History-Gradient Aided Batch Size Adaptation for Variance Reduced Algorithms

Kaiyi Ji*, Zhe Wang†, Bowen Weng‡, Yi Zhou§, Wei Zhang¶ and Yingbin Liang‖

February 17, 2020

Abstract

Variance-reduced algorithms, although achieve great theoretical performance, can run slowly in practice due to the periodic gradient estimation with a large batch of data. Batch-size adaptation thus arises as a promising approach to accelerate such algorithms. However, existing schemes either apply prescribed batch-size adaption rule or exploit the information along optimization path via additional backtracking and condition verification steps. In this paper, we propose a novel scheme, which eliminates backtracking line search but still exploits the information along optimization path by adapting the batch size via history stochastic gradients. We further theoretically show that such a scheme substantially reduces the overall complexity for popular variance-reduced algorithms SVRG and SARAH/SPIDER for both conventional nonconvex optimization and reinforcement learning problems. To this end, we develop a new convergence analysis framework to handle the dependence of the batch size on history stochastic gradients. Extensive experiments validate the effectiveness of the proposed batch-size adaptation scheme.

1 Introduction

Stochastic gradient descent (SGD) (Ghadimi & Lan, 2013) algorithms have been extensively used to efficiently solve large-scale optimization problems recently. Furthermore, various variance reduced algorithms such as SAGA (Defazio et al., 2014), SVRG (Johnson & Zhang, 2013; Reddi et al., 2016a), SARAH (Nguyen et al., 2017a,b), and SPIDER (Fang et al., 2018)/SpiderBoost (Wang et al., 2019), have been proposed to reduce the variance of SGD. Such variance reduction techniques have also been applied to policy gradient algorithms to develop SVRPG (Papini et al., 2018), SRVR-PG (Xu et al., 2019b) and SARAPO (Yuan et al., 2018) in reinforcement learning (RL). Though variance reduced algorithms have been shown to have order-level lower computational complexity than SGD (and than vanilla policy gradient in RL), they often do not perform as well as SGD in practice, largely due to the periodic large-batch gradient estimation. In fact, variance-reduced gradient estimation plays an important role only towards the later stage of the algorithm execution, and hence a promising way to accelerate variance reduced algorithms is to adaptively increase the batch size.

Two types of batch-size adaptation schemes have been proposed so far to accelerate stochastic algorithms (Smith et al., 2018; Friedlander & Schmidt, 2012; Devarakonda et al., 2017). The first approach follows a *prescribed* rule to adapt the batch size, which can be *exponential* and *polynomial* increase of batch size as in hybrid SGD (HSGD) (Friedlander & Schmidt, 2012; Zhou et al., 2018b) and *linear* increase of batch size (Zhou et al., 2018b). Moreover, Harikandeh et al. 2015 and Lei & Jordan 2019 proposed to use

*Department of Electrical and Computer Engineering, The Ohio State University; e-mail: ji.367@osu.edu
†Department of Electrical and Computer Engineering, The Ohio State University; e-mail: wang.10982@osu.edu
‡Department of Electrical and Computer Engineering, The Ohio State University; e-mail: weng.172@osu.edu
§Department of Electrical and Computer Engineering, University of Utah; e-mail: yi.zhou@utah.edu
¶Department of Mechanical Engineering, Southern University of Science and Technology; e-mail: zhangw3@sustech.edu.cn
‖Department of Electrical and Computer Engineering, The Ohio State University; e-mail: liang.889@osu.edu

exponential increase of batch size at each outer-loop iteration respectively for SVRG and for an adaptively sampled variance reduced algorithm SCSG (Lei et al., 2017).

The second approach adapts the batch size based on the information along the optimization path. For example, De et al. 2016, 2017 proposed Big Batch SGD, which adapts the batch size so that the resulting gradient and variance satisfy certain optimization properties. Since the batch size needs to be chosen *even before* its resulting gradient is calculated, the algorithm adopts the backtracking line search to iteratively check that the chosen batch size ensures the resulting gradient to satisfy a variance bound. Clearly, the backtracking step adds undesired complexity, but seems to be unavoidable, because the convergence analysis exploits the *instantaneous* variance bound.

Our contribution lies in designing an easy-to-implement scheme to adapt the batch size, which incorporates the information along the optimization path, but does not involve backtracking and condition verification. We further show by both theory and experiments that such a scheme achieves much better performance than vanilla variance reduced algorithms in both conventional optimization and RL problems.

1.1 Our Contributions

New batch-size adaptation scheme via history gradients: We propose to adapt the batch size of each epoch (i.e., each outer loop) of variance reduced algorithms SVRG and SPIDER inversely proportional to the average of stochastic gradients over each epoch, and call the algorithms as **A**daptive **ba**tch-size SVRG (AbaSVRG) and AbaSPIDER. We further apply the scheme to the variance reduced policy gradient algorithms SVRPG and SPIDER-PG (which refers to SRVR-PG in Xu et al. 2019b) in RL, and call the resulting algorithms as AbaSVRPG and AbaSPIDER-PG. These algorithms initially use small batch size (due to large gradients) and enjoy fast iteration, and then gradually increase the batch size (due to the reduced gradients) and enjoy reduced variance and stable convergence. Further technical justification is provided in Section 2.1.

Since the batch size should be set at the beginning of the epoch at which point the gradients in that epoch has not been calculated yet. It is a similar situation as in De et al. 2016, 2017, which introduced the backtracking line search to guarantee the variance bound. Here, we propose to use the average of stochastic gradients over the *preceding* epoch as an approximation of the present gradient information to avoid the complexity of backtracking line search, which we further show theoretically to still achieve guaranteed improved performance.

New convergence analysis: Since the updates in our algorithms depend on the past gradients, it becomes much more challenging to establish the provable convergence guarantee. The technical novelty of our analysis mainly lies in the following two aspects.

- We develop a new framework to analyze the convergence of variance reduced algorithms with batch size adapted to history gradients for nonconvex optimization. In particular, we bound the function values for each epoch by the average gradient in the preceding epoch due to the batch size dependence, which further facilitates the bounding of the accumulative change of the objective value over the entire execution of the algorithm. Such an analysis is quite different from the existing analysis of SVRG in Reddi et al. 2016a; Li & Li 2018 and SPIDER/SpiderBoost in Fang et al. 2018; Wang et al. 2019, which are based on guaranteeing the decrease of the objective value iterationwisely or epochwisely.

- We develop a different and simpler convergence analysis for SVRG-type algorithms than previous studies (Reddi et al., 2016a; Li & Li, 2018) for nonconvex optimization, which allows more flexible choices of parameters. More importantly, such an analysis fits well to the analysis framework we develop to handle the dependence of the adaptive batch size on the stochastic gradients in the previous epoch, whereas existing techniques do not seem to accommodate easily.

Based on the new analysis framework, we show that both AbaSVRG and AbaSPIDER for conventional nonconvex optimization and AbaSVRPG and AbaSPIDER-PG for policy optimization in RL achieve improved

Algorithm 1 AbaSVRG

Input: x_0, m, B, η, $c_\beta, c_\epsilon, \beta_1 > 0$
$\tilde{x}^0 = x_0$
for $s = 1, 2, ..., S$ **do**
  $x_0^s = \tilde{x}^{s-1}$
  Sample $\mathcal{N}_s$ from $[n]$ without replacement, where $N_s = \min\{c_\beta \sigma^2 \beta_s^{-1}, c_\epsilon \sigma^2 \epsilon^{-1}, n\}$
  $g^s = \nabla f_{\mathcal{N}_s}(\tilde{x}^{s-1})$
  Set $\beta_{s+1} = 0$
  for $t = 1, 2, ..., m$ **do**
    Sample $\mathcal{B}$ from $[n]$ with replacement
    $v_{t-1}^s = \nabla f_{\mathcal{B}}(x_{t-1}^s) - \nabla f_{\mathcal{B}}(\tilde{x}^{s-1}) + g^s$
    $x_t^s = x_{t-1}^s - \eta v_{t-1}^s$
    $\beta_{s+1} \leftarrow \beta_{s+1} + \|v_{t-1}^s\|^2/m$
  end for
  $\tilde{x}^s = x_m^s$
end for
Output: x_ζ from $\{x_{t-1}^s\}_{s\in[S],t\in[m]}$ uniformly at random

Algorithm 2 AbaSPIDER

Input: x_0, m, B, η, $c_\beta, c_\epsilon, \beta_1 > 0$
$\tilde{x}^0 = x_0$
for $s = 1, 2, ..., S$ **do**
  $x_0^s = \tilde{x}^{s-1}$
  Sample $\mathcal{N}_s$ from $[n]$ without replacement, where $N_s = \min\{c_\beta \sigma^2 \beta_s^{-1}, c_\epsilon \sigma^2 \epsilon^{-1}, n\}$
  $v_0^s = \nabla f_{\mathcal{N}_s}(\tilde{x}^{s-1})$
  $x_1^s = x_0^s - \eta v_0^s$
  Set $\beta_{s+1} = \|v_0^s\|^2/m$
  for $t = 1, 2, ..., m-1$ **do**
    Sample $\mathcal{B}$ from $[n]$ with replacement
    $v_t^s = \nabla f_{\mathcal{B}}(x_t^s) - \nabla f_{\mathcal{B}}(x_{t-1}^s) + v_{t-1}^s$
    $x_{t+1}^s = x_t^s - \eta v_t^s$
    $\beta_{s+1} \leftarrow \beta_{s+1} + \|v_t^s\|^2/m$
  end for
  $\tilde{x}^s = x_m^s$
end for
Output: x_ζ from $\{x_{t-1}^s\}_{s\in[S],t\in[m]}$ uniformly at random

complexity over their corresponding vanilla counterpart (without batch-size adaptation). The worst-case complexity of these algorithms all match the best known complexity. We also provide the convergence analysis of AbaSVRG and AbaSPIDER for nonconvex problems under the PŁ condition in Appendix C.

Experiments: We provide extensive experiments on both supervised learning and RL problems and demonstrate that the proposed adaptive batch-size scheme substantially speeds up the convergence of variance reduced algorithms.

1.2 Related Work

Variance reduced algorithms for conventional optimization. In order to improve the performance of SGD (Robbins & Monro, 1951), various variance reduced algorithms have been proposed such as SAG (Roux et al., 2012), SAGA (Defazio et al., 2014), SVRG (Allen-Zhu & Hazan, 2016; Johnson & Zhang, 2013), SARAH (Nguyen et al., 2017a,b, 2019), SNVRG (Zhou et al., 2018a), SPIDER (Fang et al., 2018), SpiderBoost (Wang et al., 2019). This paper shows that two representative algorithms SVRG and SPIDER can be equipped with the proposed adaptive batch size and attain substantial performance gain.

Variance reduced policy gradient for RL: Variance reduction methods have also been applied to policy gradient methods (S. Sutton et al., 2000) in RL. One way is to incorporate a baseline in the gradient estimator, e.g., Williams 1992; Weaver & Tao 2001; Wu et al. 2018. Optimization techniques have also been applied. For example, Papini et al. 2018; Xu et al. 2019a applied SVRG to develop stochastic variance reduced policy gradient (SVRPG) algorithm. Yuan et al. 2018 and Xu et al. 2019b applied SARAH/SPIDER to develop stochastic recursive gradient policy optimization (SARAPO) and stochastic recursive variance reduced policy gradient (SRVR-PG), respectively. Shen et al. 2019 developed Hessian aided policy gradient (HAPG). This paper shows that the batch size adaptation scheme can also be applied to variance reduced policy gradient algorithms to significantly improve their performance.

Stochastic algorithms with adaptive batch size. Adaptively changing the batch size emerges as a powerful approach for accelerating stochastic algorithms (Smith et al., 2018; Friedlander & Schmidt, 2012; Devarakonda et al., 2017). Hybrid SGD (HSGD) applies *exponential* and *polynomial* increase of batch size (Friedlander & Schmidt, 2012; Zhou et al., 2018b) and *linear* increase of batch size (Zhou et al., 2018b).

De et al. 2016, 2017 proposed Big Batch SGD with the batch size adaptive to the *instantaneous* gradient and variance information (which needs to be ensured, e.g., by backtracking line search) at each iteration. Moreover, Harikandeh et al. 2015 and Lei & Jordan 2019 proposed to use exponential increase of batch size at each outer-loop iteration respectively for SVRG and for an adaptively sampled variance reduced algorithm SCSG (Lei et al., 2017). Our algorithms adapt the batch size to history gradients, which differs from the prescribed adaptive schemes, and is easier to implement than Big Batch SGD by eliminating backtraking line search and still guarantees the convergence.

We note that a concurrent work (Sievert & Charles, 2019) also proposed an improved SGD algorithm by adapting the batch size to history gradients, but only as a practice without convergence proof. Our analysis framework is applicable to their algorithm as we show in Appendix D.

Notations. Let $\wedge$ and $\vee$ denote the minimum and the maximum. Let $[n] := \{1, ..., n\}$. For a set $\mathcal{S}$, let S be its cardinality and define $\nabla f_{\mathcal{S}}(\cdot) := \frac{1}{S}\sum_{i\in\mathcal{S}} \nabla f_i(\cdot)$.

2 Batch Size Adaptation for Nonconvex Optimization

In this section, we consider the following finite-sum optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x). \tag{P}$$

In the context of machine learning problems, each function $f_i(\cdot)$ evaluates the loss on a particular i-th data sample, and is generally nonconvex due to the complex models.

2.1 Proposed Algorithms with Batch-Size Adaptation

Two popular variance reduced algorithms to solve the optimization problem (P) are SVRG (Johnson & Zhang, 2013) and SARAH (Nguyen et al., 2017a)/SPIDER (Fang et al., 2018), which have been shown to outperform SGD. However, SVRG and SARAH/SPIDER often run slowly in practice due to the full/large-batch gradient evaluation at the beginning of each epoch. We propose a batch-size adaptation scheme to mitigate such an issue for these algorithms, and we call the corresponding algorithms as AbaSVRG and AbaSPIDER (see Algorithms 1 and 2). Note that AbaSPIDER adopts the improved version SpiderBoost (Wang et al., 2019) of the original SPIDER (Fang et al., 2018).

We take SVRG as an example to briefly explain our idea. Our analysis of SVRG shows that the decrease of the average function value over an epoch s with length m satisfies

$$\frac{\mathbb{E}(f(\tilde{x}^{s+1}) - f(\tilde{x}^s))}{m} \le -\phi\frac{\sum_{t=0}^{m-1}\mathbb{E}\|v_t\|^2}{m} + \frac{\psi I_{(N_s<n)}}{N_s}$$

where $\phi, \psi > 0$ are constants, $I_{(\cdot)}$ is the indicator function, $\tilde{x}^s$ is the snapshot in epoch s, v_t is a stochastic estimation of $\nabla f(x_t)$ within epoch s, and N_s is the batch size used at the outer-loop iteration. The above bound naturally suggests that N_s should be chosen such that the second term is at the same level as the first term, i.e., the batch size should adapt to the average stochastic gradient over the epoch, in which case the convergence guarantee follows easily. However, this is not feasible in practice, because the batch size should be chosen at the beginning of each epoch, at which point the gradients in the same loop have not been calculated yet. Such an issue was previously solved in De et al. 2016, 2017 via backtracking line search, which adds significantly additional complexity. Our main idea here is to use the stochastic gradients calculated in the previous epoch for adapting the batch size of the coming loop, and we show that such a scheme still retains the convergence guarantee and achieves improved computational complexity.

More precisely, AbaSVRG/AbaSPIDER chooses the batch size N_s at epoch s adaptively to the average β_s of stochastic gradients in the *preceding* epoch $s-1$ as given below

$$N_s = \min\{c_\beta \sigma^2 \beta_s^{-1}, c_\epsilon \sigma^2 \epsilon^{-1}, n\}, \beta_s = \frac{\sum_{t=1}^m \|v_{t-1}^{s-1}\|^2}{m},$$

where $c_\beta, c_\epsilon > 0$ are constants and σ^2 is the variance bound. As a comparison, the vanilla SVRG and SPIDER pick a *fixed* batch size to be either n or $\min\{c_\epsilon \sigma^2 \epsilon^{-1}, n\}$.

2.2 Assumptions and Definitions

We adopt the following standard assumptions (Lei et al., 2017; Reddi et al., 2016a) for convergence analysis.

Assumption 1. *The objective function in (P) satisfies:*
(1) $\nabla f_i(\cdot)$ *is L-smooth for* $i \in [n]$*, i.e., for any* $x, y \in \mathbb{R}^d$*,* $\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x - y\|$.

(2) $f(\cdot)$ *is bounded below, i.e.,* $f^* = \inf_{x \in \mathbb{R}^d} f(x) > -\infty$.

(3) $\nabla f_i(\cdot)$ *(with the index* i *uniformly randomly chosen) has bounded variance, i.e., there exists a constant* $\sigma > 0$ *such that for any* $x \in \mathbb{R}^d$*,* $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f(x)\|^2 \le \sigma^2$.

The item (3) of the bounded variance assumption is commonly adopted for proving the convergence of SGD-type algorithms (e.g., SGD (Ghadimi & Lan, 2013)) and stochastic variance reduced methods (e.g., SCSG (Lei et al., 2017)) that draw a sample batch with size less than n for gradient estimation at each outer-loop iteration.

In this paper, we use the gradient norm as the convergence criterion for nonconvex optimization.

Definition 1. *We say that* x^ζ *is an* ϵ*-accurate solution for the optimization problem (P) if* $\mathbb{E}\|\nabla f(x^\zeta)\|^2 \le \epsilon$*, where* x^ζ *is an output returned by a stochastic algorithm.*

To compare the efficiency of different stochastic algorithms, we adopt the following stochastic first-order oracle (SFO) for the analysis of the computational complexity.

Definition 2. *Given an input* $x \in \mathbb{R}^d$*, SFO randomly takes an index* $i \in [n]$ *and returns a stochastic gradient* $\nabla f_i(x)$ *such that* $\mathbb{E}_i\big[\nabla f_i(x)\big] = \nabla f(x)$.

2.3 Convergence Analysis for AbaSVRG

Since the batch size of AbaSVRG is adaptive to the *history* gradients due to the component $c_\beta \sigma^2 \beta_s^{-1}$, the existing convergence analysis for SVRG type of algorithms in Li & Li 2018; Reddi et al. 2016a does not extend easily. Here, we develop a new analysis framework for SVRG algorithms which is simpler than that in Li & Li 2018; Reddi et al. 2016a (and can be of independent interest), and enables to handle the dependence of the batch size on the stochastic gradients in the past epoch in the convergence analysis for AbaSVRG. To compare more specifically, Reddi et al. 2016a introduced a Lyapunov function $R_t^s = \mathbb{E}[f(x_t^s) + c_t\|x_t^s - \tilde{x}^{s-1}\|^2]$ and proves that R^s decreases by the accumulated gradient norms $\sum_{t=0}^{m-1} \mathbb{E}\|\nabla f(x_t^s)\|^2$ within an epoch s, and Li & Li 2018 directly showed that $\mathbb{E}f(x^s)$ decreases by $\sum_{t=0}^{m-1} \mathbb{E}\|\nabla f(x_t^s)\|^2$ using tighter bounds. Both studies adopted Young's inequality, which involves more parameters required to be carefully tuned based on the relationship between B and m. As a comparison, our analysis shows that $\mathbb{E}f(x^s)$ decreases by the accumulated *stochastic gradient* norms $\sum_{t=0}^{m-1} \mathbb{E}\|v_t^s\|^2$, and does not rely on Young's inequality and extra tuning parameters. More details about our proof can be found in Appendix E. The following theorem provides a general convergence result for AbaSVRG.

Theorem 1. *Suppose Assumption 1 is satisfied. Let $\epsilon > 0$ and $c_\beta, c_\epsilon \geq \alpha$ for certain constant $\alpha > 0$. Let $\psi = \frac{2\eta^2 L^2 m^2}{B} + \frac{2}{\alpha} + 2$ and choose β_1, α and η such that $\beta_1 \leq \epsilon S$ and $\phi = \frac{1}{2} - \frac{1}{2\alpha} - \frac{L\eta}{2} - \frac{\eta^2 L^2 m^2}{2B} > 0$, where S denotes the number of epochs. Then, the output x_ζ returned by AbaSVRG satisfies*

$$\mathbb{E}\|\nabla f(x_\zeta)\|^2 \leq \frac{\psi(f(x_0) - f^*)}{\phi\eta K} + \frac{\psi\epsilon}{\phi\alpha} + \frac{4\epsilon}{\alpha},$$

where $f^* = \inf_{x\in\mathbb{R}^d} f(x)$ and $K = Sm$ represents the total number of iterations.

Theorem 1 guarantees the convergence of AbaSVRG as long as ϕ is positive, i.e. $\frac{L\eta}{2} + \frac{\eta^2 L^2 m^2}{B} \leq \frac{1}{2} - \frac{1}{2\alpha}$, and thus allows very flexible choices of the stepsize η, the epoch length m and the mini-batch size B. Such flexibility and generality are also due to the aforementioned simpler proof that we develop for SVRG-type algorithms.

In the following corollary, we provide the complexity performance of AbaSVRG under certain choices of parameters.

Corollary 1. *Under the setting of Theorem 1, we choose the constant stepsize $\eta = \frac{1}{4L}$, the epoch length $m = \sqrt{B}$ (which B denotes the mini-batch size) and $c_\beta, c_\epsilon \geq 16$. Then, to achieve $\mathbb{E}\|\nabla f(x^\zeta)\|^2 \leq \epsilon$, the total SFO complexity of AbaSVRG is given by*

$$\underbrace{\sum_{s=1}^{S} \min\left\{\frac{c_\beta\sigma^2}{\sum_{t=1}^m \|v_{t-1}^{s-1}\|^2/m}, c_\epsilon\sigma^2\epsilon^{-1}, n\right\} + KB}_{\text{complexity of AbaSVRG}} < \underbrace{S\min\left\{c_\epsilon\sigma^2\epsilon^{-1}, n\right\} + KB}_{\text{complexity of vanilla SVRG}} = \mathcal{O}\Big(\frac{n\wedge\epsilon^{-1}}{\sqrt{B}\epsilon} + \frac{B}{\epsilon}\Big).$$

If we choose $B = n^{2/3}\wedge\epsilon^{-2/3}$, then the worst-case complexity is given by $\mathcal{O}\left(\epsilon^{-1}(n\wedge\epsilon^{-1})^{2/3}\right)$.

We make the following remarks on Corollary 1.

First, the worst-case SFO complexity under the specific choice of $B = n^{2/3}\wedge\epsilon^{-2/3}$ matches the best known result for SVRG-type algorithms. More importantly, since the adaptive component $\frac{c_\beta\sigma^2}{\sum_{t=1}^m \|v_{t-1}^{s-1}\|^2/m}$ can be much smaller than $\min\{c_\epsilon\sigma^2\epsilon^{-1}, n\}$ during the optimization process particularly in the initial stage, the actual SFO complexity of AbaSVRG can be much lower than that of SVRG with fixed batch size as well as the worst-case complexity of $\mathcal{O}\left(\frac{1}{\epsilon}(n\wedge\frac{1}{\epsilon})^{2/3}\right)$, as demonstrated in our experiments.

Second, our convergence and complexity results hold for any choice of mini-batch size B, and thus we can safely choose a small mini-batch size rather than the large one $n^{2/3}\wedge\epsilon^{-2/3}$ in the regime with large n and ϵ^{-1}. In addition, for a given B, the resulting worst-case complexity $\mathcal{O}\big(\frac{n\wedge\epsilon^{-1}}{\sqrt{B}\epsilon} + \frac{B}{\epsilon}\big)$ still matches the best known order given by ProxSVRG+ (Li & Li, 2018) for SVRG-type algorithms.

Third, Corollary 1 sets the mini-batch size $B = m^2$ to obtain the best complexity order. However, our experiments suggest that $B = m$ has better performance. Hence, we also provide analysis for this case in Appendix E.3.

2.4 Convergence Analysis for AbaSPIDER

In this subsection, we study AbaSPIDER, and compare our results with that for AbaSVRG. Note that AbaSPIDER in Algorithm 2 adopts the improved version SpiderBoost (Wang et al., 2019) of the original SPIDER (Fang et al., 2018).

The following theorem provides a general convergence result for AbaSPIDER.

Theorem 2. *Suppose Assumption 1 holds. Let $\epsilon > 0$ and $c_\beta, c_\epsilon \geq \alpha$ for certain constant $\alpha > 0$. Let $\psi = \frac{2\eta^2 L^2 m}{B} + \frac{2}{\alpha} + 2$ and choose β_1, α and η such that $\beta_1 \leq S\epsilon$ and $\phi = \frac{1}{2} - \frac{1}{2\alpha} - \frac{L\eta}{2} - \frac{\eta^2 L^2 m}{2B} > 0$. Then, the output x_ζ returned by AbaSPIDER satisfies*

$$\mathbb{E}\|\nabla f(x_\zeta)\|^2 \leq \frac{\psi(f(x_0) - f^*)}{\phi\eta K} + \frac{\psi\epsilon}{2\phi\alpha} + \frac{4\epsilon}{\alpha},$$

where $f^* = \inf_{x\in\mathbb{R}^d} f(x)$ *and* $K = Sm$.

To guarantee the convergence, AbaSPIDER allows a smaller mini-batch size B than AbaSVRG, because AbaSPIDER requires $B \geq \Theta(m\eta^2)$ (see Theorem 2) to guarantee ϕ to be positive, whereas AbaSVRG requires $B \geq \Theta(m^2\eta^2)$ (see Theorem 1). Thus, to achieve the same-level of target accuracy, AbaSPIDER uses fewer mini-batch samples than AbaSVRG, and thus achieves a lower worst-case SFO complexity, as can be seen in the following corollary.

Corollary 2. *Under the setting of Theorem 2, for any mini-batch size* $B \leq n^{1/2} \wedge \epsilon^{-1/2}$*, if we set the epoch length* $m = (n \wedge \frac{1}{\epsilon})B^{-1}$*, the stepsize* $\eta = \frac{1}{4L}\sqrt{\frac{B}{m}}$ *and* $c_\beta, c_\epsilon \geq 16$*, then to obtain an* ϵ*-accurate solution* x_ζ*, the total SFO complexity of AbaSPIDER is given by*

$$\underbrace{\sum_{s=1}^{S} \min\left\{\frac{c_\beta\sigma^2}{\sum_{t=1}^m \|v_{t-1}^{s-1}\|^2/m}, c_\epsilon\sigma^2\epsilon^{-1}, n\right\} + KB}_{\text{complexity of AbaSPIDER}}$$
$$< \underbrace{S\min\left\{c_\epsilon\sigma^2\epsilon^{-1}, n\right\} + KB}_{\text{complexity of vanilla SPIDER}} = \mathcal{O}\big(\epsilon^{-1}\big(n \wedge \epsilon^{-1}\big)^{1/2}\big).$$

Corollary 2 shows that for a *wide range* of mini-batch size B (as long as $B \leq n^{1/2} \wedge \epsilon^{-1/2}$), AbaSPIDER achieves the near-optimal worst-case complexity $\mathcal{O}\big(\frac{1}{\epsilon}(n \wedge \frac{1}{\epsilon})^{1/2}\big)$ under a proper selection of m and η. Thus, our choice of mini-batch size is much less restrictive than $B = n^{1/2} \wedge \epsilon^{-1/2}$ used by SpiderBoost (Wang et al., 2019) to achieve the optimal complexity, which can be very large in practice. More importantly, our practical complexity can be much better than those of SPIDER (Fang et al., 2018) and SpiderBoost (Wang et al., 2019) with a fixed batch size due to the batch size adaptation.

3 Batch Size Adaptation for Policy Gradient

In this section, we demonstrate an important application of our proposed batch-size adaptation scheme to variance reduced policy gradient algorithms in RL.

3.1 Problem Formulation

Consider a discrete-time Markov decision process (MDP) $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \rho\}$, where $\mathcal{S}$ denotes the state space; $\mathcal{A}$ denotes the action space; $\mathcal{P}$ denotes the Markovian transition model, $\mathcal{P}(s'|s,a)$ denotes the transition probability from state-action pair (s,a) to state s'; $\mathcal{R} \in [-R, R]$ denotes the reward function, $\mathcal{R}(s,a)$ denotes the reward at state-action pair (s,a); $\gamma \in [0,1)$ denotes the discount factor; and ρ denotes the initial state distribution. The agent's decision strategy is captured by the policy $\pi := \pi(\cdot|s)$, which represents the density function over space $\mathcal{A}$ at state s. Assume that the policy is parameterized by $\theta \in \mathbb{R}^d$. Then, the policy class can be represented as $\Pi = \{\pi_\theta | \theta \in \mathbb{R}^d\}$.

We consider a MDP problem with a finite horizon H. Then, a trajectory τ consists of a sequence of states and actions $(s_0, a_0, \cdots, s_{H-1}, a_{H-1})$ observed by following a policy π_θ and $s_0 \sim \rho$. The total reward of such a trajectory τ is given by $\mathcal{R}(\tau) = \sum_{t=0}^H \gamma^t \mathcal{R}(s_t, a_t)$. For a give policy π_θ, the corresponding expected reward is given by $J(\theta) = \mathbb{E}_{\tau\sim p(\cdot|\theta)} \mathcal{R}(\tau)$, where $p(\cdot|\theta)$ represents the probability distribution of the trajectory τ by following the policy π_θ. The goal of the problem is to find a policy that achieves the maximum accumulative reward by solving

$$\max_{\theta\in\mathbb{R}^d} J(\theta), \quad \text{where } J(\theta) = \mathop{\mathbb{E}}_{\tau\sim p(\cdot|\theta)}[\mathcal{R}(\tau)]. \tag{Q}$$

Algorithm 3 AbaSVRPG

Input: $\eta, \theta_0, \epsilon, m, \alpha, \beta > 0$
for $k = 0, 1, \ldots, K$ **do**
  if $\mathrm{mod}(k, m) = 0$ **then**
    Sample $\{\tau_i\}_{i=1}^{N}$ from $p(\cdot|\theta_k)$, where N is given by (1)
    $v_k = \frac{1}{N}\sum_{i=1}^{N} g(\tau_i|\theta_k)$
    $\tilde{\theta} = \theta_k$ and $\tilde{v} = v_k$
  else
    Draw $\{\tau_i\}_{i=1}^{B}$ samples from $p(\cdot|\theta_k)$
    $v_k = \frac{1}{B}\sum_{i=1}^{B}\left(g(\tau_i|\theta_k) - \omega(\tau_i|\theta_k, \tilde{\theta})g(\tau_i|\tilde{\theta})\right) + \tilde{v}$
  end if
  $\theta_{k+1} = \theta_k + \eta v_k$
end for
Output: θ_ξ from $\{\theta_0, \cdots, \theta_K\}$ uniformly at random.

Algorithm 4 AbaSPIDER-PG

Input: $\eta, \theta_0, \epsilon, m, \alpha, \beta > 0$
for $k = 0, 1, \ldots, K$ **do**
  if $\mathrm{mod}(k, m) = 0$ **then**
    Sample $\{\tau_i\}_{i=1}^{N}$ from $p(\cdot|\theta_k)$, where N is given by (1)
    $v_k = \frac{1}{N}\sum_{i=1}^{N} g(\tau_i|\theta_k)$
  else
    Draw $\{\tau_i\}_{i=1}^{B}$ samples from $p(\cdot|\theta_k)$.
    $v_k = \frac{1}{B}\sum_{i=1}^{B}\big(g(\tau_i|\theta_k) - \omega(\tau_i|\theta_k, \theta_{k-1})g(\tau_i|\theta_{k-1})\big) + v_{k-1}$
  end if
  $\theta_{k+1} = \theta_k + \eta v_k$
end for
Output: θ_ξ from $\{\theta_0, \cdots, \theta_K\}$ uniformly at random.

3.2 Preliminaries of Policy Gradient and Variance Reduction

Policy gradient is a popular approach to solve the problem (Q), which iteratively updates the value of θ based on the trajectory gradient of the above objective function. Since the distribution $p(\cdot|\theta)$ of τ is unknown because the MDP is unknown, policy gradient adopts the trajectory gradient (denoted by $g(\tau|\theta)$) based on the *sampled* trajectory τ for its update. Two types of commonly used trajectory gradients, namely REINFORCE (Williams, 1992) and G(PO)MDP (Baxter & Bartlett, 2001), are introduced in Appendix A. A key difference of such policy gradient algorithms from SGD in conventional optimization is that the sampling distribution $p(\cdot|\theta)$ changes as the policy parameter θ is iteratively updated, and hence trajectories here are sampled by a varying distributions during the policy gradient iteration.

This paper focuses on the following two variance reduced policy gradient algorithms, which were developed recently to improve the computational efficiency of policy gradient algorithms. First, Papini et al. 2018 proposed a stochastic variance reduced policy gradient (SVRPG) algorithm by adopting the SVRG structure in conventional optimization. In particular, SVRPG continuously adjusts the gradient estimator by *importance sampling* due to the iteratively changing trajectory distribution. Furthermore, Xu et al. 2019b applied the SARAH/SPIDER estimator in conventional optimization to develop a stochastic recursive variance reduced policy gradient (SRVR-PG) algorithm, which we refer to as SPIDER-PG in this paper.

3.3 Proposed Algorithms with Batch-size Adaptation

Both variance reduced policy gradient algorithms SVRPG and SPIDER-PG choose a large batch size N for estimating the policy gradient at the beginning of each epoch. As the result, they often run slowly in practice, and do not show significant advantage over the vanilla policy gradient algorithms. This motivates us to apply our developed batch-size adaptation scheme to reduce their computational complexity.

Thus, we propose AbaSVRPG and AbaSPIDER-PG algorithms, as outlined in Algorithms 3 and 4. More specifically, we adapt the batch size N based on the average of the trajectory gradients in the preceding epoch as

$$N = \frac{\alpha\sigma^2}{\frac{\beta}{m}\sum_{i=n_k-m}^{n_k-1}\|v_i\|^2 + \epsilon}, \tag{1}$$

where k denotes the iteration number, $n_k = \lfloor k/m \rfloor \times m$, and $\|v_{-1}\| = \cdots = \|v_{-m}\| = 0$.

3.4 Assumptions and Definitions

We take the following standard assumptions, as also adopted by Xu et al. 2019a,b; Papini et al. 2018.

Assumption 2. *The trajectory gradient $g(\tau|\theta)$ is an unbiased gradient estimator, i.e., $\mathbb{E}_{\tau\sim p(\cdot|\theta)}[g(\tau|\theta)] = \nabla J(\theta)$.*

Note that the commonly used trajectory gradients $g(\tau|\theta)$ such as REINFORCE and G(PO)MDP as given in Appendix A satisfy Assumption 2.

Assumption 3. *For any state-action pair (s,a), at any value of θ, and for $1 \le i,j \le d$, there exist constants $0 \le G, H, R < \infty$ such that $|\nabla_{\theta_i} \log \pi_\theta(a|s)| \le G$ and*

$$|\mathcal{R}(s,a)| \le R, \ \left|\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log \pi_\theta(a|s)\right| \le H.$$

Assumption 3 assumes that the reward function $\mathcal{R}$, and the gradient and Hessian of $\log \pi_\theta(a|s)$ are bounded.

Assumption 4. *The estimation variance of the trajectory gradient is bounded, i.e., there exists a constant $\sigma^2 < \infty$ such that, for any $\theta \in \mathbb{R}^d$:*

$$\mathbb{V}\mathrm{ar}[g(\tau|\theta)] = \mathop{\mathbb{E}}_{\tau\sim p(\cdot|\theta)} \|g(\tau|\theta) - \nabla J(\theta)\|^2 \le \sigma^2.$$

Since the problem (Q) in general is nonconvex, we take the following standard convergence criterion.

Definition 3. *We say that $\bar{\theta}$ is an ϵ-accurate stationary point for the problem (Q) if $\mathbb{E}\left\|\nabla J(\bar{\theta})\right\|^2 \le \epsilon$.*

To measure the computational complexity of various policy gradient methods, we take the stochastic trajectory-gradient oracle (STO) complexity as the metric, which meansures the number of trajectory-gradient computations to attain an ϵ-accurate stationary point.

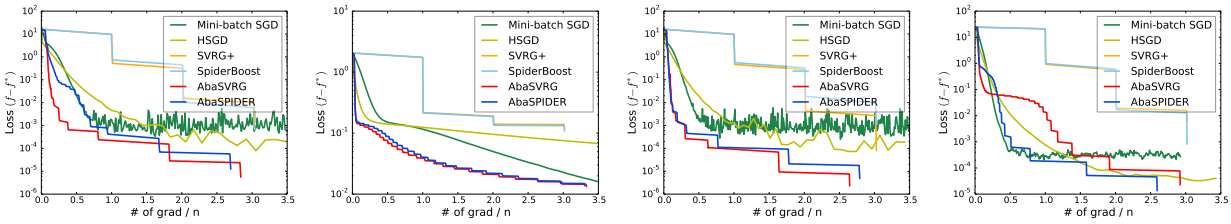


Figure 1: Comparison of various algorithms for logistic regression problem over four datasets. From left to right: a8a, ijcnn1, a9a, w8a.

3.5 Convergence Analysis for AbaSVRPG

In this subsection, we provide the convergence and complexity analysis for AbaSVRPG algorithm. Since the batch size of AbaSVRPG is adaptive to the trajectory gradients calculated in the previous epoch, we will adopt our new analysis framework to bound the change of the function value for each epoch by the trajectory gradients in the preceding epoch due to the batch size dependence. The challenge here arises due to the fact that the sampling distribution is time varying as the policy parameter is updated due to the iteration. Hence, the bound should be tightly developed in order to ensure the decrease of the accumulative change of the objective value over the entire execution of the algorithm. Such bounding procedure is very different from the convergence proofs for vanilla SVRPG in Papini et al. 2018; Xu et al. 2019a.

The following theorem characterizes the convergence of AbaSVRPG. Let $\theta^* := \arg\max_{\theta\in\mathbb{R}^d} J(\theta)$.

Theorem 3. *Suppose Assumption 2, 3, and 4 hold. Choose* $\eta = \frac{1}{2L}, m = \left(\frac{L^2\sigma^2}{Q\epsilon}\right)^{\frac{1}{3}}, B = \left(\frac{Q\sigma^4}{L^2\epsilon^2}\right)^{\frac{1}{3}}, \alpha = 48$ *and* $\beta = 6$*, where* $L > 0$ *is a Lipschitz constant given in Lemma 3 in Appendix F, and* Q *is the constant given in Lemma 5 in Appendix F. Then, the output* θ_ξ *of AbaSVRPG satisfies*

$$\mathbb{E}\left\|\nabla J(\theta_\xi)\right\|^2 \leq \frac{88L}{K+1}\left(J(\theta^*) - J(\theta_0)\right) + \frac{\epsilon}{2}, \tag{2}$$

where K *denotes the total number of iterations.*

Theorem 3 shows that the output of AbaSVRPG converges at a rate of $\mathcal{O}(1/K)$. Furthermore, the following corollary captures the overall STO complexity of AbaSVRPG and its comparison with the vanilla SVRPG.

Corollary 3. *Under the setting of Theorem 3, the overall STO complexity of AbaSVRPG to achieve* $\mathbb{E}\left\|\nabla J(\theta_\xi)\right\|^2 \leq \epsilon$ *is*

$$\underbrace{2KB + \sum_{k=0}^{n_K} \frac{\alpha\sigma^2}{\frac{\beta}{m}\sum_{i=km-m}^{km-1}\left\|v_i\right\|^2 + \epsilon}}_{\text{compelxity of AbaSVRPG}} < \underbrace{2KB + \sum_{k=0}^{n_K} \frac{\alpha\sigma^2}{\epsilon}}_{\text{compelxity of vanilla SVRPG}} = \mathcal{O}\left(\epsilon^{-5/3} + \epsilon^{-1}\right).$$

The STO complexity improves the state-of-the-art complexity of vanilla SVRPG characterized in Xu et al. 2019a, especially due to the saving samples at the initial stage. The worst-case STO complexity of AbaSVRPG is $\mathcal{O}(\epsilon^{-5/3} + \epsilon^{-1})$, which matches that of Xu et al. 2019a.

3.6 Convergence Analysis for AbaSPIDER-PG

In this section, we provide the convergence and complexity analysis for AbaSPIDER-PG algorithm.

Theorem 4. *Suppose Assumptions 2, 3, and 4 hold. Choose* $\eta = \frac{1}{2L}, m = \frac{L\sigma}{\sqrt{Q\epsilon}}, B = \frac{\sigma\sqrt{Q}}{L\sqrt{\epsilon}}, \alpha = 48$ *and* $\beta = 6$*. where* $L > 0$ *is a Lipschitz constant given in Lemma 3 in Appendix F, and* Q *is the constant given in Lemma 5 in Appendix F. Then, the output* θ_ξ *of AbaSPIDER-PG satisfies*

$$\mathbb{E}\left\|\nabla J(\theta_\xi)\right\|^2 \leq \frac{40L}{K+1}\left(J(\theta^*) - J(\theta_0)\right) + \frac{\epsilon}{2}.$$

Corollary 4. *Under the setting of Theorem 4, the total STO complexity of AbaSPIDER-PG to achieve* $\mathbb{E}\|\nabla J(\theta_\xi)\|^2 \leq \epsilon$ *is*

$$\underbrace{2KB + \sum_{k=0}^{n_K} \frac{\alpha\sigma^2}{\frac{\beta}{m}\sum_{i=km-m}^{km-1}\left\|v_i\right\|^2 + \epsilon}}_{\text{compelxity of AbaSPIDER-PG}} < \underbrace{2KB + \sum_{k=0}^{n_K} \frac{\alpha\sigma^2}{\epsilon}}_{\text{compelxity of vanilla SPIDER-PG}} = \mathcal{O}(\epsilon^{-3/2} + \epsilon^{-1}).$$

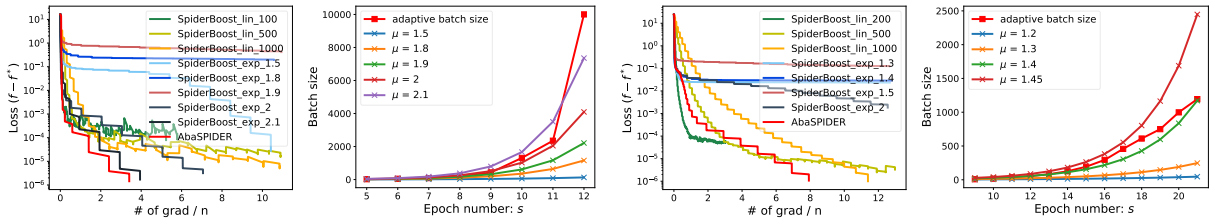
Corollary 4 shows that the worst-case STO complexity of AbaSPIDER-PG is $\mathcal{O}(\epsilon^{-3/2} + \epsilon^{-1})$, which orderwisely outperforms that of AbaSVRPG in Corollary 3, by a factor of $\mathcal{O}(\epsilon^{-1/6})$. This is due to the fact that AbaSPIDER-PG avoids the variance accumulation problem of AbaSVRPG by continuously using the gradient information from the immediate preceding step (see Appendix F.4 for more details).

4 Experiments

In this section, we compare our proposed batch-size adaptation algorithms with their corresponding vanilla algorithms in both conventional nonconvex optimization and reinforcement learning problems.

4.1 Nonconvex Optimization

We compare our proposed AbaSVRG and AbaSPIDER with the state-of-the-art algorithms including mini-batch SGD (Ghadimi & Lan, 2013), HSGD (Zhou et al., 2018b), SVRG+ (Li & Li, 2018), and SpiderBoost (Wang et al., 2019) for two nonconvex optimization problems, i.e., nonconvex logistic regression and training multi-layer neural networks. Due to the space limitations, the detailed hyper-parameter settings for all algorithms and the results on training neural networks are relegated to Appendix B.



(a) dataset: a9a (b) dataset: w8a

Figure 2: Comparison of our gradient-based adaptive batch size and exponentially and linearly increased batch sizes. For each dataset, the left figure plots loss v.s. # of gradient evaluations and the right figure plots adaptive batch size and exponentially increasing batch sizes v.s. epoch number s.

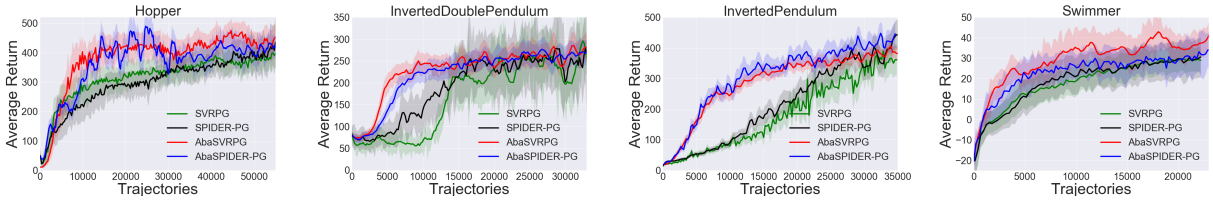


Figure 3: Comparison of various algorithms for reinforcement learning on four tasks.

We consider the following nonconvex logistic regression problem $\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \ell(w^T x_i, y_i) + \alpha \sum_{i=1}^d \frac{w_i^2}{1+w_i^2}$, where $x_i \in \mathbb{R}^d$ denote the features, $y_i \in \{\pm 1\}$ are labels, ℓ is the cross-entropy loss, and we set $\alpha = 0.1$. For this problem, we use four datasets from LIBSVM (Chang & Lin, 2011): a8a, w8a, a9a, ijcnn1.

As can be seen from Fig. 1 and Fig. 4 (in Appendix B.2), AbaSVRG and AbaSPIDER converge much faster than all other algorithms in terms of the total number of gradient evaluations (i.e., SFO complexity) on all four datasets. It can be seen that both of them take the advantage of sample-efficient SGD-like updates (due to the small batch size) at the initial stage and attain high accuracy provided by variance-reduced methods at the final stage. This is consistent with the choice of our batch-size adaptation scheme.

We then evaluate the performance of our history-gradient based batch-size adaptation scheme with the other two commonly used prescribed adaptation schemes, i.e., exponential increase of batch size $N_s = \mu^s$ and linear increase of batch size $N_s = \nu(s+1)$. Let SpiderBoost_exp_μ and SpiderBoost_lin_ν denote SpiderBoost algorithms with exponentially and linearly increasing batch sizes under parameters μ and ν, respectively. As shown in Fig. 2, our adaptive batch size scheme achieves the best performance for a9a dataset, and performs better than all other algorithms for w8a dataset except SpiderBoost_lin_200, which, however, does not converge in the high-accuracy regime. Furthermore, the performance of prescribed batch-size adaptation can be problem specific. For example, exponential increase of batch size (with $\mu = 2$ and $\mu = 2.1$) performs better than linear increase of batch size for a9a dataset, but worse for w8a dataset, whereas our scheme adapts to the optimization path, and hence performs the best in both cases.

4.2 Reinforcement Learning

We compare our proposed AbaSVRPG and AbaSPIDER-PG with vanilla SVRPG (Papini et al., 2018) and SPIDER-PG (Xu et al., 2019b) on four benchmark tasks in reinforcement learning, i.e., InvertedPendulum, InvertedDoublePendulum, Swimmer and Hopper. We apply the Gaussian policy which is constructed using a two-layer neural network (NN) with the number of hidden weights being task-dependent. We also include the setup of adaptive standard deviation. The experimental results are averaged over five trails with different random seeds, and selections of random seeds are consistent for different algorithms within each task for a fair comparison. Further details about the hyper-parameter setting and task environments are provided in Appendix B.4.

It can be seen from Figure 3 that the proposed AbaSVRPG and AbaSPIDER-PG converge much faster than the vanilla SVRPG and SPIDER-PG (without batch size adaptation) on all four tasks. Such an acceleration is more significant at the initial stage of optimization procedure due to the large trajectory gradient that suggests small batch size.

5 Conclusion

In this paper, we propose a novel scheme for adapting the batch size via history gradients, based on which we develop AbaSVRG and AbaSPIDER for conventional optimization and AbaSVRPG and AbaSPIDER-PG for reinforcement learning. We show by theory and experiments that the proposed algorithms achieve improved computational complexity than their vanilla counterparts (without batch size adaptation). Extensive experiments demonstrate the promising performances of proposed algorithms. We anticipate that such a scheme can be applied to a wide range of other stochastic algorithms to accelerate their theoretical and practical performances.

References

Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *Proc. International Conference on Machine Learning (ICML)*, pp. 699–707, 2016.

Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15(1), November 2001.

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

De, S., Yadav, A., Jacobs, D., and Goldstein, T. Big batch SGD: Automated inference using adaptive batch sizes. *arXiv preprint arXiv:1610.05792*, 2016.

De, S., Yadav, A., Jacobs, D., and Goldstein, T. Automated inference with adaptive batches. In *Proc. Artificial Intelligence and Statistics (AISTATS)*, pp. 1504–1513, 2017.

Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 1646–1654. 2014.

Devarakonda, A., Naumov, M., and Garland, M. Adabatch: adaptive batch sizes for training deep neural networks. *arXiv preprint arXiv:1712.02029*, 2017.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning (ICML)*, pp. 1329–1338, 2016.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 689–699, 2018.

Friedlander, M. P. and Schmidt, M. Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405, 2012.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Harikandeh, R., Ahmed, M. O., Virani, A., Schmidt, M., Konečnỳ, J., and Sallinen, S. Stop wasting my gradients: practical SVRG. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2251–2259, 2015.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 315–323, 2013.

Lei, L. and Jordan, M. I. On the adaptivity of stochastic gradient-based optimization. *arXiv preprint arXiv:1904.04480*, 2019.

Lei, L., Ju, C., Chen, J., and Jordan, M. I. Non-convex finite-sum optimization via SCSG methods. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2348–2358, 2017.

Li, Z. and Li, J. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5564–5574, 2018.

Nesterov, Y. and Polyak, B. T. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proc. International Conference on Machine Learning (ICML)*, pp. 2613–2621, 2017a.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. Stochastic recursive gradient algorithm for nonconvex optimization. *arXiv preprint arXiv:1705.07261*, 2017b.

Nguyen, L. M., van Dijk, M., Phan, D. T., Nguyen, P. H., Weng, T.-W., and Kalagnanam, J. R. Finite-sum smooth optimization with SARAH. *arXiv preprint arXiv:1901.07648*, 2019.

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. Stochastic variance-reduced policy gradient. In *Proc. International Conference on Machine Learning (ICML)*, pp. 4026–4035, 2018.

Polyak, B. T. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Reddi, S. J., Hefny, A., Sra, S., Poczos, B., and Smola, A. Stochastic variance reduction for nonconvex optimization. In *Proc. International Conference on Machine Learning (ICML)*, pp. 314–323, 2016a.

Reddi, S. J., Sra, S., Poczos, B., and Smola, A. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 1145–1153. 2016b.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22 (3):400–407, 1951.

Roux, N. L., Schmidt, M., and Bach, F. R. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2663–2671. 2012.

S. Sutton, R., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Proc. Neural Information Processing Systems (NIPS)*, 2000.

Shen, Z., Ribeiro, A., Hassani, H., Qian, H., and Mi, C. Hessian aided policy gradient. In *Proc. International Conference on Machine Learning (ICML)*, 2019.

Sievert, S. and Charles, Z. Improving the convergence of SGD through adaptive batch sizes. *arXiv preprint arXiv:1910.08222*, 2019.

Smith, S. L., Kindermans, P.-J., Ying, C., and Le, Q. V. Don't decay the learning rate, increase the batch size. *Proc. International Conference on Learning Representations (ICLR)*, 2018.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. SpiderBoost and momentum: Faster variance reduction algorithms. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2403–2413, 2019.

Weaver, L. and Tao, N. The optimal reward baseline for gradient-based reinforcement learning. In *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 538–545, 2001.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992.

Wu, C., Rajeswaran, A., Duan, Y., Kumar, V., Bayen, A. M., Kakade, S., Mordatch, I., and Abbeel, P. Variance reduction for policy gradient with action-dependent factorized baselines. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

Xu, P., Gao, F., and Gu, Q. An improved convergence analysis of stochastic variance-reduced policy gradient. *arXiv preprint arXiv:1905.12615*, 2019a.

Xu, P., Gao, F., and Gu, Q. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*, 2019b.

Yuan, H., Li, C. J., Tang, Y., and Zhou, Y. Policy optimization via stochastic recursive gradient algorithm. 2018.

Zhong, K., Song, Z., Jain, P., Bartlett, P. L., and Dhillon, I. S. Recovery guarantees for one-hidden-layer neural networks. In *Proc. International Conference on Machine Learning (ICML)*, pp. 4140–4149, 2017.

Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduced gradient descent for nonconvex optimization. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3921–3932, 2018a.

Zhou, P., Yuan, X., and Feng, J. New insight into hybrid stochastic gradient descent: Beyond with-replacement sampling and convexity. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1234–1243, 2018b.

Zhou, Y. and Liang, Y. Characterization of gradient dominance and regularity conditions for neural networks. *arXiv preprint arXiv:1710.06910*, 2017.

Zhou, Y., Zhang, H., and Liang, Y. Geometrical properties and accelerated gradient solvers of non-convex phase retrieval. In *54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 331–335, 2016.

Supplementary Materials

A Further Details on Policy Gradient

This paper considers the policy gradient algorithms that can adopt the following two types of trajectory gradients, namely REINFORCE (Williams, 1992) and G(PO)MDP (Baxter & Bartlett, 2001). We note that

$$\nabla J(\theta) = \nabla \mathop{\mathbb{E}}_{\tau\sim p(\cdot|\theta)}[\mathcal{R}(\tau)] = \mathop{\mathbb{E}}_{\tau\sim p(\cdot|\theta)}[\mathcal{R}(\tau)\nabla \log p(\tau|\theta)],$$

where $p(\tau|\theta) = \rho(s_0)\pi_\theta(a_0|s_0)\prod_{i=0}^{H-1}\mathcal{P}(s_{i+1}|s_i,a_i)\pi_\theta(a_{i+1}|s_{i+1})$. REINFORCE constructs the trajectory gradient as

$$g(\tau|\theta) = \underbrace{\Big(\sum_{t=0}^{H}\gamma^t\mathcal{R}(s_t,a_t) - b(s_t,a_t)\Big)\Big(\sum_{t=0}^{H}\nabla\log\pi_\theta(a_t|s_t)\Big)}_{\mathcal{R}(\tau)\nabla\log p(\tau|\theta)},$$

where $b:\mathcal{S}\times\mathcal{A}\to\mathcal{R}$ is a bias. G(PO)MDP enhances the trajectory gradient of REINFORCE by further utilizing the fact that the reward at time t does not depend on the action implemented after time t. Thus, G(PO)MDP constructs the trajectory gradient as

$$g(\tau|\theta) = \sum_{t=0}^{H}\big(\gamma^t\mathcal{R}(s_t,a_t) - b(s_t,a_t)\big)\sum_{i=0}^{t}\nabla\log\pi_\theta(a_i|s_i).$$

Note that REINFORCE and G(PO)MDP are both unbiased gradient estimators, i.e., $\mathbb{E}_{\tau\sim p(\cdot|\theta)}[g(\tau|\theta)] = \nabla J(\theta)$.

B Further Specification of Experiments and Additional Results

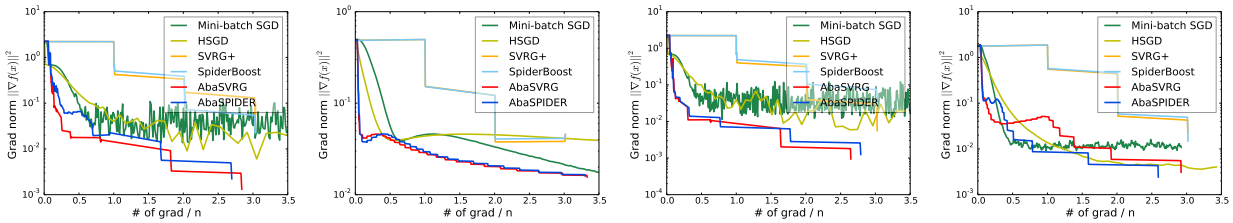
B.1 Hyper-parameter Configuration of Algorithms for Nonconvex Optimization

To implement HSGD, we follow Zhou et al. 2018b and choose the linearly increasing mini-batch size at the t^{th} iteration to be $c_b(t+1)$, and tune c_b to the best. We set the epoch length $m = 10$ for all variance-reduced algorithms. We choose the batch size to be $\min\{n, c_1\epsilon^{-1}\}$ for SVRG+ and SpiderBoost, and $\min\{n, c_1\epsilon^{-1}, c_2\beta_s^{-1}\}$ for the proposed AbaSPIDER and AbaSVRG, where $\beta_s = \frac{1}{m}\sum_{t=1}^{m}\|v_{t-1}^{s-1}\|^2$ as given in Subsection 2.1.

B.2 Additional Results for Nonconvex Logistic Regression

For logistic regression, we use four datasets: a8a ($n = 22696, d = 123$), a9a ($n = 32561, d = 123$), w8a ($n = 43793, d = 300$) and ijcnn1 ($n = 49990, d = 22$). We select the stepsize η from $\{0.1k, k = 1, 2, ..., 15\}$ and the mini-batch size B from $\{10, 28, 64, 128, 256, 512, 1024\}$ for all algorithms, and we present the best performance among these parameters. For all variance-reduced algorithms, we select constants c_1 and c_2 from $\{1, 2, 3, ..., 10\}$, and present the best performance among these parameters. For HSGD algorithm, we select c_b in its linearly increasing batch size $c_b(t+1)$ from $\{1, 5, 10, 40, 100, 400, 1000\}$, and present the best performance among these parameters.

As shown in Fig. 4, AbaSVRG and AbaSPIDER converge much faster than all other algorithms in terms of the total number of gradient evaluations on all four datasets. It can be seen that both of them take the advantage of sample-efficient SGD-like updates (due to the small batch size) at the initial stage and attain high accuracy provided by variance-reduced methods at the final stage. This is consistent with the choice of our batch-size adaptation scheme.



(a) dataset: a8a (b) dataset: ijcnn1 (c) dataset: a9a (d) dataset: w8a

Figure 4: Comparison of different algorithms for logistic regression problem on four datasets. All figures plot gradient norm v.s. # of gradient evaluations.

B.3 Results for Training Multi-Layer Neural Networks

In this subsection, we compare our proposed algorithms with other competitive algorithms as specified in Section 4.1 for training a three-layer ReLU neural network with a cross entropy loss on the dataset of MNIST ($n = 60000, d = 780$). The neural network has a size of $(d_{\text{in}}, 100, 100, d_{\text{out}})$, where d_{in} and d_{out} are the input and output dimensions and 100 is the number of neurons in the two hidden layers. We select the stepsize η from $\{10^{-4}k, k = 1, 2, ..., 15\}$ and the mini-batch size B from $\{64, 96, 128, 256, 512\}$ for all algorithms, and we present the best performance among these parameters. For all variance-reduced algorithms, we set $c_1 = 1$ and select the best c_2 from $\{10^3, 5 \times 10^3, 10^4\}$. For HSGD algorithm, we select c_b from $\{1, 10, 50, 100, 500, 1000\}$, and present the best performance among these parameters.

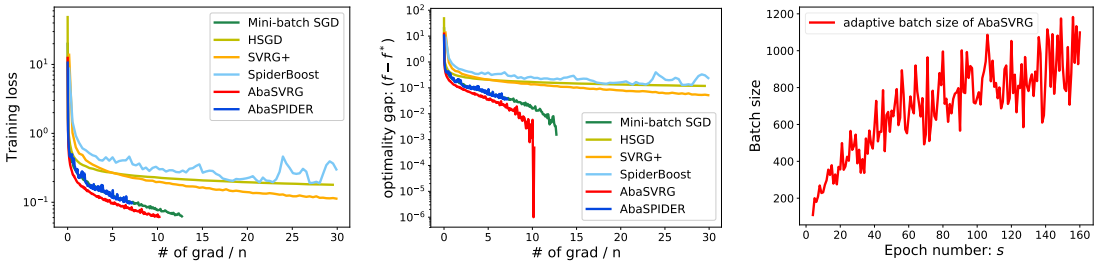


Figure 5: Comparison of various algorithms for training a three-layer neural network on MNIST.

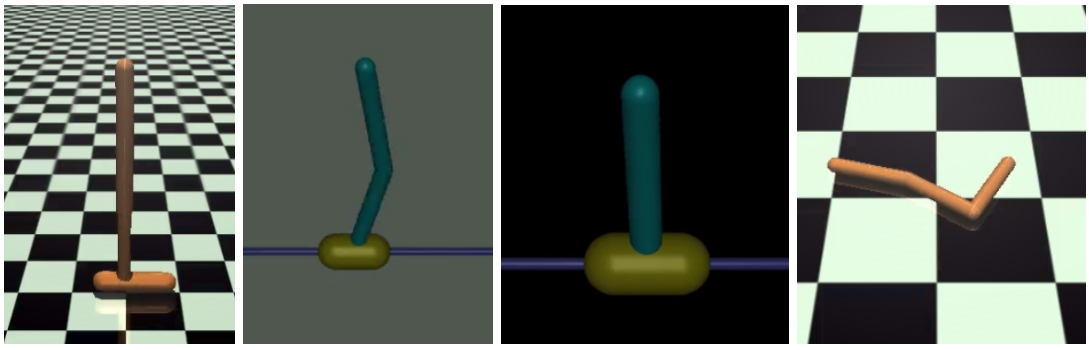
As shown in Fig. 5, our AbaSVRG achieves the best performance among all competing algorithms, and AbaSPIDER performs similarly to mini-batch SGD for decreasing training loss, but converges faster in terms of gradient norm. Interestingly, the batch-size adaptation used by AbaSVRG increases the batch size slower than both exponential and linear increase of the batch size, and its scaling is close to the *logarithmical* increase as shown in the right-most plot in Figure 5. Such an observation further demonstrates that our gradient-based batch-size adaptation scheme can also adapt to the neural network landscape with a differently (i.e., more slowly) increased batch size from that for nonconvex regression problem over a9a and w8a datasets.

B.4 Experimental Details for Reinforcement Learning

The hyper-parameters listed in Table 1 are the same among all methods on each task. For the proposed AbaSVRPG and AbaSPIDER-PG, we adopt the same hyper-parameter of $\alpha\sigma^2 = 1$ and $\beta = 1000$ in all experiments.

Table 1: Parameters used in the RL experiments

Task	InvertedPendulum	InvertedDoublePendulum	Swimmer	Hopper
Horizon	500	500	500	500
Discount Factor γ	0.99	0.99	0.99	0.99
q	10	10	10	10
N	100	100	50	50
B	20	20	20	20
ϵ	0.01	0.01	0.01	0.01
Step Size	0.001	0.001	0.0001	0.001
NN Hidden Weights	16×16	16×16	32×32	64×64
NN Activation	tanh	tanh	tanh	tanh
Baseline	No	No	Yes	Yes



(a) Hopper (b) Inverteddoublependulum (c) Invertedpendulum (d) Swimmer

Figure 6: Task Environments.

Figure 6 illustrates all task environments. The problem setup regarding each task is summarized as follows:

1. *InvertedPendulum*: A cart is moving along a track with zero friction and a pole is attached through an un-actuated joint. The pendulum is balanced by controlling the velocity of the cart. The action space is continuous with $a \in [-1, 1]$ (with -1 for pushing cart to the left and 1 for pushing cart to the right). For a single episode with time step h enumerated from 1 to 500, the episode is terminated when the pole angle $\theta_h > 0.2rad$, and otherwise a reward of value 1 is awarded.

2. *InvertedDoublePendulum*: The setup of this task is similar to that at the InvertedPendulum. The only difference is that another pendulum is added to the end of the previous pendulum through an unactuated rotational joint.

3. *Swimmer*: The agent is a 3-link robot defined in Mujoco with the state-space dimension of 13. It is actuated by two joints to swim in a viscous fluid. For a single episode with time step h enumerated from 1 to 500, the reward function encourages the agent to move forward as fast as possible while maintaining energy efficiency. That is, given forward velocity v_x and joint action a, $r(v_x, a) = v_x^2 - 10^{-4}\|a\|_2^2$.

4. *Hopper*: A two-dimensional single-legged robot is trained to hop forward. The system has the state-space dimension of 11 and action space dimension of 3. For a single episode with time step h enumerated from 1 to 600, we have forward velocity v_x and commanded action a. The episode terminates early

(before h reaches 500) when the tilting angle of upper body or the height position for center of mass drops below a certain preset threshold. The reward function encourages the agent to move forward as fast as possible in an energy efficient manner. It also gets one alive bonus for every step it survives without triggering any of the termination threshold.

For the tasks of Swimmer and Hopper, we also include the linear baseline for value function approximation (Duan et al., 2016).

C Convergence of AbaSVRG and AbaSPIDER under Local PL Geometry

Many nonconvex machine learning problems (e.g., phase retrieval (Zhou et al., 2016)) and deep learning (e.g., neural networks (Zhong et al., 2017; Zhou & Liang, 2017)) problems have been shown to satisfy the following Polyak-Łojasiewicz (PL) (i.e., gradient dominance) condition in local regions near local or global minimizers.

Definition 4 (Polyak (1963); Nesterov & Polyak (2006)). *Let $x^* = \arg\min_{x\in\mathbb{R}^d} f(x)$. Then, the function f is said to be τ-gradient dominated if for any $x \in \mathbb{R}^d$, $f(x) - f(x^*) \le \tau\|\nabla f(x)\|^2$.*

In this section, we explore whether our proposed AbaSVRG and AbaSPIDER with batch size adaptation can attain a faster linear convergence rate if the iterate enters the local PL regions. All the proofs are provided in Appendix H.

C.1 AbaSVRG: Convergence under PL Geometry without Restart

The following theorem provides the convergence and complexity for AbaSVRG under the PL condition.

Theorem 5. *Let $\eta = \frac{1}{c_\eta L}$, $B = m^2$ with $\frac{8L\tau}{c_\eta - 2} \le m < 4L\tau$, $\beta_1 \le \epsilon\big(\frac{1}{\gamma}\big)^{m(S-1)}$, and $c_\beta = c_\epsilon = \Big(2\tau + \frac{2\tau}{1-\exp(\frac{-4}{c_\eta(c_\eta-2)})}\Big) \vee \frac{16c_\eta L\tau}{m}$, where constants $c_\eta > 4$ and $\gamma = 1 - \frac{1}{8L\tau} < 1$. Then under the PL condition, the final iterate $\tilde{x}^S$ of AbaSVRG satisfies*

$$\mathbb{E}(f(\tilde{x}^S) - f(x^*)) \le \gamma^K (f(x_0) - f(x^*)) + \frac{\epsilon}{2}.$$

To obtain an ϵ-accurate solution $\tilde{x}^S$, the total number of SFO calls is given by

$$\sum_{s=1}^{S} \min\left\{ \frac{c_\beta \sigma^2}{\sum_{t=1}^m \|v_{t-1}^{s-1}\|^2/m}, c_\epsilon \sigma^2 \epsilon^{-1}, n \right\} + KB \le \mathcal{O}\Big(\Big(\frac{\tau}{\epsilon} \wedge n\Big) \log\frac{1}{\epsilon} + \tau^3 \log\frac{1}{\epsilon} \Big). \tag{3}$$

Our proof of Theorem 5 is different from and more challenging than the previous techniques developed in Reddi et al. 2016a,b; Li & Li 2018 for SVRG-type algorithms, because we need to handle the adaptive batch size N_s with the dependencies on the iterations at the previous epoch. In addition, we do not need *extra* assumptions for proving the convergence under PL condition, whereas Reddi et al. 2016b and Li & Li 2018 require $\tau \ge n^{1/3}$ and $\tau \ge n^{1/2}$, respectively. As a result, Theorem 5 can be applied to any condition number regime. For the small condition number regime where $1 \le \tau \le \Theta(n^{1/3})$, the worst-case complexity of AbaSVRG outperforms the result achieved by SVRG (Reddi et al., 2016b). Furthermore, the actual complexity of our AbaSVRG can be much lower than the worst-case complexity due to the adaptive batch size.

C.2 AbaSPIDER: Convergence under PL Geometry without Restart

The following theorem shows that AbaSPIDER achieves a linear convergence rate under the PL condition without restart. Our analysis can be of independent interest for other SPIDER-type methods.

Theorem 6. *Let* $\eta = \frac{1}{c_\eta L}$, $B = m$ *with* $\frac{8L\tau}{c_\eta - 2} \le m < 4L\tau$, $\beta_1 \le \epsilon\left(\frac{1}{\gamma}\right)^{m(S-1)}$, *and* $c_\beta = c_\epsilon = \left(2\tau + \frac{2\tau}{1-\exp(\frac{-4}{c_\eta(c_\eta-2)})}\right) \vee \frac{16 c_\eta L\tau}{m}$, *where constants* $c_\eta > 4$ *and* $\gamma = 1 - \frac{1}{8L\tau}$. *Then under the PL condition, the final iterate* $\tilde{x}^S$ *of AbaSPIDER satisfies*

$$\mathbb{E}(f(\tilde{x}^S) - f(x^*)) \le \gamma^K \mathbb{E}(f(x_0) - f(x^*)) + \frac{\epsilon}{2}.$$

To obtain an ϵ*-accurate solution* $\tilde{x}^S$, *the total number of SFO calls is given by*

$$\sum_{s=1}^{S} \min\left\{ \frac{c_\beta \sigma^2}{\sum_{t=1}^m \|v_{t-1}^{s-1}\|^2/m}, c_\epsilon \sigma^2 \epsilon^{-1}, n \right\} + KB \le \mathcal{O}\left(\left(\frac{\tau}{\epsilon} \wedge n\right) \log\frac{1}{\epsilon} + \tau^2 \log\frac{1}{\epsilon} \right).$$

As shown in Theorem 6, AbaSPIDER achieves a lower worst-case SFO complexity than AbaSVRG by a factor of $\Theta(\tau)$, and matches the best result provided by Prox-SpiderBoost-gd (Wang et al., 2019). However, Prox-SpiderBoost-gd is a variant of Prox-SpiderBoost with algorithmic modification, and has not been shown to achieve the near-optimal complexity for general nonconvex optimization. In addition, AbaSPIDER has a much lower complexity in practice due to the adaptive batch size.

D An analysis for SGD with Adaptive Mini-Batch Size

Recently, Sievert & Charles 2019 proposed an improved SGD algorithm by adapting the batch size to the gradient norms in preceding steps. However, they do not show performance guarantee for their proposed algorithm. In this section, we aim to fill this gap by providing an analysis for adaptive batch size SGD (AbaSGD) with mini-batch size depending on the stochastic gradients in the preceding m steps. as shown in Algorithm 5. To simplify notations, we set norms of the stochastic gradients before the algorithm starts to be $\|\mathbf{v}_{-1}\| = \|\mathbf{v}_{-2}\| = \cdots = \|\mathbf{v}_{-m}\| = \alpha_0$ and let $\mathbb{E}_t(\cdot) = \mathbb{E}(\cdot \,|\, \mathbf{x}_0, ..., \mathbf{x}_t)$.

Algorithm 5 AbaSGD

Input: $\mathbf{x}_0$, stepsize η, $m > 0$, $\alpha_0 > 0$.
for $t = 0, 1, ..., T$ **do**
  Set $|B_t| = \min\left\{ \frac{2\sigma^2}{\sum_{i=1}^m \|\mathbf{v}_{t-i}\|^2/m}, \frac{24\sigma^2}{\epsilon}, n \right\}$.
  if $|B_t| = n$ **then**
    Compute $\mathbf{v}_t = \nabla f(\mathbf{x}_t)$
  else
    Sample B_t from $[n]$ with replacement, and compute $\mathbf{v}_t = \nabla f_{B_t}(\mathbf{x}_t)$
  end if
  $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{v}_t$
end for
Output: choose $\mathbf{x}_\zeta$ from $\{\mathbf{x}_i\}_{i=0,...,T}$ uniformly at random

Theorem 7. *Let Assumption 1 hold,* $\epsilon > 0$ *and choose a stepsize* η *such that*

$$\phi = \eta - \frac{L\eta^2}{2} > 0.$$

Then, the output $\mathbf{x}_\zeta$ *returned by AbaSGD satisfies*

$$\mathbb{E}\|\nabla f(\mathbf{x}_\zeta)\|^2 \leq \frac{2\left(f(\mathbf{x}_0) - f^*\right) + \eta m \alpha_0^2}{2T\phi} + \frac{\eta}{12\phi}\epsilon,$$

where $f^* = \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$*, and* T *is the total number of iterations.*

Theorem 7 shows that AbaSGD achieves a $\mathcal{O}\left(\frac{1}{T}\right)$ convergence rate for nonconvex optimization by using the adaptive mini-batch size. In the following corollary, we derive the SFO complexity of AbaSGD.

Corollary 5. *Under the setting of Theorem 7, we choose the constant stepsize* $\eta = \frac{1}{2L}$*. Then, to obtain an* ϵ*-accurate solution* $\mathbf{x}_\zeta$*, the total number of iterations required by AbaSGD*

$$T = \frac{16L\left(f(\mathbf{x}_0) - f^*\right) + 4m\alpha_0^2}{\epsilon},$$

and the total number of SFO calls required by AbaSGD is given by

$$\sum_{t=0}^{T} |B_t| = \underbrace{\sum_{t=0}^{T} \min\left\{\frac{2\sigma^2}{\sum_{i=1}^{m}\|\mathbf{v}_{t-i}\|^2/m}, \frac{24\sigma^2}{\epsilon}, n\right\}}_{\text{complexity of AbaSGD}} \leq \underbrace{T\min\left\{\frac{24\sigma^2}{\epsilon}, n\right\}}_{\text{complexity of vanilla SGD}} = \mathcal{O}\left(\frac{1}{\epsilon^2} \wedge \frac{n}{\epsilon}\right).$$

Corollary 5 shows that the worst-case complexity of AbaSGD is $\mathcal{O}\left(\frac{1}{\epsilon^2} \wedge \frac{n}{\epsilon}\right)$, which is at least as good as those of SGD and GD. More importantly, the actual complexity of AbaSGD can be much lower than those of GD and SGD due to the adaptive batch size.

Technical Proofs

E Proofs for Results in Section 2

E.1 Proof of Theorem 1

To prove Theorem 1, we first establish the following lemma to upper-bound the estimation variance $\mathbb{E}_{0,s}\|\nabla f(x_{t-1}^s) - v_{t-1}^s\|^2$ for $1 \le t \le m$, where $\mathbb{E}_{t,s}(\cdot)$ denotes $\mathbb{E}(\cdot|x_0^1, x_0^2, ..., x_2^1, ..., x_t^s)$.

Lemma 1. *Let Assumption 1 hold. Then, for $1 \le t \le m$, we have*

$$\mathbb{E}_{0,s}\|\nabla f(x_{t-1}^s) - v_{t-1}^s\|^2 \le \frac{\eta^2 L^2 (t-1)}{B}\mathbb{E}_{0,s}\sum_{i=0}^{t-2}\|v_i^s\|^2 + \frac{I_{(N_s<n)}}{N_s}\sigma^2 \tag{4}$$

where $I_{(A)} = 1$ if the event A occurs and 0 otherwise, and $\sum_{i=0}^{-1}\|v_i^s\|^2 = 0$.

Proof of Lemma 1. Based on line 10 in Algorithm 1, we have, for $1 \le t \le m$,

$$\|v_{t-1}^s - \nabla f(x_{t-1}^s)\|^2 = \|\nabla f_{\mathcal{B}}(x_{t-1}^s) - \nabla f_{\mathcal{B}}(\tilde{x}^{s-1}) - \nabla f(x_{t-1}^s) + \nabla f(\tilde{x}^{s-1}) + g^s - \nabla f(\tilde{x}^{s-1})\|^2.$$

Taking the expectation $\mathbb{E}_{0,s}(\cdot)$ over the above equality yields

$$\begin{aligned}\mathbb{E}_{0,s}\|v_{t-1}^s - \nabla f(x_{t-1}^s)\|^2 &= \mathbb{E}_{0,s}\|\nabla f_{\mathcal{B}}(x_{t-1}^s) - \nabla f_{\mathcal{B}}(\tilde{x}^{s-1}) - \nabla f(x_{t-1}^s) + \nabla f(\tilde{x}^{s-1})\|^2 \\ &\quad + 2\underbrace{\mathbb{E}_{0,s}\langle \nabla f_{\mathcal{B}}(x_{t-1}^s) - \nabla f_{\mathcal{B}}(\tilde{x}^{s-1}) - \nabla f(x_{t-1}^s) + \nabla f(\tilde{x}^{s-1}), g^s - \nabla f(\tilde{x}^{s-1})\rangle}_{(*)} \\ &\quad + \mathbb{E}_{0,s}\|g^s - \nabla f(\tilde{x}^{s-1})\|^2,\end{aligned} \tag{5}$$

which, in conjunction with the fact that

$$(*) = \mathbb{E}_{x_1^s,...,x_{t-1}^s}\mathbb{E}_{t-1,s}\langle \nabla f_{\mathcal{B}}(x_{t-1}^s) - \nabla f_{\mathcal{B}}(\tilde{x}^{s-1}) - \nabla f(x_{t-1}^s) + \nabla f(\tilde{x}^{s-1}), g^s - \nabla f(\tilde{x}^{s-1})\rangle = 0$$

and letting $F_i := \nabla f_i(x_{t-1}^s) - \nabla f_i(\tilde{x}^{s-1}) - \nabla f(x_{t-1}^s) + \nabla f(\tilde{x}^{s-1})$, implies that

$$\begin{aligned}&\mathbb{E}_{0,s}\|v_{t-1}^s - \nabla f(x_{t-1}^s)\|^2 \\ &=\mathbb{E}_{0,s}\|\nabla f_{\mathcal{B}}(x_{t-1}^s) - \nabla f_{\mathcal{B}}(\tilde{x}^{s-1}) - \nabla f(x_{t-1}^s) + \nabla f(\tilde{x}^{s-1})\|^2 + \mathbb{E}_{0,s}\|g^s - \nabla f(\tilde{x}^{s-1})\|^2 \\ &=\frac{1}{B^2}\mathbb{E}_{0,s}\sum_{i\in\mathcal{B}}\|\nabla f_i(x_{t-1}^s) - \nabla f_i(\tilde{x}^{s-1}) - \nabla f(x_{t-1}^s) + \nabla f(\tilde{x}^{s-1})\|^2 + \mathbb{E}_{0,s}\|g^s - \nabla f(\tilde{x}^{s-1})\|^2 \\ &\quad + \frac{2}{B^2}\sum_{i<j,\, i,j\in\mathcal{B}}\mathbb{E}_{0,s}\langle F_i, F_j\rangle \\ &\overset{(i)}{=}\frac{1}{B}\mathbb{E}_{0,s}\|\nabla f_i(x_{t-1}^s) - \nabla f_i(\tilde{x}^{s-1}) - \nabla f(x_{t-1}^s) + \nabla f(\tilde{x}^{s-1})\|^2 + \mathbb{E}_{0,s}\|g^s - \nabla f(\tilde{x}^{s-1})\|^2 \\ &\overset{(ii)}{\le}\frac{1}{B}\mathbb{E}_{0,s}\|\nabla f_i(x_{t-1}^s) - \nabla f_i(\tilde{x}^{s-1})\|^2 + \mathbb{E}_{0,s}\|g^s - \nabla f(\tilde{x}^{s-1})\|^2 \\ &\overset{(iii)}{\le}\frac{1}{B}\mathbb{E}_{0,s}\|\nabla f_i(x_{t-1}^s) - \nabla f_i(\tilde{x}^{s-1})\|^2 + \frac{I_{(N_s<n)}}{N_s}\sigma^2\end{aligned} \tag{6}$$

where (i) follows from the fact that

$$\mathbb{E}_{0,s}\langle F_i, F_j\rangle = \mathbb{E}_{x_1^s,...,x_{t-1}^s}\left(\mathbb{E}_{t-1,s}\langle F_i, F_j\rangle\right) = \mathbb{E}_{x_1^s,...,x_{t-1}^s}\left(\langle \mathbb{E}_{t-1,s}(F_i), \mathbb{E}_{t-1,s}(F_j)\rangle\right) = 0,$$

(ii) follows from the fact that $\mathbb{E}\|y-\mathbb{E}(y)\|^2 \leq \mathbb{E}\|y\|^2$ for any $y \in \mathbb{R}^d$, and (iii) follows by combining Lemma B.2 in Lei et al. 2017 and the fact that N_s is fixed given $x_0^1, ..., x_0^s$. Then, we obtain from (6) that

$$\begin{aligned}\mathbb{E}_{0,s}\|v_{t-1}^s - \nabla f(x_{t-1}^s)\|^2 \leq & \frac{L^2}{B}\mathbb{E}_{0,s}\|x_{t-1}^s - \tilde{x}^{s-1}\| + \frac{I_{(N_s<n)}}{N_s}\sigma^2 \\ = & \frac{L^2}{B}\mathbb{E}_{0,s}\Big\|\sum_{i=0}^{t-2}(x_{i+1}^s - x_i^s)\Big\|^2 + \frac{I_{(N_s<n)}}{N_s}\sigma^2 \\ = & \frac{\eta^2 L^2}{B}\mathbb{E}_{0,s}\Big\|\sum_{i=0}^{t-2} v_i^s\Big\|^2 + \frac{I_{(N_s<n)}}{N_s}\sigma^2 \\ \overset{(i)}{\leq} & \frac{\eta^2 L^2(t-1)}{B}\mathbb{E}_{0,s}\sum_{i=0}^{t-2}\Big\|v_i^s\Big\|^2 + \frac{I_{(N_s<n)}}{N_s}\sigma^2, \end{aligned} \tag{7}$$

where (i) follows from the Cauchy–Schwartz inequality that $\|\sum_{i=1}^k a_i\|^2 \leq k\sum_{i=1}^k \|a_i\|^2$. □

Proof of Theorem 1. Based on Lemma 1, we next prove Theorem 1.

Since the objective function $f(\cdot)$ has a L-Lipschitz continuous gradient, we obtain that for $1 \leq t \leq m$,

$$\begin{aligned} f(x_t^s) \leq & f(x_{t-1}^s) + \langle \nabla f(x_{t-1}^s), x_t^s - x_{t-1}^s\rangle + \frac{L\eta^2}{2}\|v_{t-1}^s\|^2 \\ = & f(x_{t-1}^s) + \langle \nabla f(x_{t-1}^s) - v_{t-1}^s, -\eta v_{t-1}^s\rangle - \eta\|v_{t-1}^s\|^2 + \frac{L\eta^2}{2}\|v_{t-1}^s\|^2 \\ \overset{(i)}{\leq} & f(x_{t-1}^s) + \frac{\eta}{2}\|\nabla f(x_{t-1}^s) - v_{t-1}^s\|^2 + \frac{\eta}{2}\|v_{t-1}^s\|^2 - \Big(\eta - \frac{L\eta^2}{2}\Big)\|v_{t-1}^s\|^2. \end{aligned}$$

where (i) follows from the inequality that $\langle a, b\rangle \leq \frac{1}{2}(\|a\|^2 + \|b\|^2)$. Then, taking expectation $\mathbb{E}_{0,s}(\cdot)$ over the above inequality yields

$$\mathbb{E}_{0,s} f(x_t^s) \leq \mathbb{E}_{0,s} f(x_{t-1}^s) + \frac{\eta}{2}\mathbb{E}_{0,s}\|\nabla f(x_{t-1}^s) - v_{t-1}^s\|^2 - \Big(\frac{\eta}{2} - \frac{L\eta^2}{2}\Big)\mathbb{E}_{0,s}\|v_{t-1}^s\|^2. \tag{8}$$

Combining (8) and Lemma 1 yields, for $1 \leq t \leq m$

$$\begin{aligned}\mathbb{E}_{0,s} f(x_t^s) \leq & \mathbb{E}_{0,s} f(x_{t-1}^s) + \frac{\eta^3 L^2(t-1)}{2B}\mathbb{E}_{0,s}\sum_{i=0}^{t-2}\|v_i^s\|^2 + \frac{\eta I_{(N_s<n)}}{2N_s}\sigma^2 \\ & - \Big(\frac{\eta}{2} - \frac{L\eta^2}{2}\Big)\mathbb{E}_{0,s}\|v_{t-1}^s\|^2. \end{aligned}$$

Telescoping the above inequality over t from 1 to m yields

$$\begin{aligned}\mathbb{E}_{0,s} f(x_m^s) \leq & \mathbb{E}_{0,s} f(x_0^s) + \sum_{t=1}^m \frac{\eta^3 L^2(t-1)}{2B}\mathbb{E}_{0,s}\sum_{i=0}^{t-2}\|v_i^s\|^2 + \frac{\eta\sigma^2 m I_{(N_s<n)}}{2N_s} \\ & - \Big(\frac{\eta}{2} - \frac{L\eta^2}{2}\Big)\sum_{t=0}^{m-1}\mathbb{E}_{0,s}\|v_t^s\|^2 \\ \overset{(i)}{\leq} & \mathbb{E}_{0,s} f(x_0^s) + \frac{\eta^3 L^2 m^2}{2B}\mathbb{E}_{0,s}\sum_{i=0}^{m-1}\|v_i^s\|^2 + \frac{\eta\sigma^2 m I_{(N_s<n)}}{2N_s} \\ & - \Big(\frac{\eta}{2} - \frac{L\eta^2}{2}\Big)\sum_{t=0}^{m-1}\mathbb{E}_{0,s}\|v_t^s\|^2, \end{aligned} \tag{9}$$

where (i) follows from the fact that $\frac{\eta^3L^2(t-1)}{2B}\mathbb{E}_{0,s}\sum_{i=0}^{t-2}\|v_i^s\|^2 \le \frac{\eta^3L^2m}{2B}\mathbb{E}_{0,s}\sum_{i=0}^{m-1}\|v_i^s\|^2$. Recall that $N_s = \min\{c_\beta\sigma^2\beta_s^{-1}, c_\epsilon\sigma^2\epsilon^{-1}, n\}$ and $c_\beta, c_\epsilon \ge \alpha$. Then, we have

$$\frac{I_{(N_s<n)}}{N_s} \le \frac{1}{\min\{c_\beta\sigma^2\beta_s^{-1}, c_\epsilon\sigma^2\epsilon^{-1}\}} = \max\left\{\frac{\beta_s}{c_\beta\sigma^2}, \frac{\epsilon}{c_\epsilon\sigma^2}\right\} \le \max\left\{\frac{\beta_s}{\alpha\sigma^2}, \frac{\epsilon}{\alpha\sigma^2}\right\}, \tag{10}$$

which, in conjunction with (9), implies that

$$\begin{aligned}\mathbb{E}_{0,s}f(x_m^s) \overset{(i)}{\le}& \mathbb{E}_{0,s}f(x_0^s) + \frac{\eta^3L^2m^2}{2B}\mathbb{E}_{0,s}\sum_{i=0}^{m-1}\|v_i^s\|^2 + \eta m\left(\frac{\beta_s}{2\alpha} + \frac{\epsilon}{2\alpha}\right)\\ &- \Big(\frac{\eta}{2} - \frac{L\eta^2}{2}\Big)\sum_{t=0}^{m-1}\mathbb{E}_{0,s}\|v_t^s\|^2.\end{aligned} \tag{11}$$

where (i) follows from the fact that $\max(a,b) \le a+b$. Taking the expectation of (11) over $x_0^1, ..., x_0^s$, we obtain

$$\mathbb{E}f(x_m^s) \le \mathbb{E}f(x_0^s) + \frac{\eta m}{2\alpha}\mathbb{E}\beta_s + \frac{\eta m\epsilon}{2\alpha} - \Big(\frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{\eta^3L^2m^2}{2B}\Big)\sum_{t=0}^{m-1}\mathbb{E}\|v_t^s\|^2.$$

Recall that $\beta_1 \le \epsilon S$ and $\beta_s = \frac{1}{m}\sum_{t=1}^{m}\|v_{t-1}^{s-1}\|^2$ for $s = 2, ..., S$. Then, telescoping the above inequality over s from 1 to S and noting that $x_m^s = x_0^{s+1}$, we obtain

$$\begin{aligned}\mathbb{E}f(x_m^S) \le& \mathbb{E}f(x_0) - \Big(\frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{\eta^3L^2m^2}{2B}\Big)\sum_{s=1}^{S}\sum_{t=0}^{m-1}\mathbb{E}\|v_t^s\|^2 + \frac{\eta mS\epsilon}{2\alpha}\\ &+ \frac{\eta mS\epsilon}{2\alpha} + \sum_{s=2}^{S}\frac{\eta m}{2\alpha}\mathbb{E}\left(\frac{1}{m}\sum_{t=1}^{m}\|v_{t-1}^{s-1}\|^2\right)\\ \le& \mathbb{E}f(x_0) - \Big(\frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{\eta^3L^2m^2}{2B} - \frac{\eta}{2\alpha}\Big)\sum_{s=1}^{S}\sum_{t=0}^{m-1}\mathbb{E}\|v_t^s\|^2 + \frac{\eta mS\epsilon}{\alpha}.\end{aligned} \tag{12}$$

Dividing the both sides of (12) by ηSm and rearranging the terms, we obtain

$$\Big(\frac{1}{2} - \frac{1}{2\alpha} - \frac{L\eta}{2} - \frac{\eta^2L^2m^2}{2B}\Big)\frac{1}{Sm}\sum_{s=1}^{S}\sum_{t=0}^{m-1}\mathbb{E}\|v_t^s\|^2 \le \frac{f(x_0) - f^*}{\eta Sm} + \frac{\epsilon}{\alpha}, \tag{13}$$

where $f^* = \inf_{x\in\mathbb{R}^d} f(x) > -\infty$. Since the output x_ζ is chosen from $\{x_t^s\}_{t=0,...,m-1,s=1,...,S}$ uniformly at random, we have

$$\begin{aligned}Sm\,\mathbb{E}\|\nabla f(x_\zeta)\|^2 =& \sum_{s=1}^{S}\sum_{t=0}^{m-1}\mathbb{E}\|\nabla f(x_t^s)\|^2\\ \le& 2\sum_{s=1}^{S}\sum_{t=0}^{m-1}\mathbb{E}\|\nabla f(x_t^s) - v_t^s\|^2 + 2\sum_{s=1}^{S}\sum_{t=0}^{m-1}\mathbb{E}\|v_t^s\|^2\\ =& 2\sum_{s=1}^{S}\sum_{t=0}^{m-1}\mathbb{E}_{x_0^1,...,x_0^s}\left(\mathbb{E}_{0,s}\|\nabla f(x_t^s) - v_t^s\|^2\right) + 2\sum_{s=1}^{S}\sum_{t=0}^{m-1}\mathbb{E}\|v_t^s\|^2\\ \overset{(i)}{\le}& 2\sum_{s=1}^{S}\sum_{t=0}^{m-1}\mathbb{E}_{x_0^1,...,x_0^s}\left(\frac{\eta^2L^2m}{B}\mathbb{E}_{0,s}\sum_{i=0}^{m-1}\|v_i^s\|^2 + \frac{\beta_s}{\alpha} + \frac{\epsilon}{\alpha}\right) + 2\sum_{s=1}^{S}\sum_{t=0}^{m-1}\mathbb{E}\|v_t^s\|^2\end{aligned}$$

$$\leq 2\sum_{s=1}^{S}\left(\frac{\eta^2L^2m^2}{B}\mathbb{E}\sum_{i=0}^{m-1}\|v_i^s\|^2+\frac{m\beta_s}{\alpha}+\frac{m\epsilon}{\alpha}\right)+2\sum_{s=1}^{S}\sum_{t=0}^{m-1}\mathbb{E}\|v_t^s\|^2$$
$$\overset{(ii)}{\leq}\left(\frac{2\eta^2L^2m^2}{B}+\frac{2}{\alpha}+2\right)\sum_{s=1}^{S}\sum_{t=0}^{m-1}\mathbb{E}\|v_t^s\|^2+\frac{4Sm\epsilon}{\alpha} \tag{14}$$

where (i) follows from Lemma 1 and (10), and (ii) follows from the definition of β_s for $s = 1, ..., S$. Combining (13) and (14) and letting $\phi = \frac{1}{2} - \frac{1}{2\alpha} - \frac{L\eta}{2} - \frac{\eta^2L^2m^2}{2B}$ and $\psi = \frac{2\eta^2L^2m^2}{B} + \frac{2}{\alpha} + 2$, we have

$$\mathbb{E}\|\nabla f(x_\zeta)\|^2 \leq \frac{\psi(f(x_0)-f^*)}{\phi\eta Sm}+\frac{\psi\epsilon}{\phi\alpha}+\frac{4\epsilon}{\alpha}, \tag{15}$$

which finishes the proof. □

E.2 Proof of Corollary 1

Recall that $\eta = \frac{1}{4L}$, $B = m^2$ and $c_\beta, c_\epsilon \geq 16$. Then, we have $\alpha = 16$, $\phi \geq \frac{5}{16} > \frac{1}{4}$ and $\phi \leq \frac{9}{4}$ in Theorem 1, and thus

$$\mathbb{E}\|\nabla f(x_\zeta)\|^2 \leq \frac{36L(f(x_0)-f^*)}{K}+\frac{13}{16}\epsilon.$$

Thus, to achieve $\mathbb{E}\|\nabla f(x_\zeta)\|^2 < \epsilon$, AbaSVRG requires at most $192L(f(x_0) - f^*)\epsilon^{-1} = \Theta(\epsilon^{-1})$ iterations. Then, the total number of SFO calls is given by

$$\sum_{s=1}^{S}\min\{c_\beta\sigma^2\beta_s^{-1}, c_\epsilon\sigma^2\epsilon^{-1}, n\}+KB \leq S(c_\epsilon\sigma^2\epsilon^{-1}\wedge n)+KB \leq \mathcal{O}\left(\frac{\epsilon^{-1}\wedge n}{\epsilon\sqrt{B}}+\frac{B}{\epsilon}\right).$$

Furthermore, if we choose $B = n^{2/3} \wedge \epsilon^{-2/3}$, then SFO complexity of AbaSVRG becomes

$$\mathcal{O}\left(\frac{\epsilon^{-2/3}\wedge n^{2/3}}{\epsilon}\right) \leq \mathcal{O}\left(\frac{1}{\epsilon}\left(n\wedge\frac{1}{\epsilon}\right)^{2/3}\right). \tag{16}$$

E.3 Complexity under $B = m$

Corollary 6. *Let stepsize $\eta = \frac{1}{4L\sqrt{m}}$, mini-batch size $B = m$ and $c_\beta, c_\epsilon \geq 16$. Then, to obtain an ϵ-accurate solution x_ζ, the total number of SFO calls required by AbaSVRG is given by*

$$\sum_{s=1}^{S}\min\left\{\frac{c_\beta\sigma^2}{\beta_s}, \frac{c_\epsilon\sigma^2}{\epsilon}, n\right\}+KB \leq \mathcal{O}\left(\frac{n\wedge\epsilon^{-1}}{\sqrt{B}\epsilon}+\frac{B^{3/2}}{\epsilon}\right).$$

If we specially choose $B = n^{1/2} \wedge \epsilon^{-1/2}$, then the worst-case complexity is $\mathcal{O}\left(\frac{1}{\epsilon}(n\wedge\frac{1}{\epsilon})^{3/4}\right)$.

Proof. Since $\eta = \frac{1}{4L\sqrt{m}}$, $B = m$ and $c_\beta, c_\epsilon \geq 16$, we obtain $\alpha = 16$, $\phi = \frac{7}{16} - \frac{1}{8\sqrt{m}} \geq \frac{5}{16} > \frac{1}{4}$ and $\psi \leq \frac{9}{4}$ in Theorem 1, and thus

$$\mathbb{E}\|\nabla f(x_\zeta)\|^2 \leq \frac{36L\sqrt{m}(f(x_0)-f^*)}{K}+\frac{13}{16}\epsilon.$$

To achieve $\mathbb{E}\|\nabla f(x_\zeta)\|^2 < \epsilon$, AbaSVRG requires at most $192L\sqrt{m}(f(x_0) - f^*)\epsilon^{-1} = \Theta(\sqrt{m}\epsilon^{-1})$ iterations. Then, the total number of SFO calls is given by

$$\sum_{s=1}^{S}\min\{c_\beta\sigma^2\beta_s^{-1}, c_\epsilon\sigma^2\epsilon^{-1}, n\}+KB \leq S(c_\epsilon\sigma^2\epsilon^{-1}\wedge n)+KB \leq \mathcal{O}\left(\frac{\epsilon^{-1}\wedge n}{\epsilon\sqrt{B}}+\frac{B^{3/2}}{\epsilon}\right).$$

Furthermore, if we choose $B = n^{1/2} \wedge \epsilon^{-1/2}$, then the SFO complexity is $\mathcal{O}\left(\frac{1}{\epsilon}(n\wedge\frac{1}{\epsilon})^{3/4}\right)$. □

E.4 Proof of Theorem 2

In order to prove Theorem 2, we first use the following lemma to provide an upper bound on the estimation variance $\mathbb{E}_{0,s}\|\nabla f(x_t^s) - v_t^s\|^2$ for $0 \le t \le m-1$, where $\mathbb{E}_{t,s}(\cdot)$ denotes $\mathbb{E}(\cdot|x_0^1, x_0^2, ..., x_2^1, ..., x_t^s)$.

Lemma 2 (Adapted from Fang et al. 2018)**.** *Let Assumption 1 hold. Then, for* $0 \le t \le m-1$,

$$\mathbb{E}_{0,s}\|\nabla f(x_t^s) - v_t^s\|^2 \le \frac{\eta^2 L^2}{B}\sum_{i=0}^{t-1}\mathbb{E}_{0,s}\|v_i^s\|^2 + \frac{I_{(N_s<n)}}{N_s}\sigma^2. \tag{17}$$

where we define the stochastic gradients before the algorithm starts to satisfy $\sum_{i=0}^{-1}\mathbb{E}_{0,s}\|v_i^s\|^2 = 0$ *for easy presentation.*

Proof of Lemma 2. Combining A.3 and A.4 in Fang et al. 2018 yields, for $1 \le i \le m-1$,

$$\begin{aligned}\mathbb{E}_{i,s}\|\nabla f(x_i^s) - v_i^s\|^2 \le& \frac{L^2}{B}\|x_i^s - x_{i-1}^s\|^2 + \|\nabla f(x_{i-1}^s) - v_{i-1}^s\|^2\\ =& \frac{\eta^2 L^2}{B}\|v_{i-1}^s\|^2 + \|\nabla f(x_{i-1}^s) - v_{i-1}^s\|^2.\end{aligned}$$

Taking the expectation of the above inequality over $x_1^s, ..., x_i^s$, we have

$$\mathbb{E}_{0,s}\|\nabla f(x_i^s) - v_i^s\|^2 \le \frac{\eta^2 L^2}{B}\mathbb{E}_{0,s}\|v_{i-1}^s\|^2 + \mathbb{E}_{0,s}\|\nabla f(x_{i-1}^s) - v_{i-1}^s\|^2.$$

Then, telescoping the above inequality over i from 1 to t yields

$$\mathbb{E}_{0,s}\|\nabla f(x_t^s) - v_t^s\|^2 \le \frac{\eta^2 L^2}{B}\sum_{i=0}^{t-1}\mathbb{E}_{0,s}\|v_i^s\|^2 + \mathbb{E}_{0,s}\|\nabla f(x_0^s) - v_0^s\|^2. \tag{18}$$

Based on Lemma B.2 in Lei et al. 2017, we have

$$\mathbb{E}_{0,s}\|\nabla f(x_0^s) - v_0^s\|^2 \le \frac{I_{(N_s<n)}}{N_s}\sigma^2,$$

which, combined with (18), finishes the proof. □

Proof of Theorem 2. Based on Lemma 2, we now prove Theorem 2.

Since the objective function $f(\cdot)$ has a L-Lipschitz continuous gradient, we obtain that for $1 \le t \le m$,

$$\begin{aligned}f(x_t^s) \le& f(x_{t-1}^s) + \langle \nabla f(x_{t-1}^s), x_t^s - x_{t-1}^s\rangle + \frac{L\eta^2}{2}\|v_{t-1}^s\|^2\\ =& f(x_{t-1}^s) + \langle \nabla f(x_{t-1}^s) - v_{t-1}^s, -\eta v_{t-1}^s\rangle - \eta\|v_{t-1}^s\|^2 + \frac{L\eta^2}{2}\|v_{t-1}^s\|^2\\ \overset{(i)}{\le}& f(x_{t-1}^s) + \frac{\eta}{2}\|\nabla f(x_{t-1}^s) - v_{t-1}^s\|^2 + \frac{\eta}{2}\|v_{t-1}^s\|^2 - \eta\|v_{t-1}^s\|^2 + \frac{L\eta^2}{2}\|v_{t-1}^s\|^2\\ \le& f(x_{t-1}^s) + \frac{\eta}{2}\|\nabla f(x_{t-1}^s) - v_{t-1}^s\|^2 - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\|v_{t-1}^s\|^2,\end{aligned}$$

where (i) follows from the inequality that $\langle a, b\rangle \le \frac{1}{2}(\|a\|^2 + \|b\|^2)$. Then, taking expectation $\mathbb{E}_{0,s}$ over the above inequality and applying Lemma 2, we have, for $1 \le t \le m$,

$$\mathbb{E}_{0,s}f(x_t^s) \le \mathbb{E}_{0,s}f(x_{t-1}^s) + \frac{\eta}{2}\mathbb{E}_{0,s}\|\nabla f(x_{t-1}^s) - v_{t-1}^s\|^2 - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\mathbb{E}_{0,s}\|v_{t-1}^s\|^2$$

$$
\begin{aligned}
\leq & \mathbb{E}_{0,s} f(x_{t-1}^s) + \frac{\eta^3 L^2}{2B} \sum_{i=0}^{t-2} \mathbb{E}_{0,s} \|v_i^s\|^2 + \frac{I_{(N_s<n)}}{N_s} \frac{\eta\sigma^2}{2} - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right) \mathbb{E}_{0,s} \|v_{t-1}^s\|^2 \\
\overset{(i)}{\leq} & \mathbb{E}_{0,s} f(x_{t-1}^s) + \frac{\eta^3 L^2}{2B} \sum_{i=0}^{m-1} \mathbb{E}_{0,s} \|v_i^s\|^2 + \max\left\{\frac{\eta\beta_s}{2\alpha}, \frac{\eta\epsilon}{2\alpha}\right\} - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right) \mathbb{E}_{0,s} \|v_{t-1}^s\|^2
\end{aligned}
$$

where (i) follows from $t-2 < m-1$ and (10). Telescoping the above inequality over t from 1 to m and using $\max(a,b) \leq a+b$ yield

$$
\mathbb{E}_{0,s} f(x_m^s) \leq \mathbb{E}_{0,s} f(x_0^s) - \left(\frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{\eta^3 L^2 m}{2B}\right) \sum_{t=0}^{m-1} \mathbb{E}_{0,s} \|v_t^s\|^2 + \frac{\eta m \beta_s}{2\alpha} + \frac{\eta m \epsilon}{2\alpha}.
$$

Taking the expectation of the above inequality over $x_0^1, ..., x_0^s$, we obtain

$$
\mathbb{E} f(x_m^s) \leq \mathbb{E} f(x_0^s) - \left(\frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{\eta^3 L^2 m}{2B}\right) \sum_{t=0}^{m-1} \mathbb{E} \|v_t^s\|^2 + \frac{\eta m}{2\alpha} \mathbb{E}(\beta_s) + \frac{\eta m \epsilon}{2\alpha}.
$$

Recall that $\beta_1 \leq \epsilon S$ and $\beta_s = \frac{1}{m} \sum_{t=0}^{m-1} \|v_t^{s-1}\|^2$ for $s = 2, ..., S$. Then, telescoping the above inequality over s from 1 to S and noting that $x_m^s = x_0^{s+1} = \tilde{x}^s$, we have

$$
\begin{aligned}
\mathbb{E} f(\tilde{x}^S) \leq & \mathbb{E} f(x_0) - \left(\frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{\eta^3 L^2 m}{2B}\right) \sum_{s=1}^{S} \sum_{t=0}^{m-1} \mathbb{E} \|v_t^s\|^2 + \frac{\eta m S \epsilon}{2\alpha} \\
& + \frac{\eta}{2\alpha} \sum_{s=1}^{S-1} \sum_{t=0}^{m-1} \mathbb{E} \|v_t^s\|^2 \\
\leq & \mathbb{E} f(x_0) - \left(\frac{\eta}{2} - \frac{\eta}{2\alpha} - \frac{L\eta^2}{2} - \frac{\eta^3 L^2 m}{2B}\right) \sum_{s=1}^{S} \sum_{t=0}^{m-1} \mathbb{E} \|v_t^s\|^2 + \frac{\eta m S \epsilon}{2\alpha}.
\end{aligned}
$$

Dividing the both sides of the above inequality by $\eta S m$ and rearranging the terms, we obtain

$$
\left(\frac{1}{2} - \frac{1}{2\alpha} - \frac{L\eta}{2} - \frac{\eta^2 L^2 m}{2B}\right) \frac{1}{Sm} \sum_{s=1}^{S} \sum_{t=0}^{m-1} \mathbb{E} \|v_t^s\|^2 \leq \frac{f(x_0) - f^*}{\eta S m} + \frac{\epsilon}{2\alpha}. \tag{19}
$$

Since the output x_ζ is chosen from $\{x_t^s\}_{t=0,...,m-1, s=1,...,S}$ uniformly at random, we have

$$
\begin{aligned}
Sm\, \mathbb{E} \|\nabla f(x_\zeta)\|^2 = & \sum_{s=1}^{S} \sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^s)\|^2 \\
\leq & 2 \sum_{s=1}^{S} \sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^s) - v_t^s\|^2 + 2 \sum_{s=1}^{S} \sum_{t=0}^{m-1} \mathbb{E} \|v_t^s\|^2 \\
= & 2 \sum_{s=1}^{S} \sum_{t=0}^{m-1} \mathbb{E}_{x_0^1, ..., x_0^s} \left(\mathbb{E}_{0,s} \|\nabla f(x_t^s) - v_t^s\|^2\right) + 2 \sum_{s=1}^{S} \sum_{t=0}^{m-1} \mathbb{E} \|v_t^s\|^2 \\
\overset{(i)}{\leq} & 2 \sum_{s=1}^{S} \sum_{t=0}^{m-1} \mathbb{E}_{x_0^1, ..., x_0^s} \left(\frac{\eta^2 L^2}{B} \mathbb{E}_{0,s} \sum_{i=0}^{m-1} \|v_i^s\|^2 + \frac{\beta_s}{\alpha} + \frac{\epsilon}{\alpha}\right) + 2 \sum_{s=1}^{S} \sum_{t=0}^{m-1} \mathbb{E} \|v_t^s\|^2 \\
\leq & 2 \sum_{s=1}^{S} \left(\frac{\eta^2 L^2 m}{B} \mathbb{E} \sum_{i=0}^{m-1} \|v_i^s\|^2 + \frac{m\beta_s}{\alpha} + \frac{m\epsilon}{\alpha}\right) + 2 \sum_{s=1}^{S} \sum_{t=0}^{m-1} \mathbb{E} \|v_t^s\|^2
\end{aligned}
$$

$$\overset{(ii)}{\leq}\left(\frac{2\eta^2L^2m}{B}+\frac{2}{\alpha}+2\right)\sum_{s=1}^{S}\sum_{t=0}^{m-1}\mathbb{E}\|v_t^s\|^2+\frac{4Sm\epsilon}{\alpha} \tag{20}$$

where (i) follows from Lemma 2 and (10) and (ii) follows from the definition of β_s for $s=1,...,S$. Let $\phi=\frac{1}{2}-\frac{1}{2\alpha}-\frac{L\eta}{2}-\frac{\eta^2L^2m}{2B}$, $\psi=\frac{2\eta^2L^2m}{B}+\frac{2}{\alpha}+2$ and $K=Sm$. Then, combining (20) and (19), we finish the proof.

E.5 Proof of Corollary 2

Recall that $1\leq B\leq n^{1/2}\wedge\epsilon^{-1/2}$, $m=(n\wedge\frac{1}{\epsilon})B^{-1}$, $\eta=\frac{1}{4L}\sqrt{\frac{B}{m}}$ and $c_\beta,c_\epsilon\geq 16$. Then, we have $\alpha=16$, $m\geq n^{1/2}\wedge\epsilon^{-1/2}\geq B$ and $\eta\leq\frac{1}{4L}$. Thus, we obtain

$$\phi=\frac{1}{2}-\frac{1}{2\alpha}-\frac{L\eta}{2}-\frac{\eta^2L^2m}{2B}\geq\frac{5}{16}>\frac{1}{4}\text{ and }\psi\leq\frac{9}{4},$$

which, in conjunction with Theorem 2, implies that

$$\mathbb{E}\|\nabla f(x_\zeta)\|^2\leq\frac{36L\sqrt{m}(f(x_0)-f^*)}{\sqrt{B}K}+\frac{17}{32}\epsilon.$$

Thus, to achieve $\mathbb{E}\|\nabla f(x_\zeta)\|^2<\epsilon$, AbaSPIDER requires at most $\frac{384L\sqrt{m}(f(x_0)-f^*)}{5\sqrt{B}\epsilon}=\Theta\big(\frac{\sqrt{m}}{\sqrt{B}\epsilon}\big)$ iterations. Then, the total number of SFO calls is given by

$$\sum_{s=1}^{S}\min\{c_\beta\sigma^2\beta_s^{-1},c_\epsilon\sigma^2\epsilon^{-1},n\}+KB\leq S(c_\epsilon\sigma^2\epsilon^{-1}\wedge n)+KB\leq\mathcal{O}\left(\frac{\epsilon^{-1}\wedge n}{\epsilon\sqrt{mB}}+\frac{\sqrt{mB}}{\epsilon}\right),$$

which, in conjunction with $mB=n\wedge\frac{1}{\epsilon}$, finishes the proof. □

F Proofs for Results in Section 3

F.1 Useful Lemmas

In this section, we provide some useful lemmas. The following two lemmas follow directly from Assumptions in Subsection 3.4.

Lemma 3 (Papini et al. (2018)). *Under Assumptions 2 and 3, the following holds:*

(i) ∇J *is* L*- Lipschitz, i.e., for any* $\theta_1,\theta_2\in\mathbb{R}^d$*:* $\|\nabla J(\theta_1)-\nabla J(\theta_2)\|\leq L\|\theta_1-\theta_2\|$.

(ii) $g(\tau|\theta)$ *is Lipschitz continuous with Lipschitz constant* L_g*, i.e., for any trajectory* $\tau\in\mathcal{T}$*:*

$$\|g(\tau|\theta_1)-g(\tau|\theta_2)\|\leq L_g\|\theta_1-\theta_2\|.$$

(iii) $g(\tau|\theta)$ *and* $\nabla\log(p(\tau|\theta))$ *are bounded, i.e., there exist positive constants* $0\leq\Gamma,M<\infty$ *such that for any* $\tau\in\mathcal{T}$ *and* $\theta\in\Theta$*:*

$$\|\nabla\log(p(\tau|\theta))\|^2\leq M\quad\text{ and }\quad\|g(\tau|\theta)\|^2\leq\Gamma.$$

Lemma 4 (Xu et al. (2019b,a) Lemma A.1). *For any* $\theta_1,\theta_2\in\mathcal{R}^d$*, let* $\omega\big(\tau|\theta_1,\theta_2\big)=p(\tau|\theta_1)/p(\tau|\theta_2)$*. Under Assumptions 3 and 4, it holds that*

$$\mathop{\mathbb{E}}_{\tau\sim p(\cdot|\theta_1)}\left\|1-\frac{p(\tau|\theta_2)}{p(\tau|\theta_1)}\right\|^2=\mathbb{V}\mathrm{ar}\left(\omega\big(\tau|\theta_1,\theta_2\big)\right)\leq\alpha\|\theta_1-\theta_2\|_2^2,$$

where α *is a positive constant.*

The following lemma captures an important property for the trajectory gradients, and its proof follows directly from Lemma 4.

Lemma 5. *Under Assumptions 2, 3, and 4 the following inequality holds for any $\theta_1, \theta_2 \in \mathcal{R}^d$,*

$$\mathop{\mathbb{E}}_{\tau\sim p(\cdot|\theta_1)} \left\|g(\tau|\theta_1) - \omega(\tau|\theta_1,\theta_2)g(\tau|\theta_2)\right\|^2 \leq Q\left\|\theta_1-\theta_2\right\|^2,$$

where the importance sampling function $\omega(\tau|\theta_1,\theta_2) := p(\tau|\theta_2)/p(\tau|\theta_1)$, and the constant $Q := 2(L_g^2 + \Gamma\alpha)$ with constants L_g, Γ and α given in Lemmas 3 and 4.

F.2 Proof of Theorem 3

In this section, we provide the convergence analysis for AbaSVRPG. To simplify notations, we use $\mathbb{E}_k[\cdot]$ to denote the expectation operation conditioned on all the randomness before θ_k, i.e., $\mathbb{E}[\cdot|\theta_0, \cdots, \theta_k]$ and $n_k = \lfloor k/m \rfloor \times m$.

To prove the convergence of AbaSVRPG, we first present a general iteration analysis for an algorithm with the update rule taking the form of $\theta_{k+1} = \theta_k + \eta v_k$, for $k = 0, 1, \cdots$. The proof of Lemma 6 can be found in Appendix G.

Lemma 6. *Let ∇J be L-Lipschitz, and $\theta_{k+1} = \theta_{k+1} + \eta v_k$. Then, the following inequality holds:*

$$\mathbb{E}\, J(\theta_{k+1}) - \mathbb{E}\, J(\theta_k) \geq \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\mathbb{E}\left\|v_k\right\|^2 - \frac{\eta}{2}\mathbb{E}\left\|v_k - \nabla J(\theta_k)\right\|^2.$$

Since, we do not specify the exact form of v_k, Lemma 6 is applicable to various algorithms such as AbaSVRPG and AbaSPIDER-PG with the same type of update rules.

We next present the variance bound of AbaSVRPG.

Proposition 1. *Let Assuptions 2, 3, and 4 hold. Then, for $k = 0, \cdots K$, the variance of the gradient estimator v_k of AbaSVRPG can be bounded as*

$$\mathbb{E}\left\|v_k - \nabla J(\theta_k)\right\|^2 \leq (k-n_k)\frac{Q\eta^2}{B}\sum_{i=n_k}^{k}\mathbb{E}\left\|v_i\right\|^2 + \mathbb{E}\left\|v_{n_k} - \nabla J(\theta_{n_k})\right\|^2,$$

where $\|v_i\| = 0$ for $i = -1, \cdots, -m$ for simple notations.

Proof of Proposition 1. To bound the variance $\mathbb{E}\left\|v_k - \nabla J(\theta_k)\right\|^2$, it is sufficient to bound $\mathbb{E}_k\left\|v_k - \nabla J(\theta_k)\right\|^2$ since by the tower property of expectation we have $\mathbb{E}\left\|v_k - \nabla J(\theta_k)\right\|^2 = \mathbb{E}\,\mathbb{E}_k\left\|v_k - \nabla J(\theta_k)\right\|^2$. Thus, we first bound $\mathbb{E}_k\left\|v_k - \nabla J(\theta_k)\right\|^2$ for the case with $\mod(k,m) \neq 0$, and then generalize it to the case with $\mod(k,m) = 0$.

$$\begin{aligned}
&\mathbb{E}_k\left\|v_k - \nabla J(\theta_k)\right\|^2 \\
&\overset{\text{(i)}}{=} \mathbb{E}_k\left\|\frac{1}{B}\sum_{i=1}^{B} g(\tau_i|\theta_k) - \frac{1}{B}\sum_{i=1}^{B}\omega(\tau_i|\theta_k,\tilde{\theta})g(\tau_i|\tilde{\theta}) + \tilde{v} - \nabla J(\theta_k)\right\|^2 \\
&= \mathbb{E}_k\left\|\frac{1}{B}\sum_{i=1}^{B} g(\tau_i|\theta_k) - \frac{1}{B}\sum_{i=1}^{B}\omega(\tau_i|\theta_k,\tilde{\theta})g(\tau_i|\tilde{\theta}) + \nabla J(\tilde{\theta}) - \nabla J(\theta_k) + \tilde{v} - \nabla J(\tilde{\theta})\right\|^2 \\
&= \mathbb{E}_k\left\|\frac{1}{B}\sum_{i=1}^{B} g(\tau_i|\theta_k) - \frac{1}{B}\sum_{i=1}^{B}\omega(\tau_i|\theta_k,\tilde{\theta})g(\tau_i|\tilde{\theta}) + \nabla J(\tilde{\theta}) - \nabla J(\theta_k)\right\|^2 + \mathbb{E}_k\left\|\tilde{v} - \nabla J(\tilde{\theta})\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&+ 2\mathbb{E}_k \left\langle \frac{1}{B}\sum_{i=1}^{B} g(\tau_i|\theta_k) - \frac{1}{B}\sum_{i=1}^{B} \omega(\tau_i|\theta_k,\tilde{\theta}) g(\tau_i|\tilde{\theta}) + \nabla J(\tilde{\theta}) - \nabla J(\theta_k), \tilde{v} - \nabla J(\tilde{\theta}) \right\rangle \\
&\overset{\text{(ii)}}{=} \mathbb{E}_k \left\| \frac{1}{B}\sum_{i=1}^{B} g(\tau_i|\theta_k) - \frac{1}{B}\sum_{i=1}^{B} \omega(\tau_i|\theta_k,\tilde{\theta}) g(\tau_i|\tilde{\theta}) + \nabla J(\tilde{\theta}) - \nabla J(\theta_k) \right\|^2 + \mathbb{E}_k \left\| \tilde{v} - \nabla J(\tilde{\theta}) \right\|^2 \\
&\overset{\text{(iii)}}{\leq} \mathop{\mathbb{E}_k}_{\tau\sim p(\cdot|\theta_k)} \frac{1}{B} \left\| g(\tau|\theta_k) - \omega(\tau|\theta_k,\tilde{\theta}) g(\tau|\tilde{\theta}) + \nabla J(\tilde{\theta}) - \nabla J(\theta_k) \right\|^2 + \mathbb{E}_k \left\| \tilde{v} - \nabla J(\tilde{\theta}) \right\|^2 \\
&\overset{\text{(iv)}}{\leq} \mathop{\mathbb{E}_k}_{\tau\sim p(\cdot|\theta_k)} \frac{1}{B} \left\| g(\tau|\theta_k) - \omega(\tau|\theta_k,\tilde{\theta}) g(\tau|\tilde{\theta}) \right\|^2 + \mathbb{E}_k \left\| \tilde{v} - \nabla J(\tilde{\theta}) \right\|^2 \\
&\overset{\text{(v)}}{\leq} \frac{Q}{B} \left\| \theta_k - \tilde{\theta} \right\|^2 + \mathbb{E}_k \left\| \tilde{v} - \nabla J(\tilde{\theta}) \right\|^2 = \frac{Q}{B} \left\| \theta_k - \theta_{k-1} + \theta_{k-1} \cdots \theta_{n_k} \right\|^2 + \mathbb{E}_k \left\| \tilde{v} - \nabla J(\tilde{\theta}) \right\|^2 \\
&\overset{\text{(vi)}}{\leq} \frac{Q}{B}(k - n_k) \sum_{i=n_k}^{k-1} \left\| \theta_{i+1} - \theta_i \right\|^2 + \mathbb{E}_k \left\| \tilde{v} - \nabla J(\tilde{\theta}) \right\|^2,
\end{aligned}
$$

where (i) follows from the definition of v_k in Algorithm 3, (ii) follows from the fact that

$$
\mathop{\mathbb{E}}_{k} \left[\frac{1}{B}\sum_{i=1}^{B} g(\tau_i|\theta_k) - \frac{1}{B}\sum_{i=1}^{B} \omega(\tau_i|\theta_k,\tilde{\theta}) g(\tau_i|\tilde{\theta}) + \nabla J(\tilde{\theta}) - \nabla J(\theta_k) \right] = 0,
$$

and thus given $\theta_k, \cdots, \tilde{\theta}$, the expectation of the inner product is 0, (iii) follows from Lemma 7, (iv) follows from the fact that $\mathbb{V}\mathrm{ar}(X) \leq \mathbb{E}\left\|X\right\|^2$, (v) follows from Lemma 5 we provide in Appendix G, and (vi) follows from the vector inequality that $\left\|\sum_{i=1}^{m} \theta_i\right\|^2 \leq m \sum_{i=1}^{m} \left\|\theta_i\right\|^2$.

Therefore, we have

$$
\begin{aligned}
\mathbb{E}\left\|v_k - \nabla J(\theta_k)\right\|^2 &= \mathbb{E}\,\mathbb{E}_k \left\|v_k - \nabla J(\theta_k)\right\|^2 \\
&\leq \frac{Q}{B}(k - n_k) \sum_{i=n_k}^{k-1} \mathbb{E}\left\|\theta_{i+1} - \theta_i\right\|^2 + \mathbb{E}\left\| \tilde{v} - \nabla J(\tilde{\theta}) \right\|^2 \\
&\leq \frac{Q}{B}(k - n_k) \sum_{i=n_k}^{k} \mathbb{E}\left\|\theta_{i+1} - \theta_i\right\|^2 + \mathbb{E}\left\| \tilde{v} - \nabla J(\tilde{\theta}) \right\|^2 \\
&\leq \frac{Q}{B}(k - n_k)\eta^2 \sum_{i=n_k}^{k} \mathbb{E}\left\|v_i\right\|^2 + \mathbb{E}\left\| \tilde{v} - \nabla J(\tilde{\theta}) \right\|^2 \\
&\overset{\text{(i)}}{=} (k - n_k)\frac{Q\eta^2}{B} \sum_{i=n_k}^{k} \mathbb{E}\left\|v_i\right\|^2 + \mathbb{E}\left\|v_{n_k} - \nabla J(\theta_{n_k})\right\|^2,
\end{aligned}
$$

where (i) follows from the fact that at iteration k, $\tilde{v} = v_{n_k}$ and $\tilde{\theta} = \theta_{n_k}$. It is also straightforward to check that the above inequality holds for any k with $\mod(k, m) = 0$. □

Proof of Theorem 3

Since in Algorithm 3, ∇J is L-Lipschitz, and $\theta_{k+1} = \theta_{k+1} + \eta v_k$, we obtain the following inequality directly from Lemma 6:

$$
\mathbb{E}\, J(\theta_{k+1}) - \mathbb{E}\, J(\theta_k) \geq \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right) \mathbb{E}\left\|v_k\right\|^2 - \frac{\eta}{2}\mathbb{E}\left\|v_k - \nabla J(\theta_k)\right\|^2. \tag{21}
$$

By Proposition 1, we have following variance bound:

$$\mathbb{E}\left\|v_k - \nabla J(\theta_k)\right\|^2 \leq (k-n_k)\frac{Q\eta^2}{B}\sum_{i=n_k}^{k}\mathbb{E}\left\|v_i\right\|^2 + \mathbb{E}\left\|v_{n_k} - \nabla J(\theta_{n_k})\right\|^2 \tag{22}$$

Moreover, for $\mod(k,m)=0$, we obtain

$$\begin{aligned}\mathbb{E}\|v_k - \nabla f(x_k)\|^2 &= \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla g(\tau_i|\theta_k) - \nabla J(\theta_k)\right\|^2 \\ &\overset{\text{(i)}}{=} \frac{1}{N}\mathop{\mathbb{E}}_{\tau\sim p(\cdot|\theta_k)}\|\nabla g(\tau|\theta_k) - \nabla J(\theta_k)\|^2 \overset{(ii)}{\leqslant} \frac{\sigma^2}{N} \\ &\overset{(iii)}{\leqslant} \frac{\beta}{\alpha m}\sum_{i=n_k-m}^{n_k-1}\|v_i\|^2 + \frac{\epsilon}{\alpha}.\end{aligned} \tag{23}$$

where (i) follows from Lemma 7, (ii) follows from Assumption 4, and (iii) follows from the fact that

$$N = \frac{\alpha\sigma^2}{\frac{\beta}{m}\sum_{i=n_k-m}^{n_k-1}\|v_i\|^2 + \epsilon},$$

where $\alpha > 0$ and $\beta \geqslant 0$.

Plugging (23) into (22), we obtain

$$\mathbb{E}\left\|v_k - \nabla J(\theta_k)\right\|^2 \leq (k-n_k)\frac{Q\eta^2}{B}\sum_{i=n_k}^{k}\mathbb{E}\left\|v_i\right\|^2 + \frac{\beta}{\alpha m}\sum_{i=n_k-m}^{n_k-1}\|v_i\|^2 + \frac{\epsilon}{\alpha}. \tag{24}$$

Plugging (24) into (21), we obtain

$$\mathbb{E}\,J(\theta_{k+1}) - \mathbb{E}\,J(\theta_k) \geq \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\mathbb{E}\left\|v_k\right\|^2 - \frac{Q\eta^3}{2B}(k-n_k)\sum_{i=n_k}^{k}\mathbb{E}\left\|v_i\right\|^2 - \frac{\eta\beta}{2\alpha m}\sum_{i=n_k-m}^{n_k-1}\|v_i\|^2 - \frac{\epsilon\eta}{2\alpha}.$$

We note that for a given k, any iteration $n_k \leq i \leq k$ shares the same $\tilde{\theta}$, and all their corresponding n_i satisfies $n_i = n_k$. Thus, take the summation of the above inequality over k from n_k to k, we obtain

$$\begin{aligned}&\mathbb{E}J(\theta_{k+1}) - \mathbb{E}\,J(\theta_{n_k}) \\ &\geq \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\sum_{i=n_k}^{k}\mathbb{E}\left\|v_i\right\|^2 - \frac{Q\eta^3}{2B}\sum_{i=n_k}^{k}(i-n_k)\sum_{j=n_k}^{i}\mathbb{E}\left\|v_j\right\|^2 - \sum_{i=n_k}^{k}\frac{\eta\beta}{2\alpha m}\sum_{j=n_k-m}^{n_k-1}\|v_j\|^2 - \sum_{i=n_k}^{k}\frac{\epsilon\eta}{2\alpha} \\ &\geq \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\sum_{i=n_k}^{k}\mathbb{E}\left\|v_i\right\|^2 - \frac{Q\eta^3}{2B}\sum_{i=n_k}^{k}(k-n_k)\sum_{j=n_k}^{k}\mathbb{E}\left\|v_j\right\|^2 - \sum_{i=n_k}^{k}\frac{\eta\beta}{2\alpha m}\sum_{j=n_k-m}^{n_k-1}\|v_j\|^2 - \sum_{i=n_k}^{k}\frac{\epsilon\eta}{2\alpha} \\ &= \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\sum_{i=n_k}^{k}\mathbb{E}\left\|v_i\right\|^2 - \frac{Q\eta^3(k-n_k)(k-n_k+1)}{2B}\sum_{j=n_k}^{k}\mathbb{E}\left\|v_j\right\|^2 - \sum_{i=n_k}^{k}\frac{\eta\beta}{2\alpha m}\sum_{j=n_k-m}^{n_k-1}\|v_j\|^2 - \sum_{i=n_k}^{k}\frac{\epsilon\eta}{2\alpha} \\ &= \left(\frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{Q\eta^3(k-n_k)(k-n_k+1)}{2B}\right)\sum_{i=n_k}^{k}\mathbb{E}\left\|v_i\right\|^2 - \sum_{i=n_k}^{k}\frac{\eta\beta}{2\alpha m}\sum_{j=n_k-m}^{n_k-1}\|v_j\|^2 - \sum_{i=n_k}^{k}\frac{\epsilon\eta}{2\alpha} \\ &\overset{\text{(i)}}{\geq} \left(\frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{Q\eta^3m^2}{2B}\right)\sum_{i=n_k}^{k}\mathbb{E}\left\|v_i\right\|^2 - \sum_{i=n_k}^{k}\frac{\eta\beta}{2\alpha m}\sum_{j=n_k-m}^{n_k-1}\|v_j\|^2 - \sum_{i=n_k}^{k}\frac{\epsilon\eta}{2\alpha}\end{aligned}$$

$$\overset{\text{(ii)}}{=} \phi \sum_{i=n_k}^{k} \mathbb{E}\left\|v_i\right\|^2 - \frac{\eta\beta(k-n_k+1)}{2\alpha m} \sum_{i=n_k-m}^{n_k-1} \left\|v_i\right\|^2 - \sum_{i=n_k}^{k} \frac{\epsilon\eta}{2\alpha}$$
$$\overset{\text{(iii)}}{\geq} \phi \sum_{i=n_k}^{k} \mathbb{E}\left\|v_i\right\|^2 - \frac{\eta\beta}{2\alpha} \sum_{i=n_k-m}^{n_k-1} \left\|v_i\right\|^2 - \sum_{i=n_k}^{k} \frac{\epsilon\eta}{2\alpha}, \tag{25}$$

where (i) follows from the fact that $k - n_k < k - n_k + 1 \leq m$, (ii) follows from the fact that $\phi := \left(\frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{Q\eta^3 m^2}{2B}\right)$, and (iii) follows because $(k - n_k + 1)/m \leqslant 1$.

Now, we are ready to bound $J(\theta_{K+1}) - J(\theta_0)$.

$$\begin{aligned}
&\mathbb{E}J(\theta_{K+1}) - \mathbb{E}\, J(\theta_0) \\
&= \mathbb{E}\, J(\theta_{K+1}) - \mathbb{E}\, J(\theta_{n_K}) + \mathbb{E}\, J(\theta_{n_K}) \cdots + \mathbb{E}\, J(\theta_m) - \mathbb{E}\, J(\theta_0) \\
&\overset{\text{(i)}}{\geq} \phi \sum_{i=n_K}^{K} \mathbb{E}\left\|v_i\right\|^2 - \frac{\eta\beta}{2\alpha} \sum_{i=n_K-m}^{n_K-1} \left\|v_i\right\|^2 - \sum_{i=n_K}^{K} \frac{\epsilon\eta}{2\alpha} + \cdots + \phi \sum_{i=0}^{m-1} \mathbb{E}\left\|v_i\right\|^2 - \frac{\eta\beta}{2\alpha} \sum_{i=-m}^{-1} \left\|v_i\right\|^2 - \sum_{i=0}^{m} \frac{\epsilon\eta}{2\alpha} \\
&\overset{\text{(ii)}}{\geq} \phi \sum_{i=0}^{K} \mathbb{E}\left\|v_i\right\|^2 - \frac{\eta\beta}{2\alpha} \sum_{i=0}^{n_K-1} \left\|v_i\right\|^2 - \sum_{i=0}^{K} \frac{\epsilon\eta}{2\alpha} \\
&\geqslant \left(\phi - \frac{\eta\beta}{2\alpha}\right) \sum_{i=0}^{K} \mathbb{E}\left\|v_i\right\|^2 - \sum_{i=0}^{K} \frac{\epsilon\eta}{2\alpha},
\end{aligned}$$

where (i) follows from (25), and (ii) follows from the fact that we define $\|v_{-1}\| = \cdots = \|v_{-m}\| = 0$.

Thus, we obtain

$$\begin{aligned}
\left(\phi - \frac{\eta\beta}{2\alpha}\right) \sum_{i=0}^{K} \mathbb{E}\left\|v_i\right\|^2 &\leq \mathbb{E}\, J(\theta_{K+1}) - \mathbb{E}\, J(\theta_0) + \sum_{i=0}^{K} \frac{\epsilon\eta}{2\alpha} \\
&\overset{\text{(i)}}{\leq} J(\theta^*) - J(\theta_0) + \sum_{i=0}^{K} \frac{\epsilon\eta}{2\alpha},
\end{aligned}$$

where (i) follows because $\theta^* := \arg\max_{\theta \in \mathbb{R}^d} J(\theta)$. Here, we assume $\left(\phi - \frac{\eta\beta}{2\alpha}\right) > 0$ to continue our proof. Such an assumption can be satisfied by parameter tuning as shown in (32). Therefore, we obtain

$$\sum_{i=0}^{K} \mathbb{E}\left\|v_i\right\|^2 \leq \left(\phi - \frac{\eta\beta}{2\alpha}\right)^{-1} \left(J(\theta^*) - J(\theta_0) + \sum_{i=0}^{K} \frac{\epsilon\eta}{2\alpha}\right). \tag{26}$$

With (26), we next bound the gradient norm, i.e., $\|\nabla J(\theta_\xi)\|$, of the output of AbaSVRPG. Observe that

$$\mathbb{E}\|\nabla J(\theta_\xi)\|^2 = \mathbb{E}\|\nabla J(\theta_\xi) - v_\xi + v_\xi\|^2 \leq 2\mathbb{E}\|\nabla J(\theta_\xi) - v_\xi\|^2 + 2\mathbb{E}\|v_\xi\|^2. \tag{27}$$

Therefore, it is sufficient to bound the two terms on the right hand side of the above inequality. First, note that

$$\mathbb{E}\|v_\xi\|^2 \overset{\text{(i)}}{=} \frac{1}{K+1} \sum_{i=0}^{K} \mathbb{E}\|v_i\|^2 \overset{\text{(ii)}}{\leq} \left(\phi - \frac{\eta\beta}{2\alpha}\right)^{-1} \left(\frac{J(\theta^*) - J(\theta_0)}{K+1} + \frac{\epsilon\eta}{2\alpha}\right), \tag{28}$$

where (i) follows from the fact that ξ is selected uniformly at random from $\{0, \ldots, K\}$, and (ii) follows from (26). On the other hand, we observe that

$$\mathbb{E}\|\nabla J(\theta_\xi) - v_\xi\|^2$$

$$
\begin{aligned}
&\overset{\text{(i)}}{=} \frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\nabla J(\theta_k)-v_k\|^2\\
&\overset{\text{(ii)}}{\leq} \frac{1}{K+1}\sum_{k=0}^{K}\left((k-n_k)\frac{Q\eta^2}{B}\sum_{i=n_k}^{k}\mathbb{E}\left\|v_i\right\|^2+\frac{\beta}{\alpha m}\sum_{i=n_k-m}^{n_k-1}\left\|v_i\right\|^2+\frac{\epsilon}{\alpha}\right)\\
&\overset{\text{(iii)}}{=} \frac{Q\eta^2 m}{B(K+1)}\sum_{k=0}^{K}\sum_{i=n_k}^{k}\mathbb{E}\left\|v_i\right\|^2+\frac{\beta}{\alpha m(K+1)}\sum_{k=0}^{K}\sum_{i=n_k-m}^{n_k-1}\left\|v_i\right\|^2+\frac{\epsilon}{\alpha}\\
&\overset{\text{(iv)}}{=} \frac{Q\eta^2 m}{B(K+1)}\left(\sum_{k=0}^{m-1}\sum_{i=0}^{k}\mathbb{E}\left\|v_i\right\|^2+\cdots+\sum_{k=n_K}^{K}\sum_{i=n_K}^{k}\mathbb{E}\left\|v_i\right\|^2\right)+\frac{\beta}{\alpha m(K+1)}\sum_{k=0}^{K}\sum_{i=n_k-m}^{n_k-1}\left\|v_i\right\|^2+\frac{\epsilon}{\alpha}\\
&\leqslant \frac{Q\eta^2 m}{B(K+1)}\left(\sum_{k=0}^{m-1}\sum_{i=0}^{m-1}\mathbb{E}\left\|v_i\right\|^2+\cdots+\sum_{k=n_K}^{K}\sum_{i=n_K}^{K}\mathbb{E}\left\|v_i\right\|^2\right)+\frac{\beta}{\alpha m(K+1)}\sum_{k=0}^{K}\sum_{i=n_k-m}^{n_k-1}\left\|v_i\right\|^2+\frac{\epsilon}{\alpha}\\
&\overset{\text{(v)}}{\leq} \frac{Q\eta^2 m^2}{B(K+1)}\sum_{i=0}^{K}\mathbb{E}\left\|v_i\right\|^2+\frac{\beta}{\alpha m(K+1)}\sum_{k=0}^{K}\sum_{i=n_k-m}^{n_k-1}\left\|v_i\right\|^2+\frac{\epsilon}{\alpha}\\
&\overset{\text{(vi)}}{\leq} \frac{Q\eta^2 m^2}{B(K+1)}\sum_{i=0}^{K}\mathbb{E}\left\|v_i\right\|^2+\frac{\beta}{\alpha m(K+1)}\left(\sum_{k=0}^{m-1}\sum_{i=-m}^{-1}\left\|v_i\right\|^2+\cdots+\sum_{k=n_K}^{K}\sum_{i=n_K-m}^{n_K-1}\left\|v_i\right\|^2\right)+\frac{\epsilon}{\alpha}\\
&\overset{\text{(vii)}}{\leq} \frac{Q\eta^2 m^2}{B(K+1)}\sum_{i=0}^{K}\mathbb{E}\left\|v_i\right\|^2+\frac{\beta}{\alpha(K+1)}\sum_{i=-m}^{n_K-1}\left\|v_i\right\|^2+\frac{\epsilon}{\alpha}\\
&\overset{\text{(viii)}}{\leq} \frac{1}{K+1}\left(\frac{Q\eta^2 m^2}{B}+\frac{\beta}{\alpha}\right)\sum_{i=0}^{K}\mathbb{E}\left\|v_i\right\|^2+\frac{\epsilon}{\alpha}\\
&\overset{\text{(viiii)}}{\leq} \frac{1}{K+1}\left(\frac{Q\eta^2 m^2}{B}+\frac{\beta}{\alpha}\right)\left(\phi-\frac{\eta\beta}{2\alpha}\right)^{-1}\left(J(\theta^*)-J(\theta_0)+\sum_{i=0}^{K}\frac{\epsilon\eta}{2\alpha}\right)+\frac{\epsilon}{\alpha}\\
&= \left(\frac{Q\eta^2 m^2}{B}+\frac{\beta}{\alpha}\right)\left(\phi-\frac{\eta\beta}{2\alpha}\right)^{-1}\frac{(J(\theta^*)-J(\theta_0))}{K+1}+\frac{\epsilon}{\alpha}\left(1+\frac{\eta}{2}\left(\frac{Q\eta^2 m^2}{B}+\frac{\beta}{\alpha}\right)\left(\phi-\frac{\eta\beta}{2\alpha}\right)^{-1}\right) \qquad (29)
\end{aligned}
$$

where (i) follows from the fact that ξ is selected uniformly at random from $\{0,\ldots,K\}$, (ii) follows from (24), (iii) follows from the fact that $k-n_k\leq m$, (iv) follows from the fact that for $n_k\leqslant k\leqslant n_k+m-1$, $n_i=n_k$. (v) follows from $\sum_{k=n_k}^{n_k+m-1}\sum_{i=n_k}^{n_k+m-1}\mathbb{E}\left\|v_i\right\|^2=m\sum_{i=n_k}^{n_k+m-1}\mathbb{E}\left\|v_i\right\|^2$, (vi) follows from the same reasoning as in (iv), (vii) follows from $\sum_{k=n_k}^{n_k+m-1}\sum_{i=n_k-m}^{n_k-1}\mathbb{E}\left\|v_i\right\|^2=m\sum_{i=n_k-m}^{n_k-1}\mathbb{E}\left\|v_i\right\|^2$, (viii) follows from $\|v_{-1}\|=\cdots=\|v_{-m}\|=0$, and (viiii) follows from eq. (26).

Substituting (28), (29) into (27), we obtain

$$
\begin{aligned}
\mathbb{E}\|\nabla J(\theta_\xi)\|^2 \leq{}& \frac{2}{K+1}\left(1+\frac{Q\eta^2 m^2}{B}+\frac{\beta}{\alpha}\right)\left(\phi-\frac{\eta\beta}{2\alpha}\right)^{-1}(J(\theta^*)-J(\theta_0))\\
&+\frac{2\epsilon}{\alpha}\left(1+\frac{\eta}{2}\left(1+\frac{Q\eta^2 m^2}{B}+\frac{\beta}{\alpha}\right)\left(\phi-\frac{\eta\beta}{2\alpha}\right)^{-1}\right) \qquad (30)
\end{aligned}
$$

F.3 Proof of Corollary 3

Based on the parameter setting in Theorem 3 that

$$
\eta=\frac{1}{2L}, m=\left(\frac{L^2\sigma^2}{Q\epsilon}\right)^{\frac{1}{3}}, B=\left(\frac{Q\sigma^4}{L^2\epsilon^2}\right)^{\frac{1}{3}}, \alpha=48, \text{ and } \beta=6, \qquad (31)
$$

we obtain

$$\phi - \frac{\eta\beta}{2\alpha} = \left(\frac{1}{4L} - \frac{1}{8L} - \frac{1}{16L}\right) - \frac{1}{32L} = \frac{1}{32L} > 0. \tag{32}$$

Plugging (31) and (32) into (30), we obtain

$$\mathbb{E}\left\|\nabla J(\theta_\xi)\right\|^2 \leq \frac{88L}{K+1}\left(J(\theta^*) - J(\theta_0)\right) + \frac{\epsilon}{2}.$$

Hence, AbaSVRPG converges at a rate of $\mathcal{O}(1/K)$. Next, we bound the STO complexity. To acheive ϵ accuracy, we need

$$\frac{88L}{K+1}\left(J(\theta^*) - J(\theta_0)\right) \leq \frac{\epsilon}{2},$$

which gives

$$K = \frac{176L\left(J(\theta^*) - J(\theta_0)\right)}{\epsilon}.$$

We note that for $\mod(k,m) = 0$, the outer loop batch size $N = \frac{\alpha\sigma^2}{\frac{\beta}{m}\sum_{i=n_k-m}^{n_k-1}\|v_i\|^2+\epsilon} \leq \frac{\alpha\sigma^2}{\epsilon}$. Hence, the overall STO complexity is given by

$$\begin{aligned}
K \times 2B + \sum_{k=0}^{n_K} \frac{\alpha\sigma^2}{\frac{\beta}{m}\sum_{i=km-m}^{km-1}\|v_i\|^2 + \epsilon} &\leqslant K \times 2B + \sum_{k=0}^{n_K}\frac{\alpha\sigma^2}{\epsilon} \leqslant K \times 2B + \left\lceil\frac{K}{m}\right\rceil \times \frac{\alpha\sigma^2}{\epsilon} \\
&\overset{\text{(i)}}{\leq} 2KB + \frac{K}{m}\frac{\alpha\sigma^2}{\epsilon} + \frac{\alpha\sigma^2}{\epsilon} \\
&\overset{\text{(ii)}}{=} \mathcal{O}\left(\left(\frac{L\left(J(\theta^*) - J(\theta_0)\right)}{\epsilon}\right)\left(\left(\frac{Q\sigma^4}{L^2\epsilon^2}\right)^{\frac{1}{3}} + \frac{\sigma^2}{\epsilon}\left(\frac{Q\epsilon}{L^2\sigma^2}\right)^{\frac{1}{3}}\right) + \frac{\sigma^2}{\epsilon}\right) \\
&= \mathcal{O}\left(\left(\frac{L\left(J(\theta^*) - J(\theta_0)\right)}{\epsilon}\right)\left(\frac{Q\sigma^4}{L^2\epsilon^2}\right)^{\frac{1}{3}} + \frac{\sigma^2}{\epsilon}\right) \\
&= \mathcal{O}\left(\epsilon^{-5/3} + \epsilon^{-1}\right),
\end{aligned}$$

where (i) follows from the fact that $\left\lceil\frac{K}{m}\right\rceil \times N \leq \frac{KN}{m} + N$, and (ii) follows from the parameters setting of $K, B,$ and m in (31).

F.4 Proof of Theorem 4

In this section, we provide the proof of AbaSPIDER-PG. We first bound the variance of AbaSPIDER-PG given in the following proposition.

Proposition 2. *Let Assumptions 2, 3, and 4 hold. For $k = 0, \ldots K$, gradient estimator v_k of AbaSPIDER-PG satisfies*

$$\mathbb{E}\left\|v_k - \nabla J(\theta_k)\right\|^2 \leq \frac{Q\eta^2}{B}\sum_{i=n_k}^{k}\mathbb{E}\left\|v_i\right\|^2 + \mathbb{E}\left\|v_{n_k} - \nabla J(\theta_{n_k})\right\|^2.$$

Comparing Proposition 2 and Proposition 1, one can clearly see that AbaSPIDER-PG has a much smaller variance bound than AbaSVRPG, particularly as the inner loop iteration goes further (i.e., as k increases). This is because AbaSVRPG uses the initial outer loop batch gradient to construct the gradient estimator in all inner loop iterations, so that the variance in the inner loop accumulates up as the iteration goes further. In contrast, AbaSPIDER-PG avoids such a variance accumulation problem by continuously using the gradient information from the immediate preceding step, and hence has less variance during the inner loop iteration.

Proof of Proposition 2. To bound the variance $\mathbb{E}\left\|v_k - \nabla J(\theta_k)\right\|^2$, it is sufficient to bound $\mathbb{E}_k\left\|v_k - \nabla J(\theta_k)\right\|^2$, and then the tower property of expectation yields the desired result. Thus, we first bound $\mathbb{E}_k\left\|v_k - \nabla J(\theta_k)\right\|^2$ for $\mod(k,m) \neq 0$, and then generalize it to $\mod(k,m) = 0$.

$$
\begin{aligned}
&\mathbb{E}_k\left\|v_k - \nabla J(\theta_k)\right\|^2 \\
&\overset{\text{(i)}}{=} \mathbb{E}_k\left\|\frac{1}{B}\sum_{i=1}^{B} g(\tau_i|\theta_k) - \frac{1}{B}\sum_{i=1}^{B}\omega(\tau_i|\theta_k,\theta_{k-1})g(\tau_i|\theta_{k-1}) + v_{k-1} - \nabla J(\theta_k)\right\|^2 \\
&= \mathbb{E}_k\left\|\frac{1}{B}\sum_{i=1}^{B} g(\tau_i|\theta_k) - \frac{1}{B}\sum_{i=1}^{B}\omega(\tau_i|\theta_k,\theta_{k-1})g(\tau_i|\theta_{k-1}) + \nabla J(\theta_{k-1}) - \nabla J(\theta_k) + v_{k-1} - \nabla J(\theta_{k-1})\right\|^2 \\
&= \mathbb{E}_k\left\|\frac{1}{B}\sum_{i=1}^{B} g(\tau_i|\theta_k) - \frac{1}{B}\sum_{i=1}^{B}\omega(\tau_i|\theta_k,\theta_{k-1})g(\tau_i|\theta_{k-1}) + \nabla J(\theta_{k-1}) - \nabla J(\theta_k)\right\|^2 + \mathbb{E}_k\left\|v_{k-1} - \nabla J(\theta_{k-1})\right\|^2 \\
&\quad + 2\mathbb{E}_k\left\langle \frac{1}{B}\sum_{i=1}^{B} g(\tau_i|\theta_k) - \frac{1}{B}\sum_{i=1}^{B}\omega(\tau_i|\theta_k,\theta_{k-1})g(\tau_i|\theta_{k-1}) + \nabla J(\theta_{k-1}) - \nabla J(\theta_k), v_{k-1} - \nabla J(\theta_{k-1})\right\rangle \\
&\overset{\text{(ii)}}{=} \mathbb{E}_k\left\|\frac{1}{B}\sum_{i=1}^{B} g(\tau_i|\theta_k) - \frac{1}{B}\sum_{i=1}^{B}\omega(\tau_i|\theta_k,\theta_{k-1})g(\tau_i|\theta_{k-1}) + \nabla J(\theta_{k-1}) - \nabla J(\theta_k)\right\|^2 + \mathbb{E}_k\left\|v_{k-1} - \nabla J(\theta_{k-1})\right\|^2 \\
&\overset{\text{(iii)}}{\leq} \mathop{\mathbb{E}_k}_{\tau\sim p(\cdot|\theta_k)} \frac{1}{B}\left\|g(\tau|\theta_k) - \omega(\tau|\theta_k,\theta_{k-1})g(\tau|\theta_{k-1}) + \nabla J(\theta_{k-1}) - \nabla J(\theta_k)\right\|^2 + \mathbb{E}_k\left\|v_{k-1} - \nabla J(\theta_{k-1})\right\|^2 \\
&\overset{\text{(iv)}}{\leq} \mathop{\mathbb{E}_k}_{\tau\sim p(\cdot|\theta_k)} \frac{1}{B}\left\|g(\tau|\theta_k) - \omega(\tau|\theta_k,\theta_{k-1})g(\tau|\theta_{k-1})\right\|^2 + \mathbb{E}_k\left\|v_{k-1} - \nabla J(\theta_{k-1})\right\|^2 \\
&\overset{\text{(v)}}{\leq} \frac{Q}{B}\left\|\theta_k - \theta_{k-1}\right\|^2 + \mathbb{E}_k\left\|v_{k-1} - \nabla J(\theta_{k-1})\right\|^2 \\
&\overset{\text{(vi)}}{\leq} \frac{Q\eta^2}{B}\left\|v_{k-1}\right\|^2 + \mathbb{E}_k\left\|v_{k-1} - \nabla J(\theta_{k-1})\right\|^2
\end{aligned} \tag{33}
$$

where (i) follows from the definition of v_k in Algorithm 3, (ii) follows from the fact that

$$
\mathop{\mathbb{E}}_{k}\left[\frac{1}{B}\sum_{i=1}^{B} g(\tau_i|\theta_k) - \frac{1}{B}\sum_{i=1}^{B}\omega(\tau_i|\theta_k,\theta_{k-1})g(\tau_i|\theta_{k-1}) + \nabla J(\theta_{k-1}) - \nabla J(\theta_k)\right] = 0,
$$

thus given $\theta_k, \cdots, \theta_0$, the expectation of the inner product equals 0, (iii) follows from Lemma 7, (iv) follows from the fact that $\mathbb{V}\mathrm{ar}(X) \leq \mathbb{E}\left\|X\right\|^2$, (v) follows from Lemma 5, and (vi) follows because $\theta_k = \theta_{k-1} + \eta v_{k-1}$.

Therefore, we have

$$
\begin{aligned}
\mathbb{E}\left\|v_k - \nabla J(\theta_k)\right\|^2 &\overset{\text{(i)}}{=} \mathbb{E}\,\mathbb{E}_k\left\|v_k - \nabla J(\theta_k)\right\|^2 \\
&\overset{\text{(ii)}}{\leq} \frac{Q\eta^2}{B}\,\mathbb{E}\left\|v_{k-1}\right\|^2 + \mathbb{E}\left\|v_{k-1} - \nabla J(\theta_{k-1})\right\|^2,
\end{aligned} \tag{34}
$$

where (i) follows from the tower property of expectation, and (ii) follows from (33).

Telescoping (34) over k from $n_k + 1$ to k, we obtain

$$
\begin{aligned}
\mathbb{E}\left\|v_k - \nabla J(\theta_k)\right\|^2 &= \sum_{i=n_k+1}^{k} \frac{Q\eta^2}{B}\,\mathbb{E}\left\|v_{i-1}\right\|^2 + \mathbb{E}\left\|v_{n_k} - \nabla J(\theta_{n_k})\right\|^2 \\
&\leq \sum_{i=n_k}^{k} \frac{Q\eta^2}{B}\,\mathbb{E}\left\|v_i\right\|^2 + \mathbb{E}\left\|v_{n_k} - \nabla J(\theta_{n_k})\right\|^2.
\end{aligned}
$$

It is straightforward to check that the above inequality also holds for any k with $\mod(k,m) = 0$. □

Proof of Theorem 4

Since in Algorithm 4, ∇J is L-Lipschitz, and $\theta_{k+1} = \theta_{k+1} + \eta v_k$, we obtain the following inequality directly from Lemma 6:

$$\mathbb{E}\, J(\theta_{k+1}) - \mathbb{E}\, J(\theta_k) \geq \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right) \mathbb{E}\, \|v_k\|^2 - \frac{\eta}{2}\, \mathbb{E}\, \|v_k - \nabla J(\theta_k)\|^2. \tag{35}$$

By Proposition 2, we obtain

$$\mathbb{E}\, \|v_k - \nabla J(\theta_k)\|^2 \leq \sum_{i=n_k}^{k} \frac{Q\eta^2}{B}\, \mathbb{E}\, \|v_i\|^2 + \mathbb{E}\, \|v_{n_k} - \nabla J(\theta_{n_k})\|^2 \tag{36}$$

Moreover, for $\mod(k, m) = 0$, we obtain

$$\begin{aligned}
\mathbb{E}\|v_k - \nabla f(x_k)\|^2 &= \mathbb{E}\left\| \frac{1}{N} \sum_{i=1}^{N} \nabla g(\tau_i|\theta_k) - \nabla J(\theta_k) \right\|^2 \\
&\overset{\text{(i)}}{=} \frac{1}{N} \mathop{\mathbb{E}}_{\tau \sim p(\cdot|\theta_k)} \|\nabla g(\tau|\theta_k) - \nabla J(\theta_k)\|^2 \overset{(ii)}{\leqslant} \frac{\sigma^2}{N} \\
&\overset{(iii)}{\leqslant} \frac{\beta}{\alpha m} \sum_{i=n_k-m}^{n_k-1} \|v_i\|^2 + \frac{\epsilon}{\alpha},
\end{aligned} \tag{37}$$

where (i) follows from Lemma 7, (ii) follows from Assumption 4, and (iii) follows from the fact that

$$N = \frac{\alpha\sigma^2}{\frac{\beta}{m}\sum_{i=n_k-m}^{n_k-1} \|v_i\|^2 + \epsilon},$$

where $\alpha > 0$ and $\beta \geqslant 0$.

Plugging (37) into (36), we obtain

$$\mathbb{E}\, \|v_k - \nabla J(\theta_k)\|^2 \leq \frac{Q\eta^2}{B} \sum_{i=n_k}^{k} \mathbb{E}\, \|v_i\|^2 + \frac{\beta}{\alpha m} \sum_{i=n_k-m}^{n_k-1} \|v_i\|^2 + \frac{\epsilon}{\alpha}. \tag{38}$$

Plugging (38) into (35), we obtain

$$\mathbb{E}\, J(\theta_{k+1}) - \mathbb{E}\, J(\theta_k) \geq \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right) \mathbb{E}\, \|v_k\|^2 - \frac{Q\eta^3}{2B} \sum_{i=n_k}^{k} \mathbb{E}\, \|v_i\|^2 - \frac{\eta\beta}{2\alpha m} \sum_{i=n_k-m}^{n_k-1} \|v_i\|^2 - \frac{\epsilon\eta}{2\alpha}.$$

We note that for a given k, any iteration $n_k \leq i \leq k$, all their corresponding n_i satisfies $n_i = n_k$. Thus, telescoping the above inequality over k from n_k to k, we obtain

$$\begin{aligned}
&\mathbb{E} J(\theta_{k+1}) - \mathbb{E}\, J(\theta_{n_k}) \\
&\geq \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right) \sum_{i=n_k}^{k} \mathbb{E}\, \|v_i\|^2 - \frac{Q\eta^3}{2B} \sum_{i=n_k}^{k} \sum_{j=n_k}^{i} \mathbb{E}\, \|v_j\|^2 - \sum_{i=n_k}^{k} \frac{\eta\beta}{2\alpha m} \sum_{i=n_k-m}^{n_k-1} \|v_i\|^2 - \sum_{i=n_k}^{k} \frac{\epsilon\eta}{2\alpha} \\
&\geq \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right) \sum_{i=n_k}^{k} \mathbb{E}\, \|v_i\|^2 - \frac{Q\eta^3}{2B} \sum_{i=n_k}^{k} \sum_{j=n_k}^{k} \mathbb{E}\, \|v_j\|^2 - \sum_{i=n_k}^{k} \frac{\eta\beta}{2\alpha m} \sum_{i=n_k-m}^{n_k-1} \|v_i\|^2 - \sum_{i=n_k}^{k} \frac{\epsilon\eta}{2\alpha} \\
&= \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right) \sum_{i=n_k}^{k} \mathbb{E}\, \|v_i\|^2 - \frac{Q\eta^3(k - n_k + 1)}{2B} \sum_{j=n_k}^{k} \mathbb{E}\, \|v_j\|^2 - \sum_{i=n_k}^{k} \frac{\eta\beta}{2\alpha m} \sum_{i=n_k-m}^{n_k-1} \|v_i\|^2 - \sum_{i=n_k}^{k} \frac{\epsilon\eta}{2\alpha}
\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{Q\eta^3(k-n_k+1)}{2B}\right) \sum_{i=n_k}^{k} \mathbb{E}\left\|v_i\right\|^2 - \sum_{i=n_k}^{k} \frac{\eta\beta}{2\alpha m} \sum_{j=n_k-m}^{n_k-1} \left\|v_j\right\|^2 - \sum_{i=n_k}^{k} \frac{\epsilon\eta}{2\alpha} \\
&\overset{\text{(i)}}{\geq} \left(\frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{Q\eta^3 m}{2B}\right) \sum_{i=n_k}^{k} \mathbb{E}\left\|v_i\right\|^2 - \sum_{i=n_k}^{k} \frac{\eta\beta}{2\alpha m} \sum_{j=n_k-m}^{n_k-1} \left\|v_j\right\|^2 - \sum_{i=n_k}^{k} \frac{\epsilon\eta}{2\alpha} \\
&\overset{\text{(ii)}}{=} \phi \sum_{i=n_k}^{k} \mathbb{E}\left\|v_i\right\|^2 - \frac{\eta\beta(k-n_k+1)}{2\alpha m} \sum_{i=n_k-m}^{n_k-1} \left\|v_i\right\|^2 - \sum_{i=n_k}^{k} \frac{\epsilon\eta}{2\alpha} \\
&\overset{\text{(iii)}}{\geq} \phi \sum_{i=n_k}^{k} \mathbb{E}\left\|v_i\right\|^2 - \frac{\eta\beta}{2\alpha} \sum_{i=n_k-m}^{n_k-1} \left\|v_i\right\|^2 - \sum_{i=n_k}^{k} \frac{\epsilon\eta}{2\alpha},
\end{aligned} \tag{39}
$$

where (i) follows from the fact that $k - n_k + 1 \leq m$, (ii) follows from the fact that $\phi := \left(\frac{\eta}{2} - \frac{L\eta^2}{2} - \frac{Q\eta^3 m}{2B}\right)$, and (iii) follows because $(k - n_k + 1)/m \leqslant 1$.

Now, we are ready to bound $J(\theta_{K+1}) - J(\theta_0)$.

$$
\begin{aligned}
&\mathbb{E}J(\theta_{K+1}) - \mathbb{E}\, J(\theta_0) \\
&\quad = \mathbb{E}\, J(\theta_{K+1}) - \mathbb{E}\, J(\theta_{n_K}) + \mathbb{E}\, J(\theta_{n_K}) \cdots + \mathbb{E}\, J(\theta_m) - \mathbb{E}\, J(\theta_0) \\
&\quad \overset{\text{(i)}}{\geq} \phi \sum_{i=n_K}^{K} \mathbb{E}\left\|v_i\right\|^2 - \frac{\eta\beta}{2\alpha} \sum_{i=n_K-m}^{n_K-1} \left\|v_i\right\|^2 - \sum_{i=n_K}^{K} \frac{\epsilon\eta}{2\alpha} + \cdots + \phi \sum_{i=0}^{m-1} \mathbb{E}\left\|v_i\right\|^2 - \frac{\eta\beta}{2\alpha} \sum_{i=-m}^{-1} \left\|v_i\right\|^2 - \sum_{i=0}^{m} \frac{\epsilon\eta}{2\alpha} \\
&\quad \overset{\text{(ii)}}{\geq} \phi \sum_{i=0}^{K} \mathbb{E}\left\|v_i\right\|^2 - \frac{\eta\beta}{2\alpha} \sum_{i=0}^{n_K-1} \left\|v_i\right\|^2 - \sum_{i=0}^{K} \frac{\epsilon\eta}{2\alpha} \\
&\quad \geqslant \left(\phi - \frac{\eta\beta}{2\alpha}\right) \sum_{i=0}^{K} \mathbb{E}\left\|v_i\right\|^2 - \sum_{i=0}^{K} \frac{\epsilon\eta}{2\alpha},
\end{aligned}
$$

where (i) follows from (39), and (ii) follows from the fact that we define $\|v_{-1}\| = \cdots = \|v_{-m}\| = 0$. Thus, we obtain

$$
\begin{aligned}
\left(\phi - \frac{\eta\beta}{2\alpha}\right) \sum_{i=0}^{K} \mathbb{E}\left\|v_i\right\|^2 &\leq \mathbb{E}\, J(\theta_{K+1}) - \mathbb{E}\, J(\theta_0) + \sum_{i=0}^{K} \frac{\epsilon\eta}{2\alpha} \\
&\overset{\text{(i)}}{\leq} J(\theta^*) - J(\theta_0) + \sum_{i=0}^{K} \frac{\epsilon\eta}{2\alpha},
\end{aligned}
$$

where (i) follows because $\theta^* := \arg\max_{\theta \in \mathbb{R}^d} J(\theta)$. Here, we assume $\left(\phi - \frac{\eta\beta}{2\alpha}\right) > 0$ to continue our proof. Such an assumption will be satisfied by parameter tuning as shown in (32). Therefore, we obtain

$$
\sum_{i=0}^{K} \mathbb{E}\left\|v_i\right\|^2 \leq \left(\phi - \frac{\eta\beta}{2\alpha}\right)^{-1} \left(J(\theta^*) - J(\theta_0) + \sum_{i=0}^{K} \frac{\epsilon\eta}{2\alpha}\right). \tag{40}
$$

With (40), we are now able to bound the gradient norm, i.e., $\|\nabla J(\theta_\xi)\|$, of the output of AbaSPIDER-PG. Observe that

$$
\mathbb{E}\|\nabla J(\theta_\xi)\|^2 = \mathbb{E}\|\nabla J(\theta_\xi) - v_\xi + v_\xi\|^2 \leq 2\mathbb{E}\|\nabla J(\theta_\xi) - v_\xi\|^2 + 2\mathbb{E}\|v_\xi\|^2. \tag{41}
$$

Therefore, it is sufficient to bound the two terms on the right hand side of the above inequality. First, note that

$$
\mathbb{E}\|v_\xi\|^2 \overset{\text{(i)}}{=} \frac{1}{K+1} \sum_{i=0}^{K} \mathbb{E}\|v_i\|^2 \overset{\text{(ii)}}{\leq} \left(\phi - \frac{\eta\beta}{2\alpha}\right)^{-1} \left(\frac{J(\theta^*) - J(\theta_0)}{K+1} + \frac{\epsilon\eta}{2\alpha}\right), \tag{42}
$$

where (i) follows from the fact that ξ is selected uniformly at random from $\{0,\dots,K\}$, and (ii) follows from (40). On the other hand, we observe that

$$
\begin{aligned}
&\mathbb{E}\|\nabla J(\theta_\xi)-v_\xi\|^2\\
&\overset{\text{(i)}}{=}\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\nabla J(\theta_k)-v_k\|^2\\
&\overset{\text{(ii)}}{\leq}\frac{1}{K+1}\sum_{k=0}^{K}\left(\frac{Q\eta^2}{B}\sum_{i=n_k}^{k}\mathbb{E}\left\|v_i\right\|^2+\frac{\beta}{\alpha m}\sum_{i=n_k-m}^{n_k-1}\left\|v_i\right\|^2+\frac{\epsilon}{\alpha}\right)\\
&\overset{\text{(iii)}}{=}\frac{Q\eta^2}{B(K+1)}\sum_{k=0}^{K}\sum_{i=n_k}^{k}\mathbb{E}\left\|v_i\right\|^2+\frac{\beta}{\alpha m(K+1)}\sum_{k=0}^{K}\sum_{i=n_k-m}^{n_k-1}\left\|v_i\right\|^2+\frac{\epsilon}{\alpha}\\
&\overset{\text{(iv)}}{=}\frac{Q\eta^2}{B(K+1)}\left(\sum_{k=0}^{m-1}\sum_{i=0}^{k}\mathbb{E}\left\|v_i\right\|^2+\cdots+\sum_{k=n_K}^{K}\sum_{i=n_K}^{k}\mathbb{E}\left\|v_i\right\|^2\right)+\frac{\beta}{\alpha m(K+1)}\sum_{k=0}^{K}\sum_{i=n_k-m}^{n_k-1}\left\|v_i\right\|^2+\frac{\epsilon}{\alpha}\\
&\leqslant\frac{Q\eta^2}{B(K+1)}\left(\sum_{k=0}^{m-1}\sum_{i=0}^{m-1}\mathbb{E}\left\|v_i\right\|^2+\cdots+\sum_{k=n_K}^{K}\sum_{i=n_K}^{K}\mathbb{E}\left\|v_i\right\|^2\right)+\frac{\beta}{\alpha m(K+1)}\sum_{k=0}^{K}\sum_{i=n_k-m}^{n_k-1}\left\|v_i\right\|^2+\frac{\epsilon}{\alpha}\\
&\overset{\text{(v)}}{\leq}\frac{Q\eta^2 m}{B(K+1)}\sum_{i=0}^{K}\mathbb{E}\left\|v_i\right\|^2+\frac{\beta}{\alpha m(K+1)}\sum_{k=0}^{K}\sum_{i=n_k-m}^{n_k-1}\left\|v_i\right\|^2+\frac{\epsilon}{\alpha}\\
&\overset{\text{(vi)}}{\leq}\frac{Q\eta^2 m}{B(K+1)}\sum_{i=0}^{K}\mathbb{E}\left\|v_i\right\|^2+\frac{\beta}{\alpha m(K+1)}\left(\sum_{k=0}^{m-1}\sum_{i=-m}^{-1}\left\|v_i\right\|^2+\cdots+\sum_{k=n_K}^{K}\sum_{i=n_K-m}^{n_K-1}\left\|v_i\right\|^2\right)+\frac{\epsilon}{\alpha}\\
&\overset{\text{(vii)}}{\leq}\frac{Q\eta^2 m}{B(K+1)}\sum_{i=0}^{K}\mathbb{E}\left\|v_i\right\|^2+\frac{\beta}{\alpha(K+1)}\sum_{i=-m}^{n_K-1}\left\|v_i\right\|^2+\frac{\epsilon}{\alpha}\\
&\overset{\text{(viii)}}{\leq}\frac{1}{K+1}\left(\frac{Q\eta^2 m}{B}+\frac{\beta}{\alpha}\right)\sum_{i=0}^{K}\mathbb{E}\left\|v_i\right\|^2+\frac{\epsilon}{\alpha}\\
&\overset{\text{(viiii)}}{\leq}\frac{1}{K+1}\left(\frac{Q\eta^2 m}{B}+\frac{\beta}{\alpha}\right)\left(\phi-\frac{\eta\beta}{2\alpha}\right)^{-1}\left(J(\theta^*)-J(\theta_0)+\sum_{i=0}^{K}\frac{\epsilon\eta}{2\alpha}\right)+\frac{\epsilon}{\alpha}\\
&=\left(\frac{Q\eta^2 m}{B}+\frac{\beta}{\alpha}\right)\left(\phi-\frac{\eta\beta}{2\alpha}\right)^{-1}\frac{(J(\theta^*)-J(\theta_0))}{K+1}+\frac{\epsilon}{\alpha}\left(1+\frac{\eta}{2}\left(\frac{Q\eta^2 m^2}{B}+\frac{\beta}{\alpha}\right)\left(\phi-\frac{\eta\beta}{2\alpha}\right)^{-1}\right)
\end{aligned}
\tag{43}
$$

where (i) follows from the fact that ξ is selected uniformly at random from $\{0,\dots,K\}$, (ii) follows from (38), (iii) follows from the fact that $k-n_k\leq m$, (iv) follows from the fact that for $n_k\leqslant k\leqslant n_k+m-1$, $n_i=n_k$. (v) follows from $\sum_{k=n_k}^{n_k+m-1}\sum_{i=n_k}^{n_k+m-1}\mathbb{E}\left\|v_i\right\|^2=m\sum_{i=n_k}^{n_k+m-1}\mathbb{E}\left\|v_i\right\|^2$, (vi) follows from the same reasoning as in (iv), (vii) follows from $\sum_{k=n_k}^{n_k+m-1}\sum_{i=n_k-m}^{n_k-1}\mathbb{E}\left\|v_i\right\|^2=m\sum_{i=n_k-m}^{n_k-1}\mathbb{E}\left\|v_i\right\|^2$, (viii) follows from $\|v_{-1}\|=\cdots=\|v_{-m}\|=0$, and (viiii) follows from eq. (40).

Substituting (42), (43) into (41), we obtain

$$
\begin{aligned}
\mathbb{E}\|\nabla J(\theta_\xi)\|^2\leq\frac{2}{K+1}&\left(1+\frac{Q\eta^2 m}{B}+\frac{\beta}{\alpha}\right)\left(\phi-\frac{\eta\beta}{2\alpha}\right)^{-1}(J(\theta^*)-J(\theta_0))\\
&+\frac{2\epsilon}{\alpha}\left(1+\frac{\eta}{2}\left(1+\frac{Q\eta^2 m}{B}+\frac{\beta}{\alpha}\right)\left(\phi-\frac{\eta\beta}{2\alpha}\right)^{-1}\right)
\end{aligned}
\tag{44}
$$

F.5 Proof of Corollary 4

Based on the parameter setting in Theorem 4 that

$$\eta = \frac{1}{2L}, m = \frac{L\sigma}{\sqrt{Q\epsilon}}, B = \frac{\sigma\sqrt{Q}}{L\sqrt{\epsilon}}, \alpha = 48 \text{ and } \beta = 16, \tag{45}$$

we obtain

$$\phi - \frac{\eta\beta}{2\alpha} = \left(\frac{1}{4L} - \frac{1}{8L} - \frac{1}{16L}\right) - \frac{1}{32L} = \frac{1}{32L} > 0. \tag{46}$$

Plugging (45) and (46) into (44), we obtain

$$\mathbb{E}\left\|\nabla J(\theta_\xi)\right\|^2 \leq \frac{88L}{K+1}\left(J(\theta^*) - J(\theta_0)\right) + \frac{\epsilon}{2}$$

To obtain ϵ accuracy, we need

$$\frac{88L}{K+1}\left(J(\theta^*) - J(\theta_0)\right) \leq \frac{\epsilon}{2},$$

which gives

$$K = \frac{176L\left(J(\theta^*) - J(\theta_0)\right)}{\epsilon}.$$

We note that for $\mod(k, m) = 0$, the outer loop batch size $N = \frac{\alpha\sigma^2}{\frac{\beta}{m}\sum_{i=n_k-m}^{n_k-1}\|v_i\|^2+\epsilon} \leq \frac{\alpha\sigma^2}{\epsilon}$. Hence, the overall STO complexity is given by

$$\begin{aligned}
K \times 2B + \sum_{k=0}^{n_K} \frac{\alpha\sigma^2}{\frac{\beta}{m}\sum_{i=km-m}^{km-1}\left\|v_i\right\|^2 + \epsilon} &\leqslant K \times 2B + \sum_{k=0}^{n_K} \frac{\alpha\sigma^2}{\epsilon} \leqslant K \times 2B + \left\lceil \frac{K}{m} \right\rceil \times \frac{\alpha\sigma^2}{\epsilon} \\
&\overset{\text{(i)}}{\leq} 2KB + \frac{K}{m}\frac{\alpha\sigma^2}{\epsilon} + \frac{\alpha\sigma^2}{\epsilon} \\
&\overset{\text{(ii)}}{=} \mathcal{O}\left(\left(\frac{L\left(J(\theta^*) - J(\theta_0)\right)}{\epsilon}\right)\left(\frac{\sigma\sqrt{Q}}{L\sqrt{\epsilon}} + \frac{\sigma^2}{\epsilon}\frac{\sqrt{Q\epsilon}}{L\sigma}\right) + \frac{\sigma^2}{\epsilon}\right) \\
&= \mathcal{O}\left(\left(\frac{L\left(J(\theta^*) - J(\theta_0)\right)}{\epsilon}\right)\left(\frac{\sigma\sqrt{Q}}{L\sqrt{\epsilon}}\right) + \frac{\sigma^2}{\epsilon}\right) \\
&= \mathcal{O}\left(\epsilon^{-3/2} + \epsilon^{-1}\right).
\end{aligned}$$

where (i) follows from the fact that $\left\lceil \frac{K}{m} \right\rceil \times N \leq \frac{KN}{m} + N$, and (ii) follows from the parameters setting of $K, B,$ and m in (45).

G Proof of Technical Lemmas

G.1 Proof of Lemma 3

(i), (ii), (iii) follow from Lemma B.2, Lemma B.3 and Lemma B.4 in Papini et al. 2018, respectively.

G.2 Proof of Lemma 5

Note that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\tau\sim p(\cdot|\theta_1)} \left\|g(\tau|\theta_1) - \omega(\tau|\theta_1,\theta_2)g(\tau|\theta_2)\right\|^2 \\
&\quad= \mathop{\mathbb{E}}_{\tau\sim p(\cdot|\theta_1)} \left\|g(\tau|\theta_1) - g(\tau|\theta_2) + g(\tau|\theta_2) - \omega(\tau|\theta_1,\theta_2)g(\tau|\theta_2)\right\|^2 \\
&\quad\overset{\text{(i)}}{=} \mathop{\mathbb{E}}_{\tau\sim p(\cdot|\theta_1)} 2\left\|g(\tau|\theta_1) - g(\tau|\theta_2)\right\|^2 + \mathop{\mathbb{E}}_{\tau\sim p(\cdot|\theta_1)} 2\left\|g(\tau|\theta_2) - \omega(\tau|\theta_1,\theta_2)g(\tau|\theta_2)\right\|^2 \\
&\quad\overset{\text{(ii)}}{\leq} 2L_g^2\left\|\theta_1-\theta_2\right\|^2 + \mathop{\mathbb{E}}_{\tau\sim p(\cdot|\theta_1)} 2\left\|g(\tau|\theta_2)\right\|^2\left\|1-\omega(\tau|\theta_1,\theta_2)\right\|^2 \\
&\quad\overset{\text{(iii)}}{\leq} 2L_g^2\left\|\theta_1-\theta_2\right\|^2 + 2\Gamma\alpha\left\|\theta_1-\theta_2\right\|^2 = 2(L_g^2+\Gamma\alpha)\left\|\theta_1-\theta_2\right\|^2 \overset{\text{(iv)}}{=} Q\left\|\theta_1-\theta_2\right\|^2,
\end{aligned}$$

where (i) follows from the fact that $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$, (ii) follows from item (ii) in Lemma 3, and (iii) follows from item (iii) in Lemma 3 and Lemma 4. Then, the proof is complete.

G.3 Proof of Lemma 6

We derive the following lower bound

$$\begin{aligned}
J(\theta_{k+1}) - J(\theta_k) &\overset{\text{(i)}}{\geq} \langle \nabla J(\theta_k), \theta_{k+1}-\theta_k\rangle - \frac{L}{2}\left\|\theta_{k+1}-\theta_k\right\|^2 \\
&\overset{\text{(ii)}}{=} \eta\left\langle \nabla J(\theta_k), v_k\right\rangle - \frac{L\eta^2}{2}\left\|v_k\right\|^2 \\
&= \eta\left\langle \nabla J(\theta_k) - v_k + v_k, v_k\right\rangle - \frac{L\eta^2}{2}\left\|v_k\right\|^2 \\
&= \eta\left\|v_k\right\|^2 + \eta\left\langle \nabla J(\theta_k) - v_k, v_k\right\rangle - \frac{L\eta^2}{2}\left\|v_k\right\|^2 \\
&\overset{\text{(iii)}}{\geq} \eta\left\|v_k\right\|^2 - \eta\frac{\left\|v_k-\nabla J(\theta_k)\right\|^2 + \left\|v_k\right\|^2}{2} - \frac{L\eta^2}{2}\left\|v_k\right\|^2 \\
&= \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\left\|v_k\right\|^2 - \frac{\eta}{2}\left\|\nabla J(v_k) - \theta_k\right\|^2,
\end{aligned}$$

where (i) follows from the fact that ∇J is L-Lipschitz, (ii) follows from the update rule $x_{k+1} = x_k + \eta v_k$, and (iii) follows from Young's inequality. Taking the expectation over the entire random process on both sides, we obtain the desired result.

G.4 Proof of Lemma 7

Lemma 7. *Let $X, X_1, \cdots, X_n$ be independent and identically distributed (i.i.d.) random variables with mean $\mathbb{E}[X]$, then, the following equation holds:*

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}\,X\right\|^2 = \frac{\mathbb{E}\left\|X - \mathbb{E}\,X\right\|^2}{n}$$

Proof. Standard calculation yields

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}\,X\right\|^2 = \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}\,X)\right\|^2 = \frac{1}{n^2}\mathbb{E}\left\|\sum_{i=1}^n (X_i - \mathbb{E}\,X)\right\|^2$$

$$
\begin{aligned}
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left\langle X_i - \mathbb{E}\,X, X_j - \mathbb{E}\,X\right\rangle \\
&\overset{\text{(i)}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\langle X_i - \mathbb{E}\,X, X_i - \mathbb{E}\,X\right\rangle \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\|X_i - \mathbb{E}\,X\right\|^2 \\
&\overset{\text{(ii)}}{=} \frac{\mathbb{E}\left\|X - \mathbb{E}\,X\right\|^2}{n},
\end{aligned}
$$

where (i) follows from the fact that $X_1, \cdots, X_n$ are i.i.d. random variables such that if $i \neq j$, $\mathbb{E}\left\langle X_i - \mathbb{E}\,X, X_j - \mathbb{E}\,X\right\rangle = 0$, and (ii) follows from the fact that for i.i.d. random variables $\mathbb{E}\left\|X - \mathbb{E}\,X\right\|^2 = \mathbb{E}\left\|X_1 - \mathbb{E}\,X\right\|^2 \cdots = \mathbb{E}\left\|X_n - \mathbb{E}\,X\right\|^2$. □

H Proofs for Results in Appendix C

H.1 Proof for Theorem 5

To simplify notations, we let $c_\beta = c_\epsilon = \alpha = \left(2\tau + \frac{2\tau}{1-\exp(\frac{-4}{c_\eta(c_\eta-2)})}\right) \vee \frac{16 c_\eta L\tau}{m}$.

Since the objective function $f(\cdot)$ has a L-Lipschitz continuous gradient, we obtain that for $1 \leq t \leq m$,

$$
\begin{aligned}
f(x_t^s) \leq& f(x_{t-1}^s) + \langle \nabla f(x_{t-1}^s), x_t^s - x_{t-1}^s\rangle + \frac{L\eta^2}{2}\|v_{t-1}^s\|^2 \\
=& f(x_{t-1}^s) + \frac{\eta}{2}\|\nabla f(x_{t-1}^s) - v_{t-1}^s\|^2 - \frac{\eta}{2}\|\nabla f(x_{t-1}^s)\|^2 - \frac{\eta}{2}\|v_{t-1}^s\|^2 + \frac{L\eta^2}{2}\|v_{t-1}^s\|^2,
\end{aligned}
$$

which, in conjunction with the PL condition that $\|\nabla f(x_{t-1}^s)\|^2 \geq \frac{1}{\tau}(f(x_{t-1}^s) - f(x^*))$, implies that

$$
f(x_t^s) - f(x^*) \leq \left(1 - \frac{\eta}{2\tau}\right)(f(x_{t-1}^s) - f(x^*)) - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\|v_{t-1}^s\|^2 + \frac{\eta}{2}\|\nabla f(x_{t-1}^s) - v_{t-1}^s\|^2.
$$

Recall that $\mathbb{E}_{t,s}(\cdot)$ denotes $\mathbb{E}(\cdot|x_0^1, x_0^2, ..., x_2^1, ..., x_t^s)$. Then, taking expectation $\mathbb{E}_{0,s}(\cdot)$ over the above inequality yields, for $1 \leq t \leq m$,

$$
\begin{aligned}
\mathbb{E}_{0,s}(f(x_t^s) - f(x^*)) \leq& \left(1 - \frac{\eta}{2\tau}\right)\mathbb{E}_{0,s}(f(x_{t-1}^s) - f(x^*)) - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\mathbb{E}_{0,s}\|v_{t-1}^s\|^2 \\
&+ \frac{\eta}{2}\mathbb{E}_{0,s}\|\nabla f(x_{t-1}^s) - v_{t-1}^s\|^2,
\end{aligned} \tag{47}
$$

which, in conjunction with Lemma 1, implies that

$$
\begin{aligned}
\mathbb{E}_{0,s}(f(x_t^s) - f(x^*)) \leq& \left(1 - \frac{\eta}{2\tau}\right)\mathbb{E}_{0,s}(f(x_{t-1}^s) - f(x^*)) - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\mathbb{E}_{0,s}\|v_{t-1}^s\|^2 \\
&+ \frac{\eta^3 L^2(t-1)}{2B}\mathbb{E}_{0,s}\sum_{i=0}^{t-2}\|v_i^s\|^2 + \frac{\eta I_{(N_s<n)}}{2N_s}\sigma^2.
\end{aligned}
$$

Let $\gamma := 1 - \frac{\eta}{2\tau}$. Then, telescoping the above inequality over t from 1 to m and using the fact that $t - 1 < m$, we have

$$
\mathbb{E}_{0,s}(f(x_m^s) - f(x^*)) \leq \gamma^m \mathbb{E}_{0,s}(f(x_0^s) - f(x^*)) - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}_{0,s}\|v_t^s\|^2
$$

$$+\frac{\eta^3L^2m}{2B}\sum_{t=0}^{m-2}\gamma^{m-2-t}\mathbb{E}_{0,s}\sum_{i=0}^{t}\|v_i^s\|^2+\left(\sum_{t=0}^{m-1}\gamma^t\right)\frac{\eta I_{(N_s<n)}}{2N_s}\sigma^2. \tag{48}$$

Note that $\gamma^{m-1-t}\geq\gamma^m$ for $0\leq t\leq m-1$ and $\sum_{t=0}^{m-1}\gamma^t=\frac{1-\gamma^m}{1-\gamma}\leq\frac{1}{1-\gamma}=\frac{2\tau}{\eta}$. Then, we obtain from (48) that

$$\begin{aligned}
\mathbb{E}_{0,s}(f(x_m^s)-f(x^*))\leq&\gamma^m\mathbb{E}_{0,s}(f(x_0^s)-f(x^*))-\left(\frac{\eta}{4}-\frac{L\eta^2}{2}\right)\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}_{0,s}\|v_t^s\|^2\\
&-\frac{\eta}{4}\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}_{0,s}\|v_t^s\|^2+\frac{\eta^3L^2m}{2B}\mathbb{E}_{0,s}\sum_{i=0}^{m-1}\|v_i^s\|^2\left(\sum_{t=0}^{m-2}\gamma^{m-2-t}\right)\\
&+\frac{\tau I_{(N_s<n)}}{N_s}\sigma^2\\
\leq&\gamma^m\mathbb{E}_{0,s}(f(x_0^s)-f(x^*))-\left(\frac{\eta}{4}-\frac{L\eta^2}{2}\right)\gamma^m\sum_{t=0}^{m-1}\mathbb{E}_{0,s}\|v_t^s\|^2\\
&-\frac{\eta}{4}\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}_{0,s}\|v_t^s\|^2+\frac{\eta^2L^2m\tau}{B}\mathbb{E}_{0,s}\sum_{i=0}^{m-1}\|v_i^s\|^2+\frac{\tau I_{(N_s<n)}}{N_s}\sigma^2\\
\leq&\gamma^m\mathbb{E}_{0,s}(f(x_0^s)-f(x^*))-\left(\left(\frac{\eta}{4}-\frac{L\eta^2}{2}\right)\gamma^m-\frac{\eta^2L^2m\tau}{B}\right)\sum_{t=0}^{m-1}\mathbb{E}_{0,s}\|v_t^s\|^2\\
&+\frac{\tau I_{(N_s<n)}}{N_s}\sigma^2-\frac{\eta}{4}\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}_{0,s}\|v_t^s\|^2.
\end{aligned}\tag{49}$$

Recall $\eta=\frac{1}{c_\eta L}$ $(c_\eta>4)$, $\frac{8L\tau}{c_\eta-2}\leq m<4L\tau$ and $B=m^2$. Then, we have

$$\begin{aligned}
\left(\frac{\eta}{4}-\frac{L\eta^2}{2}\right)\gamma^m=\eta\left(\frac{1}{4}-\frac{1}{2c_\eta}\right)\left(1-\frac{1}{2c_\eta\tau L}\right)^m>\eta\left(\frac{1}{4}-\frac{1}{2c_\eta}\right)\left(1-\frac{1}{2m}\right)^m\\
\overset{(i)}{\geq}\frac{\eta}{2}\left(\frac{1}{4}-\frac{1}{2c_\eta}\right)\geq\frac{\eta^2L^2\tau}{m}=\frac{\eta^2L^2m\tau}{B},
\end{aligned}\tag{50}$$

where (i) follows from the fact that $\left(1-\frac{1}{2m}\right)^m\geq\frac{1}{2}$ for $m\geq1$. Recall $c_\beta=c_\epsilon=\alpha$ and $N_s=\min\{c_\beta\sigma^2\beta_s^{-1},c_\epsilon\sigma^2\epsilon^{-1},n\}$. Then, combining (10), (49) and (50) yields

$$\begin{aligned}
\mathbb{E}_{0,s}(f(x_m^s)-f(x^*))\leq&\gamma^m\mathbb{E}_{0,s}(f(x_0^s)-f(x^*))+\frac{\tau I_{(N_s<n)}}{N_s}\sigma^2-\frac{\eta}{4}\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}_{0,s}\|v_t^s\|^2.\\
\leq&\gamma^m\mathbb{E}_{0,s}(f(x_0^s)-f(x^*))+\tau\left(\frac{\beta_s}{\alpha}+\frac{\epsilon}{\alpha}\right)-\frac{\eta}{4}\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}_{0,s}\|v_t^s\|^2.
\end{aligned}$$

Further taking expectation of the above inequality over $x_0^1,...,x_0^s$, we obtain

$$\mathbb{E}(f(x_m^s)-f(x^*))\leq\gamma^m\mathbb{E}(f(x_0^s)-f(x^*))+\frac{\tau}{\alpha}\mathbb{E}\beta_s+\frac{\tau\epsilon}{\alpha}-\frac{\eta}{4}\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}\|v_t^s\|^2$$

Recall that $\beta_1\leq\epsilon\left(\frac{1}{\gamma}\right)^{m(S-1)}$ and $\beta_s=\frac{1}{m}\sum_{t=1}^{m}\|v_{t-1}^{s-1}\|^2$ for $s\geq2$. Then, telescoping the above inequality over s from 1 to S yields

$$\mathbb{E}(f(x_m^S)-f(x^*))\leq\gamma^{Sm}\mathbb{E}(f(x_0)-f(x^*))+\sum_{s=1}^{S-1}\gamma^{m(S-1-s)}\frac{\tau}{\alpha m}\sum_{t=0}^{m-1}\mathbb{E}\|v_t^s\|^2$$

$$
\begin{aligned}
&+\gamma^{m(S-1)}\frac{\tau\beta_1}{\alpha}+\sum_{s=1}^{S}\gamma^{m(S-s)}\frac{\tau\epsilon}{\alpha}-\frac{\eta}{4}\sum_{s=1}^{S}\gamma^{m(S-s)}\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}\|v_t^s\|^2\\
&\overset{(i)}{\leq}\gamma^K\mathbb{E}(f(x_0)-f(x^*))-\left(\frac{\eta}{4}\gamma^{2m}-\frac{\tau}{\alpha m}\right)\sum_{s=1}^{S-1}\gamma^{m(S-1-s)}\sum_{t=0}^{m-1}\mathbb{E}\|v_t^s\|^2\\
&+\left(1+\frac{1}{1-\exp(-\frac{4}{c_\eta(c_\eta-2)})}\right)\frac{\tau\epsilon}{\alpha},
\end{aligned}
\tag{51}
$$

where (i) follows from the fact that $\gamma^{m-1-t}\geq\gamma^m$ for $0\leq t\leq m-1$, $\gamma^{m(S-1)}\leq 1$, $\sum_{s=1}^{S}\gamma^{m(S-s)}\leq\frac{1}{1-\gamma^m}$ and

$$
\gamma^m=\left(1-\frac{1}{2c_\eta\tau L}\right)^m\leq\left(1-\frac{4}{c_\eta(c_\eta-2)m}\right)^m\leq\exp\left(-\frac{4}{c_\eta(c_\eta-2)}\right).
$$

Since $\alpha=\left(2\tau+\frac{2\tau}{1-\exp(-\frac{4}{c_\eta(c_\eta-2)})}\right)\vee\frac{16c_\eta L\tau}{m}$, we have

$$
\left(1+\frac{1}{1-\exp(-\frac{4}{c_\eta(c_\eta-2)})}\right)\frac{\tau\epsilon}{\alpha}\leq\frac{1}{2},\quad\frac{\eta}{4}\gamma^{2m}\overset{(i)}{>}\frac{1}{16c_\eta L}\geq\frac{\tau}{\alpha m}
\tag{52}
$$

where (i) follows from (50) that $\gamma^m\geq\left(1-\frac{1}{2m}\right)^m\geq\frac{1}{2}$. Note that $x_m^S=\tilde{x}^S$. Then, combining (52) and (51) yields

$$
\mathbb{E}(f(\tilde{x}^S)-f(x^*))\leq\gamma^K(f(x_0)-f(x^*))+\frac{\epsilon}{2}.
\tag{53}
$$

Let $K=(2c_\eta\tau L-1)\log\left(\frac{2(f(x_0)-f(x^*))}{\epsilon}\right)$. Then, we have

$$
\gamma^K(f(x_0)-f(x^*))=\exp\left[(2c_\eta\tau L-1)\log\frac{1}{\gamma}\log\left(\frac{\epsilon}{2(f(x_0)-f(x^*))}\right)\right](f(x_0)-f(x^*))\overset{(i)}{\leq}\frac{\epsilon}{2},
$$

where (i) follows from the fact that $\log\frac{1}{\gamma}=\log\left(1+\frac{1}{2c_\eta\tau L-1}\right)\leq\frac{1}{2c_\eta\tau L-1}$. Thus, the total number of SFO calls is

$$
\begin{aligned}
\sum_{s=1}^{S}\min\left\{\frac{c_\beta}{\beta_s},\frac{c_\epsilon}{\epsilon},n\right\}+KB\leq&\mathcal{O}\left(\left(\frac{c_\epsilon}{\epsilon}\wedge n\right)\frac{\tau}{m}\log\frac{1}{\epsilon}+B\tau\log\frac{1}{\epsilon}\right)\\
\overset{(i)}{\leq}&\mathcal{O}\left(\left(\frac{\tau}{\epsilon}\wedge n\right)\log\frac{1}{\epsilon}+\tau^3\log\frac{1}{\epsilon}\right),
\end{aligned}
$$

where (i) follows from the fact that $m=\Theta(\tau)$ and $c_\epsilon=\Theta(\tau)$.

H.2 Proof of Theorem 6

To simplify notations, we let $c_\beta=c_\epsilon=\alpha=\left(2\tau+\frac{2\tau}{1-\exp(\frac{-4}{c_\eta(c_\eta-2)})}\right)\vee\frac{16c_\eta L\tau}{m}$.

Using an approach similar to (47), we have, for $1\leq t\leq m$

$$
\begin{aligned}
\mathbb{E}_{0,s}(f(x_t^s)-f(x^*))\leq&\left(1-\frac{\eta}{2\tau}\right)\mathbb{E}_{0,s}(f(x_{t-1}^s)-f(x^*))-\left(\frac{\eta}{2}-\frac{L\eta^2}{2}\right)\mathbb{E}_{0,s}\|v_{t-1}^s\|^2\\
&+\frac{\eta}{2}\mathbb{E}_{0,s}\|\nabla f(x_{t-1}^s)-v_{t-1}^s\|^2,
\end{aligned}
$$

which, in conjunction with Lemma 2, implies that

$$\mathbb{E}_{0,s}(f(x_t^s)-f(x^*)) \le \left(1-\frac{\eta}{2\tau}\right)\mathbb{E}_{0,s}(f(x_{t-1}^s)-f(x^*)) - \left(\frac{\eta}{2}-\frac{L\eta^2}{2}\right)\mathbb{E}_{0,s}\|v_{t-1}^s\|^2 + \frac{\eta^3L^2}{2B}\sum_{i=0}^{t-2}\mathbb{E}_{0,s}\|v_i^s\|^2 + \frac{\eta I_{(N_s<n)}}{2N_s}\sigma^2.$$

Let $\gamma := 1-\frac{\eta}{2\tau}$. Then, telescoping the above inequality over t from 1 to m yields

$$\mathbb{E}_{0,s}(f(x_m^s)-f(x^*)) \le \gamma^m\mathbb{E}_{0,s}(f(x_0^s)-f(x^*)) - \left(\frac{\eta}{2}-\frac{L\eta^2}{2}\right)\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}_{0,s}\|v_t^s\|^2 + \frac{\eta^3L^2}{2B}\sum_{t=0}^{m-2}\gamma^{m-2-t}\sum_{i=0}^{t}\mathbb{E}_{0,s}\|v_i^s\|^2 + \left(\sum_{t=0}^{m-1}\gamma^t\right)\frac{\eta I_{(N_s<n)}}{2N_s}\sigma^2,$$

which, in conjunction with $\sum_{t=0}^{m-1}\gamma^t = \frac{1-\gamma^m}{1-\gamma} \le \frac{1}{1-\gamma} = \frac{2\tau}{\eta}$ and $\gamma^{m-1-t} \ge \gamma^m$ for $0 \le t \le m-1$, implies that

$$\begin{aligned}
\mathbb{E}_{0,s}(f(x_m^s)-f(x^*)) \le& \gamma^m\mathbb{E}_{0,s}(f(x_0^s)-f(x^*)) - \left(\frac{\eta}{4}-\frac{L\eta^2}{2}\right)\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}_{0,s}\|v_t^s\|^2 \\
&- \frac{\eta}{4}\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}_{0,s}\|v_t^s\|^2 + \frac{\eta^3L^2}{2B}\left(\sum_{t=0}^{m-2}\gamma^{m-2-t}\right)\mathbb{E}_{0,s}\sum_{i=0}^{m-1}\|v_i^s\|^2 \\
&+ \frac{\tau I_{(N_s<n)}}{N_s}\sigma^2 \\
\le& \gamma^m\mathbb{E}_{0,s}(f(x_0^s)-f(x^*)) - \left(\frac{\eta}{4}-\frac{L\eta^2}{2}\right)\gamma^m\sum_{t=0}^{m-1}\mathbb{E}_{0,s}\|v_t^s\|^2 \\
&- \frac{\eta}{4}\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}_{0,s}\|v_t^s\|^2 + \frac{\eta^2L^2\tau}{B}\mathbb{E}_{0,s}\sum_{i=0}^{m-1}\|v_i^s\|^2 + \frac{\tau I_{(N_s<n)}}{N_s}\sigma^2 \\
\le& \gamma^m\mathbb{E}_{0,s}(f(x_0^s)-f(x^*)) - \left(\left(\frac{\eta}{4}-\frac{L\eta^2}{2}\right)\gamma^m - \frac{\eta^2L^2\tau}{B}\right)\sum_{t=0}^{m-1}\mathbb{E}_{0,s}\|v_t^s\|^2 \\
&+ \frac{\tau I_{(N_s<n)}}{N_s}\sigma^2 - \frac{\eta}{4}\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}_{0,s}\|v_t^s\|^2. \qquad (54)
\end{aligned}$$

Recall that $\eta = \frac{1}{c_\eta L}$ with $c_\eta > 4$ and $B = m$ with $\frac{8L\tau}{c_\eta-2} \le m < 4L\tau$. Then, we have

$$\begin{aligned}
\left(\frac{\eta}{4}-\frac{L\eta^2}{2}\right)\gamma^m &= \eta\left(\frac{1}{4}-\frac{1}{2c_\eta}\right)\left(1-\frac{1}{2c_\eta\tau L}\right)^m > \eta\left(\frac{1}{4}-\frac{1}{2c_\eta}\right)\left(1-\frac{1}{2m}\right)^m \\
&\ge \frac{\eta}{2}\left(\frac{1}{4}-\frac{1}{2c_\eta}\right) \ge \frac{\eta^2L^2\tau}{m} = \frac{\eta^2L^2\tau}{B}, \qquad (55)
\end{aligned}$$

which, combined with (54) and (10) , implies that

$$\mathbb{E}_{0,s}(f(x_m^s)-f(x^*)) \le \gamma^m\mathbb{E}_{0,s}(f(x_0^s)-f(x^*)) + \tau\left(\frac{\beta_s}{\alpha}+\frac{\epsilon}{\alpha}\right) - \frac{\eta}{4}\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}_{0,s}\|v_t^s\|^2.$$

Taking the expectation of the above inequality $x_0^1, ..., x_0^s$, we obtain

$$\mathbb{E}(f(x_m^s)-f(x^*)) \le \gamma^m\mathbb{E}(f(x_0^s)-f(x^*)) + \frac{\tau}{\alpha}\mathbb{E}\beta_s + \frac{\tau\epsilon}{\alpha} - \frac{\eta}{4}\sum_{t=0}^{m-1}\gamma^{m-1-t}\mathbb{E}\|v_t^s\|^2.$$

Recall $\beta_1 \leq \epsilon\left(\frac{1}{\gamma}\right)^{m(S-1)}$ and $\beta_s = \frac{1}{m}\sum_{t=0}^{m-1}\|v_t^{s-1}\|^2$ for $s = 2, ..., S$. Then, telescoping the above inequality over s from 1 to S and using an approach similar to (51), we have

$$\begin{aligned}\mathbb{E}(f(x_m^S) - f(x^*)) \leq & \gamma^K \mathbb{E}(f(x_0) - f(x^*)) - \left(\frac{\eta}{4}\gamma^{2m} - \frac{\tau}{\alpha m}\right)\sum_{s=1}^{S-1}\gamma^{m(S-1-s)}\sum_{t=0}^{m-1}\mathbb{E}\|v_t^s\|^2 \\ & + \left(1 + \frac{1}{1 - \exp(-\frac{4}{c_\eta(c_\eta - 2)})}\right)\frac{\tau\epsilon}{\alpha},\end{aligned}$$

which, in conjunction with (52), yields

$$\mathbb{E}(f(x_m^S) - f(x^*)) \leq \gamma^K \mathbb{E}(f(x_0) - f(x^*)) + \frac{\epsilon}{2}. \tag{56}$$

Let $K = (2c_\eta \tau L - 1)\log\left(\frac{2(f(x_0) - f(x^*))}{\epsilon}\right)$. Then, we have $\gamma^K(f(x_0) - f(x^*)) \leq \frac{\epsilon}{2}$. Thus, the total number of SFO calls is given by

$$\begin{aligned}\sum_{s=1}^{S}\min\left\{\frac{c_\beta}{\beta_s}, \frac{c_\epsilon}{\epsilon}, n\right\} + KB \leq & \mathcal{O}\left(\left(\frac{c_\epsilon}{\epsilon} \wedge n\right)\frac{\tau}{m}\log\frac{1}{\epsilon} + B\tau\log\frac{1}{\epsilon}\right) \\ \overset{(i)}{\leq} & \mathcal{O}\left(\left(\frac{\tau}{\epsilon} \wedge n\right)\log\frac{1}{\epsilon} + \tau^2\log\frac{1}{\epsilon}\right),\end{aligned}$$

where (i) follows from the fact that $B = m = \Theta(\tau)$ and $c_\epsilon = \Theta(\tau)$.

I Proofs for Results in Appendix D

I.1 Proof of Theorem 7

Since the gradient ∇f is L-Lipschitz, we obtain that, for $t \geq 0$,

$$\begin{aligned}
f(x_{t+1}) \leq & f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\
\overset{(i)}{=} & f(x_t) - \eta \langle \nabla f(x_t), v_t \rangle + \frac{L\eta^2}{2}\|v_t\|^2 \\
= & f(x_t) - \eta \langle \nabla f(x_t) - v_t + v_t, v_t \rangle + \frac{L\eta^2}{2}\|v_t\|^2 \\
= & f(x_t) - \eta \langle \nabla f(x_t) - v_t, v_t \rangle - \eta \|v_t\|^2 + \frac{L\eta^2}{2}\|v_t\|^2 \\
= & f(x_t) - \eta \langle \nabla f(x_t) - v_t, v_t - \nabla f(x_t) + \nabla f(x_t) \rangle - \Big(\eta - \frac{L\eta^2}{2}\Big)\|v_t\|^2 \\
= & f(x_t) - \eta \langle \nabla f(x_t) - v_t, v_t - \nabla f(x_t) \rangle - \eta \langle \nabla f(x_t) - v_t, \nabla f(x_t) \rangle - \Big(\eta - \frac{L\eta^2}{2}\Big)\|v_t\|^2 \\
= & f(x_t) + \eta \|\nabla f(x_t) - v_t\|^2 - \eta \langle \nabla f(x_t) - v_t, \nabla f(x_t) \rangle - \Big(\eta - \frac{L\eta^2}{2}\Big)\|v_t\|^2
\end{aligned}$$

where (i) follows from the fact that $x_{t+1} = x_t - \eta v_t$. Then, taking expectation $\mathbb{E}(\cdot)$ over the above inequality yields

$$\begin{aligned}
\mathbb{E} f(x_{t+1}) \leq & \mathbb{E} f(x_t) + \eta \mathbb{E}\|\nabla f(x_t) - v_t\|^2 - \eta \mathbb{E}\langle \nabla f(x_t) - v_t, \nabla f(x_t) \rangle - \Big(\eta - \frac{L\eta^2}{2}\Big)\mathbb{E}\|v_t\|^2 \\
\overset{(i)}{=} & \mathbb{E} f(x_t) + \eta \mathbb{E}\|\nabla f(x_t) - v_t\|^2 - \Big(\eta - \frac{L\eta^2}{2}\Big)\mathbb{E}\|v_t\|^2 \\
= & \mathbb{E} f(x_t) - \Big(\eta - \frac{L\eta^2}{2}\Big)\mathbb{E}\|v_t\|^2 + \eta \mathbb{E}\|\nabla f(x_t) - v_t\|^2,
\end{aligned} \tag{57}$$

where (i) follows from $\mathbb{E}\langle \nabla f(x_t) - v_t, \nabla f(x_t) \rangle = \mathbb{E}_{x_0,\ldots,x_t}\left(\mathbb{E}_t \langle \nabla f(x_t) - v_t, \nabla f(x_t) \rangle\right) = 0$.

Next, we upper-bound $\mathbb{E}\|\nabla f(x_t) - v_t\|^2$. For the case when $|B_t| < n$, we have

$$\begin{aligned}
\mathbb{E}\|\nabla f(x_t) - v_t\|^2 = & \mathbb{E}\left\|\nabla f(x_t) - \frac{1}{|B_t|}\sum_{i \in B_t} \nabla f_i(x_t)\right\|^2 = \mathbb{E}\left\|\frac{1}{|B_t|}\sum_{i \in B_t} (\nabla f(x_t) - \nabla f_i(x_t))\right\|^2 \\
= & \mathbb{E}\frac{1}{|B_t|^2}\left\|\sum_{i \in B_t} (\nabla f(x_t) - \nabla f_i(x_t))\right\|^2 \\
= & \mathbb{E}\frac{1}{|B_t|^2}\sum_{i \in B_t}\sum_{j \in B_t} \langle \nabla f(x_t) - \nabla f_i(x_t), \nabla f(x_t) - \nabla f_j(x_t) \rangle \\
= & \mathbb{E}_{x_0,\ldots,x_t}\left(\mathbb{E}_t \frac{1}{|B_t|^2}\sum_{i \in B_t}\sum_{j \in B_t} \langle \nabla f(x_t) - \nabla f_i(x_t), \nabla f(x_t) - \nabla f_j(x_t) \rangle\right) \\
= & \mathbb{E}_{x_0,\ldots,x_t}\frac{1}{|B_t|^2}\sum_{i \in B_t}\sum_{j \in B_t} \mathbb{E}_t \langle \nabla f(x_t) - \nabla f_i(x_t), \nabla f(x_t) - \nabla f_j(x_t) \rangle \\
\overset{(i)}{=} & \mathbb{E}_{x_0,\ldots,x_t}\frac{1}{|B_t|^2}\sum_{i \in B_t} \mathbb{E}_t \|\nabla f(x_t) - \nabla f_i(x_t)\|^2 \overset{(ii)}{\leq} \mathbb{E}\frac{\sigma^2}{|B_t|},
\end{aligned}$$

where (i) follows from $\mathbb{E}_t \nabla f_i(x_t) = \nabla f(x_t)$, and $\mathbb{E}_t \langle \nabla f(x_t) - \nabla f_i(x_t), \nabla f(x_t) - \nabla f_j(x_t) \rangle = 0$ for $i \neq j$, and (ii) follows from item (3) in Assumption 1. For the case when $|B_t| = n$, we have $v_t = \nabla f(x_t)$, and thus

$\mathbb{E}\|\nabla f(x_t) - v_t\|^2 = 0$. Combining the above two cases, we have

$$\mathbb{E}\|\nabla f(x_t) - v_t\|^2 \leq \mathbb{E}\left(\frac{I_{(|B_t|<n)}}{|B_t|}\sigma^2\right). \tag{58}$$

Plugging (58) into (57), we obtain

$$\Big(\eta - \frac{L\eta^2}{2}\Big)\mathbb{E}\|v_t\|^2 \leq \mathbb{E}f(x_t) - \mathbb{E}f(x_{t+1}) + \mathbb{E}\frac{I_{(|B_t|<n)}}{|B_t|}\eta\sigma^2.$$

Telescoping the above inequality over t from 0 to T yields

$$\sum_{t=0}^{T}\Big(\eta - \frac{L\eta^2}{2}\Big)\mathbb{E}\|v_t\|^2 \leq \mathbb{E}f(x_0) - \mathbb{E}f(x_{T+1}) + \sum_{t=0}^{T}\mathbb{E}\frac{I_{(|B_t|<n)}}{|B_t|}\eta\sigma^2. \tag{59}$$

Next, we upper-bound $\sum_{t=0}^{T}\mathbb{E}\big(\frac{I_{(|B_t|<n)}}{|B_t|}\eta\sigma^2\big)$ in the above inequality through the following steps.

$$\begin{aligned}
\sum_{t=0}^{T}\mathbb{E}\frac{I_{(|B_t|<n)}}{|B_t|}\eta\sigma^2 &\overset{(i)}{\leq} \sum_{t=0}^{T}\mathbb{E}\left(\frac{\sum_{i=1}^{m}\|v_{t-i}\|^2}{2m\sigma^2} + \frac{\epsilon}{24\sigma^2}\right)\eta\sigma^2 \\
&= \frac{\eta}{2m}\sum_{t=0}^{T}\sum_{i=1}^{m}\mathbb{E}\|v_{t-i}\|^2 + \sum_{t=0}^{T}\frac{\eta\epsilon}{24} \\
&= \frac{\eta}{2m}\sum_{t=1}^{T}\sum_{i=1}^{\min\{m,t\}}\mathbb{E}\|v_{t-i}\|^2 + \frac{\eta}{2m}\sum_{t=0}^{\min\{m-1,T\}}\sum_{i=t+1}^{m}\mathbb{E}\|v_{t-i}\|^2 + \sum_{t=0}^{T}\frac{\eta\epsilon}{24} \\
&\overset{(ii)}{\leq} \frac{\eta}{2m}\sum_{t=1}^{T}\sum_{i=1}^{\min\{m,t\}}\mathbb{E}\|v_{t-i}\|^2 + \frac{\eta}{2m}\sum_{t=0}^{m-1}\sum_{i=t+1}^{m}\mathbb{E}\|v_{-1}\|^2 + \sum_{t=0}^{T}\frac{\eta\epsilon}{24} \\
&= \frac{\eta}{2m}\sum_{t=1}^{T}\sum_{i=1}^{\min\{m,t\}}\mathbb{E}\|v_{t-i}\|^2 + \frac{\eta m}{2}\|v_{-1}\|^2 + \sum_{t=0}^{T}\frac{\eta\epsilon}{24} \\
&= \frac{\eta}{2m}\sum_{i=0}^{T-1}\mathbb{E}\|v_i\|^2\sum_{t=i+1}^{\min\{i+m,T\}}1 + \frac{\eta m}{2}\|v_{-1}\|^2 + \sum_{t=0}^{T}\frac{\eta\epsilon}{24} \\
&\leq \frac{\eta}{2}\sum_{i=0}^{T}\mathbb{E}\|v_i\|^2 + \frac{\eta m}{2}\alpha_0^2 + \sum_{t=0}^{T}\frac{\eta\epsilon}{24}
\end{aligned} \tag{60}$$

where (i) follows from the definition of $|B_t|$, (ii) follows from the fact that $\|v_{-1}\| = \|v_{-2}\| = \cdots = \|v_{-m}\| = \alpha_0$.

Plugging (60) into (59), we obtain

$$\sum_{t=0}^{T}\Big(\eta - \frac{L\eta^2}{2}\Big)\mathbb{E}\|v_t\|^2 \leq \mathbb{E}f(x_0) - \mathbb{E}f(x_{T+1}) + \frac{\eta}{2}\sum_{i=0}^{T}\mathbb{E}\|v_i\|^2 + \frac{\eta m}{2}\alpha_0^2 + \sum_{t=0}^{T}\frac{\eta\epsilon}{24},$$

which further yields

$$\begin{aligned}
\sum_{t=0}^{T}\Big(\frac{\eta}{2} - \frac{L\eta^2}{2}\Big)\mathbb{E}\|v_t\|^2 &\leq \mathbb{E}f(x_0) - \mathbb{E}f(x_{T+1}) + \frac{\eta m}{2}\alpha_0^2 + \sum_{t=0}^{T}\frac{\eta\epsilon}{24} \\
&\leq f(x_0) - f^* + \frac{\eta m}{2}\alpha_0^2 + \sum_{t=0}^{T}\frac{\eta\epsilon}{24}.
\end{aligned} \tag{61}$$

Recall that $\phi := \left(\eta - \frac{L\eta^2}{2}\right) > 0$. Then, we obtain from (61) that

$$\sum_{t=0}^{T} \mathbb{E}\|v_t\|^2 \leq \frac{2\left(f(x_0) - f^*\right) + \eta m \alpha_0^2}{2\phi} + \frac{(T+1)\eta\epsilon}{24\phi}. \tag{62}$$

Recall that the output x_ζ is chosen from $\{x_t\}_{t=0,\ldots,T}$ uniformly at random. Then, based on (62), we have

$$\begin{aligned}
\mathbb{E}\left\|\nabla f(x_\zeta)\right\|^2 &= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left\|\nabla f(x_t)\right\|^2 = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left\|\mathbb{E}_t v_t\right\|^2 \overset{(i)}{\leq} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left(\mathbb{E}_t \left\|v_t\right\|^2\right) \\
&= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left\|v_t\right\|^2 \overset{(ii)}{\leq} \frac{2\left(f(x_0) - f^*\right) + \eta m \alpha_0^2}{2T\phi} + \frac{(T+1)\eta\epsilon}{24T\phi} \\
&\leq \frac{2\left(f(x_0) - f^*\right) + \eta m \alpha_0^2}{2T\phi} + \frac{\eta\epsilon}{12\phi},
\end{aligned}$$

where (i) follows from the Jensen's inequality, and (ii) follows from (62).

I.2 Proof of Corollary 5

Since $\eta = \frac{1}{2L}$, have

$$\phi = \left(\eta - \frac{L\eta^2}{2}\right) = \frac{1}{8L} > 0.$$

Then, plugging $\eta = \frac{1}{2L}, \phi = \frac{1}{8L}$ and $T = \left(16L\left(f(x_0) - f^*\right) + 4m\alpha_0^2\right)\epsilon^{-1}$ in Theorem 7, we have

$$\mathbb{E}\left\|\nabla f(x_\zeta)\right\|^2 \leq \frac{8L\left(f(x_0) - f^*\right) + 2m\alpha_0^2}{T} + \frac{\epsilon}{3} \leq \frac{5}{6}\epsilon \leq \epsilon.$$

Thus, the total SFO calls required by AbaSGD is given by

$$\sum_{t=0}^{T} |B_t| = \sum_{t=0}^{T} \min\left\{\frac{2\sigma^2}{\sum_{i=1}^{m}\|v_{t-i}\|^2/m}, \frac{24\sigma^2}{\epsilon}, n\right\} \leq (T+1)\left(\frac{24\sigma^2}{\epsilon} \wedge n\right) = \mathcal{O}\left(\frac{1}{\epsilon^2} \wedge \frac{n}{\epsilon}\right).$$