g_t^{(s)},x_t^{(s)}-x^*\right\rangle+\left(a^{(s)}\right)^2\left\langle\nabla f(\overline{x}_t^{(s)}),u^{(s-1)}-\overline{x}_t^{(s)}\right\rangle\right]\\
&-\mathbb{E}\left[\frac{A_{t-1}^{(s)}}{2\beta}\left\|\nabla f(\overline{x}_t^{(s)})-\nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right]\\
&+\mathbb{E}\left[\left(A_t^{(s)}-\left(a^{(s)}\right)^2\right)\left(h(\overline{x}_t^{(s)})-h(x^*)\right)-A_{t-1}^{(s)}\left(h(\overline{x}_{t-1}^{(s)})-h(x^*)\right)\right].
\end{aligned}$$

□

To analyze the stochastic regret, we split the inner product as follows

$$\left\langle g_t^{(s)}, x_t^{(s)} - x^* \right\rangle = \left\langle g_t^{(s)}, z_t^{(s)} - x^* \right\rangle + \left\langle g_t^{(s)} - g_{t-1}^{(s)}, x_t^{(s)} - z_t^{(s)} \right\rangle + \left\langle g_{t-1}^{(s)}, x_t^{(s)} - z_t^{(s)} \right\rangle.$$

For each term we give a bound as stated in Lemma A.5.

Lemma A.5. *For any $s \geq 1$ all iterations $t \in [T_s]$, we have*

$$\begin{aligned} a^{(s)} \left\langle g_{t-1}^{(s)}, x_t^{(s)} - z_t^{(s)} \right\rangle \leq{}& \frac{\gamma_{t-1}^{(s)}}{2} \left\| z_{t-1}^{(s)} - z_t^{(s)} \right\|^2 - \frac{\gamma_{t-1}^{(s)}}{2} \left\| z_{t-1}^{(s)} - x_t^{(s)} \right\|^2 - \frac{\gamma_{t-1}^{(s)}}{2} \left\| x_t^{(s)} - z_t^{(s)} \right\|^2 \\ &+ a^{(s)} \left(h(z_t^{(s)}) - h(x_t^{(s)}) \right). \end{aligned}$$

$$\begin{aligned} a^{(s)} \left\langle g_t^{(s)}, z_t^{(s)} - x^* \right\rangle \leq{}& \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2} \left\| x_t^{(s)} - x^* \right\|^2 + \frac{\gamma_{t-1}^{(s)}}{2} \left\| z_{t-1}^{(s)} - x^* \right\|^2 - \frac{\gamma_t^{(s)}}{2} \left\| z_t^{(s)} - x^* \right\|^2 \\ &- \frac{\gamma_{t-1}^{(s)}}{2} \left\| z_t^{(s)} - z_{t-1}^{(s)} \right\|^2 - \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2} \left\| x_t^{(s)} - z_t^{(s)} \right\|^2 \\ &+ a^{(s)} \left(h(x^*) - h(z_t^{(s)}) \right). \end{aligned}$$

$$a^{(s)} \left\langle g_t^{(s)} - g_{t-1}^{(s)}, x_t^{(s)} - z_t^{(s)} \right\rangle \leq \frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}} \left\| g_t^{(s)} - g_{t-1}^{(s)} \right\|^2 + \frac{\gamma_t^{(s)}}{2} \left\| x_t^{(s)} - z_t^{(s)} \right\|^2.$$

Proof. Since $x_t^{(s)} = \arg\min_{x \in \mathcal{X}} \left\{ a^{(s)} \left\langle g_{t-1}^{(s)}, x \right\rangle + a^{(s)} h(x) + \frac{\gamma_{t-1}^{(s)}}{2} \left\| x - z_{t-1}^{(s)} \right\|^2 \right\}$, by the optimality condition of $x_t^{(s)}$, we have

$$\left\langle a^{(s)} g_{t-1}^{(s)} + a^{(s)} h'(x_t^{(s)}) + \gamma_{t-1}^{(s)} \left(x_t^{(s)} - z_{t-1}^{(s)} \right), x_t^{(s)} - z_t^{(s)} \right\rangle \leq 0,$$

where $h'(x_t^{(s)}) \in \partial h(x_t^{(s)})$ is a subgradient of h at $x_t^{(s)}$. We rearrange the above inequality and obtain

$$\begin{aligned} a^{(s)} \left\langle g_{t-1}^{(s)}, x_t^{(s)} - z_t^{(s)} \right\rangle &\leq \gamma_{t-1}^{(s)} \left\langle x_t^{(s)} - z_{t-1}^{(s)}, z_t^{(s)} - x_t^{(s)} \right\rangle + a^{(s)} \left\langle h'(x_t^{(s)}), z_t^{(s)} - x_t^{(s)} \right\rangle \\ &\overset{(a)}{\leq} \gamma_{t-1}^{(s)} \left\langle x_t^{(s)} - z_{t-1}^{(s)}, z_t^{(s)} - x_t^{(s)} \right\rangle + a^{(s)} \left(h(z_t^{(s)}) - h(x_t^{(s)}) \right) \\ &\overset{(b)}{=} \frac{\gamma_{t-1}^{(s)}}{2} \left\| z_{t-1}^{(s)} - z_t^{(s)} \right\|^2 - \frac{\gamma_{t-1}^{(s)}}{2} \left\| z_{t-1}^{(s)} - x_t^{(s)} \right\|^2 - \frac{\gamma_{t-1}^{(s)}}{2} \left\| x_t^{(s)} - z_t^{(s)} \right\|^2 \\ &\quad + a^{(s)} \left(h(z_t^{(s)}) - h(x_t^{(s)}) \right), \end{aligned}$$

where (a) follows from the convexity of h and the fact that $h'(x_t^{(s)}) \in \partial h(x_t^{(s)})$, and (b) is due to the identity $\langle a, b \rangle = \frac{1}{2} \left(\|a+b\|^2 - \|a\|^2 - \|b\|^2 \right)$.

Using the optimality condition of $z_t^{(s)}$, we have

$$\left\langle a^{(s)} g_t^{(s)} + a^{(s)} h'(z_t^{(s)}) + \gamma_{t-1}^{(s)} \left(z_t^{(s)} - z_{t-1}^{(s)} \right) + \left(\gamma_t^{(s)} - \gamma_{t-1}^{(s)} \right) \left(z_t^{(s)} - x_t^{(s)} \right), z_t^{(s)} - x^* \right\rangle \leq 0$$

where $h'(z_t^{(s)}) \in \partial h(z_t^{(s)})$ is a subgradient of h at $z_t^{(s)}$. We rearrange the above inequality and obtain

$$
\begin{aligned}
a^{(s)} \left\langle g_t^{(s)}, z_t^{(s)} - x^* \right\rangle &\leq \gamma_{t-1}^{(s)} \left\langle z_t^{(s)} - z_{t-1}^{(s)}, x^* - z_t^{(s)} \right\rangle + \left(\gamma_t^{(s)} - \gamma_{t-1}^{(s)}\right) \left\langle z_t^{(s)} - x_t^{(s)}, x^* - z_t^{(s)} \right\rangle \\
&\quad + a^{(s)} \left\langle h'(z_t^{(s)}), x^* - z_t^{(s)} \right\rangle \\
&\overset{(c)}{\leq} \gamma_{t-1}^{(s)} \left\langle z_t^{(s)} - z_{t-1}^{(s)}, x^* - z_t^{(s)} \right\rangle + \left(\gamma_t^{(s)} - \gamma_{t-1}^{(s)}\right) \left\langle z_t^{(s)} - x_t^{(s)}, x^* - z_t^{(s)} \right\rangle \\
&\quad + a^{(s)} \left(h(x^*) - h(z_t^{(s)})\right) \\
&\overset{(d)}{=} \frac{\gamma_{t-1}^{(s)}}{2} \left[\left\| z_{t-1}^{(s)} - x^* \right\|^2 - \left\| z_t^{(s)} - x^* \right\|^2 - \left\| z_t^{(s)} - z_{t-1}^{(s)} \right\|^2 \right] \\
&\quad + \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2} \left[\left\| x_t^{(s)} - x^* \right\|^2 - \left\| z_t^{(s)} - x^* \right\|^2 - \left\| x_t^{(s)} - z_t^{(s)} \right\|^2 \right] \\
&\quad + a^{(s)} \left(h(x^*) - h(z_t^{(s)})\right) \\
&= \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2} \left\| x_t^{(s)} - x^* \right\|^2 + \frac{\gamma_{t-1}^{(s)}}{2} \left\| z_{t-1}^{(s)} - x^* \right\|^2 - \frac{\gamma_t^{(s)}}{2} \left\| z_t^{(s)} - x^* \right\|^2 \\
&\quad - \frac{\gamma_{t-1}^{(s)}}{2} \left\| z_t^{(s)} - z_{t-1}^{(s)} \right\|^2 - \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2} \left\| x_t^{(s)} - z_t^{(s)} \right\|^2 \\
&\quad + a^{(s)} \left(h(x) - h(z_t^{(s)})\right),
\end{aligned}
$$

where (c) follows from the convexity of h and the fact that $h'(z_t^{(s)}) \in \partial h(z_t^{(s)})$, and (d) is due to the identity $\langle a, b \rangle = \frac{1}{2}\left(\|a+b\|^2 - \|a\|^2 - \|b\|^2\right)$.

For the third inequality, we have

$$
\begin{aligned}
a^{(s)} \left\langle g_t^{(s)} - g_{t-1}^{(s)}, x_t^{(s)} - z_t^{(s)} \right\rangle &\overset{(e)}{\leq} a^{(s)} \left\| g_t^{(s)} - g_{t-1}^{(s)} \right\| \left\| x_t^{(s)} - z_t^{(s)} \right\| \\
&\overset{(f)}{\leq} \frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}} \left\| g_t^{(s)} - g_{t-1}^{(s)} \right\|^2 + \frac{\gamma_t^{(s)}}{2} \left\| x_t^{(s)} - z_t^{(s)} \right\|^2.
\end{aligned}
$$

where (e) is by the Cauchy–Schwarz inequality, (f) is by Young's inequality. □

With above results, we obtain the descent lemma for one iteration. A key idea to remove $\left(a^{(s)}\right)^2$ from the coefficient of $\left(F(\overline{x}_t^{(s)}) - F(x^*)\right)$ is to split the term $\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}} \left\| g_t^{(s)} - g_{t-1}^{(s)} \right\|^2$ into $\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}} - \frac{\left(a^{(s)}\right)^2}{16\beta}\right) \left\| g_t^{(s)} - g_{t-1}^{(s)} \right\|^2 + \frac{\left(a^{(s)}\right)^2}{16\beta} \left\| g_t^{(s)} - g_{t-1}^{(s)} \right\|^2$ and apply the VR lemma for the second term.

Lemma A.6. *For all epochs $s \geq 1$ and all iterations $t \in [T_s]$, we have*

$$
\begin{aligned}
&\mathbb{E}\left[\left(A_t^{(s)} - \left(a^{(s)}\right)^2\right)\left(F(\overline{x}_t^{(s)}) - F(x^*)\right) - A_{t-1}^{(s)}\left(F(\overline{x}_{t-1}^{(s)}) - F(x^*)\right)\right] \\
\leq &\mathbb{E}\left[\frac{\gamma_{t-1}^{(s)}}{2}\left\|z_{t-1}^{(s)} - x^*\right\|^2 - \frac{\gamma_t^{(s)}}{2}\left\|z_t^{(s)} - x^*\right\|^2 + \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2\right] \\
&+ \mathbb{E}\left[\left(a^{(s)}\right)^2\left\langle \nabla f(\overline{x}_t^{(s)}), u^{(s-1)} - \overline{x}_t^{(s)}\right\rangle\right] \\
&+ \mathbb{E}\left[\frac{\left(a^{(s)}\right)^2}{2}\left(f(u^{(s-1)}) - f(\overline{x}_t^{(s)}) - \left\langle \nabla f(\overline{x}_t^{(s)}), u^{(s-1)} - \overline{x}_t^{(s)}\right\rangle\right)\right] \\
&+ \mathbb{E}\left[\frac{\left(a^{(s)}\right)^2}{2}\left(f(u^{(s-1)}) - f(\overline{x}_{t-1}^{(s)}) - \left\langle \nabla f(\overline{x}_{t-1}^{(s)}), u^{(s-1)} - \overline{x}_{t-1}^{(s)}\right\rangle\right)\right] \\
&+ \mathbb{E}\left[\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}} - \frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2 + \left(\frac{\left(a^{(s)}\right)^2}{8\beta} - \frac{A_{t-1}^{(s)}}{2\beta}\right)\left\|\nabla f(\overline{x}_t^{(s)}) - \nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right] \\
&+ \mathbb{E}\left[\left(a^{(s)}\right)^2\left(h(u^{(s-1)}) - h(\overline{x}_t^{(s)})\right)\right].
\end{aligned}
$$

Proof. By Lemma A.5, we can bound $a^{(s)}\left\langle g_t^{(s)}, x_t^{(s)} - x^*\right\rangle$ as follows

$$
\begin{aligned}
a^{(s)}\left\langle g_t^{(s)}, x_t^{(s)} - x^*\right\rangle &\leq \frac{\gamma_{t-1}^{(s)}}{2}\left\|z_{t-1}^{(s)} - x^*\right\|^2 - \frac{\gamma_t^{(s)}}{2}\left\|z_t^{(s)} - x^*\right\|^2 + \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2 - \frac{\gamma_{t-1}^{(s)}}{2}\left\|z_{t-1}^{(s)} - x_t^{(s)}\right\|^2 \\
&\quad + a^{(s)}\left(h(x^*) - h(x_t^{(s)})\right) + \frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2. \\
&\leq \frac{\gamma_{t-1}^{(s)}}{2}\left\|z_{t-1}^{(s)} - x^*\right\|^2 - \frac{\gamma_t^{(s)}}{2}\left\|z_t^{(s)} - x^*\right\|^2 + \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2 \\
&\quad + a^{(s)}\left(h(x^*) - h(x_t^{(s)})\right) + \frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2.
\end{aligned}
$$

Combining the above result with Lemma A.4, we know

$$
\begin{aligned}
&\mathbb{E}\left[\left(A_t^{(s)} - \left(a^{(s)}\right)^2\right)\left(F(\overline{x}_t^{(s)}) - F(x^*)\right) - A_{t-1}^{(s)}\left(F(\overline{x}_{t-1}^{(s)}) - F(x^*)\right)\right] \\
\leq &\mathbb{E}\left[\frac{\gamma_{t-1}^{(s)}}{2}\left\|z_{t-1}^{(s)} - x^*\right\|^2 - \frac{\gamma_t^{(s)}}{2}\left\|z_t^{(s)} - x^*\right\|^2 + \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2\right] \\
&+ \mathbb{E}\left[\left(a^{(s)}\right)^2\left\langle \nabla f(\overline{x}_t^{(s)}), u^{(s-1)} - \overline{x}_t^{(s)}\right\rangle\right] \\
&+ \mathbb{E}\left[\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2 - \frac{A_{t-1}^{(s)}}{2\beta}\left\|\nabla f(\overline{x}_t^{(s)}) - \nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right] \\
&+ \mathbb{E}\left[\left(A_t^{(s)} - \left(a^{(s)}\right)^2\right)\left(h(\overline{x}_t^{(s)}) - h(x^*)\right) - A_{t-1}^{(s)}\left(h(\overline{x}_{t-1}^{(s)}) - h(x^*)\right) + a^{(s)}\left(h(x^*) - h(x_t^{(s)})\right)\right]. \qquad (4)
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\left(A_t^{(s)} - \left(a^{(s)}\right)^2\right)\left(h(\overline{x}_t^{(s)}) - h(x^*)\right) - A_{t-1}^{(s)}\left(h(\overline{x}_{t-1}^{(s)}) - h(x^*)\right) + a^{(s)}\left(h(x^*) - h(x_t^{(s)})\right)\\
=&\left(A_t^{(s)} - \left(a^{(s)}\right)^2\right)\left(h(\overline{x}_t^{(s)}) - h(x^*)\right) + \left(a^{(s)}\right)^2\left(h(u^{(s-1)}) - h(x^*)\right)\\
&- \left(a^{(s)}\right)^2\left(h(u^{(s-1)}) - h(x^*)\right) - A_{t-1}^{(s)}\left(h(\overline{x}_{t-1}^{(s)}) - h(x^*)\right) - a^{(s)}\left(h(x_t^{(s)}) - h(x^*)\right)\\
\overset{(a)}{\leq}&\left(A_t^{(s)} - \left(a^{(s)}\right)^2\right)\left(h(\overline{x}_t^{(s)}) - h(x^*)\right) + \left(a^{(s)}\right)^2\left(h(u^{(s-1)}) - h(x^*)\right)\\
&- A_t^{(s)}(h(\overline{x}_t^{(s)}) - h(x^*))\\
=&\left(a^{(s)}\right)^2\left(h(u^{(s-1)}) - h(\overline{x}_t^{(s)})\right),
\end{aligned}
\tag{5}
$$

where (a) is by the convexity of h and $A_t^{(s)} = A_{t-1}^{(s)} + a^{(s)} + \left(a^{(s)}\right)^2$. Plugging in (5) into (4), we know

$$
\begin{aligned}
&\mathbb{E}\left[\left(A_t^{(s)} - \left(a^{(s)}\right)^2\right)\left(F(\overline{x}_t^{(s)}) - F(x^*)\right) - A_{t-1}^{(s)}\left(F(\overline{x}_{t-1}^{(s)}) - F(x^*)\right)\right]\\
\leq&\mathbb{E}\left[\frac{\gamma_{t-1}^{(s)}}{2}\left\|z_{t-1}^{(s)} - x^*\right\|^2 - \frac{\gamma_t^{(s)}}{2}\left\|z_t^{(s)} - x^*\right\|^2 + \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2\right]\\
&+\mathbb{E}\left[\left(a^{(s)}\right)^2\left\langle\nabla f(\overline{x}_t^{(s)}), u^{(s-1)} - \overline{x}_t^{(s)}\right\rangle\right]\\
&+\mathbb{E}\left[\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2 - \frac{A_{t-1}^{(s)}}{2\beta}\left\|\nabla f(\overline{x}_t^{(s)}) - \nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right]\\
&+\mathbb{E}\left[\left(a^{(s)}\right)^2\left(h(u^{(s-1)}) - h(\overline{x}_t^{(s)})\right)\right].
\end{aligned}
$$

Now for $\mathbb{E}\left[\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2\right]$, when $1 < t < T_s$, we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2\right] \leq&\ \mathbb{E}\left[4\left\|\nabla f(\overline{x}_t^{(s)}) - g_t^{(s)}\right\|^2 + 4\left\|\nabla f(\overline{x}_{t-1}^{(s)}) - g_{t-1}^{(s)}\right\|^2\right]\\
&+\mathbb{E}\left[2\left\|\nabla f(\overline{x}_t^{(s)}) - \nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right]\\
\overset{(b)}{\leq}&\ \mathbb{E}\left[8\beta\left(f(u^{(s-1)}) - f(\overline{x}_t^{(s)}) - \left\langle\nabla f(\overline{x}_t^{(s)}), u^{(s-1)} - \overline{x}_t^{(s)}\right\rangle\right)\right]\\
&+\mathbb{E}\left[8\beta\left(f(u^{(s-1)}) - f(\overline{x}_{t-1}^{(s)}) - \left\langle\nabla f(\overline{x}_{t-1}^{(s)}), u^{(s-1)} - \overline{x}_{t-1}^{(s)}\right\rangle\right)\right]\\
&+\mathbb{E}\left[2\left\|\nabla f(\overline{x}_t^{(s)}) - \nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right],
\end{aligned}
\tag{6}
$$

where (b) is by Lemma A.2 for all $1 < t < T_s$. When $t = 1$, note that both $\left\|\nabla f(\overline{x}_{t-1}^{(s)}) - g_{t-1}^{(s)}\right\|^2$ and $f(u^{(s-1)}) - f(\overline{x}_{t-1}^{(s)}) - \left\langle\nabla f(\overline{x}_{t-1}^{(s)}), u^{(s-1)} - \overline{x}_{t-1}^{(s)}\right\rangle$ are zero by our definition $\overline{x}_0^{(s)} = u^{(s-1)}$ and $\nabla f(\overline{x}_0^{(s)}) = \nabla f(\overline{x}_{T_{s-1}}^{(s-1)}) = g_{T_{s-1}}^{(s-1)} = g_0^{(s)}$, which means the above inequality is still true. When $t = T_s$, note that $\left\|\nabla f(\overline{x}_t^{(s)}) - g_t^{(s)}\right\|^2 = 0$ and $f(u^{(s-1)}) - f(\overline{x}_t^{(s)}) - \left\langle\nabla f(\overline{x}_t^{(s)}), u^{(s-1)} - \overline{x}_t^{(s)}\right\rangle$ is always non-negative due to the convexity of f. So the above inequality also holds in this case. Now we conlclude the above inequality is right for $t \in [T_s]$.

Splitting $\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2$ into $\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}} - \frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2 + \frac{\left(a^{(s)}\right)^2}{16\beta}\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2$ and applying (6) to

$\frac{\left(a^{(s)}\right)^2}{16\beta}\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left(A_t^{(s)}-\left(a^{(s)}\right)^2\right)\left(F(\overline{x}_t^{(s)})-F(x^*)\right)-A_{t-1}^{(s)}\left(F(\overline{x}_{t-1}^{(s)})-F(x^*)\right)\right]\\
\leq&\mathbb{E}\left[\frac{\gamma_{t-1}^{(s)}}{2}\left\|z_{t-1}^{(s)}-x^*\right\|^2-\frac{\gamma_t^{(s)}}{2}\left\|z_t^{(s)}-x^*\right\|^2+\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2\right]\\
&+\mathbb{E}\left[\left(a^{(s)}\right)^2\left\langle\nabla f(\overline{x}_t^{(s)}),u^{(s-1)}-\overline{x}_t^{(s)}\right\rangle\right]\\
&+\mathbb{E}\left[\frac{\left(a^{(s)}\right)^2}{2}\left(f(u^{(s-1)})-f(\overline{x}_t^{(s)})-\left\langle\nabla f(\overline{x}_t^{(s)}),u^{(s-1)}-\overline{x}_t^{(s)}\right\rangle\right)\right]\\
&+\mathbb{E}\left[\frac{\left(a^{(s)}\right)^2}{2}\left(f(u^{(s-1)})-f(\overline{x}_{t-1}^{(s)})-\left\langle\nabla f(\overline{x}_{t-1}^{(s)}),u^{(s-1)}-\overline{x}_{t-1}^{(s)}\right\rangle\right)\right]\\
&+\mathbb{E}\left[\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}-\frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2+\left(\frac{\left(a^{(s)}\right)^2}{8\beta}-\frac{A_{t-1}^{(s)}}{2\beta}\right)\left\|\nabla f(\overline{x}_t^{(s)})-\nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right]\\
&+\mathbb{E}\left[\left(a^{(s)}\right)^2\left(h(u^{(s-1)})-h(\overline{x}_t^{(s)})\right)\right].
\end{aligned}
$$

□

A.2. Single epoch progress and final output

Even though Lemma A.6 looks somewhat more convoluted, when we sum up over all iterations in one epoch, many terms are canceled out nicely and we obtain the following lemma that states the progress of the function value in one epoch. The trick is to set the value for each term at the end of one epoch equal to its value in the next one, with an exception for $A_{T_{s-1}}^{(s-1)}$. Due to the accumulation of the term $\left(F(u^{(s-1)})-F(x^*)\right)$ throughout the epoch, we will set $A_0^{(s)}=A_{T_{s-1}}^{(s-1)}-T_s\left(a^{(s)}\right)^2$.

Lemma A.7. *For all epochs $s\geq 1$, if*

$$
\left(a^{(s)}\right)^2\leq 4A_{t-1}^{(s)},\forall t\in[T_s].
$$

We have

$$
\begin{aligned}
&\mathbb{E}\left[A_{T_s}^{(s)}\left(F(u^{(s)})-F(x^*)\right)-A_{T_{s-1}}^{(s-1)}\left(F(u^{(s-1)})-F(x^*)\right)\right]\\
\leq&\mathbb{E}\left[\frac{\gamma_0^{(s)}}{2}\left\|z_0^{(s)}-x^*\right\|^2-\frac{\gamma_0^{(s+1)}}{2}\left\|z_0^{(s+1)}-x^*\right\|^2+\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2\right]\\
&+\mathbb{E}\left[\sum_{t=1}^{T_s}\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}-\frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2\right].
\end{aligned}
$$

Proof. Using Lemma A.6, we know

$$
\begin{aligned}
&\sum_{t=1}^{T_s}\mathbb{E}\left[\left(A_t^{(s)}-\left(a^{(s)}\right)^2\right)\left(F(\overline{x}_t^{(s)})-F(x^*)\right)-A_{t-1}^{(s)}\left(F(\overline{x}_{t-1}^{(s)})-F(x^*)\right)\right]\\
\le&\sum_{t=1}^{T_s}\mathbb{E}\left[\frac{\gamma_{t-1}^{(s)}}{2}\left\|z_{t-1}^{(s)}-x^*\right\|^2-\frac{\gamma_t^{(s)}}{2}\left\|z_t^{(s)}-x^*\right\|^2+\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2\right]\\
&+\sum_{t=1}^{T_s}\mathbb{E}\left[\left(a^{(s)}\right)^2\left\langle\nabla f(\overline{x}_t^{(s)}),u^{(s-1)}-\overline{x}_t^{(s)}\right\rangle+\left(a^{(s)}\right)^2\left(h(u^{(s-1)})-h(\overline{x}_t^{(s)})\right)\right]\\
&+\sum_{t=1}^{T_s}\mathbb{E}\left[\frac{\left(a^{(s)}\right)^2}{2}\left(f(u^{(s-1)})-f(\overline{x}_t^{(s)})-\left\langle\nabla f(\overline{x}_t^{(s)}),u^{(s-1)}-\overline{x}_t^{(s)}\right\rangle\right)\right]\\
&+\sum_{t=1}^{T_s}\mathbb{E}\left[\frac{\left(a^{(s)}\right)^2}{2}\left(f(u^{(s-1)})-f(\overline{x}_{t-1}^{(s)})-\left\langle\nabla f(\overline{x}_{t-1}^{(s)}),u^{(s-1)}-\overline{x}_{t-1}^{(s)}\right\rangle\right)\right]\\
&+\sum_{t=1}^{T_s}\mathbb{E}\left[\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}-\frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2+\left(\frac{\left(a^{(s)}\right)^2}{8\beta}-\frac{A_{t-1}^{(s)}}{2\beta}\right)\left\|\nabla f(\overline{x}_t^{(s)})-\nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right]\\
\overset{(a)}{=}&\mathbb{E}\left[\frac{\gamma_0^{(s)}}{2}\left\|z_0^{(s)}-x^*\right\|^2-\frac{\gamma_0^{(s+1)}}{2}\left\|z_0^{(s+1)}-x^*\right\|^2+\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2\right]\\
&+\mathbb{E}\left[\sum_{t=1}^{T_s-1}\left(a^{(s)}\right)^2\left(f(u^{(s-1)})-f(\overline{x}_t^{(s)})\right)+\left(a^{(s)}\right)^2\left(h(u^{(s-1)})-h(\overline{x}_t^{(s)})\right)\right]\\
&+\mathbb{E}\left[\frac{\left(a^{(s)}\right)^2}{2}\left(f(u^{(s-1)})-f(\overline{x}_{T_s}^{(s)})+\left\langle\nabla f(\overline{x}_{T_s}^{(s)}),u^{(s-1)}-\overline{x}_{T_s}^{(s)}\right\rangle\right)\right]\\
&+\mathbb{E}\left[\sum_{t=1}^{T_s}\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}-\frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2+\left(\frac{\left(a^{(s)}\right)^2}{8\beta}-\frac{A_{t-1}^{(s)}}{2\beta}\right)\left\|\nabla f(\overline{x}_t^{(s)})-\nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right]\\
\overset{(b)}{\le}&\mathbb{E}\left[\frac{\gamma_0^{(s)}}{2}\left\|z_0^{(s)}-x^*\right\|^2-\frac{\gamma_0^{(s+1)}}{2}\left\|z_0^{(s+1)}-x^*\right\|^2+\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2\right]\\
&+\mathbb{E}\left[\sum_{t=1}^{T_s}\left(a^{(s)}\right)^2\left(f(u^{(s-1)})-f(\overline{x}_t^{(s)})+\left(a^{(s)}\right)^2\left(h(u^{(s-1)})-h(\overline{x}_t^{(s)})\right)\right)\right]\\
&+\mathbb{E}\left[\sum_{t=1}^{T_s}\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}-\frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2+\left(\frac{\left(a^{(s)}\right)^2}{8\beta}-\frac{A_{t-1}^{(s)}}{2\beta}\right)\left\|\nabla f(\overline{x}_t^{(s)})-\nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right]\\
\overset{(c)}{=}&\mathbb{E}\left[\frac{\gamma_0^{(s)}}{2}\left\|z_0^{(s)}-x^*\right\|^2-\frac{\gamma_0^{(s+1)}}{2}\left\|z_0^{(s+1)}-x^*\right\|^2+\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2\right]\\
&+\mathbb{E}\left[\sum_{t=1}^{T_s}\left(a^{(s)}\right)^2\left(F(u^{(s-1)})-F(\overline{x}_t^{(s)})\right)\right]\\
&+\mathbb{E}\left[\sum_{t=1}^{T_s}\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}-\frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2+\left(\frac{\left(a^{(s)}\right)^2}{8\beta}-\frac{A_{t-1}^{(s)}}{2\beta}\right)\left\|\nabla f(\overline{x}_t^{(s)})-\nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right].
\end{aligned}
\tag{7}
$$

where (a) is due to $z_0^{(s+1)}=z_{T_s}^{(s)}$, $\gamma_0^{(s+1)}=\gamma_{T_s}^{(s)}$, $\overline{x}_0^{(s)}=u^{(s-1)}$, (b) is by the convexity of f

$$
\left\langle\nabla f(\overline{x}_{T_s}^{(s)}),u^{(s-1)}-\overline{x}_{T_s}^{(s)}\right\rangle\le f(u^{(s-1)})-f(\overline{x}_{T_s}^{(s)}),
$$

(c) is by the definition of $F = f + h$. By adding $\sum_{t=1}^{T_s} \left(a^{(s)}\right)^2 \left(F(\overline{x}_t^{(s)}) - F(x^*)\right)$ to both sides of 7. we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T_s} A_t^{(s)} \left(F(\overline{x}_t^{(s)}) - F(x^*)\right) - A_{t-1}^{(s)} \left(F(\overline{x}_{t-1}^{(s)}) - F(x^*)\right)\right] \\
\leq &\mathbb{E}\left[\frac{\gamma_0^{(s)}}{2}\left\|z_0^{(s)} - x^*\right\|^2 - \frac{\gamma_0^{(s+1)}}{2}\left\|z_0^{(s+1)} - x^*\right\|^2 + \sum_{t=1}^{T_s} \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2\right] \\
&+ \mathbb{E}\left[T_s \left(a^{(s)}\right)^2 \left(F(u^{(s-1)}) - F(x^*)\right)\right] \\
&+ \mathbb{E}\left[\sum_{t=1}^{T_s}\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}} - \frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2 + \left(\frac{\left(a^{(s)}\right)^2}{8\beta} - \frac{A_{t-1}^{(s)}}{2\beta}\right)\left\|\nabla f(\overline{x}_t^{(s)}) - \nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right].
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T_s} A_t^{(s)} \left(F(\overline{x}_t^{(s)}) - F(x^*)\right) - A_{t-1}^{(s)} \left(F(\overline{x}_{t-1}^{(s)}) - F(x^*)\right)\right] \\
= &\mathbb{E}\left[A_{T_s}^{(s)} \left(F(\overline{x}_{T_s}^{(s)}) - F(x^*)\right) - A_0^{(s)} \left(F(\overline{x}_0^{(s)}) - F(x^*)\right)\right] \\
\overset{(d)}{=} &\mathbb{E}\left[A_{T_s}^{(s)} \left(F(u^{(s)}) - F(x^*)\right) - A_0^{(s)} \left(F(u^{(s-1)}) - F(x^*)\right)\right],
\end{aligned}
$$

where (d) is due to the definition $u^{(s)} = \overline{x}_{T_s}^{(s)}$ and $\overline{x}_0^{(s)} = u^{(s-1)}$. Finally we have

$$
\begin{aligned}
&\mathbb{E}\left[A_{T_s}^{(s)} \left(F(u^{(s)}) - F(x^*)\right) - \left(A_0^{(s)} + T_s \left(a^{(s)}\right)^2\right) \left(F(u^{(s-1)}) - F(x^*)\right)\right] \\
\leq &\mathbb{E}\left[\frac{\gamma_0^{(s)}}{2}\left\|z_0^{(s)} - x^*\right\|^2 - \frac{\gamma_0^{(s+1)}}{2}\left\|z_0^{(s+1)} - x^*\right\|^2 + \sum_{t=1}^{T_s} \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2\right] \\
&+ \mathbb{E}\left[\sum_{t=1}^{T_s}\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}} - \frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2 + \left(\frac{\left(a^{(s)}\right)^2}{8\beta} - \frac{A_{t-1}^{(s)}}{2\beta}\right)\left\|\nabla f(\overline{x}_t^{(s)}) - \nabla f(\overline{x}_{t-1}^{(s)})\right\|^2\right].
\end{aligned}
$$

Combining the fact $A_0^{(s)} = A_{T_{s-1}}^{(s-1)} - T_s \left(a^{(s)}\right)^2$ and our condition $\left(a^{(s)}\right)^2 \leq 4A_{t-1}^{(s)}$, we get the desired result. □

The telescoping sum on the LSH allows us to obtain the guarantee for the final output $u^{(S)}$.

Lemma A.8. *For all $S \geq 1$, assume we have*

$$
\left(a^{(s)}\right)^2 \leq 4A_{t-1}^{(s)}, \forall t \in [T_s], \forall s \in [S].
$$

Then

$$
\begin{aligned}
&\mathbb{E}\left[A_{T_S}^{(S)} \left(F(u^{(S)}) - F(x^*)\right)\right] \\
\leq &A_{T_0}^{(0)} \left(F(u^{(0)}) - F(x^*)\right) + \frac{\gamma}{2}\left\|u^{(0)} - x^*\right\|^2 \\
&+ \mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s} \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2 + \left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}} - \frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2\right].
\end{aligned}
$$

Proof. Note that our assumptions satisfy the requirements for Lemma A.7, by Applying Lemma A.7 and make the

telescoping sum from $s=1$ to S, we obtain

$$\begin{aligned}
&\mathbb{E}\left[A_{T_S}^{(S)}\left(F(u^{(S)})-F(x^*)\right)\right]\\
\leq&A_{T_0}^{(0)}\left(F(u^{(0)})-F(x^*)\right)+\mathbb{E}\left[\frac{\gamma_0^{(s)}}{2}\left\|z_0^{(s)}-x^*\right\|^2-\frac{\gamma_0^{(S+1)}}{2}\left\|z_0^{(S+1)}-x^*\right\|^2\right]\\
&+\mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}-\frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2\right]\\
\leq&A_{T_0}^{(0)}\left(F(u^{(0)})-F(x^*)\right)+\frac{\gamma_0^{(1)}}{2}\left\|z_0^{(s)}-x^*\right\|^2\\
&+\mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}-\frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2\right]\\
=&A_{T_0}^{(0)}\left(F(u^{(0)})-F(x^*)\right)+\frac{\gamma}{2}\left\|u^{(0)}-x^*\right\|^2\\
&+\mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}-\frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2\right],
\end{aligned}$$

where we use $\gamma_0^{(1)}=\gamma$ and $z_0^{(1)}=u^{(0)}$. □

A.3. Bound for the residual term

We turn to bound the term

$$\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}-\frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2$$

This follows the standard analysis used to bound the residual term in adaptive methods. We first admit Lemma A.10 to give the final bound for this term.

Lemma A.9. *If $\mathcal{X}$ is a compact convex set with diameter D, we have*

$$\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}-\frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2\leq\frac{8\beta\left(D^4+2\eta^4\right)}{\eta^2}$$

Proof. It follows that

$$\begin{aligned}
&\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}-\frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2\\
\overset{(a)}{\leq}&\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}D^2+\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}-\frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2\\
\overset{(b)}{=}&\frac{\gamma_{T_s}^{(s)}-\gamma_0^{(1)}}{2}D^2+\sum_{s=1}^{S}\sum_{t=1}^{T_s}\left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}}-\frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2\\
=&\frac{\gamma_{T_s}^{(s)}-\gamma_0^{(1)}}{2}D^2-\frac{D^4}{16\beta\left(D^4+2\eta^4\right)}\sum_{s=1}^{S}\sum_{t=1}^{T_s}\left(a^{(s)}\right)^2\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2\\
&+\sum_{s=1}^{S}\sum_{t=1}^{T_s}\left(\frac{1}{2\gamma_t^{(s)}}-\frac{\eta^4}{8\beta\left(D^4+2\eta^4\right)}\right)\left(a^{(s)}\right)^2\left\|g_t^{(s)}-g_{t-1}^{(s)}\right\|^2\\
\overset{(c)}{\leq}&\frac{8\beta\left(D^4+2\eta^4\right)}{\eta^2}
\end{aligned}$$

where (a) is by $\gamma_t^{(s)} \geq \gamma_{t-1}^{(s)}$ and $\left\|x_t^{(s)} - x^*\right\| \leq D$, (b) is by noticing $\gamma_0^{(s+1)} = \gamma_{T_s}^{(s)}$, (c) is by Lemma A.10. □

Lemma A.10. *Under our update rule of $\gamma_t^{(s)}$, we have*

$$\frac{D^2}{2}\left(\gamma_{T_S}^{(S)} - \gamma_0^{(1)}\right) - \frac{D^4}{16\beta\left(D^4+2\eta^4\right)}\sum_{s=1}^{S}\sum_{t=1}^{T_s}\left(a^{(s)}\right)^2\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2 \leq \frac{4\beta\left(D^4+2\eta^4\right)}{\eta^2}$$

$$\sum_{s=1}^{S}\sum_{t=1}^{T_s}\left(\frac{1}{2\gamma_t^{(s)}} - \frac{\eta^4}{8\beta\left(D^4+2\eta^4\right)}\right)\left(a^{(s)}\right)^2\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2 \leq \frac{4\beta\left(D^4+2\eta^4\right)}{\eta^2}$$

Proof. For simplicity, $g_0^{(1)}, g_1^{(1)}, \ldots, g_{T_1}^{(1)} = g_0^{(2)}, g_1^{(2)}, \ldots, g_{T_2}^{(2)}, \ldots$ as $(g_k)_{k\geq 0}$ and $\gamma_0^{(1)}, \gamma_1^{(1)}, \ldots, \gamma_{T_1}^{(1)} = \gamma_0^{(2)}, \gamma_1^{(2)}, \ldots, \gamma_{T_2}^{(2)}, \ldots$ as $(\gamma_k)_{k\geq 0}$. For $k \geq 1$, assume that $g_t^{(s)}$ is the element that correspond to g_k, and let $a_k = a^{(s)}$. Then we can write $\gamma_k = \frac{1}{\eta}\sqrt{\eta^2\gamma_{k-1}^2 + a_k^2\left\|g_k - g_{k-1}\right\|^2}$. By writing $\eta^2\gamma_k^2 = \eta^2\gamma_{k-1}^2 + a_k^2\left\|g_k - g_{k-1}\right\|^2$ we obtain $\eta^2\gamma_k^2 = \eta^2\gamma_0^2 + \sum_{t=1}^k a_t^2\left\|g_t - g_{t-1}\right\|^2$ and hence $\gamma_k = \frac{1}{\eta}\sqrt{\eta^2\gamma_0 + \sum_{t=1}^k a_t^2\left\|g_t - g_{t-1}\right\|^2}$.

For 1). Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ we have $\gamma_k \leq \gamma_0 + \frac{1}{\eta}\sqrt{\sum_{t=1}^k a_t^2\left\|g_t - g_{t-1}\right\|^2}$. Therefore

$$\begin{aligned}
&\frac{D^2}{2}\left(\gamma_k - \gamma_0\right) - \frac{D^4}{16\beta\left(D^4+2\eta^4\right)}\sum_{t=1}^{k} a_t^2\left\|g_t - g_{t-1}\right\|^2 \\
\leq &\frac{D^2}{2\eta}\sqrt{\sum_{t=1}^{k} a_t^2\left\|g_t - g_{t-1}\right\|^2} - \frac{D^4}{16\beta\left(D^4+2\eta^4\right)}\sum_{t=1}^{k} a_t^2\left\|g_t - g_{t-1}\right\|^2 \\
\overset{(a)}{\leq} &\frac{4\beta\left(D^4+2\eta^4\right)}{\eta^2}
\end{aligned}$$

where for (a) we use $ax - bx^2 \leq \frac{a^2}{4b}$.

For 2). Let τ be the last index such that $\gamma_\tau \leq \frac{4\beta\left(D^4+2\eta^4\right)}{\eta^4}$ or $\tau = -1$ if $\gamma_0 > \frac{4\beta\left(D^4+2\eta^4\right)}{\eta^4}$. If $\tau \leq 0$ we have $\sum_{t=1}^k\left(\frac{1}{2\gamma_t} - \frac{\eta^4}{8\beta(D^4+2\eta^4)}\right)a_t^2\left\|g_t - g_{t-1}\right\|^2 \leq 0$ for all k. Assume $\tau > 0$

$$\begin{aligned}
&\sum_{t=1}^{k}\left(\frac{1}{2\gamma_t} - \frac{\eta^4}{8\beta\left(D^4+2\eta^4\right)}\right)a_t^2\left\|g_t - g_{t-1}\right\|^2 \\
\leq &\sum_{t=1}^{\tau}\frac{1}{2\gamma_t}a_t^2\left\|g_t - g_{t-1}\right\|^2 \\
= &\eta^2\sum_{t=1}^{\tau}\frac{\gamma_t^2 - \gamma_{t-1}^2}{2\gamma_t} \\
= &\eta^2\sum_{t=1}^{\tau}\frac{\left(\gamma_t - \gamma_{t-1}\right)\left(\gamma_t + \gamma_{t-1}\right)}{2\gamma_t} \\
\overset{(b)}{\leq} &\eta^2\sum_{t=1}^{\tau}\left(\gamma_t - \gamma_{t-1}\right) \\
\leq &\eta^2\gamma_\tau \\
\overset{(c)}{\leq} &\frac{4\beta\left(D^4+2\eta^4\right)}{\eta^2}
\end{aligned}$$

where (b) is due to $\gamma_{t-1} \leq \gamma_t$, (c) is by the definition of τ. □

Finally we give an explicit choice for the parameters to satisfy all conditions and give the final necessary bound.

A.4. Parameter choice and bound

The following lemma states the bound for the coefficients.

Lemma A.11. *Under the choice of parameters in Theorem A.1, $\forall s \geq 1$, we have*

$$\left(a^{(s)}\right)^2 < 4A_0^{(s)}$$

and

$$A_{T_s}^{(s)} \geq \begin{cases} n(4n)^{-0.5^s} & 1 \leq s \leq s_0 \\ \frac{n}{4c}(s - s_0)^2 & s_0 < s \end{cases}$$

Proof. As a reminder, we choose the parameters as follows, where $c = \frac{3}{2}$ and $s = s_0 = \lceil \log_2 \log_2 4n \rceil$

$$\begin{aligned} a^{(s)} &= \begin{cases} (4n)^{-0.5^s} & 1 \leq s \leq s_0 \\ \frac{s - s_0 - 1 + c}{2c} & s_0 < s \end{cases}, \\ T_s &= n, \\ A_{T_0}^{(0)} &= \frac{5}{4}. \end{aligned}$$

The idea in this choice is that we divide the time into two phases in which the convergence behaves differently. In the first phase, $A_{T_s}^{(s)}$ quickly gets to $\Omega(n)$ and we can set the coefficients for the checkpoint relatively small. In the second phase, to achieve the optimal $\sqrt{\frac{n\beta}{\epsilon}}$ rate, $A_{T_s}^{(s)} = \Omega(n^2)$. In this phase, we need to be more conservative and set the coefficients for the checkpoint large. We analyze the two phases separately.

First we show by induction that for $1 \leq s \leq s_0$,

$$A_0^{(s)} = 1 + n \sum_{k=0}^{s-2} (4n)^{-0.5^k}, \tag{8}$$

$$A_{T_s}^{(s)} = 1 + n \sum_{k=0}^{s} (4n)^{-0.5^k}. \tag{9}$$

Indeed, we have

$$\begin{aligned} A_0^{(1)} &= A_{T_0}^{(0)} - T_1 \left(a^{(1)}\right)^2 \overset{(a)}{=} \frac{5}{4} - n(4n)^{-1} = \frac{5}{4} - \frac{1}{4} = 1, \\ A_{T_1}^{(1)} &= A_0^{(1)} + T_1 \left(a^{(1)} + \left(a^{(1)}\right)^2\right) \overset{(b)}{=} 1 + n\left((4n)^{-0.5} + (4n)^{-1}\right), \end{aligned}$$

where (a) and (b) are both by plugging in $a^{(1)} = (4n)^{-0.5}$ and $T_1 = n$. Supposed that 8 and 9 hold for all $k \leq s < s_0$. For

$k = s+1 \le s_0$, we have

$$\begin{aligned}
A_0^{(s+1)} &= A_{T_s}^{(s)} - T_{s+1}\left(a^{(s+1)}\right)^2 \\
&\overset{(c)}{=} \left(1 + n\sum_{k=0}^{s}(4n)^{-0.5^k}\right) - n(4n)^{-0.5^s} \\
&= 1 + n\sum_{k=0}^{s-1}(4n)^{-0.5^k}, \\
A_{T_{s+1}}^{(s+1)} &= A_0^{(s+1)} + T_{s+1}\left(a^{(s+1)} + \left(a^{(s+1)}\right)^2\right) \\
&\overset{(d)}{=} \left(1 + n\sum_{k=0}^{s-1}(4n)^{-0.5^k}\right) + n\left((4n)^{-0.5^{s+1}} + (4n)^{-0.5^s}\right) \\
&= 1 + n\sum_{k=0}^{s+1}(4n)^{-0.5^k},
\end{aligned}$$

where (c) is by plugging $a^{(s+1)} = (4n)^{-0.5^{s+1}}$, $T_{s+1} = n$ and the assumption on $A_{T_s}^{(s)}$, (d) is by plugging $a^{(s+1)} = (4n)^{-0.5^{s+1}}$ and $T_{s+1} = n$. Now the induction is completed. From this we can see that $A_0^{(s)} \ge 1 > \frac{\left(a^{(s)}\right)^2}{4}$ and $A_{T_s}^{(s)} > n(4n)^{-0.5^s}$.

Next, for $s > s_0$, we show by induction that

$$A_0^{(s)} > \frac{n}{2} + \frac{n}{4c}(s - s_0 - 2 + 2c)(s - s_0 - 1) - \frac{n}{4c^2}(s - s_0 - 1 + c)^2, \tag{10}$$

$$A_{T_s}^{(s)} > \frac{n}{2} + \frac{n}{4c}(s - s_0 - 1 + 2c)(s - s_0). \tag{11}$$

Indeed we have $A_{T_{s_0}}^{(s_0)} = 1 + n\sum_{k=0}^{s_0}(4n)^{-0.5^k} > n(4n)^{-0.5^{s_0}} \ge n(4n)^{-0.5^{\log_2\log_2 4n}} = \frac{n}{2}$. Hence

$$\begin{aligned}
A_0^{(s_0+1)} &= A_{T_{s_0}}^{(s_0)} - T_{s_0+1}\left(a^{(s_0+1)}\right)^2 \\
&\overset{(e)}{\ge} \frac{n}{2} - \frac{n}{4}, \\
A_{T_{s_0+1}}^{(s_0+1)} &= A_0^{(s_0+1)} + T_{s_0+1}\left(\left(a^{(s_0+1)}\right) + \left(a^{(s_0+1)}\right)^2\right) \\
&\overset{(f)}{\ge} \frac{n}{2} + n\left(\frac{1}{2} + \frac{1}{4}\right) \\
&> \frac{n}{2} + \frac{n}{2},
\end{aligned}$$

where (e) and (f) are both by $a^{(s_0+1)} = \frac{1}{2}$, $T_{s_0+1} = n$. Supposed that 10 and 11 hold for all $s_0 < k \le s$. For $k = s+1$

we have

$$\begin{aligned}
A_0^{(s+1)} &= A_{T_s}^{(s)} - T_{s+1}\left(a^{(s+1)}\right)^2 \\
&\overset{(g)}{>} \frac{n}{2} + \frac{n}{4c}(s - s_0 - 1 + 2c)(s - s_0) - n\left(\frac{s - s_0 + c}{2c}\right)^2 \\
&= \frac{n}{2} + \frac{n}{4c}(s - s_0 - 1 + 2c)(s - s_0) - \frac{n}{4c^2}(s - s_0 + c)^2, \\
A_{T_{s+1}}^{(s+1)} &= A_0^{(s+1)} + T_{s+1}\left(a^{(s+1)} + \left(a^{(s+1)}\right)^2\right) \\
&= A_{T_s}^{(s)} + T_{s+1}a^{(s+1)} \\
&\overset{(h)}{>} \frac{n}{2} + \frac{n}{4c}(s - s_0 - 1 + 2c)(s - s_0) + \frac{n}{2c}(s - s_0 + c) \\
&= \frac{n}{2} + \frac{n}{4c}(s - s_0 + 2c)(s - s_0 + 1),
\end{aligned}$$

where (g) and (h) are both due to $T_{s+1} = n$, $a^{(s+1)} = \frac{s-s_0+c}{2c}$ and the assumption on $A_{T_s}^{(s)}$. Now the induction is completed. We can see that if $c = \frac{3}{2}$. we have

$$\begin{aligned}
A_0^{(s)} &> n\left(\frac{1}{2} + \frac{(s - s_0 - 1 + c)^2}{4c}\left(1 - \frac{1}{c}\right) - \frac{s - s_0 - 1 + c^2}{4c}\right) \\
&= n\left(\frac{1}{2} + \frac{(s - s_0 - 1 + c)^2}{12c} - \frac{s - s_0 - 1 + c^2}{4c}\right) \\
&= n\left(\frac{1}{2} + \frac{(s - s_0 - 1 + c)^2}{16c^2} + \frac{(s - s_0 - 1 + c)^2}{24c} - \frac{s - s_0 - 1 + c^2}{4c}\right) \\
&= n\left(\frac{(s - s_0 - 1 + c)^2}{16c^2} + \frac{(s - s_0 - 1)^2 - 3(s - s_0 - 1) + (c^2 - 6c^2 + 12c)}{24c}\right) \\
&> \frac{(s - s_0 - 1 + c)^2}{16c^2} = \frac{\left(a^{(s)}\right)^2}{4}
\end{aligned}$$

and $A_{T_s}^{(s)} > \frac{n}{4c}(s - s_0)^2$. □

A.5. Putting all together

We are now ready to put everything together and complete the proof of Theorem A.1.

Proof. (Theorem A.1) From Lemma A.11, we know $\left(a^{(s)}\right)^2 < 4A_0^{(s)}$ for any $s \geq 1$, which implies for any $s \geq 1$, $t \in [T_s]$

$$\left(a^{(s)}\right)^2 < 4A_{t-1}^{(s)}.$$

Combining our parameters, we can find the requirements for Lemma A.8 are satisfied, which will give us

$$\begin{aligned}
\mathbb{E}\left[A_{T_S}^{(S)}\left(F(u^{(S)}) - F(x^*)\right)\right] \leq A_{T_0}^{(0)}\left(F(u^{(0)}) - F(x^*)\right) + \frac{\gamma}{2}\left\|u^{(0)} - x^*\right\|^2 \\
+ \mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2 + \left(\frac{\left(a^{(s)}\right)^2}{2\gamma_t^{(s)}} - \frac{\left(a^{(s)}\right)^2}{16\beta}\right)\left\|g_t^{(s)} - g_{t-1}^{(s)}\right\|^2\right].
\end{aligned}$$

By using Lemma A.9, we know

$$\begin{aligned}\mathbb{E}\left[A_{T_S}^{(S)}\left(F(u^{(S)})-F(x^*)\right)\right] &\le A_{T_0}^{(0)}\left(F(u^{(0)})-F(x^*)\right)+\frac{\gamma}{2}\left\|u^{(0)}-x^*\right\|^2+\frac{8\beta\left(D^4+2\eta^4\right)}{\eta^2}\\ &\overset{(a)}{\le} \frac{5}{4}\left(F(u^{(0)})-F(x^*)\right)+\frac{\gamma}{2}\left\|u^{(0)}-x^*\right\|^2+\frac{8\beta\left(D^4+2\eta^4\right)}{\eta^2}\\ \Rightarrow \mathbb{E}\left[F(u^{(S)})-F(x^*)\right] &\le \frac{V}{2A_{T_S}^{(S)}}\\ &\overset{(b)}{\le}\begin{cases}\frac{2V}{(4n)^{1-0.5^S}} & 1\le S\le s_0\\ \frac{2cV}{n(S-s_0)^2} & s_0<S\end{cases}\end{aligned}$$

where (a) is by plugging in $A_{T_0}^{(0)}=\frac{5}{4}$, (b) is by A.11. □

- If $\epsilon \ge \frac{V}{n}$, we choose $S=\lceil \log_2\log_2 \frac{4V}{\epsilon}\rceil \le \lceil \log_2\log_2 4n\rceil = s_0$, so we have

$$\begin{aligned}\mathbb{E}\left[F(u^{(S)})-F(x^*)\right] &\le \frac{2V}{(4n)^{1-0.5^S}}\\ &\overset{(c)}{\le} \frac{2V}{\left(\frac{4V}{\epsilon}\right)^{1-0.5^S}}\\ &= \frac{\epsilon}{2\left(\frac{4V}{\epsilon}\right)^{-0.5^S}}\\ &\overset{(d)}{\le} \epsilon,\end{aligned}$$

  where (c) is by $n\ge\frac{V}{\epsilon}$, (d) is by $\left(\frac{4V}{\epsilon}\right)^{-0.5^S}=\left(\frac{4V}{\epsilon}\right)^{-0.5^{\lceil \log_2\log_2\frac{4V}{\epsilon}\rceil}}\ge\left(\frac{4V}{\epsilon}\right)^{-0.5^{\log_2\log_2\frac{4V}{\epsilon}}}=\frac{1}{2}$. Note that the final full gradient computation in the last epoch is not needed, therefore the number of individual gradient evaluations is

$$\begin{aligned}\#grads &= n+\sum_{s=1}^{S-1}\left(2(T_s-1)+n\right)+2(T_S-1)\\ &< 3nS\\ &= 3n\left\lceil \log_2\log_2\frac{4V}{\epsilon}\right\rceil\\ &= \mathcal{O}\left(n\log\log\frac{V}{\epsilon}\right).\end{aligned}$$

- If $\epsilon < \frac{V}{n}$, we choose $S=s_0+\left\lceil\sqrt{\frac{2cV}{n\epsilon}}\right\rceil \ge s_0+1$, so we have

$$\begin{aligned}\mathbb{E}\left[F(u^{(S)})-F(x^*)\right] &\le \frac{2cV}{n(S-s_0)^2}\\ &= \frac{2cV}{n\left(\left\lceil\sqrt{\frac{2cV}{n\epsilon}}\right\rceil\right)^2}\\ &\le \frac{2cV}{n\left(\sqrt{\frac{2cV}{n\epsilon}}\right)^2}\\ &= \epsilon.\end{aligned}$$

The number of individual gradient evaluations is

$$\begin{aligned}\#grads &= n + \sum_{s=1}^{S-1} \left(2(T_s - 1) + n\right) + 2(T_S - 1) \\ &< 3nS \\ &= 3ns_0 + 3n(S - s_0) \\ &= 3n\left\lceil \log_2 \log_2 4n \right\rceil + 3n\left\lceil \sqrt{\frac{2cV}{n\epsilon}} \right\rceil \\ &= \mathcal{O}\left(n \log\log n + \sqrt{\frac{nV(z)}{\epsilon}}\right).\end{aligned}$$

B. Analysis of algorithm 2

In this section, we analyze Algorithm 2 and prove the following convergence guarantee:

Theorem B.1. *(Convergence of AdaVRAG) Define* $s_0 = \lceil \log_2 \log_2 4n \rceil$, $c = \frac{3+\sqrt{33}}{4}$. *Suppose we set the parameters of Algorithm 2 as follows:*

$$\begin{aligned}a^{(s)} &= \begin{cases} 1 - (4n)^{-0.5^s} & 1 \le s \le s_0 \\ \frac{c}{s - s_0 + 2c} & s_0 < s \end{cases}, \\ q^{(s)} &= \begin{cases} \frac{1}{\left(1 - a^{(s)}\right)a^{(s)}} & 1 \le s \le s_0 \\ \frac{8\left(2 - a^{(s)}\right)a^{(s)}}{3(1 - a^{(s)})} & s_0 < s \end{cases}, \\ T_s &= n.\end{aligned}$$

Suppose that $\mathcal{X}$ *is a compact convex set with diameter* D *and we set* $\eta = \Theta(D)$. *Addtionally, we assume that* $2\eta^2 > D^2$ *if Option I is used for setting the step size. The number of individual gradient evaluations to achieve a solution* $u^{(S)}$ *such that* $\mathbb{E}\left[F(u^{(S)}) - F(x^*)\right] \le \epsilon$ *for Algorithm 2 is*

$$\#grads = \begin{cases} \mathcal{O}\left(n \log\log \frac{V}{\epsilon}\right) & \epsilon \ge \frac{V}{n} \\ \mathcal{O}\left(n \log\log n + \sqrt{\frac{nV}{\epsilon}}\right) & \epsilon < \frac{V}{n} \end{cases},$$

where

$$V = \begin{cases} \frac{1}{2}(F(u^{(0)}) - F(x^*)) + \gamma \left\|u^{(0)} - x^*\right\|^2 + \left[\beta - \left(1 - \frac{D^2}{2\eta^2}\right)\gamma\right]^+ \left(D^2 + 2(\eta^2 + D^2)\log \frac{\frac{2\eta^2\beta}{2\eta^2 - D^2}}{\gamma}\right) & \textit{for Option I} \\ \frac{1}{2}(F(u^{(0)}) - F(x^*)) + \gamma \left\|u^{(0)} - x^*\right\|^2 + \eta^2 \left(\frac{D^2}{\eta^2} + \beta - \gamma\right)^+ \left(\frac{2D^2}{\eta^2} + \beta - \gamma\right) & \textit{for Option II} \end{cases}.$$

B.1. Single epoch progress and final output

We first analyze the progress in function value made in a single iteration of an epoch. The analysis is done in a standard way by combining the smoothness and convexity of f, the convexity of h and the optimality condition of $x_t^{(s)}$.

Lemma B.2. *For all epochs* $s \ge 1$ *and all iterations* $t \in [T_s]$, *we have*

$$\begin{aligned}\mathbb{E}\left[F(\overline{x}_t^{(s)}) - F(x^*)\right] \le\ & \mathbb{E}\left[\left(1 - a^{(s)}\right)\left(F(u^{(s-1)}) - F(x^*)\right)\right] \\ & + \mathbb{E}\left[\frac{\gamma_{t-1}^{(s)} q^{(s)} a^{(s)}}{2}\left(\left\|x_{t-1}^{(s)} - x^*\right\|^2 - \left\|x_t^{(s)} - x^*\right\|^2\right)\right] \\ & + \mathbb{E}\left[\left(\frac{\beta\left(2 - a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1 - a^{(s)}\right)} - \frac{\gamma_{t-1}^{(s)} q^{(s)} a^{(s)}}{2}\right)\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right].\end{aligned}$$

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left[f(\overline{x}_t^{(s)}) - f(\overline{x}_{t-1}^{(s)})\right] \\
\overset{(a)}{\leq}&\mathbb{E}\left[\left\langle \nabla f(\overline{x}_{t-1}^{(s)}), \overline{x}_t^{(s)} - \overline{x}_{t-1}^{(s)}\right\rangle + \frac{\beta}{2}\left\|\overline{x}_t^{(s)} - \overline{x}_{t-1}^{(s)}\right\|^2\right] \\
=&\mathbb{E}\left[\left\langle g_t^{(s)}, \overline{x}_t^{(s)} - \overline{x}_{t-1}^{(s)}\right\rangle + \left\langle \nabla f(\overline{x}_{t-1}^{(s)}) - g_t^{(s)}, \overline{x}_t^{(s)} - \overline{x}_{t-1}^{(s)}\right\rangle + \frac{\beta}{2}\left\|\overline{x}_t^{(s)} - \overline{x}_{t-1}^{(s)}\right\|^2\right],
\end{aligned}$$

where (a) is due to f being β-smooth. Using Cauchy–Schwarz inequality and Young's inequality ($ab \leq \frac{\lambda}{2}a^2 + \frac{1}{2\lambda}b^2$ with $\lambda > 0$) we have

$$\begin{aligned}
&\left\langle \nabla f(\overline{x}_{t-1}^{(s)}) - g_t^{(s)}, \overline{x}_t^{(s)} - \overline{x}_{t-1}^{(s)}\right\rangle \\
\leq&\left\|\nabla f(\overline{x}_{t-1}^{(s)}) - g_t^{(s)}\right\|\left\|\overline{x}_t^{(s)} - \overline{x}_{t-1}^{(s)}\right\| \\
\leq&\frac{1-a^{(s)}}{2\beta}\left\|\nabla f(\overline{x}_{t-1}^{(s)}) - g_t^{(s)}\right\|^2 + \frac{\beta}{2(1-a^{(s)})}\left\|\overline{x}_t^{(s)} - \overline{x}_{t-1}^{(s)}\right\|^2,
\end{aligned}$$

also note that

$$\overline{x}_t^{(s)} - \overline{x}_{t-1}^{(s)} = \left(a^{(s)}x_t^{(s)} + (1-a^{(s)})u^{(s-1)}\right) - \left(a^{(s)}x_{t-1}^{(s)} + (1-a^{(s)})u^{(s-1)}\right) = a^{(s)}\left(x_t^{(s)} - x_{t-1}^{(s)}\right).$$

Hence, we obtain

$$\begin{aligned}
&\mathbb{E}\left[f(\overline{x}_t^{(s)}) - f(\overline{x}_{t-1}^{(s)})\right] \\
\leq&\mathbb{E}\left[\left\langle g_t^{(s)}, a^{(s)}\left(x_t^{(s)} - x_{t-1}^{(s)}\right)\right\rangle + \frac{1-a^{(s)}}{2\beta}\left\|\nabla f(\overline{x}_{t-1}^{(s)}) - g_t^{(s)}\right\|^2 + \frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)}\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right] \\
\overset{(b)}{\leq}&\mathbb{E}\left[\left\langle g_t^{(s)}, a^{(s)}\left(x_t^{(s)} - x_{t-1}^{(s)}\right)\right\rangle + \left(1-a^{(s)}\right)\left(f(u^{(s-1)}) - f(\overline{x}_{t-1}^{(s)}) - \langle \nabla f(\overline{x}_{t-1}^{(s)}), u^{(s-1)} - \overline{x}_{t-1}^{(s)}\rangle\right)\right] \\
&+\mathbb{E}\left[\frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)}\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right] \\
=&\mathbb{E}\left[\left\langle g_t^{(s)}, a^{(s)}\left(x_t^{(s)} - x^*\right)\right\rangle + \left\langle g_t^{(s)}, a^{(s)}\left(x^* - x_{t-1}^{(s)}\right)\right\rangle - \left\langle \nabla f(\overline{x}_{t-1}^{(s)}), \left(1-a^{(s)}\right)\left(u^{(s-1)} - \overline{x}_{t-1}^{(s)}\right)\right\rangle\right] \\
&+\mathbb{E}\left[\left(1-a^{(s)}\right)\left(f(u^{(s-1)}) - f(\overline{x}_{t-1}^{(s)})\right)\right] + \mathbb{E}\left[\frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)}\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right] \\
\overset{(c)}{=}&\mathbb{E}\left[\left\langle g_t^{(s)}, a^{(s)}\left(x_t^{(s)} - x^*\right)\right\rangle + \left\langle \nabla f(\overline{x}_{t-1}^{(s)}), a^{(s)}\left(x^* - x_{t-1}^{(s)}\right) - \left(1-a^{(s)}\right)\left(u^{(s-1)} - \overline{x}_{t-1}^{(s)}\right)\right\rangle\right] \\
&+\mathbb{E}\left[\left(1-a^{(s)}\right)\left(f(u^{(s-1)}) - f(\overline{x}_{t-1}^{(s)})\right)\right] + \mathbb{E}\left[\frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)}\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right] \\
\overset{(d)}{=}&\mathbb{E}\left[\left\langle g_t^{(s)}, a^{(s)}\left(x_t^{(s)} - x^*\right)\right\rangle + \left\langle \nabla f(\overline{x}_{t-1}^{(s)}), a^{(s)}\left(x^* - \overline{x}_{t-1}^{(s)}\right)\right\rangle + \left(1-a^{(s)}\right)\left(f(u^{(s-1)}) - f(\overline{x}_{t-1}^{(s)})\right)\right] \\
&+\mathbb{E}\left[\frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)}\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right] \\
\overset{(e)}{\leq}&\mathbb{E}\left[\left\langle g_t^{(s)}, a^{(s)}\left(x_t^{(s)} - x^*\right)\right\rangle + \frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)}\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right. \\
&\left.+\left(1-a^{(s)}\right)f(u^{(s-1)}) + a^{(s)}f(x^*) - f(\overline{x}_{t-1}^{(s)})\right],
\end{aligned} \tag{12}$$

where (b) is by Lemma A.2, (c) is because of

$$\mathbb{E}\left[\left\langle g_t^{(s)}, a^{(s)}\left(x^* - x_{t-1}^{(s)}\right)\right\rangle\right] = \mathbb{E}\left[\left\langle \nabla f(\overline{x}_{t-1}^{(s)}), a^{(s)}\left(x^* - x_{t-1}^{(s)}\right)\right\rangle\right],$$

(d) is by $\overline{x}_{t-1}^{(s)} = a^{(s)}x_{t-1}^{(s)} + \left(1-a^{(s)}\right)u^{(s-1)}$, (e) is due to the convexity of f which implies

$$\left\langle \nabla f(\overline{x}_{t-1}^{(s)}), a^{(s)}\left(x^* - \overline{x}_{t-1}^{(s)}\right)\right\rangle \le a^{(s)}\left(f(x^*) - f(\overline{x}_{t-1}^{(s)})\right).$$

By adding $\mathbb{E}\left[f(\overline{x}_{t-1}^{(s)}) - f(x^*)\right]$ to both sides of (12), we obtain

$$\begin{aligned}
&\mathbb{E}\left[f(\overline{x}_t^{(s)}) - f(x^*)\right] \\
\le&\mathbb{E}\left[\left\langle g_t^{(s)}, a^{(s)}\left(x_t^{(s)} - x^*\right)\right\rangle + \frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)}\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2 + \left(1-a^{(s)}\right)\left(f(u^{(s-1)}) - f(x^*)\right)\right]. \qquad (13)
\end{aligned}$$

Next, we upper bound the inner product $\left\langle g_t^{(s)}, a^{(s)}\left(x_t^{(s)} - x^*\right)\right\rangle$. By the optimality condition of $x_t^{(s)}$, we have

$$\left\langle g_t^{(s)} + h'(x_t^{(s)}) + \gamma_{t-1}^{(s)}q^{(s)}\left(x_t^{(s)} - x_{t-1}^{(s)}\right), x_t^{(s)} - x^*\right\rangle \le 0,$$

where $h'(x_t^{(s)}) \in \partial h(x_t^{(s)})$ is a subgradient of h at $x_t^{(s)}$. We rearrange the above inequality and obtain

$$\begin{aligned}
&a^{(s)}\left\langle g_t^{(s)}, x_t^{(s)} - x^*\right\rangle \\
\le&a^{(s)}\left\langle h'(x_t^{(s)}) + \gamma_{t-1}^{(s)}q^{(s)}\left(x_t^{(s)} - x_{t-1}^{(s)}\right), x^* - x_t^{(s)}\right\rangle \\
\overset{(f)}{\le}&a^{(s)}\left(h(x^*) - h(x_t^{(s)})\right) + a^{(s)}\gamma_{t-1}^{(s)}q^{(s)}\left\langle x_t^{(s)} - x_{t-1}^{(s)}, x^* - x_t^{(s)}\right\rangle \\
\overset{(g)}{=}&a^{(s)}\left(h(x^*) - h(x_t^{(s)})\right) + \frac{a^{(s)}\gamma_{t-1}^{(s)}q^{(s)}}{2}\left(\left\|x_{t-1}^{(s)} - x^*\right\|^2 - \left\|x_t^{(s)} - x^*\right\|^2 - \left\|x_{t-1}^{(s)} - x_t^{(s)}\right\|^2\right), \qquad (14)
\end{aligned}$$

where (f) follows from the convexity of h and the fact that $h'(x_t^{(s)}) \in \partial h(x_t^{(s)})$, and (g) is due to the identity $\langle a, b\rangle = \frac{1}{2}\left(\|a+b\|^2 - \|a\|^2 - \|b\|^2\right)$.

We plug in (14) into (13), and obtain

$$\begin{aligned}
&\mathbb{E}\left[f(\overline{x}_t^{(s)}) - f(x^*)\right] \\
\le&\mathbb{E}\left[\left(1-a^{(s)}\right)\left(f(u^{(s-1)}) - f(x^*)\right) + a^{(s)}\left(h(x^*) - h(x_t^{(s)})\right)\right] \\
&+\mathbb{E}\left[\frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\left(\left\|x_{t-1}^{(s)} - x^*\right\|^2 - \left\|x_t^{(s)} - x^*\right\|^2\right) + \left(\frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)} - \frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\right)\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right] \\
\overset{(h)}{=}&\mathbb{E}\left[\left(1-a^{(s)}\right)\left(F(u^{(s-1)}) - F(x^*)\right) + h(x^*) - a^{(s)}h(x_t^{(s)}) - \left(1-a^{(s)}\right)h(u^{(s-1)})\right] \\
&+\mathbb{E}\left[\frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\left(\left\|x_{t-1}^{(s)} - x^*\right\|^2 - \left\|x_t^{(s)} - x^*\right\|^2\right) + \left(\frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)} - \frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\right)\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right] \\
\overset{(i)}{\le}&\mathbb{E}\left[\left(1-a^{(s)}\right)\left(F(u^{(s-1)}) - F(x^*)\right) + h(x^*) - h(\overline{x}_t^{(s)})\right] \\
&+\mathbb{E}\left[\frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\left(\left\|x_{t-1}^{(s)} - x^*\right\|^2 - \left\|x_t^{(s)} - x^*\right\|^2\right) + \left(\frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)} - \frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\right)\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right],
\end{aligned}$$

where (h) is by the definition of $F = f + h$, and (i) is by the convexity of h which implies

$$h(\overline{x}_t^{(s)}) = h\left(a^{(s)}x_t^{(s)} + (1-a^{(s)})u^{(s-1)}\right) \le a^{(s)}h(x_t^{(s)}) + (1-a^{(s)})h(u^{(s-1)}).$$

Now we move the term $\mathbb{E}\left[h(x^*) - h(\overline{x}_t^{(s)})\right]$ to the LHS, and obtain

$$\begin{aligned}
&\mathbb{E}\left[F(\overline{x}_t^{(s)}) - F(x^*)\right] \\
\le&\mathbb{E}\left[\left(1-a^{(s)}\right)\left(F(u^{(s-1)}) - F(x^*)\right) + \frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\left(\left\|x_{t-1}^{(s)} - x^*\right\|^2 - \left\|x_t^{(s)} - x^*\right\|^2\right)\right] \\
&+\mathbb{E}\left[\left(\frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)} - \frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\right)\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right].
\end{aligned}$$

□

By Lemma B.2, if $\frac{1}{T_s}\sum_{t=1}^{T_s}\overline{x}_t^{(s)}$ is defined as a new chekpoint like what we do in Algorithm 2, the following guarantee for the function value progress in one epoch comes up immediately by the convexity of F.

Lemma B.3. *For all epochs $s \ge 1$, we have*

$$\begin{aligned}
\mathbb{E}\left[F(u^{(s)}) - F(x^*)\right] \le&\ \mathbb{E}\left[\left(1-a^{(s)}\right)\left(F(u^{(s-1)}) - F(x^*)\right)\right] \\
&+\mathbb{E}\left[\frac{1}{T_s}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\left(\left\|x_{t-1}^{(s)} - x^*\right\|^2 - \left\|x_t^{(s)} - x^*\right\|^2\right)\right] \\
&+\mathbb{E}\left[\frac{1}{T_s}\sum_{t=1}^{T_s}\left(\frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)} - \frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\right)\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right].
\end{aligned}$$

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left[F(u^{(s)}) - F(x^*)\right] \\
\overset{(a)}{\le}&\mathbb{E}\left[\frac{1}{T_s}\sum_{t=1}^{T_s}\left(F(\overline{x}_t^{(s)}) - F(x^*)\right)\right] \\
\overset{(b)}{\le}&\mathbb{E}\left[\left(1-a^{(s)}\right)\left(F(u^{(s-1)}) - F(x^*)\right)\right] \\
&+\mathbb{E}\left[\frac{1}{T_s}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\left(\left\|x_{t-1}^{(s)} - x^*\right\|^2 - \left\|x_t^{(s)} - x^*\right\|^2\right)\right] \\
&+\mathbb{E}\left[\frac{1}{T_s}\sum_{t=1}^{T_s}\left(\frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)} - \frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\right)\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right],
\end{aligned}$$

where (a) is by the convexity of F and the definition of $u^{(s)} = \frac{1}{T_s}\sum_{t=1}^{T_s}\overline{x}_t^{(s)}$, and (b) is by Lemma B.2. □

Lemma A.8 is a quite general result without any assumptions on any parameters. To ensure that we can make the telescoping sum over the function value part, and also to simplify the term besides the function value part, we need some specific conditions on our parameters to be satisfied, which is stated in Lemma B.4. With these extra conditions, we can finally find the following guarantee for the function value gap of the final output $u^{(S)}$.

Lemma B.4. *For all $S \ge 1$, if the parameters satisfy*

$$\frac{\left(2-a^{(s)}\right)a^{(s)}}{1-a^{(s)}} \le q^{(s)}, \forall s \in [S]$$

and

$$\frac{(1-a^{(s+1)})T_{s+1}}{q^{(s+1)}a^{(s+1)}} \leq \frac{T_s}{q^{(s)}a^{(s)}}, \forall s \in [S-1].$$

then we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{T_S}{q^{(S)}a^{(S)}}(F(u^{(S)}) - F(x^*))\right] \\
\leq &\frac{(1-a^{(1)})T_1}{q^{(1)}a^{(1)}}(F(u^{(0)}) - F(x^*)) \\
&+ \mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_{t-1}^{(s)} - x^*\right\|^2 - \frac{\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2 + \frac{\beta - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right].
\end{aligned}$$

Proof. If $\frac{\left(2-a^{(s)}\right)a^{(s)}}{1-a^{(s)}} \leq q^{(s)}$ for any $s \in [S]$, by using Lemma B.3, we know

$$\begin{aligned}
&\mathbb{E}\left[F(u^{(s)}) - F(x^*)\right] \\
\leq &\mathbb{E}\left[\left(1-a^{(s)}\right)\left(F(u^{(s-1)}) - F(x^*)\right)\right] \\
&+ \mathbb{E}\left[\frac{1}{T_s}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\left(\left\|x_{t-1}^{(s)} - x^*\right\|^2 - \left\|x_t^{(s)} - x^*\right\|^2\right)\right] \\
&+ \mathbb{E}\left[\frac{1}{T_s}\sum_{t=1}^{T_s}\left(\frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)} - \frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\right)\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right] \\
\leq &\mathbb{E}\left[(1-a^{(s)})(F(u^{(s-1)}) - F(x^*))\right] \\
&+ \mathbb{E}\left[\frac{q^{(s)}a^{(s)}}{T_s}\left(\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_{t-1}^{(s)} - x^*\right\|^2 - \frac{\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2 + \frac{\beta - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right)\right]
\end{aligned}$$

Now multiply both sides by $\frac{T_s}{q^{(s)}a^{(s)}}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{T_s}{q^{(s)}a^{(s)}}(F(u^{(s)}) - F(x^*))\right] \\
\leq &\mathbb{E}\left[\frac{(1-a^{(s)})T_s}{q^{(s)}a^{(s)}}(F(u^{(s-1)}) - F(x^*))\right] \\
&+ \mathbb{E}\left[\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_{t-1}^{(s)} - x^*\right\|^2 - \frac{\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2 + \frac{\beta - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right].
\end{aligned}$$

If $\frac{(1-a^{(s+1)})T_{s+1}}{q^{(s+1)}a^{(s+1)}} \leq \frac{T_s}{q^{(s)}a^{(s)}}$ is satisfied for any $s \in [S-1]$, we can make the telescoping sum from $s=1$ to S to get

$$\begin{aligned}
&\mathbb{E}\left[\frac{T_S}{q^{(S)}a^{(S)}}(F(u^{(S)}) - F(x^*))\right] \\
\leq &\frac{(1-a^{(1)})T_1}{q^{(1)}a^{(1)}}(F(u^{(0)}) - F(x^*)) \\
&+ \mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_{t-1}^{(s)} - x^*\right\|^2 - \frac{\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x^*\right\|^2 + \frac{\beta - \gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right].
\end{aligned}$$

□

B.2. Bound for the residual term

By the analysis in the previous subsection, we get an upper bound for the function value gap of $u^{(S)}$ involving $F(u^{(0)}) - F(x^*)$ and

$$\mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_{t-1}^{(s)}-x^*\right\|^2-\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\right]. \tag{15}$$

In this subsection we will show how to bound 15 under the compact assumption of $\mathcal{X}$. Before giving the detailed analysis of the two different update options, we first state the following lemma to simplify 15.

Lemma B.5. *If $\gamma_t^{(s)} \geq \gamma_{t-1}^{(s)}$ for any $s \in [S]$, $t \in [T_s]$ and $\mathcal{X}$ is a compact convex set with diameter D, then we have*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_{t-1}^{(s)}-x^*\right\|^2-\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\right]\\
\leq&\frac{\gamma}{2}\left\|u^{(0)}-x^*\right\|^2+\mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}D^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\right].
\end{aligned}$$

Proof. It follows that

$$\begin{aligned}
&\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_{t-1}^{(s)}-x^*\right\|^2-\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\\
=&\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_{t-1}^{(s)}-x^*\right\|^2-\frac{\gamma_t^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\\
\overset{(a)}{\leq}&\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_{t-1}^{(s)}-x^*\right\|^2-\frac{\gamma_t^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}D^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\\
=&\sum_{s=1}^{S}\left(\frac{\gamma_0^{(s)}}{2}\left\|x_0^{(s)}-x^*\right\|^2-\frac{\gamma_{T_s}^{(s)}}{2}\left\|x_{T_s}^{(s)}-x^*\right\|^2+\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}D^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\right)\\
\overset{(b)}{=}&\sum_{s=1}^{S}\left(\frac{\gamma_0^{(s)}}{2}\left\|x_0^{(s)}-x^*\right\|^2-\frac{\gamma_0^{(s+1)}}{2}\left\|x_0^{(s+1)}-x^*\right\|^2+\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}D^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\right)\\
=&\frac{\gamma_0^{(1)}}{2}\left\|x_0^{(1)}-x^*\right\|^2-\frac{\gamma_0^{(S+1)}}{2}\left\|x_0^{(S+1)}-x^*\right\|^2+\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}D^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\\
\leq&\frac{\gamma_0^{(1)}}{2}\left\|x_0^{(1)}-x^*\right\|^2+\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}D^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\\
\overset{(c)}{=}&\frac{\gamma}{2}\left\|u^{(0)}-x^*\right\|^2+\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}D^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2,
\end{aligned}$$

where (a) is due to $\gamma_t^{(s)} \geq \gamma_{t-1}^{(s)}$ and $\left\|x_t^{(s)}-x^*\right\| \leq D$, (b) follows from the definition of $x_0^{(s+1)} = x_{T_s}^{(s)}$ and $\gamma_0^{(s+1)} = \gamma_{T_s}^{(s)}$, (c) is by the definition of $x_0^{(1)} = u^{(0)}$ and $\gamma_0^{(1)} = \gamma$. Now Taking expectations with both sides yields what we want. □

With the above result, we can show the bound of 15 under Option I and Option II respectively. There are two key common parts in our analysis, the first one is to notice that we can reduce the doubly indexed sequence $\left\{x_t^{(s)}\right\}$ and $\left\{\gamma_t^{(s)}\right\}$ into two singly indexed sequences, which are much easier to bound. The second technique is to define a hitting time τ to upper bound $\gamma_t^{(s)}$. Read our proof for the details.

Lemma B.6. *For Option I, if $\mathcal{X}$ is a compact convex set with diameter D and $2\eta^2 > D^2$, we have*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_{t-1}^{(s)}-x^*\right\|^2-\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\right]\\
\leq&\frac{\gamma}{2}\left\|u^{(0)}-x^*\right\|^2+\left[\frac{\beta}{2}-\left(\frac{1}{2}-\frac{D^2}{4\eta^2}\right)\gamma\right]^+\left(D^2+2(\eta^2+D^2)\log\frac{\frac{2\eta^2\beta}{2\eta^2-D^2}}{\gamma}\right).
\end{aligned}$$

Proof. For Option I, by the definition of $\gamma_t^{(s)}$, we have

$$\gamma_t^{(s)}\geq\gamma_{t-1}^{(s)},\forall s\in[S],t\in[T_s].$$

By requiring that $\mathcal{X}$ is a compact convex set with diameter D, we can apply Lemma B.5 and obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_{t-1}^{(s)}-x^*\right\|^2-\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\right]\\
\leq&\frac{\gamma}{2}\left\|u^{(0)}-x^*\right\|^2+\mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}D^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\right].
\end{aligned}\tag{16}$$

Note that the last element $x_{T_s}^{(s)}$ (resp. $\gamma_{T_s}^{(s)}$) in the s-th epoch is just the start element $x_0^{(s+1)}$ (resp. $\gamma_0^{(s+1)}$) in the $(s+1)$-th epoch, which means we can consider the doubly indexed sequences $\{x_t^{(s)}\}$ and $\{\gamma_t^{(s)}\}$ as two singly indexed sequences $\{x_t', t\geq 0\}$ and $\{\gamma_t', t\geq 0, \gamma_0'=\gamma\}$ with the reformulated update rule as follows

$$\gamma_t'=\gamma_{t-1}'\sqrt{1+\frac{\left\|x_t'-x_{t-1}'\right\|^2}{\eta^2}}.$$

Besides, by defining $T'=\sum_{s=1}^{S}T_s$, we have

$$\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}D^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2=\sum_{t=1}^{T'}\frac{\gamma_t'-\gamma_{t-1}'}{2}D^2+\frac{\beta-\gamma_{t-1}'}{2}\left\|x_t'-x_{t-1}'\right\|^2.$$

Note that we require $2\eta^2>D^2$, so if $\gamma\geq\frac{2\eta^2\beta}{2\eta^2-D^2}\Leftrightarrow\frac{\beta}{2}-\left(\frac{1}{2}-\frac{D^2}{4\eta^2}\right)\gamma\leq 0\Rightarrow\frac{\beta}{2}-\left(\frac{1}{2}-\frac{D^2}{4\eta^2}\right)\gamma_{t-1}'\leq 0$, by using the reformulated update rule, we have

$$\begin{aligned}
&\sum_{t=1}^{T'}\frac{\gamma_t'-\gamma_{t-1}'}{2}D^2+\frac{\beta-\gamma_{t-1}'}{2}\left\|x_t'-x_{t-1}'\right\|^2\\
=&\sum_{t=1}^{T'}\frac{(\gamma_t')^2-(\gamma_{t-1}')^2}{2(\gamma_t'+\gamma_{t-1}')}D^2+\frac{\beta-\gamma_{t-1}'}{2}\left\|x_t'-x_{t-1}'\right\|^2\\
=&\sum_{t=1}^{T'}\frac{(\gamma_{t-1}')^2D^2}{2\eta^2(\gamma_t'+\gamma_{t-1}')}\left\|x_t'-x_{t-1}'\right\|^2+\frac{\beta-\gamma_{t-1}'}{2}\left\|x_t'-x_{t-1}'\right\|^2\\
\overset{(a)}{\leq}&\sum_{t=1}^{T'}\left(\frac{\gamma_{t-1}'}{4\eta^2}D^2+\frac{\beta-\gamma_{t-1}'}{2}\right)\left\|x_t'-x_{t-1}'\right\|^2\\
=&\sum_{t=1}^{T'}\left[\frac{\beta}{2}-\left(\frac{1}{2}-\frac{D^2}{4\eta^2}\right)\gamma_{t-1}'\right]\left\|x_t'-x_{t-1}'\right\|^2\\
\leq&0,
\end{aligned}$$

where (a) is by $\gamma_t' \geq \gamma_{t-1}'$. Now we assume $\gamma < \frac{2\eta^2\beta}{2\eta^2 - D^2}$, define

$$\tau = \max\left\{t \in [T'], \gamma_{t-1}' < \frac{2\eta^2\beta}{2\eta^2 - D^2}\right\}.$$

By our assumption on γ, we know $\tau \geq 1$, Combining the reformulated update rule, we have

$$\begin{aligned}
&\sum_{t=1}^{T'} \frac{\gamma_t' - \gamma_{t-1}'}{2} D^2 + \frac{\beta - \gamma_{t-1}'}{2} \left\|x_t' - x_{t-1}'\right\|^2 \\
\leq &\sum_{t=1}^{T'} \left[\frac{\beta}{2} - \left(\frac{1}{2} - \frac{D^2}{4\eta^2}\right)\gamma_{t-1}'\right] \left\|x_t' - x_{t-1}'\right\|^2 \\
\leq &\sum_{t=1}^{\tau} \left[\frac{\beta}{2} - \left(\frac{1}{2} - \frac{D^2}{4\eta^2}\right)\gamma_{t-1}'\right] \left\|x_t' - x_{t-1}'\right\|^2 \\
\overset{(b)}{\leq} &\left[\frac{\beta}{2} - \left(\frac{1}{2} - \frac{D^2}{4\eta^2}\right)\gamma\right] \sum_{t=1}^{\tau} \left\|x_t' - x_{t-1}'\right\|^2 \\
\overset{(c)}{\leq} &\left[\frac{\beta}{2} - \left(\frac{1}{2} - \frac{D^2}{4\eta^2}\right)\gamma\right] \left(D^2 + \sum_{t=1}^{\tau-1} \left\|x_t' - x_{t-1}'\right\|^2\right) \\
\overset{(d)}{=} &\left[\frac{\beta}{2} - \left(\frac{1}{2} - \frac{D^2}{4\eta^2}\right)\gamma\right] \left(D^2 + \sum_{t=1}^{\tau-1} \eta^2 \frac{(\gamma_t')^2 - (\gamma_{t-1}')^2}{(\gamma_{t-1}')^2}\right) \\
\overset{(e)}{\leq} &\left[\frac{\beta}{2} - \left(\frac{1}{2} - \frac{D^2}{4\eta^2}\right)\gamma\right] \left(D^2 + \left(\eta^2 + D^2\right) \sum_{t=1}^{\tau-1} \frac{(\gamma_t')^2 - (\gamma_{t-1}')^2}{(\gamma_t')^2}\right) \\
\overset{(f)}{\leq} &\left[\frac{\beta}{2} - \left(\frac{1}{2} - \frac{D^2}{4\eta^2}\right)\gamma\right] \left(D^2 + 2\left(\eta^2 + D^2\right) \sum_{t=1}^{\tau-1} \log \frac{\gamma_t'}{\gamma_{t-1}'}\right) \\
= &\left[\frac{\beta}{2} - \left(\frac{1}{2} - \frac{D^2}{4\eta^2}\right)\gamma\right] \left(D^2 + 2\left(\eta^2 + D^2\right) \log \frac{\gamma_{\tau-1}'}{\gamma}\right) \\
\overset{(g)}{\leq} &\left[\frac{\beta}{2} - \left(\frac{1}{2} - \frac{D^2}{4\eta^2}\right)\gamma\right] \left(D^2 + 2\left(\eta^2 + D^2\right) \log \frac{\frac{2\eta^2\beta}{2\eta^2 - D^2}}{\gamma}\right),
\end{aligned}$$

where (b) is by $\gamma_{t-1}' \geq \gamma$, (c) is by $\left\|x_\tau' - x_{\tau-1}'\right\| \leq D$, (d) is by the reformulated update rule, (e) is due to

$$\gamma_t' = \gamma_{t-1}' \sqrt{1 + \frac{\left\|x_t' - x_{t-1}'\right\|^2}{\eta^2}} \leq \gamma_{t-1}' \sqrt{1 + \frac{D^2}{\eta^2}},$$

(f) is by the inequality $1 - \frac{1}{x^2} \leq \log x^2 = 2\log x$, (g) is by the definition of τ.

Combining two cases of γ, we obtain the bound

$$\begin{aligned}
\sum_{s=1}^{S}\sum_{t=1}^{T_s} \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2} D^2 + \frac{\beta - \gamma_{t-1}^{(s)}}{2} \left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2 &= \sum_{t=1}^{T'} \frac{\gamma_t' - \gamma_{t-1}'}{2} D^2 + \frac{\beta - \gamma_{t-1}'}{2} \left\|x_t' - x_{t-1}'\right\|^2 \\
&\leq \left[\frac{\beta}{2} - \left(\frac{1}{2} - \frac{D^2}{4\eta^2}\right)\gamma\right]^+ \left(D^2 + 2\left(\eta^2 + D^2\right) \log \frac{\frac{2\eta^2\beta}{2\eta^2 - D^2}}{\gamma}\right).
\end{aligned} \tag{17}$$

By plugging in (17) into (16), we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_{t-1}^{(s)}-x^*\right\|^2-\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\right]\\ \leq&\frac{\gamma}{2}\left\|u^{(0)}-x^*\right\|^2+\left[\frac{\beta}{2}-\left(\frac{1}{2}-\frac{D^2}{4\eta^2}\right)\gamma\right]^+\left(D^2+2\left(\eta^2+D^2\right)\log\frac{\frac{2\eta^2\beta}{2\eta^2-D^2}}{\gamma}\right).\end{aligned}$$

□

Lemma B.7. *For Option II, if $\mathcal{X}$ is a compact set with diameter D, we have*

$$\begin{aligned}&\mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_{t-1}^{(s)}-x^*\right\|^2-\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\right]\\ \leq&\frac{\gamma}{2}\left\|u^{(0)}-x^*\right\|^2+\frac{\eta^2}{2}\left(\frac{D^2}{\eta^2}+\beta-\gamma\right)^+\left(\frac{2D^2}{\eta^2}+\beta-\gamma\right).\end{aligned}$$

Proof. For Option II, by the definition of $\gamma_t^{(s)}$, we have

$$\gamma_t^{(s)}\geq\gamma_{t-1}^{(s)},\forall s\in[S],t\in[T_s].$$

By requiring that $\mathcal{X}$ is a compact convex set with diameter D, we can apply Lemma B.5 and obtain

$$\begin{aligned}&\mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_{t-1}^{(s)}-x^*\right\|^2-\frac{\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x^*\right\|^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\right]\\ \leq&\frac{\gamma}{2}\left\|u^{(0)}-x^*\right\|^2+\mathbb{E}\left[\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}D^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\right].\end{aligned}\tag{18}$$

Note that the last element $x_{T_s}^{(s)}$ (resp. $\gamma_{T_s}^{(s)}$) in the s-th epoch is just the starting element $x_0^{(s+1)}$ (resp. $\gamma_0^{(s+1)}$) in the $(s+1)$-th epoch, which means we can consider the doubly indexed sequences $\{x_t^{(s)}\}$ and $\{\gamma_t^{(s)}\}$ as two singly indexed sequences $\{x_t',t\geq0\}$ and $\{\gamma_t',t\geq0,\gamma_0'=\gamma\}$ with the reformulated update rule as follows

$$\gamma_t'=\gamma_{t-1}'+\frac{\left\|x_t'-x_{t-1}'\right\|^2}{\eta^2}.$$

Besides, by defining $T'=\sum_{s=1}^{S}T_s$, we have

$$\sum_{s=1}^{S}\sum_{t=1}^{T_s}\frac{\gamma_t^{(s)}-\gamma_{t-1}^{(s)}}{2}D^2+\frac{\beta-\gamma_{t-1}^{(s)}}{2}\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2=\sum_{t=1}^{T'}\frac{\gamma_t'-\gamma_{t-1}'}{2}D^2+\frac{\beta-\gamma_{t-1}'}{2}\left\|x_t'-x_{t-1}'\right\|^2.$$

If $\gamma\geq\frac{D^2}{\eta^2}+\beta\Leftrightarrow\frac{D^2}{2\eta^2}+\frac{\beta-\gamma}{2}\leq0\Rightarrow\frac{D^2}{2\eta^2}+\frac{\beta-\gamma_{t-1}'}{2}\leq0$, by using the reformulated update rule, we have

$$\begin{aligned}\sum_{t=1}^{T'}\frac{\gamma_t'-\gamma_{t-1}'}{2}D^2+\frac{\beta-\gamma_{t-1}'}{2}\left\|x_t'-x_{t-1}'\right\|^2=&\sum_{t=1}^{T'}\left(\frac{D^2}{2\eta^2}+\frac{\beta-\gamma_{t-1}'}{2}\right)\left\|x_t'-x_{t-1}'\right\|^2\\ \leq&0.\end{aligned}$$

Now we assume $\gamma<\frac{D^2}{\eta^2}+\beta$. Define

$$\tau=\max\left\{t\in[T'],\gamma_{t-1}'<\frac{D^2}{\eta^2}+\beta\right\}.$$

By our assumption on γ, we know $\tau \geq 1$. Combining the reformulated update rule, we have

$$
\begin{aligned}
&\sum_{t=1}^{T'} \frac{\gamma_t' - \gamma_{t-1}'}{2} D^2 + \frac{\beta - \gamma_{t-1}'}{2} \left\|x_t' - x_{t-1}'\right\|^2 \\
=&\sum_{t=1}^{T'} \left(\frac{D^2}{2\eta^2} + \frac{\beta - \gamma_{t-1}'}{2}\right) \left\|x_t' - x_{t-1}'\right\|^2 \\
\leq&\sum_{t=1}^{\tau} \left(\frac{D^2}{2\eta^2} + \frac{\beta - \gamma_{t-1}'}{2}\right) \left\|x_t' - x_{t-1}'\right\|^2 \\
\overset{(a)}{\leq}&\sum_{t=1}^{\tau} \left(\frac{D^2}{2\eta^2} + \frac{\beta - \gamma}{2}\right) \left\|x_t' - x_{t-1}'\right\|^2 \\
\overset{(b)}{=}&\sum_{t=1}^{\tau} \left(\frac{D^2}{2\eta^2} + \frac{\beta - \gamma}{2}\right) \eta^2 \left(\gamma_t' - \gamma_{t-1}'\right) \\
=&\left(\frac{D^2}{2\eta^2} + \frac{\beta - \gamma}{2}\right) \eta^2 \left(\gamma_\tau' - \gamma\right) \\
\overset{(c)}{=}&\left(\frac{D^2}{2\eta^2} + \frac{\beta - \gamma}{2}\right) \eta^2 \left(\gamma_{\tau-1}' + \frac{\left\|x_\tau' - x_{\tau-1}'\right\|^2}{\eta^2} - \gamma\right) \\
\overset{(d)}{\leq}&\left(\frac{D^2}{2\eta^2} + \frac{\beta - \gamma}{2}\right) \eta^2 \left(2\frac{D^2}{\eta^2} + \beta - \gamma\right) \\
=&\frac{\eta^2}{2} \left(\frac{D^2}{\eta^2} + \beta - \gamma\right) \left(\frac{2D^2}{\eta^2} + \beta - \gamma\right),
\end{aligned}
$$

where (a) is by the fact $\gamma_{t-1}' \geq \gamma$, (b) and (c) are by the reformulated update rule, (d) is by the definition of τ and $\left\|x_\tau' - x_{\tau-1}'\right\| \leq D$.

Combining two cases of γ, we obtain the bound

$$
\begin{aligned}
&\sum_{s=1}^{S} \sum_{t=1}^{T_s} \frac{\gamma_t^{(s)} - \gamma_{t-1}^{(s)}}{2} D^2 + \frac{\beta - \gamma_{t-1}^{(s)}}{2} \left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2 \\
=&\sum_{t=1}^{T'} \frac{\gamma_t' - \gamma_{t-1}'}{2} D^2 + \frac{\beta - \gamma_{t-1}'}{2} \left\|x_t' - x_{t-1}'\right\|^2 \\
\leq&\frac{\eta^2}{2} \left(\frac{D^2}{\eta^2} + \beta - \gamma\right)^+ \left(\frac{2D^2}{\eta^2} + \beta - \gamma\right).
\end{aligned} \tag{19}
$$

By plugging in (19) into (18), we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{S} \sum_{t=1}^{T_s} \frac{\gamma_{t-1}^{(s)}}{2} \left\|x_{t-1}^{(s)} - x^*\right\|^2 - \frac{\gamma_{t-1}^{(s)}}{2} \left\|x_t^{(s)} - x^*\right\|^2 + \frac{\beta - \gamma_{t-1}^{(s)}}{2} \left\|x_t^{(s)} - x_{t-1}^{(s)}\right\|^2\right] \\
\leq&\frac{\gamma}{2} \left\|u^{(0)} - x^*\right\|^2 + \frac{\eta^2}{2} \left(\frac{D^2}{\eta^2} + \beta - \gamma\right)^+ \left(\frac{2D^2}{\eta^2} + \beta - \gamma\right).
\end{aligned}
$$

□

B.3. Parameter bound

Combining the previous two parts analysis on the function value gap and the residual term, we already can see the bound for $F(u^{(S)}) - F(x^*)$. However, we need to make sure that our choice stated in Theorem B.1 indeed satisfies the conditions used in previous lemmas, besides, we also need to give the bounds for our choice explicitly. The following two lemmas can help us to do this.

Lemma B.8. *Under the choice of parameters in Theorem B.1, $\forall s \geq 1$, we have the following facts*

$$\begin{aligned} a^{(s_0)} &\leq \frac{1}{2}, \\ \frac{\left(2-a^{(s)}\right)a^{(s)}}{1-a^{(s)}} &\leq q^{(s)}, \\ \frac{(1-a^{(s+1)})T_{s+1}}{q^{(s+1)}a^{(s+1)}} &\leq \frac{T_s}{q^{(s)}a^{(s)}}. \end{aligned}$$

Proof. Under the choice of parameters in Theorem B.1, the first inequality follows that

$$a^{(s_0)} = 1-(4n)^{-0.5^{s_0}} \leq 1-(4n)^{-0.5^{\log_2\log_2 4n}} = \frac{1}{2}.$$

For the second inequality, note that

$$\frac{\left(2-a^{(s)}\right)a^{(s)}}{\left(1-a^{(s)}\right)q^{(s)}} = \begin{cases} \left(2-a^{(s)}\right)\left(a^{(s)}\right)^2 & 1 \leq s \leq s_0 \\ \frac{3}{8} & s_0 < s \end{cases}.$$

By noticing $\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2 \leq a^{(s)} \leq 1$, the inequality $\frac{\left(2-a^{(s)}\right)a^{(s)}}{1-a^{(s)}} \leq q^{(s)}$ becomes true immediately.

For the third inequality, note that we have $T_s \equiv n$, we only need to prove for any $s \geq 1$, there is

$$\frac{1-a^{(s+1)}}{q^{(s+1)}a^{(s+1)}} \leq \frac{1}{q^{(s)}a^{(s)}}.$$

We consider the following three cases:

- For $1 \leq s \leq s_0-1$, note that $\left(1-a^{(s+1)}\right)^2 = (4n)^{-0.5^s} = 1-a^{(s)}$, $q^{(s)} = \frac{1}{\left(1-a^{(s)}\right)a^{(s)}}$. We know

$$\begin{aligned} \frac{1-a^{(s+1)}}{q^{(s+1)}a^{(s+1)}} &= (1-a^{(s+1)})^2 \\ &= 1-a^{(s)} \\ &= \frac{1}{q^{(s)}a^{(s)}}. \end{aligned}$$

- For $s = s_0$, note that $a^{(s_0+1)} = \frac{c}{1+2c} = \frac{9-\sqrt{33}}{8}$, $q^{(s_0+1)} = \frac{8(2-a^{(s_0+1)})a^{(s_0+1)}}{3(1-a^{(s_0+1)})}$ we have

$$\begin{aligned} \frac{1-a^{(s_0+1)}}{q^{(s_0+1)}a^{(s_0+1)}} &= \frac{3(1-a^{(s_0+1)})^2}{8(2-a^{(s_0+1)})\left(a^{(s_0+1)}\right)^2} \\ &= \frac{1}{2} \\ &\overset{(a)}{\leq} 1-a^{(s_0)} \\ &\overset{(b)}{=} \frac{1}{q^{(s_0)}a^{(s_0)}}, \end{aligned}$$

  where (a) is by $a^{(s_0)} \leq \frac{1}{2}$, (b) is by $q^{(s_0)} = \frac{1}{\left(1-a^{(s_0)}\right)a^{(s_0)}}$.

- For $s \geq s_0+1$, note that $q^{(s)} = \frac{8(2-a^{(s)})a^{(s)}}{3\left(1-a^{(s)}\right)}$, by plugging in $q^{(s)}$, we only need to show

$$\frac{(1-a^{(s+1)})^2}{(a^{(s+1)})^2(2-a^{(s+1)})} \leq \frac{1-a^{(s)}}{(a^{(s)})^2(2-a^{(s)})}.$$

Plug in $a^{(s)} = \frac{c}{s-s_0+2c}$, the above inequality is equivalent to

$$(2(s-s_0)+3c)(s-s_0+1+2c)(s-s_0+1+c)^2 \le (2(s-s_0)+2+3c)(s-s_0+c)(s-s_0+2c)^2.$$

Let $y = s - s_0 \ge 1$, we need to show

$$(2y+3c)(y+1+2c)(y+1+c)^2 \le (2y+2+3c)(y+c)(y+2c)^2$$

is true for $y \ge 1$. People can check when $c = \frac{3+\sqrt{33}}{4}$, the above inequality is right for $y \ge 1$.

□

Lemma B.9. *Under the choice of parameters in Theorem B.1, $\forall s \ge 1$, we have the following bounds*

$$\frac{(1-a^{(1)})T_1}{q^{(1)}a^{(1)}} = \frac{1}{4}$$

and

$$\frac{q^{(s)}a^{(s)}}{T_s} \le \begin{cases} \frac{4}{(4n)^{1-0.5^s}} & 1 \le s \le s_0 \\ \frac{2(5+\sqrt{33})c^2}{3n(s-s_0+2c)^2} & s_0 < s \end{cases}.$$

Proof. Note that $a^{(1)} = 1 - \frac{1}{2\sqrt{n}}$, $T_1 = n$, $q^{(1)} = \frac{1}{(1-a^{(1)})a^{(1)}}$, plugging in these values, we obtain

$$\begin{aligned} \frac{(1-a^{(1)})T_1}{q^{(1)}a^{(1)}} &= (1-a^{(1)})^2 T_1 \\ &= \frac{1}{4} \end{aligned}$$

- For $1 \le s \le s_0$, note that $q^{(s)} = \frac{1}{(1-a^{(s)})a^{(s)}}$ in our choice, so we know

$$\begin{aligned} \frac{q^{(s)}a^{(s)}}{T_s} &= \frac{1}{T_s(1-a^{(s)})} \\ &\overset{(a)}{=} \frac{4}{(4n)^{1-0.5^s}} \end{aligned}$$

  where (a) is by plugging in $T_s = n$ and $a^{(s)} = 1 - (4n)^{-0.5^s}$.

- For $s > s_0$, note that $q^{(s)} = \frac{8(2-a^{(s)})a^{(s)}}{3(1-a^{(s)})}$ we have

$$\begin{aligned} \frac{q^{(s)}a^{(s)}}{T_s} &= \frac{8(2-a^{(s)})(a^{(s)})^2}{3T_s(1-a^{(s)})} \\ &\overset{(b)}{=} \frac{8(2-a^{(s)})(a^{(s)})^2}{3n(1-a^{(s)})} \\ &\overset{(c)}{\le} \frac{2(5+\sqrt{33})c^2}{3n(s-s_0+2c)^2}, \end{aligned}$$

  where (b) is by plugging in$T_s = n$, (c) is by noticing $\frac{2-a^{(s)}}{1-a^{(s)}} \le \frac{2-a^{(s_0+1)}}{1-a^{(s_0+1)}} = \frac{5+\sqrt{33}}{4}$ for $s > s_0$, and plug in $a^{(s)} = \frac{c}{s-s_0+2c}$.

□

B.4. Putting all together

We are now ready to put everything together and complete the proof of Theorem B.1.

Proof. (Theorem B.1) By Lemma B.8, $\forall s \geq 1$, we have

$$\frac{\left(2-a^{(s)}\right)a^{(s)}}{1-a^{(s)}} \leq q^{(s)},$$
$$\frac{(1-a^{(s+1)})T_{s+1}}{q^{(s+1)}a^{(s+1)}} \leq \frac{T_s}{q^{(s)}a^{(s)}}.$$

Hence all the conditions for Lemma B.4 are satisfied. Besides, we assume $\mathcal{X}$ is a compact convex set with diameter D, which satisfies the requirements for Lemma B.7 and B.6.

1. For Option I, by Lemma B.4 and B.6

$$\mathbb{E}\left[\frac{T_S}{q^{(S)}a^{(S)}}(F(u^{(S)})-F(x^*))\right] \leq \frac{(1-a^{(1)})T_1}{q^{(1)}a^{(1)}}(F(u^{(0)})-F(x^*)) + \frac{\gamma}{2}\left\|u^{(0)}-x^*\right\|^2 + \left[\frac{\beta}{2}-\left(\frac{1}{2}-\frac{D^2}{4\eta^2}\right)\gamma\right]^+\left(D^2+2(\eta^2+D^2)\log\frac{\frac{2\eta^2\beta}{2\eta^2-D^2}}{\gamma}\right).$$

2. For Option II, by Lemma B.4 and B.7

$$\mathbb{E}\left[\frac{T_S}{q^{(S)}a^{(S)}}(F(u^{(S)})-F(x^*))\right] \leq \frac{(1-a^{(1)})T_1}{q^{(1)}a^{(1)}}(F(u^{(0)})-F(x^*)) + \frac{\gamma}{2}\left\|u^{(0)}-x^*\right\|^2 + \frac{\eta^2}{2}\left(\frac{D^2}{\eta^2}+\beta-\gamma\right)^+\left(\frac{2D^2}{\eta^2}+\beta-\gamma\right).$$

Plugging in the bound $\frac{(1-a^{(1)})T_1}{q^{(1)}a^{(1)}} = \frac{1}{4}$ from Lemma B.9, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{T_S}{q^{(S)}a^{(S)}}(F(u^{(S)})-F(x^*))\right] &\leq \frac{V}{2} \\
\Rightarrow \mathbb{E}\left[F(u^{(S)})-F(x^*)\right] &\leq \frac{q^{(S)}a^{(S)}V}{2T_S} \\
&\overset{(a)}{\leq} \begin{cases} \frac{2V}{(4n)^{1-0.5^S}} & 1 \leq S \leq s_0 \\ \frac{(5+\sqrt{33})c^2V}{3n(S-s_0+2c)^2} & s_0 < S \end{cases},
\end{aligned}$$

where (a) is by the bound for $\frac{q^{(S)}a^{(S)}}{T_S}$ from Lemma B.9.

- If $\epsilon \geq \frac{V}{n}$, we choose $S = \lceil \log_2 \log_2 \frac{4V}{\epsilon} \rceil \leq \lceil \log_2 \log_2 4n \rceil = s_0$, so we have

$$\begin{aligned}
\mathbb{E}\left[F(u^{(S)})-F(x^*)\right] &\leq \frac{2V}{(4n)^{1-0.5^S}} \\
&\overset{(b)}{\leq} \frac{2V}{\left(\frac{4V}{\epsilon}\right)^{1-0.5^S}} \\
&= \frac{\epsilon}{2\left(\frac{4V}{\epsilon}\right)^{-0.5^S}} \\
&\overset{(c)}{\leq} \epsilon,
\end{aligned}$$

where (b) is by $n \geq \frac{V}{\epsilon}$, (c) is by $\left(\frac{4V}{\epsilon}\right)^{-0.5^S} = \left(\frac{4V}{\epsilon}\right)^{-0.5^{\left\lceil \log_2 \log_2 \frac{4V}{\epsilon} \right\rceil}} \geq \left(\frac{4V}{\epsilon}\right)^{-0.5^{\log_2 \log_2 \frac{4V}{\epsilon}}} = \frac{1}{2}$. The number of individual gradient evaluations is

$$\begin{aligned}
\#grads &= nS + \sum_{s=1}^{S} 2T_s \\
&= 3nS \\
&= 3n \left\lceil \log_2 \log_2 \frac{4V}{\epsilon} \right\rceil \\
&= \mathcal{O}\left(n \log\log \frac{V}{\epsilon}\right).
\end{aligned}$$

- If $\epsilon < \frac{V}{n}$, we choose $S = s_0 + \left\lceil c\left(\sqrt{\frac{(5+\sqrt{33})V}{3n\epsilon}} - \frac{15}{8}\right)\right\rceil \geq s_0 + \left\lceil c\left(\sqrt{\frac{5+\sqrt{33}}{3}} - \frac{15}{8}\right)\right\rceil = s_0 + 1$, so we have

$$\begin{aligned}
\mathbb{E}\left[F(u^{(S)}) - F(x^*)\right] &\leq \frac{(5+\sqrt{33})c^2 V}{3n(S - s_0 + 2c)^2} \\
&= \frac{(5+\sqrt{33})c^2 V}{3n\left(s_0 + \left\lceil c\left(\sqrt{\frac{(5+\sqrt{33})V}{3n\epsilon}} - \frac{15}{8}\right)\right\rceil + 2c\right)^2} \\
&\leq \frac{(5+\sqrt{33})c^2 V}{3n\left(c\sqrt{\frac{(5+\sqrt{33})V}{3n\epsilon}} + \frac{c}{8}\right)^2} \\
&\leq \frac{(5+\sqrt{33})c^2 V}{3n\left(c\sqrt{\frac{(5+\sqrt{33})V}{3n\epsilon}}\right)^2} \\
&= \epsilon.
\end{aligned}$$

The number of individual gradient evaluations is

$$\begin{aligned}
\#grads &= nS + \sum_{s=1}^{S} 2T_s \\
&= 3nS \\
&= 3ns_0 + 3n(S - s_0) \\
&= 3n\left\lceil \log_2 \log_2 4n \right\rceil + 3n\left\lceil c\left(\sqrt{\frac{(5+\sqrt{33})V}{3n\epsilon}} - \frac{15}{8}\right)\right\rceil \\
&= \mathcal{O}\left(n \log\log n + \sqrt{\frac{nV}{\epsilon}}\right).
\end{aligned}$$

□

C. AdaVRAE for known β

In this section, we give a non-adaptive version of our algorithm AdaVRAE. The algorithm is shown in Algorithm 3. The only change is in the step size: we set $\gamma_t^{(s)} = 8\beta$ for all epochs s and iterations t. The analysis readily extends to show the following convergence guarantee:

Algorithm 3 VRAE

Input: initial point $u^{(0)}$, smoothness parameter β.
Parameters: $\{a^{(s)}\}$, $\{T_s\}$, $A_{T_0}^{(0)} > 0$
$\overline{x}_0^{(1)} = z_0^{(1)} = u^{(0)}$, compute $\nabla f(u^{(0)})$
for $s = 1$ **to** S**:**
  $A_0^{(s)} = A_{T_{s-1}}^{(s-1)} - T_s \left(a^{(s)}\right)^2$
  for $t = 1$ **to** T_s**:**
    $x_t^{(s)} = \arg\min_{x\in\mathcal{X}} \left\{ a^{(s)} \left\langle g_{t-1}^{(s)}, x \right\rangle + a^{(s)} h(x) + 4\beta \left\| x - z_{t-1}^{(s)} \right\|^2 \right\}$
    Let $A_t^{(s)} = A_{t-1}^{(s)} + a^{(s)} + \left(a^{(s)}\right)^2$
    $\overline{x}_t^{(s)} = \frac{1}{A_t^{(s)}} \left(A_{t-1}^{(s)} \overline{x}_{t-1}^{(s)} + a^{(s)} x_t^{(s)} + \left(a^{(s)}\right)^2 u^{(s-1)} \right)$
    if $t \neq T_s$**:**
      Pick $i_t^{(s)} \sim \text{Uniform}\left([n]\right)$
      $g_t^{(s)} = \nabla f_{i_t^{(s)}}(\overline{x}_t^{(s)}) - \nabla f_{i_t^{(s)}}(u^{(s-1)}) + \nabla f(u^{(s-1)})$
    else:
      $g_t^{(s)} = \nabla f(\overline{x}_t^{(s)})$
    $z_t^{(s)} = \arg\min_{z\in\mathcal{X}} \left\{ a^{(s)} \left\langle g_t^{(s)}, z \right\rangle + a^{(s)} h(z) + 4\beta \left\| z - z_{t-1}^{(s)} \right\|^2 \right\}$
  $u^{(s)} = \overline{x}_0^{(s+1)} = \overline{x}_{T_s}^{(s)}$, $z_0^{(s+1)} = z_{T_s}^{(s)}$, $g_0^{(s+1)} = g_{T_s}^{(s)}$
return $u^{(S)}$

Theorem C.1. *Let $s_0 = \lceil \log_2 \log_2 4n \rceil$, $c = \frac{3}{2}$. If we choose parameters as follows*

$$a^{(s)} = \begin{cases} (4n)^{-0.5^s} & 1 \le s \le s_0 \\ \frac{s - s_0 - 1 + c}{2c} & s_0 < s \end{cases},$$
$$T_s = n,$$
$$A_{T_0}^{(0)} = \frac{5}{4}.$$

The number of gradient evaluations to achieve a solution $u^{(S)}$ such that $\mathbb{E}\left[F(u^{(S)}) - F(x^*)\right] \le \epsilon$ for Algorithm 3 is

$$\#grads = \begin{cases} \mathcal{O}\left(n \log\log \frac{V}{\epsilon}\right) & \text{if } \epsilon \ge \frac{V}{n} \\ \mathcal{O}\left(n \log\log n + \sqrt{\frac{Vn}{\epsilon}}\right) & \text{if } \epsilon < \frac{V}{n} \end{cases}$$

where $V = \frac{5}{2}\left(F(u^{(0)}) - F(x^*)\right) + 8\beta \left\| u^{(0)} - x^* \right\|^2$.

Proof. Note that Algorithm 3 is essentially the same as Algorithm 1 by choosing $\gamma_t^{(s)} \equiv 8\beta$ with no other changes. Hence the requirements for Lemma A.8 still hold. So we can obtain

$$\mathbb{E}\left[A_{T_S}^{(S)} \left(F(u^{(S)}) - F(x^*) \right) \right] \le A_{T_0}^{(0)} \left(F(u^{(0)}) - F(x^*) \right) + 4\beta \left\| u^{(0)} - x^* \right\|^2.$$

Then by the similar proof in Theorem A.1, we get the desired result. □

D. AdaVRAG for known β

In this section, we give a non-adaptive version of our algorithm AdaVRAG. The algorithm is shown in Algorithm 4. VRAG admits the following convergence guarantee:

Algorithm 4 VRAG

Input: initial point $u^{(0)}$, smoothness parameter β
Parameters: $\{a^{(s)}\}$ where $a^{(s)} \in (0,1)$, $\{T_s\}$
$x_0^{(1)} = u^{(0)}$
for $s = 1$ **to** S**:**
  $\overline{x}_0^{(s)} = a^{(s)} x_0^{(s)} + (1 - a^{(s)}) u^{(s-1)}$, calculate $\nabla f(u^{(s-1)})$
  for $t = 1$ **to** T_s**:**
    Pick $i_t^{(s)} \sim$ Uniform $([n])$
    $g_t^{(s)} = \nabla f_{i_t^{(s)}}(\overline{x}_{t-1}^{(s)}) - \nabla f_{i_t^{(s)}}(u^{(s-1)}) + \nabla f(u^{(s-1)})$
    $x_t^{(s)} = \arg\min_{x \in \mathcal{X}} \left\{ \left\langle g_t^{(s)}, x \right\rangle + h(x) + \frac{\beta\left(2-a^{(s)}\right)a^{(s)}}{2(1-a^{(s)})} \left\| x - x_{t-1}^{(s)} \right\|^2 \right\}$
    $\overline{x}_t^{(s)} = a^{(s)} x_t^{(s)} + (1 - a^{(s)}) u^{(s-1)}$
  $u^{(s)} = \frac{1}{T_s} \sum_{t=1}^{T_s} \overline{x}_t^{(s)}$, $x_0^{(s+1)} = x_{T_s}^{(s)}$
return $u^{(S)}$

Theorem D.1. *(Convergence of VRAG) Define* $s_0 = \lceil \log_2 \log_2 4n \rceil$, $c = \frac{3+\sqrt{33}}{4}$. *If we choose the parameters as follows*

$$a^{(s)} = \begin{cases} 1 - (4n)^{-0.5^s} & 1 \le s \le s_0 \\ \frac{c}{s - s_0 + 2c} & s_0 < s \end{cases},$$
$$T_s = n.$$

The number of individual gradient evaluations to achieve a solution $u^{(S)}$ *such that* $\mathbb{E}\left[F(u^{(S)}) - F(x^*)\right] \le \epsilon$ *for Algorithm 4 is*

$$\#grads = \begin{cases} \mathcal{O}\left(n \log\log \frac{V}{\epsilon}\right) & \epsilon \ge \frac{V}{n} \\ \mathcal{O}\left(n \log\log n + \sqrt{\frac{nV}{\epsilon}}\right) & \epsilon < \frac{V}{n} \end{cases},$$

where

$$V = \frac{1}{2}(F(u^{(0)}) - F(x^*)) + \beta \left\| u^{(0)} - x^* \right\|^2.$$

Before giving the proof of Theorem C.1, we state some intuition on our parameter choice. Note that by defining the following two auxiliary sequences

$$q^{(s)} = \begin{cases} \frac{1}{\left(1-a^{(s)}\right)a^{(s)}} & 1 \le s \le s_0 \\ \frac{\left(2-a^{(s)}\right)a^{(s)}}{1-a^{(s)}} & s_0 < s \end{cases},$$
$$\gamma_{t-1}^{(s)} = \frac{\beta\left(2 - a^{(s)}\right) a^{(s)}}{(1 - a^{(s)}) q^{(s)}}, \forall t \in [T_s],$$

the update rule of $x_t^{(s)}$ in every epoch in Algorithm 4 is equivalent to the update rule of $x_t^{(s)}$ in every epoch in Algorithm 2. Since $\gamma_{t-1}^{(s)}$ is a constant in the corresponding epoch now, we will use $\gamma^{(s)}$ without the subscript to simplify the notation. The above argument means that we can apply Lemma B.3 directly to obtain the following lemma.

Lemma D.2. *For all epochs* $s \ge 1$, *we have*

$$\mathbb{E}\left[F(u^{(s)}) - F(x^*)\right] \le \mathbb{E}\left[\left(1 - a^{(s)}\right)\left(F(u^{(s-1)}) - F(x^*)\right) + \frac{\gamma^{(s)} q^{(s)} a^{(s)}}{2T_s}\left(\left\| x_0^{(s)} - x^* \right\|^2 - \left\| x_0^{(s+1)} - x^* \right\|^2\right)\right].$$

Proof. By applying Lemma B.3, we know

$$\begin{aligned}
&\mathbb{E}\left[F(u^{(s)})-F(x^*)\right]\\
\leq&\mathbb{E}\left[\left(1-a^{(s)}\right)\left(F(u^{(s-1)})-F(x^*)\right)\right]\\
&+\mathbb{E}\left[\frac{1}{T_s}\sum_{t=1}^{T_s}\frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\left(\left\|x_{t-1}^{(s)}-x^*\right\|^2-\left\|x_t^{(s)}-x^*\right\|^2\right)\right]\\
&+\mathbb{E}\left[\frac{1}{T_s}\sum_{t=1}^{T_s}\left(\frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2\left(1-a^{(s)}\right)}-\frac{\gamma_{t-1}^{(s)}q^{(s)}a^{(s)}}{2}\right)\left\|x_t^{(s)}-x_{t-1}^{(s)}\right\|^2\right]\\
\overset{(a)}{=}&\mathbb{E}\left[\left(1-a^{(s)}\right)\left(F(u^{(s-1)})-F(x^*)\right)+\frac{\gamma^{(s)}q^{(s)}a^{(s)}}{2T_s}\sum_{t=1}^{T_s}\left\|x_{t-1}^{(s)}-x^*\right\|^2-\left\|x_t^{(s)}-x^*\right\|^2\right]\\
=&\mathbb{E}\left[\left(1-a^{(s)}\right)\left(F(u^{(s-1)})-F(x^*)\right)+\frac{\gamma^{(s)}q^{(s)}a^{(s)}}{2T_s}\left(\left\|x_0^{(s)}-x^*\right\|^2-\left\|x_{T_s}^{(s)}-x^*\right\|^2\right)\right]\\
\overset{(b)}{=}&\mathbb{E}\left[\left(1-a^{(s)}\right)\left(F(u^{(s-1)})-F(x^*)\right)+\frac{\gamma^{(s)}q^{(s)}a^{(s)}}{2T_s}\left(\left\|x_0^{(s)}-x^*\right\|^2-\left\|x_0^{(s+1)}-x^*\right\|^2\right)\right],
\end{aligned}$$

where (a) is by $\gamma_{t-1}^{(s)}q^{(s)}=\frac{\beta\left(2-a^{(s)}\right)a^{(s)}}{1-a^{(s)}}$ and $\gamma^{(s)}=\gamma_{t-1}^{(s)},\forall t\in[T_s]$, (b) is by $x_0^{(s+1)}=x_{T_s}^{(s)}$. □

Now if we still multiply both sides by $\frac{T_s}{q^{(s)}a^{(s)}}$, we need to ensure that $\gamma^{(s)}$ can help us to make a telescoping sum. However, this is not always true. So we need some different conditions as stated in the following lemma to obtain a bound for the function value gap of $u^{(S)}$. The new bound for the function value gap of $u^{(S)}$ for Algorithm 4 is as follows.

Lemma D.3. *If $\forall s\neq s_0$, we have*

$$\begin{aligned}
a^{(s+1)}&\leq a^{(s)},\\
\frac{(1-a^{(s+1)})T_{s+1}}{q^{(s+1)}a^{(s+1)}}&\leq\frac{T_s}{q^{(s)}a^{(s)}}.
\end{aligned}$$

Additionally, for s_0, assume we have

$$\frac{\left(1-a^{(s_0+1)}\right)^2T_{s_0+1}}{\left(2-a^{(s_0+1)}\right)\left(a^{(s_0+1)}\right)^2}\leq\frac{(1-a^{(s_0)})T_{s_0}}{\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2}.$$

Then for $S\leq s_0$,

$$\mathbb{E}\left[\frac{T_S}{q^{(S)}a^{(S)}}\left(F(u^{(S)})-F(x^*)\right)\right]\leq\frac{\left(1-a^{(1)}\right)T_1}{q^{(1)}a^{(1)}}\left(F(u^{(0)})-F(x^*)\right)+\frac{\beta}{2}\left\|u^{(0)}-x^*\right\|^2.$$

For $S>s_0$,

$$\mathbb{E}\left[\frac{\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2T_S}{q^{(S)}a^{(S)}}\left(F(u^{(S)})-F(x^*)\right)\right]\leq\frac{\left(1-a^{(1)}\right)T_1}{q^{(1)}a^{(1)}}\left(F(u^{(0)})-F(x^*)\right)+\frac{\beta}{2}\left\|u^{(0)}-x^*\right\|^2$$

Proof. By applying Lemma D.2 and multiply both sides by $\frac{T_s}{q^{(s)}a^{(s)}}$,we have

$$\mathbb{E}\left[\frac{T_s}{q^{(s)}a^{(s)}}\left(F(u^{(s)})-F(x^*)\right)\right]\leq\mathbb{E}\left[\frac{\left(1-a^{(s)}\right)T_s}{q^{(s)}a^{(s)}}\left(F(u^{(s-1)})-F(x^*)\right)+\frac{\gamma^{(s)}}{2}\left(\left\|x_0^{(s)}-x^*\right\|^2-\left\|x_0^{(s+1)}-x^*\right\|^2\right)\right].$$

For $S \leq s_0$

$$
\begin{aligned}
&\mathbb{E}\left[\frac{T_S}{q^{(S)}a^{(S)}}\left(F(u^{(S)})-F(x^*)\right)\right]\\
\leq&\mathbb{E}\left[\frac{\left(1-a^{(1)}\right)T_1}{q^{(1)}a^{(1)}}\left(F(u^{(0)})-F(x^*)\right)+\sum_{s=1}^{S}\frac{\gamma^{(s)}}{2}\left(\left\|x_0^{(s)}-x^*\right\|^2-\left\|x_0^{(s+1)}-x^*\right\|^2\right)\right]\\
\overset{(a)}{=}&\frac{\left(1-a^{(1)}\right)T_1}{q^{(1)}a^{(1)}}\left(F(u^{(0)})-F(x^*)\right)+\mathbb{E}\left[\sum_{s=1}^{S}\frac{\beta\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2}{2}\left(\left\|x_0^{(s)}-x^*\right\|^2-\left\|x_0^{(s+1)}-x^*\right\|^2\right)\right]\\
=&\frac{\left(1-a^{(1)}\right)T_1}{q^{(1)}a^{(1)}}\left(F(u^{(0)})-F(x^*)\right)+\frac{\beta\left(2-a^{(1)}\right)\left(a^{(1)}\right)^2}{2}\left\|x_0^{(1)}-x^*\right\|^2\\
&+\mathbb{E}\left[\sum_{s=1}^{S-1}\frac{\beta\left[\left(2-a^{(s+1)}\right)\left(a^{(s+1)}\right)^2-\left(2-a^{(s)}\right)\left(a^{(s)}\right)^2\right]}{2}\left(\left\|x_0^{(s+1)}-x^*\right\|^2\right)\right]\\
&-\mathbb{E}\left[\frac{\beta\left(2-a^{(S)}\right)\left(a^{(S)}\right)^2}{2}\left(\left\|x_0^{(S+1)}-x^*\right\|^2\right)\right]\\
\overset{(b)}{\leq}&\frac{\left(1-a^{(1)}\right)T_1}{q^{(1)}a^{(1)}}\left(F(u^{(0)})-F(x^*)\right)+\frac{\beta}{2}\left\|u^{(0)}-x^*\right\|^2\\
&-\mathbb{E}\left[\frac{\beta\left(2-a^{(S)}\right)\left(a^{(S)}\right)^2}{2}\left(\left\|x_0^{(S+1)}-x^*\right\|^2\right)\right],
\end{aligned}
$$

where (a) is by the definition of $\gamma^{(s)}$ when $s \leq s_0$, (b) is by $\left(2-a^{(1)}\right)\left(a^{(1)}\right)^2 \leq 1$ and $x_0^{(1)} = u^{(0)}$, additionally, note that our assumption $a^{(s+1)} \leq a^{(s)} \Rightarrow \left(2-a^{(s+1)}\right)\left(a^{(s+1)}\right)^2 \leq \left(2-a^{(s)}\right)\left(a^{(s)}\right)^2$.

For $S > s_0$, we can also make the telescoping sum from $s = s_0+1$ to S by a similar argument to get

$$
\mathbb{E}\left[\frac{T_S}{q^{(S)}a^{(S)}}\left(F(u^{(S)})-F(x^*)\right)\right] \leq \mathbb{E}\left[\frac{\left(1-a^{(s_0+1)}\right)T_{s_0+1}}{q^{(s_0+1)}a^{(s_0+1)}}\left(F(u^{(s_0)})-F(x^*)\right)+\frac{\beta}{2}\left\|x_0^{(s_0+1)}-x^*\right\|^2\right].
$$

Multiplying both sides by $\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2 T_S}{q^{(S)}a^{(S)}}\left(F(u^{(S)})-F(x^*)\right)\right]\\
\leq&\mathbb{E}\left[\frac{\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2\left(1-a^{(s_0+1)}\right)T_{s_0+1}}{q^{(s_0+1)}a^{(s_0+1)}}\left(F(u^{(s_0)})-F(x^*)\right)+\frac{\beta\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2}{2}\left\|x_0^{(s_0+1)}-x^*\right\|^2\right]\\
\overset{(c)}{=}&\mathbb{E}\left[\frac{\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2\left(1-a^{(s_0+1)}\right)^2 T_{s_0+1}}{\left(2-a^{(s_0+1)}\right)\left(a^{(s_0+1)}\right)^2}\left(F(u^{(s_0)})-F(x^*)\right)+\frac{\beta\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2}{2}\left\|x_0^{(s_0+1)}-x^*\right\|^2\right],
\end{aligned}
$$

where (c) is by the definition $q^{(s_0+1)} = \frac{\left(2-a^{(s_0+1)}\right)a^{(s_0+1)}}{1-a^{(s_0+1)}}$. Note that by our assumption

$$
\begin{aligned}
\frac{\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2\left(1-a^{(s_0+1)}\right)^2 T_{s_0+1}}{\left(2-a^{(s_0+1)}\right)\left(a^{(s_0+1)}\right)^2} &\leq (1-a^{(s_0)})T_{s_0}\\
&= \frac{T_{s_0}}{q^{(s_0)}a^{(s_0)}},
\end{aligned}
$$

so we know

$$\begin{aligned}&\mathbb{E}\left[\frac{\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2 T_S}{q^{(S)}a^{(S)}}\left(F(u^{(S)})-F(x^*)\right)\right]\\ \leq&\mathbb{E}\left[\frac{T_{s_0}}{q^{(s_0)}a^{(s_0)}}\left(F(u^{(s_0)})-F(x^*)\right)+\frac{\beta\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2}{2}\left\|x_0^{(s_0+1)}-x^*\right\|^2\right].\end{aligned}$$

Now combining

$$\begin{aligned}\mathbb{E}\left[\frac{T_{s_0}}{q^{(s_0)}a^{(s_0)}}\left(F(u^{(s_0)})-F(x^*)\right)\right]\leq&\frac{\left(1-a^{(1)}\right)T_1}{q^{(1)}a^{(1)}}\left(F(u^{(0)})-F(x^*)\right)+\frac{\beta}{2}\left\|u^{(0)}-x^*\right\|^2\\ &-\mathbb{E}\left[\frac{\beta\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2}{2}\left(\left\|x_0^{(s_0+1)}-x^*\right\|^2\right)\right],\end{aligned}$$

we have

$$\mathbb{E}\left[\frac{\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2 T_S}{q^{(S)}a^{(S)}}\left(F(u^{(S)})-F(x^*)\right)\right]\leq\frac{\left(1-a^{(1)}\right)T_1}{q^{(1)}a^{(1)}}\left(F(u^{(0)})-F(x^*)\right)+\frac{\beta}{2}\left\|u^{(0)}-x^*\right\|^2.$$

□

Using the above new lemma w.r.t. the function value gap of $u^{(S)}$, we finally can give the proof of Theorem D.1.

Proof. (Theorem D.1) Note that by our choice $a^{(s+1)}\leq a^{(s)}$ is true for any $s\neq s_0$. Besides, our parameters $\left\{a^{(s)}\right\}$ and $\left\{q^{(s)}\right\}$ are totally the same as the choice in Theorem B.1 when $s\leq s_0$. Hence we know

$$\frac{(1-a^{(s+1)})T_{s+1}}{q^{(s+1)}a^{(s+1)}}\leq\frac{T_s}{q^{(s)}a^{(s)}}$$

is still true for $s\leq s_0-1$. For $s\geq s_0+1$, note that our new $\left\{q^{(s)}\right\}$ are only different from the choice in Theorem B.1 by a constant, which implies

$$\frac{(1-a^{(s+1)})T_{s+1}}{q^{(s+1)}a^{(s+1)}}\leq\frac{T_s}{q^{(s)}a^{(s)}}$$

also holds for $s\geq s_0+1$. Besides, we can show

$$\frac{\left(1-a^{(s_0+1)}\right)^2 T_{s_0+1}}{\left(2-a^{(s_0+1)}\right)\left(a^{(s_0+1)}\right)^2}\leq\frac{(1-a^{(s_0)})T_{s_0}}{\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2}$$

is true by plugging in the value of $a^{(s_0+1)}=\frac{c}{1+2c}$ and noticing that $a^{(s_0)}\leq\frac{1}{2}$. Hence all the conditions for Lemma D.3 are satisfied, then we know for $S\leq s_0$,

$$\mathbb{E}\left[\frac{T_S}{q^{(S)}a^{(S)}}\left(F(u^{(S)})-F(x^*)\right)\right]\leq\frac{\left(1-a^{(1)}\right)T_1}{q^{(1)}a^{(1)}}\left(F(u^{(0)})-F(x^*)\right)+\frac{\beta}{2}\left\|u^{(0)}-x^*\right\|^2.$$

For $S>s_0$,

$$\mathbb{E}\left[\frac{\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2 T_S}{q^{(S)}a^{(S)}}\left(F(u^{(S)})-F(x^*)\right)\right]\leq\frac{\left(1-a^{(1)}\right)T_1}{q^{(1)}a^{(1)}}\left(F(u^{(0)})-F(x^*)\right)+\frac{\beta}{2}\left\|u^{(0)}-x^*\right\|^2.$$

By noticing

$$\begin{aligned}a^{(s_0)}&=1-(4n)^{-0.5^{s_0}}\\ &\geq 1-(4n)^{-0.5^{(\log_2\log_2 4n)+1}}\\ &=1-\frac{1}{\sqrt{2}}\\ \Rightarrow\left(2-a^{(s_0)}\right)\left(a^{(s_0)}\right)^2&\geq\frac{2-\sqrt{2}}{4}\end{aligned}$$

and

$$\frac{\left(1-a^{(1)}\right)T_1}{q^{(1)}a^{(1)}}=\frac{1}{4},$$

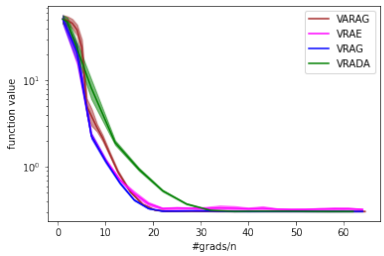
combining the fact that our new $\left\{q^{(s)}\right\}$ for $S>s_0$ have the same order of the choice in Theorem B.1. Following a similar proof, we can arrive the desired result. □

E. Hyperparameter choices and additional results

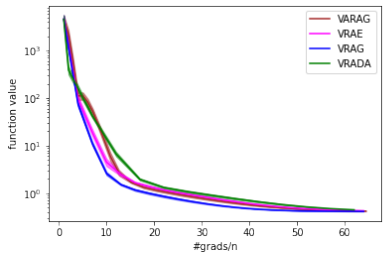
Table 2 reports the hyperparameter choices used in the experiments. VRAG and VRAE are the non-adaptive versions our algorithms (Algorithms 3 and 4). We set their step sizes via a hyperparameter search as described in Section 3. Figures 5, 6, 7, 8 give the experimental evaluation of our non-adaptive algorithms.

Table 2. Hyperparameters used in the experiments

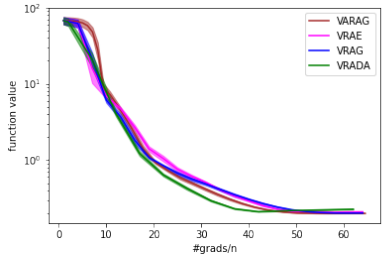
Dataset	Loss	SVRG	SVRG^{++}	VARAG	VRADA	VRAG	VRAE
a1a	logistic	0.5	0.5	1	1	1	1
	squared	0.01	0.05	0.05	0.1	0.1	0.05
	huber	0.05	0.1	0.1	0.5	0.1	0.1
mushrooms	logistic	0.5	1	1	1	1	1
	squared	0.01	0.01	0.05	0.1	0.05	0.01
	huber	0.05	0.1	0.1	0.1	0.1	0.05
w8a	logistic	0.1	1	1	100	1	5
	squared	0.01	0.01	0.01	100	0.05	0.05
	huber	0.01	0.1	0.1	100	0.1	0.5
phishing	logistic	50	100	100	100	100	100
	squared	0.05	0.5	1	1	1	1
	huber	0.5	1	1	5	5	5



(a) Logistic loss

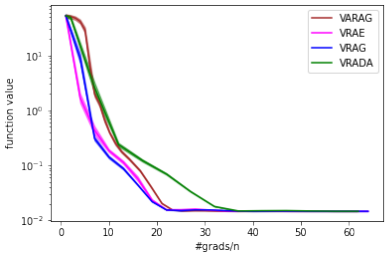


(b) Squared loss

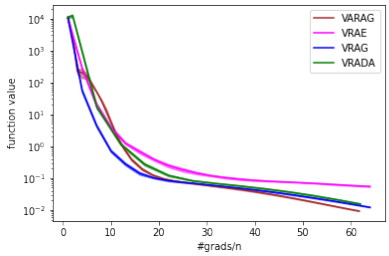


(c) Huber loss

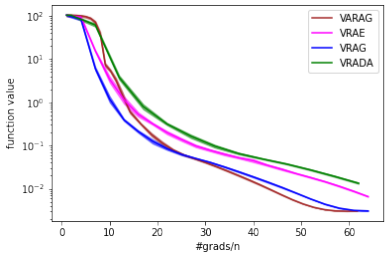
Figure 5. a1a



(a) Logistic loss

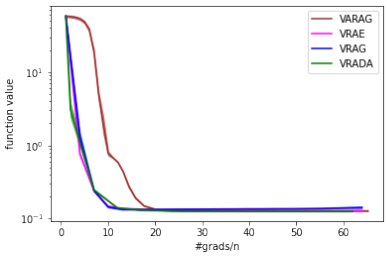


(b) Squared loss

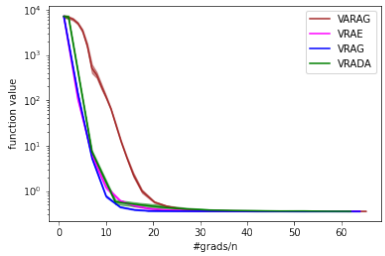


(c) Huber loss

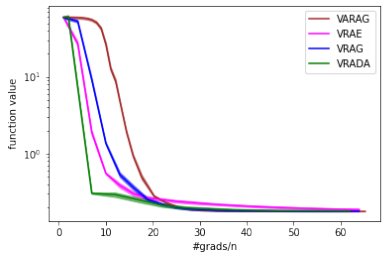
Figure 6. mushrooms



(a) Logistic loss

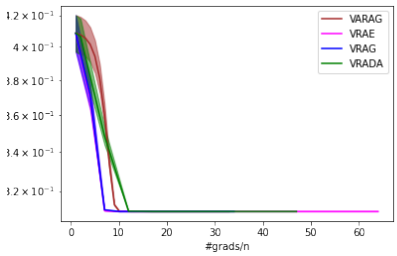


(b) Squared loss

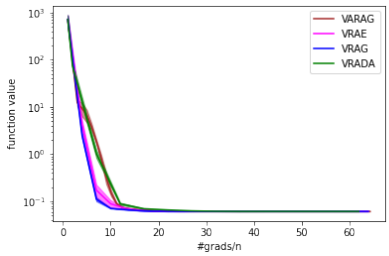


(c) Huber loss

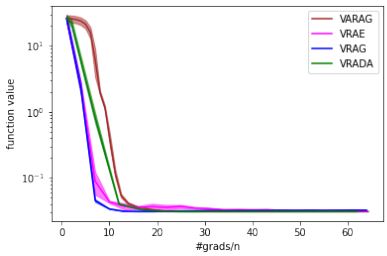
Figure 7. w8a



(a) Logistic loss



(b) Squared loss



(c) Huber loss

Figure 8. phishing