
Efficient Sign-Based Optimization: Accelerating Convergence via Variance Reduction

Wei Jiang¹, Sifan Yang^{1,2}, Wenhao Yang^{1,2}, Lijun Zhang^{1,3,2,*}

¹National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

²School of Artificial Intelligence, Nanjing University, Nanjing, China

³Pazhou Laboratory (Huangpu), Guangzhou, China

{jiangw, yangsf, yangwh, zhanglj}@lamda.nju.edu.cn

Abstract

Sign stochastic gradient descent (signSGD) is a communication-efficient method that transmits only the sign of stochastic gradients for parameter updating. Existing literature has demonstrated that signSGD can achieve a convergence rate of $\mathcal{O}(d^{1/2}T^{-1/4})$, where d represents the dimension and T is the iteration number. In this paper, we improve this convergence rate to $\mathcal{O}(d^{1/2}T^{-1/3})$ by introducing the Sign-based Stochastic Variance Reduction (SSVR) method, which employs variance reduction estimators to track gradients and leverages their signs to update. For finite-sum problems, our method can be further enhanced to achieve a convergence rate of $\mathcal{O}(m^{1/4}d^{1/2}T^{-1/2})$, where m denotes the number of component functions. Furthermore, we investigate the heterogeneous majority vote in distributed settings and introduce two novel algorithms that attain improved convergence rates of $\mathcal{O}(d^{1/2}T^{-1/2} + dn^{-1/2})$ and $\mathcal{O}(d^{1/4}T^{-1/4})$ respectively, outperforming the previous results of $\mathcal{O}(dT^{-1/4} + dn^{-1/2})$ and $\mathcal{O}(d^{3/8}T^{-1/8})$, where n represents the number of nodes. Numerical experiments across different tasks validate the effectiveness of our proposed methods.

1 Introduction

This paper investigates the stochastic optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}), \quad (1)$$

where $f : \mathbb{R}^d \mapsto \mathbb{R}$ is a smooth and non-convex function. We assume that only noisy estimations of the gradient $\nabla f(\mathbf{x})$ can be accessed, represented as $\nabla f(\mathbf{x}; \xi)$, where ξ is a random sample drawn from a stochastic oracle such that $\mathbb{E}[\nabla f(\mathbf{x}; \xi)] = \nabla f(\mathbf{x})$.

The most well-known method for problem (1) is stochastic gradient descent (SGD), which performs $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t; \xi_t)$ for each iteration, where ξ_t is the sample used in the t -th iteration, and η is the learning rate. It has been proved that the SGD method can obtain a convergence rate of $\mathcal{O}(T^{-1/4})$ [Ghadimi and Lan, 2013], where T is the iteration number. Recently, sign stochastic gradient descent (signSGD) method [Seide et al., 2014, Bernstein et al., 2018] has become popular in the machine learning community, which uses the sign of the stochastic gradient to update, i.e.,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \text{Sign}(\nabla f(\mathbf{x}_t; \xi_t)).$$

This method can largely reduce the communication overhead in distributed environments, and prior research [Bernstein et al., 2018, 2019] has established that signSGD can achieve a convergence rate

*Lijun Zhang is the corresponding author.

of $\mathcal{O}(d^{1/2}T^{-1/4})$ measured in terms of the l_1 -norm. Since the $\mathcal{O}(T^{-1/4})$ rate is already optimal for SGD methods when measured in the l_2 -norm [Arjevani et al., 2023], we can not further improve the dependence on T for signSGD method, considering that $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$ for any \mathbf{x} .² However, it is also known that variance reduction techniques can further enhance the convergence rate to $\mathcal{O}(T^{-1/3})$, under a slightly stronger assumption of average smoothness [Fang et al., 2018, Wang et al., 2019, Cutkosky and Orabona, 2019]. This leads to a natural question: *Can the convergence of sign-based methods be further improved by employing variance reduction techniques along with the average smoothness assumption?* We respond affirmatively by introducing the Sign-based Stochastic Variance Reduction (SSVR) method. By integrating variance reduction technique [Cutkosky and Orabona, 2019] with sign operations, we achieve an improved convergence rate of $\mathcal{O}(d^{1/2}T^{-1/3})$ measured in the l_1 -norm, matching the optimal rates in terms of T for stochastic variance reduction methods [Fang et al., 2018, Li et al., 2021, Arjevani et al., 2023].

Furthermore, we investigate a special case of problem (1), in which the objective function exhibits a finite-sum structure:

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{x}), \quad (2)$$

where each $f_i(\cdot)$ is smooth and non-convex. This problem has been extensively studied in stochastic optimization [Zhang et al., 2013, Defazio et al., 2014, Fang et al., 2018], but is less explored with sign-based methods. Previous literature proposes signSVRG [Chzhen and Schechtman, 2023] method to deal with the finite-sum problem, which achieves a convergence rate of $\mathcal{O}(m^{1/2}d^{1/2}T^{-1/2})$. However, its dependence on m is sub-optimal, failing to match the $\mathcal{O}(m^{1/4}T^{-1/2})$ lower bound [Fang et al., 2018, Li et al., 2021] for problem (2). To address this gap, we propose the SSVR-FS algorithm, which periodically computes the exact gradient [Zhang et al., 2013, Johnson and Zhang, 2013] and incorporates it into the variance reduction estimator. In this way, we can achieve an improved convergence rate of $\mathcal{O}(m^{1/4}d^{1/2}T^{-1/2})$ for finite-sum problems.

Finally, sign-based methods are especially favorable in distributed settings, where the parameter server aggregates gradient signs from each worker through majority vote [Bernstein et al., 2018], allowing 1-bit compression of communication in both directions. Existing literature [Bernstein et al., 2018, 2019] has proved that signSGD can obtain a convergence rate of $\mathcal{O}(d^{1/2}T^{-1/4})$ for majority vote in homogeneous settings, where the data across nodes is uniformly distributed or identical. For the more challenging heterogeneous setting, in which data distribution can vary significantly across nodes, existing methods can only achieve convergence rates of $\mathcal{O}(dT^{-1/4} + dn^{-1/2})$ [Sun et al., 2023] and $\mathcal{O}(d^{3/8}T^{-1/8})^3$ [Jin et al., 2023], where n denotes the number of nodes. Note that the first rate indicates that the gradient does not converge to zero as T approaches infinity, and the second one suffers from a high sample complexity. To address these limitations, we first introduce our basic SSVR-MV method, which employs variance reduction estimators to track gradients and replaces the sign operation in each worker as a stochastic unbiased sign operation. This practice ensures 1-bit compression and unbiased estimation at the same time, and the newly proposed method can obtain an improved convergence rate of $\mathcal{O}(d^{1/2}T^{-1/2} + dn^{-1/2})$. By further substituting the sign operation in the parameter server with another stochastic unbiased sign operation, our method can further achieve a convergence rate of $\mathcal{O}(d^{1/4}T^{-1/4})$, which converges to zero as T increases.

In summary, compared with existing methods, this paper makes the following contributions:

- For stochastic non-convex functions, we develop a sign-based variance reduction algorithm to achieve an improved convergence rate of $\mathcal{O}(d^{1/2}T^{-1/3})$, surpassing the $\mathcal{O}(d^{1/2}T^{-1/4})$ rate for signSGD methods.
- For non-convex finite-sum optimization, we further improve the our proposed method to obtain an enhanced convergence rate of $\mathcal{O}(m^{1/4}d^{1/2}T^{-1/2})$, which is better than the $\mathcal{O}(m^{1/2}d^{1/2}T^{-1/2})$ convergence rate for SignSVRG method.

²Note that most lower bounds are measured in the l_2 -norm, and there are no existing lower bounds for the l_1 -norm to the best of our knowledge. However, since $\|\mathbf{x}\|_1 \approx \sqrt{d} \|\mathbf{x}\|_2$ for dense vector \mathbf{x} , the additional $\mathcal{O}(\sqrt{d})$ term is acceptable when considering the l_1 -norm. This is also the case for our methods (measured in the l_1 -norm) and the lower bounds for variance reduction methods (for l_2 -norm).

³The original convergence rate is measured under the squared l_2 -norm, and we convert it to the rate under the l_2 -norm criterion for a fair comparison.

Table 1: Summary of results for sign-based algorithms. Here, stochastic indicates problem (1), finite-sum represents problem (2), N is the number of stochastic gradient calls, and m is the number of component functions. Note that some rates are measured under squared l_1 - or l_2 -norm, and we convert them to l_1 - or l_2 -norm for a fair comparison.

Method	Setting	Measure	Convergence rate
signSGD [Bernstein et al., 2018]	stochastic	l_1 -norm	$\mathcal{O}\left(\frac{d^{1/2}}{N^{1/4}}\right)$
EF-signSGD [Karimireddy et al., 2019]	stochastic	l_2 -norm	$\mathcal{O}\left(\frac{1}{N^{1/4}}\right)$
signSGD-SIM [Sun et al., 2023]	stochastic	l_1 -norm	$\mathcal{O}\left(\frac{d}{N^{1/4}}\right)$
SignSVRG [Chzhen and Schechtman, 2023]	finite-sum	l_1 -norm	$\mathcal{O}\left(\frac{d^{1/2}m^{1/2}}{N^{1/2}}\right)$
SignRVR/SignRVM [Qin et al., 2023]	finite-sum	l_1 -norm	$\mathcal{O}\left(\frac{d^{1/2}m^{1/2}}{N^{1/2}}\right)$
Theorem 1	stochastic	l_1 -norm	$\mathcal{O}\left(\frac{d^{1/2}}{N^{1/3}}\right)$
Theorem 2	finite-sum	l_1 -norm	$\mathcal{O}\left(\frac{d^{1/2}m^{1/4}}{N^{1/2}}\right)$

Table 2: Summary of results for sign-based algorithms under the majority vote setting, where n is the number of workers. Some rates are measured under squared l_1 - or l_2 -norm, and we convert them to l_1 - or l_2 -norm for a fair comparison.

Method	Heterogeneous	Measure	Convergence rate
signSGD [Bernstein et al., 2018]	✗	l_1 -norm	$\mathcal{O}\left(\frac{d^{1/2}}{T^{1/4}}\right)$
Signum [Bernstein et al., 2019]	✗	l_1 -norm	$\mathcal{O}\left(\frac{d^{1/2}}{T^{1/4}}\right)$
SSDM [Safaryan and Richtarik, 2021]	✓	l_2 -norm	$\mathcal{O}\left(\frac{d^{1/2}}{T^{1/4}}\right)$
Sto-signSGD [Jin et al., 2023]	✓	l_2 -norm	$\mathcal{O}\left(\frac{d^{3/8}}{T^{1/8}}\right)$
MV-sto-signSGD-SIM [Sun et al., 2023]	✓	l_1 -norm	$\mathcal{O}\left(\frac{d}{T^{1/4}} + \frac{d}{n^{1/2}}\right)$
Theorem 3	✓	l_1 -norm	$\mathcal{O}\left(\frac{d^{1/2}}{T^{1/2}} + \frac{d}{n^{1/2}}\right)$
Theorem 4	✓	l_2 -norm	$\mathcal{O}\left(\frac{d^{1/4}}{T^{1/4}}\right)$

- We also investigate sign-based variance reduction methods with heterogeneous majority vote in distributed settings. The proposed algorithms can obtain the convergence rates of $\mathcal{O}(d^{1/2}T^{-1/2} + dn^{-1/2})$ and $\mathcal{O}(d^{1/4}T^{-1/4})$, which outperform the previous results of $\mathcal{O}(dT^{-1/4} + dn^{-1/2})$ and $\mathcal{O}(d^{3/8}T^{-1/8})$, respectively.

We compare our results with existing methods in Table 1 and Table 2, and validate the effectiveness of our method via numerical experiments in Section 5.

2 Related work

This section provides an overview of the existing literature on signSGD methods and stochastic variance reduction techniques.

2.1 SignSGD and its variants

The idea of only transmitting the sign information of the stochastic gradient traces back to the 1-bit SGD algorithm, introduced by Seide et al. [2014]. Despite the biased nature of the sign operation, Bernstein et al. [2018] demonstrated that signSGD achieves a convergence rate of $\mathcal{O}(d^{1/2}T^{-1/4})$ by using large batch sizes in each iteration. Despite the theoretical assurance, Karimireddy et al. [2019] highlighted that signSGD may not converge to the optimal solutions for convex functions and could suffer from poor generalization without large batches. To address these issues, they proposed the

EF-signSGD method, which integrates error feedback into signSGD to correct errors introduced by the sign operation. Instead of requiring unbiased stochastic gradients in previous literature, Safaryan and Richtarik [2021] assumed that the signs of the stochastic gradient are the same as those of true gradient with a probability greater than $1/2$. Under this assumption, they demonstrated that signSGD can obtain a similar convergence rate but does not require large batches anymore. Recently, Sun et al. [2023] proposed the signSGD-SIM method, which incorporates the momentum into the signSGD, achieving a convergence rate of $\mathcal{O}(dT^{-1/4})$ with constant batch sizes and an improved convergence of $\mathcal{O}(d^{3/2}T^{-2/7})$ with second-order smoothness.

To deal with the finite-sum problems, Chzhen and Schechtman [2023] developed SignSVRG algorithm, which combines SVRG [Johnson and Zhang, 2013] method with signSGD and achieves a convergence rate of $\mathcal{O}(d^{1/2}m^{1/2}T^{-1/2})$, where m is the number of component functions. More recently, Qin et al. [2023] further investigate signSGD with random reshuffling, achieving a convergence rate of $\mathcal{O}(m^{-1/2}T^{-1/2} \log(mT) + \|\sigma\|_1)$, where σ is the variance bound of stochastic gradients. By leveraging variance-reduced gradients and momentum updates, they further propose the SignRVR and SignRVM methods, both achieving the convergence rate of $\mathcal{O}(d^{1/2}m^{1/2}T^{-1/2})$.

In distributed settings, sign-based methods with majority vote are also widely investigated. Bernstein et al. [2018, 2019] first indicated that signSGD and its momentum variant Signum can enable 1-bit compression of worker-server communication, obtaining the $\mathcal{O}(d^{1/2}T^{-1/4})$ convergence rates in the homogeneous environment. For the more challenging heterogeneous settings, SSDM method [Safaryan and Richtarik, 2021] attains the same $\mathcal{O}(d^{1/2}T^{-1/4})$ convergence rate, but the information sent back to the server is not a sign information anymore. To remedy this issue, Sto-signSGD algorithm [Jin et al., 2023] is proposed, equipped with a convergence rate of $\mathcal{O}(d^{3/4}T^{-1/4})$ measured in squared l_2 -norm. More recently, Sun et al. [2023] introduced the MV-signSGD-SIM algorithm and demonstrated a convergence rate of $\mathcal{O}(dT^{-1/4} + dn^{-1/2})$, which could be further enhanced to $\mathcal{O}(d^{3/2}T^{-2/7} + dn^{-1/2})$ under second-order smoothness conditions, where n denotes the number of nodes in the distributed system.

2.2 Stochastic variance reduction methods

Stochastic variance reduction methods have gained significant attention in the optimization community in recent years. Among the pioneering approaches, the stochastic average gradient (SAG) method [Roux et al., 2012] and Stochastic Dual Coordinate Ascent (SDCA) algorithm [Shalev-Shwartz and Zhang, 2013] utilize a memory of previous gradients to ensure variance reduction, achieving linear convergence for strongly convex functions. To circumvent the need for storing gradients, the stochastic variance reduced gradient (SVRG) [Zhang et al., 2013, Johnson and Zhang, 2013] recalculates the full gradient periodically to enhance the accuracy of gradient estimators, maintaining linear convergence for strongly convex functions. Inspired by SAG and SVRG, Defazio et al. [2014] introduced the SAGA algorithm, which not only provides superior convergence rates but also supports proximal regularization. Subsequently, the stochastic recursive gradient algorithm (SARAH) [Nguyen et al., 2017] employs a simple recursive approach to update gradient estimators, ensuring better convergence for smooth convex functions.

For non-convex optimization, inspired by the SVRG algorithm, many methods [Reddi et al., 2016, Lei et al., 2017, Zhou et al., 2020] employ variance reduction to design their algorithms and provide the corresponding convergence guarantees. More recent well-known advancements include the SPIDER [Fang et al., 2018] and SpiderBoost [Wang et al., 2019] methods, which improved the $\mathcal{O}(T^{-1/4})$ convergence rate of traditional SGD to $\mathcal{O}(T^{-1/3})$ under the average smoothness assumption. The convergence rate can be further improved to $\mathcal{O}(m^{1/4}T^{-1/2})$ for problems with a finite-sum structure, where m represents the number of component functions. However, these methods typically require a huge batch size to ensure convergence. To avoid this limitation, the stochastic recursive momentum (STORM) method [Cutkosky and Orabona, 2019] introduces a momentum-based updating and an adaptive learning rate based on the stochastic gradients, achieving a convergence rate of $\tilde{\mathcal{O}}(T^{-1/3})$ without necessitating large batches. More recently, variance reduction techniques are widely employed in more complex problems to improve the existing convergence rates, such as compositional optimization [Wang et al., 2017a,b, Yuan et al., 2019, Jiang et al., 2022a, 2023], multi-level optimization [Chen et al., 2021, Zhang and Xiao, 2021, Jiang et al., 2022b, 2024b], adaptive algorithms [Kavis et al., 2022, Jiang et al., 2024a], and distributionally robust optimization [Yu et al., 2024].

Algorithm 1 SSVR

- 1: **Input:** time step T , initial point \mathbf{x}_1
 - 2: **for** time step $t = 1$ **to** T **do**
 - 3: Draw a batch of samples $\{\xi_t^1, \dots, \xi_t^{B_1}\}$
 - 4: Compute $\mathbf{v}_t = \frac{1}{B_1} \sum_{k=1}^{B_1} \nabla f(\mathbf{x}_t; \xi_t^k) + (1 - \beta) \left(\mathbf{v}_{t-1} - \frac{1}{B_1} \sum_{k=1}^{B_1} \nabla f(\mathbf{x}_{t-1}; \xi_t^k) \right)$
 - 5: Update the decision variable: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \text{Sign}(\mathbf{v}_t)$
 - 6: **end for**
 - 7: Select τ uniformly at random from $\{1, \dots, T\}$
 - 8: Return \mathbf{x}_τ
-

3 The proposed methods

In this section, we present the proposed methods for the expectation case, i.e., problem (1), and the finite-sum structure, i.e., problem (2), respectively, along with corresponding theoretical guarantees.

3.1 Sign-based stochastic variance reduction

In this subsection, we introduce our Sign-based Stochastic Variance Reduction (SSVR) method for problem (1). One crucial step in stochastic optimization is to track the gradient of the objective function. Here, we use a variance reduction gradient estimator \mathbf{v}_t to evaluate the overall gradient $\nabla f(\mathbf{x}_t)$. In the first iteration ($t = 1$), the estimator is defined as $\mathbf{v}_1 = \frac{1}{B_0} \sum_{k=1}^{B_0} \nabla f(\mathbf{x}_1; \xi_1^k)$, where B_0 is the batch size used in the first iteration. For subsequent iterations ($t \geq 2$), \mathbf{v}_t is updated in the style of STORM [Cutkosky and Orabona, 2019], i.e.,

$$\mathbf{v}_t = \frac{1}{B_1} \sum_{k=1}^{B_1} \nabla f(\mathbf{x}_t; \xi_t^k) + (1 - \beta) \left(\mathbf{v}_{t-1} - \frac{1}{B_1} \sum_{k=1}^{B_1} \nabla f(\mathbf{x}_{t-1}; \xi_t^k) \right),$$

where β represents the momentum parameter and B_1 is the batch size. This method ensures that the expectation of the estimation error $\mathbb{E}[\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|^2]$ would be reduced gradually. After obtaining the gradient estimator \mathbf{v}_t , we update the decision variable using the sign of \mathbf{v}_t :

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \text{Sign}(\mathbf{v}_t).$$

The whole algorithm is outlined in Algorithm 1. Next, we introduce the following assumptions for our SSVR method, which are standard and commonly adopted in the analysis of variance reduction methods and stochastic non-convex optimization [Fang et al., 2018, Wang et al., 2019, Cutkosky and Orabona, 2019, Li et al., 2021].

Assumption 1 (Average smoothness)

$$\mathbb{E}_\xi \left[\|\nabla f(\mathbf{x}; \xi) - \nabla f(\mathbf{y}; \xi)\|^2 \right] \leq L^2 \|\mathbf{x} - \mathbf{y}\|^2.$$

Assumption 2 (Bounded variance)

$$\mathbb{E}_\xi \left[\|\nabla f(\mathbf{x}; \xi) - \nabla f(\mathbf{x})\|^2 \right] \leq \sigma^2.$$

With the above assumptions, we can obtain the theoretical guarantee for our method as stated below.

Theorem 1 Under Assumptions 1 and 2, by setting $\beta = \mathcal{O}(\frac{1}{T^{2/3}})$, $\eta = \mathcal{O}(\frac{1}{d^{1/2}T^{2/3}})$, $B_0 = \mathcal{O}(T^{1/3})$, and $B_1 = \mathcal{O}(1)$, our SSVR method ensures:

$$\mathbb{E}[\|\nabla f(\mathbf{x}_\tau)\|_1] \leq \mathcal{O}\left(\frac{d^{1/2}}{T^{1/3}}\right).$$

Remark: This convergence rate surpasses the $\mathcal{O}(d^{1/2}T^{-1/4})$ rate achieved by previous sign-based methods [Bernstein et al., 2018, 2019], and it also outperforms the $\mathcal{O}(d^{3/2}T^{-2/7})$ convergence rate under the second-order smoothness condition [Sun et al., 2023]. Specifically, to ensure that $\mathbb{E}[\|\nabla f(\mathbf{x}_\tau)\|_1] \leq \epsilon$, our method requires a sample complexity of $\mathcal{O}(d^{3/2}\epsilon^{-3})$, which is much better than the $\mathcal{O}(d^2\epsilon^{-4})$ and $\mathcal{O}(d^{21/4}\epsilon^{-7/2})$ complexities of previous approaches.

Algorithm 2 SSVR for Finite-Sum (SSVR-FS)

```
1: Input: time step  $T$ , initial point  $\mathbf{x}_1$ 
2: for time step  $t = 1$  to  $T$  do
3:   if  $t \bmod I == 0$  then
4:     Set  $\tau = t$  and compute  $\nabla f(\mathbf{x}_\tau) = \frac{1}{m} \sum_{i=1}^m \nabla f_i(\mathbf{x}_\tau)$ 
5:   end if
6:   Sample  $i_t$  randomly from  $\{1, 2, \dots, m\}$ 
7:   Compute gradient estimator  $\mathbf{v}_t$  according to equation (3)
8:   Update the decision variable:  $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \text{Sign}(\mathbf{v}_t)$ 
9: end for
10: Select  $\varphi$  uniformly at random from  $\{1, \dots, T\}$ 
11: Return  $\mathbf{x}_\varphi$ 
```

3.2 Sign-based stochastic variance reduction for finite-sum structure

We now extend our SSVR method to deal with the finite-sum structure in problem (2). In this context, we introduce the following assumption for each component function, which is standard and widely adopted in existing literature [Fang et al., 2018, Wang et al., 2019, Li et al., 2021].

Assumption 3 (Smoothness) For each $i \in \{1, 2, \dots, m\}$, the gradient functions satisfy:

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|.$$

To handle the finite-sum problems, we retain the core structure of our SSVR method while incorporating elements from the SVRG [Zhang et al., 2013, Johnson and Zhang, 2013] approach. Specifically, we compute a full batch gradient at the first step and every I iteration, i.e.,

$$\nabla f(\mathbf{x}_\tau) = \frac{1}{m} \sum_{i=1}^m \nabla f_i(\mathbf{x}_\tau).$$

For other iterations, we randomly select an index i_t from the set $\{1, 2, \dots, m\}$ and construct a variance reduction gradient estimator \mathbf{v}_t as follows:

$$\mathbf{v}_t = \underbrace{\nabla f_{i_t}(\mathbf{x}_t)}_{\text{STORM estimator}} + (1 - \beta)(\mathbf{v}_{t-1} - \nabla f_{i_t}(\mathbf{x}_{t-1})) - \underbrace{\beta(\nabla f_{i_t}(\mathbf{x}_\tau) - \nabla f(\mathbf{x}_\tau))}_{\text{error correction}}. \quad (3)$$

The first two terms of \mathbf{v}_t align with the STORM estimator, and the last term measures the difference of past gradients between the selected component function $\nabla f_{i_t}(\mathbf{x}_\tau)$ and the overall objective $\nabla f(\mathbf{x}_\tau)$. Note that the STORM estimator employs the component gradient $\nabla f_{i_t}(\mathbf{x}_t)$ to track the overall gradient $\nabla f(\mathbf{x}_t)$, which leads to an estimation error due to the gap between the component function and the overall objective. This gap can be effectively mitigated by the error correction term we introduced in equation (3). With such a design, we can obtain a better gradient estimation of the overall gradient, and ensure that the estimation error $\mathbb{E}[\|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|^2]$ can be reduced gradually. After computing \mathbf{v}_t , we utilize its sign information to update the decision variable. The detailed procedure is outlined in Algorithm 2. Next, we present the theoretical convergence for this method.

Theorem 2 Under Assumption 3, by setting $\beta = \mathcal{O}(\frac{1}{m})$, $I = m$, and $\eta = \mathcal{O}(\frac{1}{m^{1/4}d^{1/2}T^{1/2}})$, our algorithm ensures:

$$\mathbb{E}[\|\nabla F(\mathbf{x}_\varphi)\|_1] \leq \mathcal{O}\left(\frac{m^{1/4}d^{1/2}}{T^{1/2}}\right).$$

Remark: To ensure $\mathbb{E}[\|\nabla F(\mathbf{x}_\varphi)\|_1] \leq \epsilon$, the sample complexity is $\mathcal{O}(m + \frac{d\sqrt{m}}{\epsilon^2})$, which improves over the $\mathcal{O}(\frac{dm}{\epsilon^2})$ complexity of the previous SignSVRG method [Chzhen and Schechtman, 2023].

4 Sign-based stochastic variance reduction with majority vote

Sign-based methods are advantageous in distributed settings for their low communication overhead, as they can only transmit sign information between nodes via majority vote. This section explores sign-based stochastic methods with majority vote, a typical example of distributed learning extensively

Algorithm 3 SSVR-MV

- 1: **Input:** time step T , initial point \mathbf{x}_1
 - 2: **for** time step $t = 1$ **to** T **do**
 - 3: **On** node $j \in \{1, 2, \dots, n\}$:
 - 4: Draw sample ξ_t^j and compute $\mathbf{v}_t^j = \nabla f_j(\mathbf{x}_t; \xi_t^j) + (1 - \beta) \left(\mathbf{v}_{t-1}^j - \nabla f_j(\mathbf{x}_{t-1}; \xi_t^j) \right)$
 - 5: Option 1: Send $S_R(\mathbf{v}_t^j)$ to the parameter server, where $R = 4G$
 - 6: Option 2: Send $S_G(\hat{\mathbf{v}}_t^j)$ to the parameter server, where $\hat{\mathbf{v}}_t^j = \Pi_G[\mathbf{v}_t^j]$
 - 7: **On** parameter server:
 - 8: Option 1: Send $\mathbf{v}_t = \text{Sign} \left(\frac{1}{n} \sum_{j=1}^n S_R(\mathbf{v}_t^j) \right)$ to all nodes
 - 9: Option 2: Send $\mathbf{v}_t = S_1 \left(\frac{1}{n} \sum_{j=1}^n S_G(\hat{\mathbf{v}}_t^j) \right)$ to all nodes
 - 10: **On** node $j \in \{1, 2, \dots, n\}$:
 - 11: Update the decision variable $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \mathbf{v}_t$
 - 12: **end for**
 - 13: Select τ uniformly at random from $\{1, \dots, T\}$
 - 14: **Return** \mathbf{x}_τ
-

studied in previous sign-based algorithms [Bernstein et al., 2018, 2019, Safaryan and Richtarik, 2021, Sun et al., 2023]. To begin with, we investigate the following distributed learning task:

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) := \frac{1}{n} \sum_{j=1}^n f_j(\mathbf{x}), \quad f_j(\mathbf{x}) = \mathbb{E}_{\xi^j \sim \mathcal{D}_j} [f_j(\mathbf{x}; \xi^j)], \quad (4)$$

where \mathcal{D}_j represents the data distribution for node j , and $f_j(\mathbf{x})$ is the corresponding loss function. Some previous studies [Bernstein et al., 2018, 2019] investigate the homogeneous setting, which assumes the data across each node is uniformly distributed or identical, ensuring that $\mathbb{E}[f_i(\mathbf{x})] = f(\mathbf{x})$. In contrast, this paper considers the more challenging heterogeneous setting [Jin et al., 2023, Sun et al., 2023], where data distributions can vary significantly across nodes.

For sign-based methods in distributed settings, each node j computes a gradient estimator \mathbf{v}_t^j and transmits its sign, i.e., $\text{Sign}(\mathbf{v}_t^j)$, to the parameter server. Note that the server can not directly send the aggregate information $\sum_{j=1}^n \text{Sign}(\mathbf{v}_t^j)$ back to each node, since it loses binary characteristic after summation. A natural solution is to apply another sign operation to update the decision variable as:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \text{Sign} \left(\frac{1}{n} \sum_{j=1}^n \text{Sign}(\mathbf{v}_t^j) \right).$$

This process is called majority vote [Bernstein et al., 2018], as each worker votes on the sign of the gradient, with the server tallying these votes and broadcasting the decision back to the nodes. However, the sign operation introduces bias in the estimation, and employing it twice can significantly amplify this bias, particularly in a heterogeneous environment. Previous analysis [Chen et al., 2020] indicates that signSGD fails to converge in the heterogeneous setting. To deal with this problem, we introduce an unbiased sign operation $S_R(\cdot)$, which is defined below.

Definition 1 For any vector \mathbf{v} with $\|\mathbf{v}\|_\infty \leq R$, define the function mapping $S_R(\mathbf{v})$ as:

$$[S_R(\mathbf{v})]_k = \begin{cases} +1, & \text{with probability } \frac{1}{2} + \frac{[\mathbf{v}]_k}{2R}, \\ -1, & \text{with probability } \frac{1}{2} - \frac{[\mathbf{v}]_k}{2R}. \end{cases} \quad (5)$$

Remark: This operation provides an unbiased estimation of \mathbf{v}/R , such that $\mathbb{E}[S_R(\mathbf{v})] = \mathbf{v}/R$. It is worth noting that the function mapping is valid when $\|\mathbf{v}\|_\infty \leq R$, since the probability should always fall within $[0, 1]$. For this purpose, we need to further assume that the gradient is bounded.

Utilizing this unbiased sign operation, we can update the decision variable as:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \text{Sign} \left(\frac{1}{n} \sum_{j=1}^n S_R(\mathbf{v}_t^j) \right).$$

After applying $S_R(\cdot)$, the output is a sign information, which can be transported between nodes efficiently. The complete algorithm, named SSVR with majority vote (SSVR-MV), is described in Algorithm 3 (with *Option 1*). Note that in Step 4, we set $\mathbf{v}_1^j = \nabla f(\mathbf{x}_1; \xi_1^j)$ when $t = 1$. Next, we present the convergence guarantee for the proposed algorithm with the following assumption.

Assumption 4 For each node j , the stochastic gradient is bounded by G in the infinity norm, such that $\|\nabla f_j(\mathbf{x}; \xi)\|_\infty \leq G$.

Theorem 3 Under Assumptions 1, 2 and 4, by setting $\beta = \frac{1}{2}$ and $\eta = \mathcal{O}(\frac{1}{T^{1/2}d^{1/2}})$, our SSVR-MV method (with *Option 1*) ensures:

$$\mathbb{E} [\|\nabla f(\mathbf{x}_\tau)\|_1] \leq \mathcal{O} \left(\frac{d^{1/2}}{T^{1/2}} + \frac{d}{n^{1/2}} \right).$$

Remark: Our rate is better than the previous result of $\mathcal{O}(dT^{-1/4} + dn^{-1/2})$, and also outperforms the rate of $\mathcal{O}(d^{3/2}T^{-2/7} + dn^{-1/2})$ under the second-order smoothness [Sun et al., 2023].

Although the above convergence rate is superior to previous results, we have to note that the gradient does not converge to zero even as $T \rightarrow \infty$. To address this issue, we propose replacing another sign operation with the $S_1(\cdot)$ mapping, as defined in equation (5) with $R = 1$. Additionally, in our prior analysis, we ensured that each \mathbf{v}_t^j is bounded by assuming the stochastic gradient is bounded and using a constant β . Here, we instead suppose that the true gradient is bounded, as detailed below.

Assumption 4' For each node j , the gradient is bounded such that $\|\nabla f_j(\mathbf{x})\| \leq G$.

Remark: This assumption is weaker than the one used by Sun et al. [2023], which assumes all stochastic gradients are bounded, i.e., $\|\nabla f_j(\mathbf{x}; \xi)\| \leq G$.

To ensure each gradient estimator is bounded, we employ a projection operation $\hat{\mathbf{v}}_t^j = \Pi_G[\mathbf{v}_t^j]$, where Π_G denotes the projection onto a ball of radius G . This allows us to utilize an unbiased sign mapping $S_G(\hat{\mathbf{v}}_t^j)$ before transmission to the parameter server. The revised algorithm is presented in Algorithm 3 (with *Option 2*), and the modifications lie in Steps 6 and 9. We now present the convergence guarantee for this modified approach below.

Theorem 4 Under Assumptions 1, 2 and 4', by setting $\beta = \mathcal{O}(\frac{1}{T^{1/2}})$ and $\eta = \mathcal{O}(\frac{1}{d^{1/2}T^{1/2}})$, our SSVR-MV method (with *Option 2*) ensures:

$$\mathbb{E} [\|\nabla f(\mathbf{x}_\tau)\|] \leq \mathcal{O} \left(\frac{d^{1/4}}{T^{1/4}} \right).$$

Remark: This rate converges to zero as $T \rightarrow \infty$, and offers a significant improvement over the previous results of $\mathcal{O}(d^{3/8}T^{-1/8})$ [Jin et al., 2023]. Our result is also better than the $\mathcal{O}(d^{1/2}T^{-1/4})$ convergence rate obtained by Safaryan and Richtarik [2021], whose algorithm requires transmitting $\sum_{j=1}^n \text{sign}(\mathbf{v}_t^j)$ back to all nodes, which is actually not sign information anymore.

5 Experiments

In this section, we assess the performance of the proposed methods through numerical experiments. We first evaluate the SSVR and SSVR-FS algorithms within the centralized setting, and then assess the performance of SSVR-MV method in the distributed learning environment. All experiments are conducted on NVIDIA 3090 GPUs.

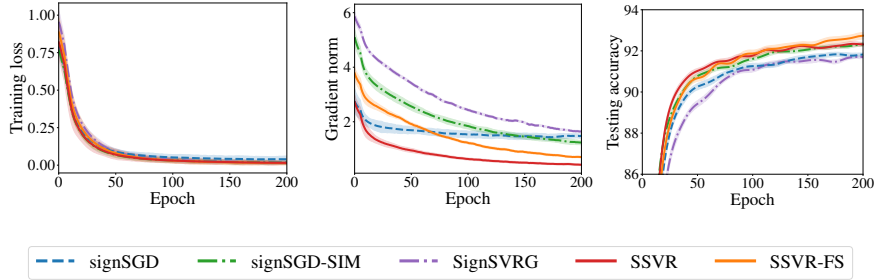


Figure 1: Results for CIFAR-10 dataset in the centralized environment.

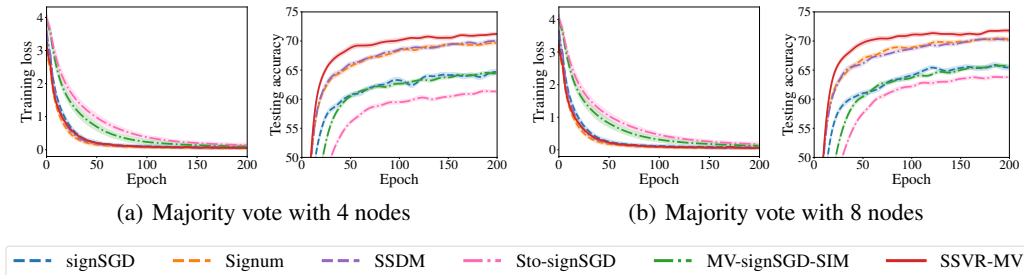


Figure 2: Results for CIFAR-100 dataset in the distributed environment.

5.1 Evaluation of SSVR and SSVR-FS methods in the centralized environment

To begin with, we conduct numerical experiments on multi-class image classification tasks to validate the effectiveness of our proposed methods. Concretely, we train a ResNet18 model [He et al., 2016] on the CIFAR-10 dataset [Krizhevsky, 2009]. We compare the performance of our SSVR and SSVR-FS methods against signSGD [Bernstein et al., 2018], signSGD-SIM [Sun et al., 2023], and SignSVRG [Chzhen and Schechtman, 2023]. For hyper-parameter tuning, we either follow the recommendations from the original papers or employ a grid search to determine the best settings. Specifically, the momentum parameter β is searched from the set $\{0.1, 0.5, 0.9, 0.99\}$, and the learning rate is fine-tuned within the range of $\{1e-5, 1e-4, 1e-3, 1e-2, 1e-1\}$.

Results. The training loss, gradient norm, and testing accuracy are presented in Figure 1, with curves averaged over five runs. We observe that all methods exhibit a rapid decrease in training losses, with our methods showing a more pronounced reduction in the gradient norm. In terms of testing accuracy, our SSVR algorithm outperforms other sign-based methods, and our SSVR-FS method achieves superior accuracy in the final epochs.

5.2 Evaluation of SSVR-MV method in the distributed learning

Subsequently, we conduct experiments to evaluate the effectiveness of the SSVR-MV method in the distributed environment. Specifically, we train a ResNet50 model [He et al., 2016] on the CIFAR-100 dataset [Krizhevsky, 2009] with 4 and 8 nodes respectively. We compare the performance of our method against signSGD (with majority vote) [Bernstein et al., 2018], Signum (with majority vote) [Bernstein et al., 2019], SSDM [Safaryan and Richtarik, 2021], Sto-signSGD [Jin et al., 2023], and MV-signSGD-SIM [Sun et al., 2023]. The hyper-parameter tuning follows the same methodology as in the centralized environment experiment.

Results. We plot the training loss and testing accuracy in Figure 2, with all curves averaged over five runs. The results indicate that the training loss of our SSVR-MV algorithm decreases rapidly, and our method obtains higher testing accuracy compared to other methods, both in experiments with 4 nodes and 8 nodes.

6 Conclusion

In this paper, we explore sign-based stochastic variance reduction (SSVR) methods, which use only the sign information of variance-reduced estimators to update decision variables. The proposed method achieves an improved convergence rate of $\mathcal{O}(d^{1/2}T^{-1/3})$, surpassing the $\mathcal{O}(d^{1/2}T^{-1/4})$ convergence rate of signSGD methods. When applied to finite-sum problems, this rate can be further enhanced to $\mathcal{O}(m^{1/4}d^{1/2}T^{-1/2})$, which is also better than the $\mathcal{O}(m^{1/2}d^{1/2}T^{-1/2})$ convergence rate of SignSVRG. Finally, we investigate the SSVR method in distributed settings and devise novel algorithms to attain convergence rates of $\mathcal{O}(d^{1/2}T^{-1/2} + dn^{-1/2})$ and $\mathcal{O}(d^{1/4}T^{-1/4})$, which improve upon the previous results of $\mathcal{O}(dT^{-1/4} + dn^{-1/2})$ and $\mathcal{O}(d^{3/8}T^{-1/8})$ respectively.

Acknowledgements

This work was partially supported by NSFC (62122037), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX24_0231).

References

- Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. E. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023.
- J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, pages 560–569, 2018.
- J. Bernstein, J. Zhao, K. Azizzadenesheli, and A. Anandkumar. signSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations*, 2019.
- T. Chen, Y. Sun, and W. Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021.
- X. Chen, T. Chen, H. Sun, S. Z. Wu, and M. Hong. Distributed training with heterogeneous data: Bridging median- and mean-based algorithms. In *Advances in Neural Information Processing Systems 33*, pages 21616–21626, 2020.
- Z. Chen, Y. Zhou, Y. Liang, and Z. Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *Proceedings of the 40th International Conference on Machine Learning*, pages 5396–5427, 2023.
- E. Chzhen and S. Schechtman. SignSVRG: fixing SignSGD via variance reduction. *ArXiv e-prints*, arXiv:2305.13187, 2023.
- A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems 32*, pages 15210–15219, 2019.
- A. Defazio, F. R. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems 27*, pages 1646–1654, 2014.
- C. Fang, C. J. Li, Z. Lin, and T. Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems 31*, pages 689–699, 2018.
- M. Faw, I. Tziotis, C. Caramanis, A. Mokhtari, S. Shakkottai, and R. Ward. The power of adaptivity in SGD: Self-tuning step sizes with unbounded gradients and affine variance. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pages 313–355, 2022.
- S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- W. Jiang, G. Li, Y. Wang, L. Zhang, and T. Yang. Multi-block-single-probe variance reduced estimator for coupled compositional optimization. In *Advances in Neural Information Processing Systems 35*, pages 32499–32511, 2022a.
- W. Jiang, B. Wang, Y. Wang, L. Zhang, and T. Yang. Optimal algorithms for stochastic multi-level compositional optimization. In *Proceedings of the 39th International Conference on Machine Learning*, pages 10195–10216, 2022b.
- W. Jiang, J. Qin, L. Wu, C. Chen, T. Yang, and L. Zhang. Learning unnormalized statistical models via compositional optimization. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15105–15124, 2023.
- W. Jiang, S. Yang, Y. Wang, and L. Zhang. Adaptive variance reduction for stochastic optimization under weaker assumptions. In *Advances in Neural Information Processing Systems 37*, 2024a.
- W. Jiang, S. Yang, W. Yang, Y. Wang, Y. Wan, and L. Zhang. Projection-free variance reduction methods for stochastic constrained multi-level compositional optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pages 21962–21987, 2024b.
- R. Jin, Y. Huang, X. He, H. Dai, and T. Wu. Stochastic-Sign SGD for federated learning with theoretical guarantees. *ArXiv e-prints*, arXiv:2002.10940, 2023.
- R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pages 315–323, 2013.
- S. P. Karimireddy, Q. Rebjock, S. Stich, and M. Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3252–3261, 2019.
- A. Kavis, S. Skoulakis, K. Antonakopoulos, L. T. Dadi, and V. Cevher. Adaptive stochastic variance reduction for non-convex finite-sum minimization. In *Advances in Neural Information Processing Systems 35*, pages 23524–23538, 2022.
- A. Krizhevsky. Learning multiple layers of features from tiny images. *Masters Thesis, Department of Computer Science, University of Toronto*, 2009.
- L. Lei, C. Ju, J. Chen, and M. I. Jordan. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, volume 30, pages 2345–2355, 2017.
- Z. Li, H. Bao, X. Zhang, and P. Richtarik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6286–6295, 2021.
- L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takac. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2613–2621, 2017.
- Z. Qin, Z. Liu, and P. Xu. Convergence of sign-based random reshuffling algorithms for nonconvex optimization. *ArXiv e-prints*, arXiv:2310.15976, 2023.
- S. J. Reddi, A. Hefny, S. Sra, B. Póczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 314–323, 2016.
- N. L. Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems 25*, pages 2672–2680, 2012.
- M. Safaryan and P. Richtarik. Stochastic sign descent methods: New algorithms and better theory. In *Proceedings of the 38th International Conference on Machine Learning*, pages 9224–9234, 2021.

- F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Conference of the International Speech Communication Association*, pages 1058–1062, 2014.
- S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(16):567–599, 2013.
- T. Sun, Q. Wang, D. Li, and B. Wang. Momentum ensures convergence of SIGNSGD under weaker assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 33077–33099, 2023.
- M. Wang, E. X. Fang, and H. Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2): 419–449, 2017a.
- M. Wang, J. Liu, and E. X. Fang. Accelerating stochastic composition optimization. *Journal of Machine Learning Research*, 18:105:1–105:23, 2017b.
- Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh. SpiderBoost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems 32*, pages 2406–2416, 2019.
- D. Yu, Y. Cai, W. Jiang, and L. Zhang. Efficient algorithms for empirical group distributionally robust optimization and beyond. In *Proceedings of the 41st International Conference on Machine Learning*, pages 57384–57414, 2024.
- H. Yuan, X. Lian, C. J. Li, J. Liu, and W. Hu. Efficient smooth non-convex stochastic compositional optimization via stochastic recursive gradient descent. In *Advances in Neural Information Processing Systems 32*, pages 14905–14916, 2019.
- J. Zhang and L. Xiao. Multilevel composite stochastic optimization via nested variance reduction. *SIAM Journal on Optimization*, 31(2):1131–1157, 2021.
- L. Zhang, M. Mahdavi, and R. Jin. Linear convergence with condition number independent access of full gradients. In *Advances in Neural Information Processing Systems 26*, pages 980–988, 2013.
- D. Zhou, P. Xu, and Q. Gu. Stochastic nested variance reduction for nonconvex optimization. *Journal of Machine Learning Research*, 21(103):1–63, 2020.

A Proof of Theorem 1

Firstly, note that Assumption 1 indicates that the objective function $f(\mathbf{x})$ is also L -smooth [Li et al., 2021]. Given this property, we have that

$$\begin{aligned}
& f(\mathbf{x}_{t+1}) \\
& \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
& \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), -\eta \text{Sign}(\mathbf{v}_t) \rangle + \frac{\eta^2 L}{2} \|\text{Sign}(\mathbf{v}_t)\|^2 \\
& \leq f(\mathbf{x}_t) + \eta \langle \nabla f(\mathbf{x}_t), \text{Sign}(\nabla f(\mathbf{x}_t)) - \text{Sign}(\mathbf{v}_t) \rangle - \eta \langle \nabla f(\mathbf{x}_t), \text{Sign}(\nabla f(\mathbf{x}_t)) \rangle + \frac{\eta^2 Ld}{2} \quad (6) \\
& = f(\mathbf{x}_t) + \eta \langle \nabla f(\mathbf{x}_t), \text{Sign}(\nabla f(\mathbf{x}_t)) - \text{Sign}(\mathbf{v}_t) \rangle - \eta \|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 Ld}{2} \\
& \leq f(\mathbf{x}_t) + 2\eta\sqrt{d} \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| - \eta \|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 Ld}{2},
\end{aligned}$$

where the last inequality is due to the fact that

$$\begin{aligned}
& \langle \nabla f(\mathbf{x}_t), \text{Sign}(\nabla f(\mathbf{x}_t)) - \text{Sign}(\mathbf{v}_t) \rangle \\
& = \sum_{i=1}^d \langle [\nabla f(\mathbf{x}_t)]_i, \text{Sign}([\nabla f(\mathbf{x}_t)]_i) - \text{Sign}([\mathbf{v}_t]_i) \rangle \\
& \leq \sum_{i=1}^d 2 |[\nabla f(\mathbf{x}_t)]_i| \cdot \mathbb{I}(\text{Sign}([\nabla f(\mathbf{x}_t)]_i) \neq \text{Sign}([\mathbf{v}_t]_i)) \\
& \leq \sum_{i=1}^d 2 |[\nabla f(\mathbf{x}_t)]_i - [\mathbf{v}_t]_i| \cdot \mathbb{I}(\text{Sign}([\nabla f(\mathbf{x}_t)]_i) \neq \text{Sign}([\mathbf{v}_t]_i)) \quad (7) \\
& \leq \sum_{i=1}^d 2 |[\nabla f(\mathbf{x}_t)]_i - [\mathbf{v}_t]_i| \\
& = 2 \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|_1 \\
& \leq 2\sqrt{d} \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|.
\end{aligned}$$

Summing up and rearranging the equation (6), we derive:

$$\begin{aligned}
& \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] \\
& \leq \frac{f(\mathbf{x}_1) - f(\mathbf{x}_{T+1})}{\eta T} + 2\sqrt{d} \cdot \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| \right] + \frac{\eta Ld}{2} \quad (8) \\
& \leq \frac{\Delta_f}{\eta T} + 2\sqrt{d} \cdot \sqrt{\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|^2 \right]} + \frac{\eta Ld}{2}
\end{aligned}$$

where we define $\Delta_f = f(\mathbf{x}_1) - f_*$, and the second inequality is due to Jensen's Inequality.

Next, we can bound the term $\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|^2 \right]$ as follows.

$$\begin{aligned}
& \mathbb{E} \left[\|\nabla f(\mathbf{x}_{t+1}) - \mathbf{v}_{t+1}\|^2 \right] \\
&= \mathbb{E} \left[\left\| (1-\beta)\mathbf{v}_t + \frac{1}{B_1} \sum_{k=1}^{B_1} \nabla f(\mathbf{x}_{t+1}; \xi_{t+1}^k) - (1-\beta) \frac{1}{B_1} \sum_{k=1}^{B_1} \nabla f(\mathbf{x}_t; \xi_{t+1}^k) - \nabla f(\mathbf{x}_{t+1}) \right\|^2 \right] \\
&= \mathbb{E} \left[\left\| (1-\beta)(\mathbf{v}_t - \nabla f(\mathbf{x}_t)) + \beta \frac{1}{B_1} \sum_{k=1}^{B_1} (\nabla f(\mathbf{x}_t; \xi_{t+1}^k) - \nabla f(\mathbf{x}_t)) \right. \right. \\
&\quad \left. \left. + \left(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t+1}) + \frac{1}{B_1} \sum_{k=1}^{B_1} \nabla f(\mathbf{x}_{t+1}; \xi_{t+1}^k) - \frac{1}{B_1} \sum_{k=1}^{B_1} \nabla f(\mathbf{x}_t; \xi_{t+1}^k) \right) \right\|^2 \right] \\
&\leq (1-\beta)^2 \mathbb{E} \left[\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|^2 \right] + 2\beta^2 \mathbb{E} \left[\left\| \frac{1}{B_1} \sum_{k=1}^{B_1} \nabla f(\mathbf{x}_t; \xi_{t+1}^k) - \nabla f(\mathbf{x}_t) \right\|^2 \right] \\
&\quad + 2\mathbb{E} \left[\left\| \frac{1}{B_1} \sum_{k=1}^{B_1} (\nabla f(\mathbf{x}_{t+1}; \xi_{t+1}^k) - \nabla f(\mathbf{x}_t; \xi_{t+1}^k)) \right\|^2 \right] \\
&\leq (1-\beta) \mathbb{E} \left[\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|^2 \right] + \frac{2\beta^2 \sigma^2}{B_1} + \frac{2L^2 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2}{B_1} \\
&\leq (1-\beta) \mathbb{E} \left[\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|^2 \right] + \frac{2\beta^2 \sigma^2}{B_1} + \frac{2L^2 \eta^2 d}{B_1},
\end{aligned}$$

where the first inequality is due to the fact $\mathbb{E} \left[\left(\beta \frac{1}{B_1} \sum_{k=1}^{B_1} (\nabla f(\mathbf{x}_t; \xi_{t+1}^k) - \nabla f(\mathbf{x}_t)) + \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t+1}) + \frac{1}{B_1} \sum_{k=1}^{B_1} (\nabla f(\mathbf{x}_{t+1}; \xi_{t+1}^k) - \nabla f(\mathbf{x}_t; \xi_{t+1}^k)) \right) \right] = 0$, and $(a+b)^2 \leq 2a^2 + 2b^2$.

Summing up and noticing that we use a batch size of B_0 in the first iteration, we can ensure

$$\begin{aligned}
\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|^2 \right] &\leq \frac{\mathbb{E} \left[\|\mathbf{v}_1 - \nabla f(\mathbf{x}_1)\|^2 \right]}{\beta T} + \frac{2\sigma^2 \beta}{B_1} + \frac{2L^2 \eta^2 d}{\beta B_1} \\
&\leq \frac{\sigma^2}{B_0 \beta T} + \frac{2\sigma^2 \beta}{B_1} + \frac{2L^2 \eta^2 d}{\beta B_1}
\end{aligned} \tag{9}$$

Incorporating the above into equation (8) and setting that $\beta = \mathcal{O}(T^{-2/3})$, $\eta = \mathcal{O}(d^{-1/2} T^{-2/3})$, $B_0 = \mathcal{O}(T^{1/3})$, $B_1 = \mathcal{O}(1)$, we observe:

$$\begin{aligned}
\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] &\leq \frac{\Delta_f}{\eta T} + 2\sqrt{d} \cdot \sqrt{\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|^2 \right]} + \frac{\eta L d}{2} \\
&\leq \frac{\Delta_f}{\eta T} + 2\sqrt{d} \cdot \sqrt{\frac{\sigma^2}{B_0 \beta T} + \frac{2\sigma^2 \beta}{B_1} + \frac{2L^2 \eta^2 d}{\beta B_1}} + \frac{\eta L d}{2} \\
&= \mathcal{O} \left(\frac{(\Delta_f + \sigma + L) d^{1/2}}{T^{1/3}} \right) \\
&= \mathcal{O} \left(\frac{d^{1/2}}{T^{1/3}} \right),
\end{aligned}$$

which finishes the proof of Theorem 1.

B Proof of Theorem 2

To improve the convergence rate for finite-sum structures, we can reuse the results of equation (8), but bound the term $\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|^2 \right]$ differently. Since $\mathbf{v}_t = (1-\beta)\mathbf{v}_{t-1} + \beta \mathbf{h}_t + (1-$

$\beta)(\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}_{t-1}))$, where $\mathbf{h}_t = \nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}_\tau) + \nabla f(\mathbf{x}_\tau)$, we have:

$$\begin{aligned}
& \mathbb{E} \left[\|\nabla f(\mathbf{x}_{t+1}) - \mathbf{v}_{t+1}\|^2 \right] \\
&= \mathbb{E} \left[\|(1-\beta)\mathbf{v}_t + \beta\mathbf{h}_{t+1} + (1-\beta)(\nabla f_{i_{t+1}}(\mathbf{x}_{t+1}) - \nabla f_{i_{t+1}}(\mathbf{x}_t)) - \nabla f(\mathbf{x}_{t+1})\|^2 \right] \\
&= \mathbb{E} \left[\|(1-\beta)(\mathbf{v}_t - \nabla f(\mathbf{x}_t)) + \beta(\mathbf{h}_{t+1} - \nabla f(\mathbf{x}_{t+1})) \right. \\
&\quad \left. + (1-\beta)(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t+1}) + \nabla f_{i_{t+1}}(\mathbf{x}_{t+1}) - \nabla f_{i_{t+1}}(\mathbf{x}_t))\|^2 \right] \\
&\leq (1-\beta)^2 \mathbb{E} [\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|^2] + 2\beta^2 \mathbb{E} [\|\mathbf{h}_{t+1} - \nabla f(\mathbf{x}_{t+1})\|^2] \\
&\quad + 2(1-\beta)^2 \mathbb{E} [\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t+1}) + \nabla f_{i_{t+1}}(\mathbf{x}_{t+1}) - \nabla f_{i_{t+1}}(\mathbf{x}_t)\|^2] \\
&\leq (1-\beta)^2 \mathbb{E} [\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|^2] + 2\beta^2 \mathbb{E} [\|\mathbf{h}_{t+1} - \nabla f(\mathbf{x}_{t+1})\|^2] \\
&\quad + 2(1-\beta)^2 \mathbb{E} [\|\nabla f_{i_{t+1}}(\mathbf{x}_{t+1}) - \nabla f_{i_{t+1}}(\mathbf{x}_t)\|^2] \\
&\leq (1-\beta) \mathbb{E} [\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|^2] + 2\beta^2 \mathbb{E} [\|\mathbf{h}_{t+1} - \nabla f(\mathbf{x}_{t+1})\|^2] + 2L^2 \mathbb{E} [\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2] \\
&\leq (1-\beta) \mathbb{E} [\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|^2] + 2\beta^2 L^2 \mathbb{E} [\|\mathbf{x}_{t+1} - \mathbf{x}_\tau\|^2] + 2L^2 \mathbb{E} [\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2],
\end{aligned}$$

where the last inequality is due to the fact that:

$$\begin{aligned}
\mathbb{E} [\|\mathbf{h}_{t+1} - \nabla f(\mathbf{x}_{t+1})\|^2] &= \mathbb{E} [\|\nabla f_{i_{t+1}}(\mathbf{x}_{t+1}) - \nabla f_{i_{t+1}}(\mathbf{x}_\tau) + \nabla f(\mathbf{x}_\tau) - \nabla f(\mathbf{x}_{t+1})\|^2] \\
&\leq \mathbb{E} [\|\nabla f_{i_{t+1}}(\mathbf{x}_{t+1}) - \nabla f_{i_{t+1}}(\mathbf{x}_\tau)\|^2] \\
&\leq L^2 \mathbb{E} [\|\mathbf{x}_{t+1} - \mathbf{x}_\tau\|^2].
\end{aligned}$$

By rearranging and summing up, we establish:

$$\begin{aligned}
& \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|^2 \right] \\
&\leq \frac{\mathbb{E} [\|\mathbf{v}_1 - \nabla f(\mathbf{x}_1)\|^2]}{\beta T} + 2\beta L^2 \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\mathbf{x}_{t+1} - \mathbf{x}_\tau\|^2 \right] + \frac{2L^2}{\beta} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \right] \\
&\leq 2\beta I^2 L^2 \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \right] + \frac{2L^2}{\beta} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \right] \\
&\leq 2L^2 \left(\beta I^2 + \frac{1}{\beta} \right) \eta^2 d,
\end{aligned}$$

where we use full batch in the first iteration, and the second inequality is due to the fact that

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^T \|\mathbf{x}_{t+1} - \mathbf{x}_\tau\|^2 &= \frac{1}{T} \sum_{t=1}^T \left\| \sum_{i=\tau}^t (\mathbf{x}_{i+1} - \mathbf{x}_i) \right\|^2 \leq \frac{1}{T} \sum_{t=1}^T I \sum_{i=\tau}^t \|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2 \\
&\leq \frac{I^2}{T} \sum_{t=1}^T \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2.
\end{aligned}$$

Incorporate the above into equation (8) and setting $I = m, \beta = \mathcal{O}\left(\frac{1}{m}\right)$, and $\eta = \mathcal{O}\left(\frac{1}{m^{1/4}d^{1/2}T^{1/2}}\right)$, we refine the bound as:

$$\begin{aligned}
\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] &\leq \frac{\Delta_f}{\eta T} + 2\sqrt{d} \cdot \sqrt{\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\|^2 \right]} + \frac{\eta L d}{2} \\
&\leq \frac{\Delta_f}{\eta T} + 2\sqrt{d} \cdot \sqrt{2L^2 \left(\beta I^2 + \frac{1}{\beta} \right) \eta^2 d + \frac{\eta L d}{2}} \\
&= \mathcal{O} \left(\frac{(\Delta_f + L) m^{1/4} d^{1/2}}{T^{1/2}} \right) \\
&= \mathcal{O} \left(\frac{m^{1/4} d^{1/2}}{T^{1/2}} \right).
\end{aligned}$$

C Proof of Theorem 3

By setting $\beta = \frac{1}{2}$, we can ensure that $\|\mathbf{v}_t^j\|_\infty \leq R = 4G$, since

$$\begin{aligned} \|\mathbf{v}_t^j\|_\infty &= \left\| (1-\beta)\mathbf{v}_{t-1}^j + \nabla f(\mathbf{x}_t; \xi_t^j) - (1-\beta)f(\mathbf{x}_{t-1}; \xi_t^j) \right\|_\infty \\ &\leq (1-\beta) \|\mathbf{v}_{t-1}^j\|_\infty + (2-\beta)G \\ &\leq (1-\beta)^{t-1} \|\mathbf{v}_1^j\|_\infty + (2-\beta)G \sum_{s=1}^t (1-\beta)^{t-s} \\ &\leq G + \frac{(2-\beta)G}{\beta} \leq 4G = R. \end{aligned}$$

Since the overall objective function $f(\mathbf{x})$ is L -smooth, we have the following:

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &\leq f(\mathbf{x}_t) - \eta \left\langle \nabla f(\mathbf{x}_t), \text{Sign} \left(\frac{1}{n} \sum_{j=1}^n \mathbf{S}_R(\mathbf{v}_t^j) \right) \right\rangle + \frac{\eta^2 L d}{2} \\ &\quad - \eta \langle \nabla f(\mathbf{x}_t), \text{Sign}(\nabla f(\mathbf{x}_t)) \rangle + \frac{\eta^2 L d}{2} \\ &= f(\mathbf{x}_t) + \eta \left\langle \nabla f(\mathbf{x}_t), \text{Sign}(\nabla f(\mathbf{x}_t)) - \text{Sign} \left(\frac{1}{n} \sum_{j=1}^n \mathbf{S}_R(\mathbf{v}_t^j) \right) \right\rangle - \eta \|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 L d}{2} \\ &\leq f(\mathbf{x}_t) + 2\eta R \sqrt{d} \left\| \frac{\nabla f(\mathbf{x}_t)}{R} - \frac{1}{n} \sum_{j=1}^n \mathbf{S}_R(\mathbf{v}_t^j) \right\| - \eta \|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 L d}{2}, \end{aligned} \tag{10}$$

where the last inequality is because of

$$\begin{aligned} &\left\langle \nabla f(\mathbf{x}_t), \text{Sign}(\nabla f(\mathbf{x}_t)) - \text{Sign} \left(\frac{1}{n} \sum_{j=1}^n \mathbf{S}_R(\mathbf{v}_t^j) \right) \right\rangle \\ &= \sum_{i=1}^d \left\langle [\nabla f(\mathbf{x}_t)]_i, \text{Sign}([\nabla f(\mathbf{x}_t)]_i) - \text{Sign} \left(\left[\frac{1}{n} \sum_{j=1}^n \mathbf{S}_R(\mathbf{v}_t^j) \right]_i \right) \right\rangle \\ &\leq \sum_{i=1}^d 2R |[\nabla f(\mathbf{x}_t)]_i / R| \cdot \mathbb{I} \left(\text{Sign}([\nabla f(\mathbf{x}_t)]_i) \neq \text{Sign} \left(\left[\frac{1}{n} \sum_{j=1}^n \mathbf{S}_R(\mathbf{v}_t^j) \right]_i \right) \right) \\ &\leq \sum_{i=1}^d 2R \left| \frac{[\nabla f(\mathbf{x}_t)]_i}{R} - \left[\frac{1}{n} \sum_{j=1}^n \mathbf{S}_R(\mathbf{v}_t^j) \right]_i \right| \cdot \mathbb{I} \left(\text{Sign}([\nabla f(\mathbf{x}_t)]_i) \neq \text{Sign} \left(\left[\frac{1}{n} \sum_{j=1}^n \mathbf{S}_R(\mathbf{v}_t^j) \right]_i \right) \right) \\ &\leq \sum_{i=1}^d 2R \left| \frac{[\nabla f(\mathbf{x}_t)]_i}{R} - \left[\frac{1}{n} \sum_{j=1}^n \mathbf{S}_R(\mathbf{v}_t^j) \right]_i \right| \\ &= 2R \left\| \frac{\nabla f(\mathbf{x}_t)}{R} - \frac{1}{n} \sum_{j=1}^n \mathbf{S}_R(\mathbf{v}_t^j) \right\|_1 \leq 2R \sqrt{d} \left\| \frac{\nabla f(\mathbf{x}_t)}{R} - \frac{1}{n} \sum_{j=1}^n \mathbf{S}_R(\mathbf{v}_t^j) \right\|. \end{aligned} \tag{11}$$

Rearranging and taking the expectation over equation (10), we have:

$$\begin{aligned}
& \mathbb{E} [f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t)] \\
& \leq 2\eta R\sqrt{d}\mathbb{E} \left[\left\| \frac{\nabla f(\mathbf{x}_t)}{R} - \frac{1}{n} \sum_{j=1}^n \mathbf{S}_R(\mathbf{v}_t^j) \right\| \right] - \eta \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|_1] + \frac{\eta^2 Ld}{2} \\
& \leq 2\eta R\sqrt{d}\mathbb{E} \left[\left\| \frac{\nabla f(\mathbf{x}_t)}{R} - \frac{1}{nR} \sum_{j=1}^n \mathbf{v}_t^j \right\| \right] + 2\eta R\sqrt{d}\mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \left(\mathbf{S}_R(\mathbf{v}_t^j) - \frac{\mathbf{v}_t^j}{R} \right) \right\| \right] \\
& \quad - \eta \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|_1] + \frac{\eta^2 Ld}{2} \\
& \leq 2\eta\sqrt{d}\mathbb{E} \left[\left\| \nabla f(\mathbf{x}_t) - \frac{1}{n} \sum_{j=1}^n \mathbf{v}_t^j \right\| \right] + 2\eta R\sqrt{d} \sqrt{\mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \left(\mathbf{S}_R(\mathbf{v}_t^j) - \frac{\mathbf{v}_t^j}{R} \right) \right\|^2 \right]} \\
& \quad - \eta \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|_1] + \frac{\eta^2 Ld}{2} \tag{12} \\
& \leq 2\eta\sqrt{d}\mathbb{E} \left[\left\| \nabla f(\mathbf{x}_t) - \frac{1}{n} \sum_{j=1}^n \mathbf{v}_t^j \right\| \right] + 2\eta R\sqrt{d} \sqrt{\frac{1}{n^2} \sum_{j=1}^n \mathbb{E} \left[\left\| \left(\mathbf{S}_R(\mathbf{v}_t^j) - \frac{\mathbf{v}_t^j}{R} \right) \right\|^2 \right]} \\
& \quad - \eta \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|_1] + \frac{\eta^2 Ld}{2} \\
& \leq 2\eta\sqrt{d}\mathbb{E} \left[\left\| \nabla f(\mathbf{x}_t) - \frac{1}{n} \sum_{j=1}^n \mathbf{v}_t^j \right\| \right] + 2\eta R\sqrt{d} \sqrt{\frac{1}{n^2} \sum_{j=1}^n \mathbb{E} \left[\left\| \mathbf{S}_R(\mathbf{v}_t^j) \right\|^2 \right]} \\
& \quad - \eta \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|_1] + \frac{\eta^2 Ld}{2} \\
& \leq 2\eta\sqrt{d}\mathbb{E} \left[\left\| \nabla f(\mathbf{x}_t) - \frac{1}{n} \sum_{j=1}^n \mathbf{v}_t^j \right\| \right] + \frac{2\eta dR}{\sqrt{n}} - \eta \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|_1] + \frac{\eta^2 Ld}{2},
\end{aligned}$$

where the third inequality is due to the fact that $(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$, and the fourth inequality is because of $\mathbb{E} \left[\mathbf{S} \left(\mathbf{v}_t^j \right) \right] = \frac{\mathbf{v}_t^j}{R}$, as well as the \mathbf{S} operation in each node is independent.

Rearranging the terms and summing up, we have:

$$\begin{aligned}
\frac{1}{T} \sum_{i=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|_1] & \leq \frac{\Delta_f}{\eta T} + 2\sqrt{d}\mathbb{E} \left[\frac{1}{T} \sum_{i=1}^T \left\| \nabla f(\mathbf{x}_t) - \frac{1}{n} \sum_{j=1}^n \mathbf{v}_t^j \right\| \right] + \frac{2dR}{\sqrt{n}} + \frac{\eta Ld}{2} \\
& \leq \frac{\Delta_f}{\eta T} + 2\sqrt{d} \sqrt{\mathbb{E} \left[\frac{1}{T} \sum_{i=1}^T \left\| \nabla f(\mathbf{x}_t) - \frac{1}{n} \sum_{j=1}^n \mathbf{v}_t^j \right\|^2 \right]} + \frac{2dR}{\sqrt{n}} + \frac{\eta Ld}{2},
\end{aligned}$$

where the last inequality is due to Jensen's inequality.

For each worker j , we have the following according to the definition of \mathbf{v}_t^j :

$$\begin{aligned}
\mathbf{v}_{t+1}^j - \nabla f_j(\mathbf{x}_{t+1}) & = (1 - \beta) \left(\mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t) \right) + \beta \left(\nabla f_j(\mathbf{x}_{t+1}; \xi_{t+1}^j) - \nabla f_j(\mathbf{x}_{t+1}) \right) \\
& \quad + (1 - \beta) \left(\nabla f_j(\mathbf{x}_{t+1}; \xi_{t+1}^j) - \nabla f_j(\mathbf{x}_t; \xi_{t+1}^j) + \nabla f_j(\mathbf{x}_t) - \nabla f_j(\mathbf{x}_{t+1}) \right).
\end{aligned}$$

Summing over $\{n\}$ and noting that $\nabla f(\mathbf{x}) = \frac{1}{n} \sum_{j=1}^n \nabla f_j(\mathbf{x})$, we can obtain:

$$\begin{aligned} & \frac{1}{n} \sum_{j=1}^n \mathbf{v}_{t+1}^j - \nabla f(\mathbf{x}_{t+1}) = \frac{1}{n} \sum_{j=1}^n \left(\mathbf{v}_{t+1}^j - \nabla f_j(\mathbf{x}_{t+1}) \right) \\ & = (1-\beta) \frac{1}{n} \sum_{j=1}^n \left(\mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t) \right) + \beta \frac{1}{n} \sum_{j=1}^n \left(\nabla f_j(\mathbf{x}_{t+1}; \xi_{t+1}^j) - \nabla f_j(\mathbf{x}_{t+1}) \right) \\ & \quad + (1-\beta) \frac{1}{n} \sum_{j=1}^n \left(\nabla f_j(\mathbf{x}_{t+1}; \xi_{t+1}^j) - \nabla f_j(\mathbf{x}_t; \xi_{t+1}^j) + \nabla f_j(\mathbf{x}_t) - \nabla f_j(\mathbf{x}_{t+1}) \right). \end{aligned}$$

Then we have

$$\begin{aligned} & \mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \mathbf{v}_{t+1}^j - \nabla f(\mathbf{x}_{t+1}) \right\|^2 \right] \\ & \leq (1-\beta)^2 \mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \left(\mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t) \right) \right\|^2 \right] + 2\beta^2 \frac{1}{n^2} \sum_{j=1}^n \mathbb{E} \left[\left\| \nabla f_j(\mathbf{x}_{t+1}; \xi_{t+1}^j) - \nabla f_j(\mathbf{x}_{t+1}) \right\|^2 \right] \\ & \quad + 2(1-\beta)^2 \frac{1}{n^2} \sum_{j=1}^n \mathbb{E} \left[\left\| \nabla f_j(\mathbf{x}_{t+1}; \xi_{t+1}^j) - \nabla f_j(\mathbf{x}_t; \xi_{t+1}^j) \right\|^2 \right] \\ & \leq (1-\beta) \mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \left(\mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t) \right) \right\|^2 \right] + \frac{2\beta^2 \sigma^2}{n} + \frac{2L^2}{n} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ & \leq (1-\beta) \mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \mathbf{v}_t^j - \nabla f(\mathbf{x}_t) \right\|^2 \right] + \frac{2\beta^2 \sigma^2}{n} + \frac{2L^2 \eta^2 d}{n}. \end{aligned}$$

By summing up and rearranging, we observe

$$\begin{aligned} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \left\| \frac{1}{n} \sum_{j=1}^n \mathbf{v}_t^j - \nabla f(\mathbf{x}_t) \right\|^2 \right] & \leq \frac{\mathbb{E} \left[\left\| \frac{1}{n} \sum_{j=1}^n \mathbf{v}_1^j - \nabla f(\mathbf{x}_1) \right\|^2 \right]}{\beta T} + \frac{2\sigma^2 \beta}{n} + \frac{2L^2 \eta^2 d}{n\beta} \\ & \leq \frac{\sigma^2}{n\beta T} + \frac{2\sigma^2 \beta}{n} + \frac{2L^2 \eta^2 d}{n\beta}. \end{aligned} \quad (13)$$

Finally, by setting $\beta = \frac{1}{2}$ and $\eta = \mathcal{O}(T^{-1/2} d^{-1/2})$, we ensure that

$$\begin{aligned} \frac{1}{T} \sum_{i=1}^T \|\nabla f(\mathbf{x}_t)\|_1 & \leq \frac{\Delta_f}{\eta T} + \frac{2dR}{\sqrt{n}} + \frac{\eta Ld}{2} + 2\sqrt{d} \sqrt{\mathbb{E} \left[\frac{1}{T} \sum_{i=1}^T \left\| \nabla f(\mathbf{x}_t) - \frac{1}{n} \sum_{j=1}^n \mathbf{v}_t^j \right\|^2 \right]} \\ & \leq \frac{\Delta_f}{\eta T} + \frac{2dR}{\sqrt{n}} + \frac{\eta Ld}{2} + 2\sqrt{d} \sqrt{\frac{\sigma^2}{n\beta T} + \frac{2\sigma^2 \beta}{n} + \frac{2L^2 \eta^2 d}{n\beta}} \\ & = \mathcal{O} \left(\frac{d^{1/2}}{T^{1/2}} + \frac{d}{n^{1/2}} \right). \end{aligned}$$

D Proof of Theorem 4

Due to the fact that the overall objective function $f(\mathbf{x})$ is L -smooth, we have the following:

$$\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&\leq f(\mathbf{x}_t) - \eta \left\langle \nabla f(\mathbf{x}_t), S_1 \left(\frac{1}{n} \sum_{j=1}^n S_G(\hat{\mathbf{v}}_t^j) \right) \right\rangle + \frac{\eta^2 Ld}{2} \\
&= f(\mathbf{x}_t) + \eta \left\langle \nabla f(\mathbf{x}_t), \nabla f(\mathbf{x}_t) - S_1 \left(\frac{1}{n} \sum_{j=1}^n S_G(\hat{\mathbf{v}}_t^j) \right) \right\rangle - \eta \|\nabla f(\mathbf{x}_t)\|^2 + \frac{\eta^2 Ld}{2}.
\end{aligned}$$

Taking expectations leads to:

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t)] \\
&\leq \eta \mathbb{E} \left[\left\langle \nabla f(\mathbf{x}_t), \frac{1}{G} \nabla f(\mathbf{x}_t) - S_1 \left(\frac{1}{n} \sum_{j=1}^n S_G(\hat{\mathbf{v}}_t^j) \right) \right\rangle \right] - \frac{\eta}{G} \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|^2] + \frac{\eta^2 Ld}{2} \\
&\leq \eta \mathbb{E} \left[\left\langle \nabla f(\mathbf{x}_t), \frac{1}{G} \nabla f(\mathbf{x}_t) - \frac{1}{n} \sum_{j=1}^n S_G(\hat{\mathbf{v}}_t^j) \right\rangle \right] - \frac{\eta}{G} \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|^2] + \frac{\eta^2 Ld}{2} \\
&\leq \eta \mathbb{E} \left[\left\langle \nabla f(\mathbf{x}_t), \frac{1}{G} \nabla f(\mathbf{x}_t) - \frac{1}{nG} \sum_{j=1}^n \hat{\mathbf{v}}_t^j \right\rangle \right] - \frac{\eta}{G} \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|^2] + \frac{\eta^2 Ld}{2} \tag{14} \\
&\leq \eta \mathbb{E} \left[\frac{1}{2G} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2G} \left\| \nabla f(\mathbf{x}_t) - \frac{1}{n} \sum_{j=1}^n \hat{\mathbf{v}}_t^j \right\|^2 \right] - \frac{\eta}{G} \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|^2] + \frac{\eta^2 Ld}{2} \\
&\leq \frac{\eta}{2G} \mathbb{E} \left[\left\| \nabla f(\mathbf{x}_t) - \frac{1}{n} \sum_{j=1}^n \hat{\mathbf{v}}_t^j \right\|^2 \right] - \frac{\eta}{2G} \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|^2] + \frac{\eta^2 Ld}{2}
\end{aligned}$$

Rearranging the terms and summing up:

$$\begin{aligned}
\frac{1}{T} \sum_{i=1}^T \mathbb{E} \|\nabla f(\mathbf{x}_t)\|^2 &\leq \frac{2\Delta_f G}{\eta T} + \mathbb{E} \left[\frac{1}{T} \sum_{i=1}^T \left\| \nabla f(\mathbf{x}_t) - \frac{1}{n} \sum_{j=1}^n \hat{\mathbf{v}}_t^j \right\|^2 \right] + \eta LdG \\
&\leq \frac{2\Delta_f G}{\eta T} + \mathbb{E} \left[\frac{1}{n} \sum_{j=1}^n \frac{1}{T} \sum_{i=1}^T \left\| \nabla f_j(\mathbf{x}_t) - \hat{\mathbf{v}}_t^j \right\|^2 \right] + \eta LdG \\
&\leq \frac{2\Delta_f G}{\eta T} + \mathbb{E} \left[\frac{1}{n} \sum_{j=1}^n \frac{1}{T} \sum_{i=1}^T \left\| \nabla f_j(\mathbf{x}_t) - \Pi_G [\mathbf{v}_t^j] \right\|^2 \right] + \eta LdG \\
&\leq \frac{2\Delta_f G}{\eta T} + \mathbb{E} \left[\frac{1}{n} \sum_{j=1}^n \frac{1}{T} \sum_{i=1}^T \left\| \nabla f_j(\mathbf{x}_t) - \mathbf{v}_t^j \right\|^2 \right] + \eta LdG.
\end{aligned}$$

where the last inequality is due to the non-expansive property of the projection operation.

For each worker j , according to the definition of \mathbf{v}_t^j , we have:

$$\begin{aligned}
\mathbf{v}_{t+1}^j - \nabla f_j(\mathbf{x}_{t+1}) &= (1 - \beta) \left(\mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t) \right) + \beta \left(\nabla f_j(\mathbf{x}_{t+1}; \xi_{t+1}^j) - \nabla f_j(\mathbf{x}_{t+1}) \right) \\
&\quad + (1 - \beta) \left(\nabla f_j(\mathbf{x}_{t+1}; \xi_{t+1}^j) - \nabla f_j(\mathbf{x}_t; \xi_{t+1}^j) + \nabla f_j(\mathbf{x}_t) - \nabla f_j(\mathbf{x}_{t+1}) \right).
\end{aligned}$$

Then we have

$$\begin{aligned}
& \mathbb{E} \left[\left\| \mathbf{v}_{t+1}^j - \nabla f_j(\mathbf{x}_{t+1}) \right\|^2 \right] \\
& \leq (1 - \beta)^2 \mathbb{E} \left[\left\| \mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t) \right\|^2 \right] + 2\beta^2 \mathbb{E} \left[\left\| \left(\nabla f_j(\mathbf{x}_{t+1}; \xi_{t+1}^j) - \nabla f_j(\mathbf{x}_{t+1}) \right) \right\|^2 \right] \\
& \quad + 2(1 - \beta)^2 \mathbb{E} \left[\left\| \left(\nabla f_j(\mathbf{x}_{t+1}; \xi_{t+1}^j) - \nabla f_j(\mathbf{x}_t; \xi_{t+1}^j) \right) \right\|^2 \right] \\
& \leq (1 - \beta) \mathbb{E} \left[\left\| \mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t) \right\|^2 \right] + 2\beta^2 \sigma^2 + 2L^2 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
& \leq (1 - \beta) \mathbb{E} \left[\left\| \mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t) \right\|^2 \right] + 2\beta^2 \sigma^2 + 2L^2 \eta^2 d.
\end{aligned}$$

As a result, we know that

$$\mathbb{E} \left[\frac{1}{n} \sum_{j=1}^n \frac{1}{T} \sum_{t=1}^T \left\| \mathbf{v}_t^j - \nabla f_j(\mathbf{x}_t) \right\|^2 \right] \leq \frac{\sigma^2}{\beta T} + 2\sigma^2 \beta + \frac{2L^2 \eta^2 d}{\beta}.$$

Finally, combining the above and setting that $\beta = \mathcal{O}\left(\frac{1}{T^{1/2}}\right)$, $\eta = \mathcal{O}\left(\frac{1}{d^{1/2}T^{1/2}}\right)$, we obtain the final bound:

$$\begin{aligned}
\mathbb{E} \left[\frac{1}{T} \sum_{i=1}^T \|\nabla f(\mathbf{x}_t)\| \right] & \leq \sqrt{\mathbb{E} \left[\frac{1}{T} \sum_{i=1}^T \|\nabla f(\mathbf{x}_t)\|^2 \right]} \\
& \leq \sqrt{\frac{2\Delta_f G}{\eta T} + \eta L d G + \frac{\sigma^2}{\beta T} + 2\sigma^2 \beta + \frac{2L^2 \eta^2 d}{\beta}} \\
& \leq \mathcal{O} \left(\sqrt{\frac{d^{1/2} (\Delta_f + L) + \sigma^2 + L^2}{T^{1/2}}} \right) \\
& = \mathcal{O} \left(\frac{d^{1/4}}{T^{1/4}} \right).
\end{aligned}$$

E Results under weaker assumptions

In this section, we demonstrate that our proposed methods can maintain similar convergence rates under less stringent assumptions — expected α -symmetric generalized-smoothness [Chen et al., 2023] and affine variance [Faw et al., 2022]. We first detail these relaxed assumptions below.

Assumption 1' (*Expected α -symmetric generalized-smoothness*)

$$\mathbb{E}_\xi \left[\|\nabla f(\mathbf{x}; \xi) - \nabla f(\mathbf{y}; \xi)\|^2 \right] \leq \|\mathbf{x} - \mathbf{y}\|^2 \mathbb{E}_\xi \left[\left(L_0 + L_1 \max_{\theta \in [0,1]} \|\nabla f(\mathbf{x}_\theta; \xi)\|^\alpha \right)^2 \right],$$

where $\mathbf{x}_\theta := \theta \mathbf{x} + (1 - \theta) \mathbf{y}$, and $0 \leq \alpha \leq 1$.

Assumption 2' (*Affine variance*)

$$\mathbb{E}_\xi \left[\|\nabla f(\mathbf{x}; \xi) - \nabla f(\mathbf{x})\|^2 \right] \leq \Gamma^2 \|\nabla f(\mathbf{x})\|^2 + \Lambda^2.$$

Remark: Assumption 1' can be reduced to standard average smoothness (Assumption 1) when $L_1 = 0$. Note that α -symmetric generalized-smooth functions not only include asymmetric and Hessian-based generalized-smooth functions, but also contain high-order polynomials and exponential functions [Chen et al., 2023]. Moreover, affine variance is also weaker than Assumption 2 and can be reduced to it when $\Gamma = 0$.

We then demonstrate that these relaxed conditions are sufficient for our algorithms to achieve the same convergence rate.

Theorem 5 Under Assumptions 1' and 2', by setting that $\beta = \mathcal{O}(\frac{d^{1/3}}{T^{2/3}})$, $\eta = \mathcal{O}(\frac{1}{d^{1/6}T^{2/3}})$, $B_0 = \mathcal{O}(1)$, $B_1 = \mathcal{O}(d)$, and denoting N as the samples used, our SSVR method guarantees:

$$\mathbb{E} [\|\nabla f(\mathbf{x}_\tau)\|_1] \leq \mathcal{O}\left(\frac{d^{1/2}}{N^{1/3}}\right).$$

Furthermore, we introduce the following relaxed assumption for the finite-sum problem.

Assumption 3' (Generalized smoothness) For each $i \in \{1, 2, \dots, m\}$, we have

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\| \left(L_0 + L_1 \max_{\theta \in [0,1]} \|\nabla f(\mathbf{x}_\theta)\|^\alpha \right),$$

where $\mathbf{x}_\theta := \theta\mathbf{x} + (1 - \theta)\mathbf{y}$, and $0 \leq \alpha \leq 1$.

This assumption is weaker than the standard Assumption 3. We validate that our SSVR-FS algorithm can still achieve similar convergence under this relaxed condition.

Theorem 6 Under Assumption 3', by setting $\eta = \mathcal{O}(\min\{\frac{1}{m^{1/4}d^{1/2}T^{1/2}}, \frac{1}{md}\})$, $\beta = \mathcal{O}(\frac{1}{m})$, and $I = m$, our SSVR-FS algorithm ensures:

$$\mathbb{E} [\|\nabla F(\mathbf{x}_\varphi)\|_1] \leq \mathcal{O}\left(\frac{m^{1/4}d^{1/2}}{T^{1/2}} + \frac{md}{T}\right).$$

Remark: When the iteration number T is large, the dominant term becomes $\mathcal{O}(m^{1/4}d^{1/2}T^{-1/2})$, which aligns with the results in Theorem 2.

E.1 Proof of Theorem 5

We first present some useful tools for analysis. According to Proposition 4 in Chen et al. [2023], Assumption 1' leads to the following lemmas.

Lemma 1 For $\alpha \in (0, 1)$, we have:

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle \\ &\quad + \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 (K_0 + K_1 \|\nabla f(\mathbf{x}_t)\|^\alpha + 2K_2 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^{1-\alpha}), \end{aligned}$$

where $K_0 := L_0(2^{\frac{\alpha}{1-\alpha}} + 1)$, $K_1 := L_1 \cdot 2^{\frac{\alpha}{1-\alpha}} \cdot 3^\alpha$, $K_2 := L_1^{\frac{1}{1-\alpha}} \cdot 2^{\frac{\alpha}{1-\alpha}} \cdot 3^\alpha (1 - \alpha)^{\frac{\alpha}{1-\alpha}}$.

For $\alpha = 1$, we also have:

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle \\ &\quad + \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 (L_0 + L_1 \|\nabla f(\mathbf{x}_t)\|) \exp(L_1 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|), \end{aligned}$$

Similarly, according to the Proposition 4 in Chen et al. [2023], we have the following guarantees.

Lemma 2 For $\alpha \in (0, 1)$, we have:

$$\mathbb{E}_\xi \|\nabla f(\mathbf{x}; \xi) - \nabla f(\mathbf{y}; \xi)\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2 (\bar{K}_0 + \bar{K}_1 \mathbb{E}_\xi \|\nabla f(\mathbf{y}; \xi)\|^\alpha + \bar{K}_2 \|\mathbf{x} - \mathbf{y}\|^{1-\alpha})^2.$$

where $\bar{K}_0 = 2^{\frac{2-\alpha}{1-\alpha}} L_0$, $\bar{K}_1 = 2^{\frac{2-\alpha}{1-\alpha}} L_1$, $\bar{K}_2 = (5L_1)^{\frac{1}{1-\alpha}}$.

For $\alpha = 1$, we also have:

$$\mathbb{E}_\xi \|\nabla f(\mathbf{x}; \xi) - \nabla f(\mathbf{y}; \xi)\|^2 \leq 2\|\mathbf{x} - \mathbf{y}\|^2 (L_0^2 + 2L_1^2 \mathbb{E}_\xi \|\nabla f(\mathbf{y}; \xi)\|^2) \exp(12L_1^2 \|\mathbf{x} - \mathbf{y}\|^2).$$

Then, we can begin our proof. For $\alpha \in (0, 1)$, according to Lemma 1, by setting $\eta \leq d^{-\frac{1}{2}}$, we have:

$$\begin{aligned}
& f(\mathbf{x}_{t+1}) \\
& \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle \\
& \quad + \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 (K_0 + K_1 \|\nabla f(\mathbf{x}_t)\|^\alpha + 2K_2 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^{1-\alpha}) \\
& \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), -\eta \text{Sign}(\mathbf{v}_t) \rangle \\
& \quad + \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 (K_0 + K_1 (1 + \|\nabla f(\mathbf{x}_t)\|) + 2K_2) \\
& \leq f(\mathbf{x}_t) + \eta \langle \nabla f(\mathbf{x}_t), \text{Sign}(\nabla f(\mathbf{x}_t)) - \text{Sign}(\mathbf{v}_t) \rangle - \eta \langle \nabla f(\mathbf{x}_t), \text{Sign}(\nabla f(\mathbf{x}_t)) \rangle \\
& \quad + \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 (K_0 + K_1 (1 + \|\nabla f(\mathbf{x}_t)\|) + 2K_2) \\
& = f(\mathbf{x}_t) + \eta \langle \nabla f(\mathbf{x}_t), \text{Sign}(\nabla f(\mathbf{x}_t)) - \text{Sign}(\mathbf{v}_t) \rangle - \eta \|\nabla f(\mathbf{x}_t)\|_1 \\
& \quad + \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 ((K_0 + K_1 + 2K_2) + K_1 \|\nabla f(\mathbf{x}_t)\|) \\
& \leq f(\mathbf{x}_t) + 2\eta\sqrt{d} \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| - \eta \|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 d}{2} ((K_0 + K_1 + 2K_2) + K_1 \|\nabla f(\mathbf{x}_t)\|),
\end{aligned}$$

where the second inequality is due to the fact that $\alpha < 1$ and $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \leq \eta^2 d \leq 1$.

Rearranging and summing up, we then have

$$\begin{aligned}
\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] & \leq \frac{\Delta_f}{\eta T} + 2\sqrt{d} \cdot \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| \right] \\
& \quad + \frac{\eta d (K_0 + K_1 + 2K_2)}{2} + \frac{\eta d K_1}{2T} \mathbb{E} \left[\sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\| \right].
\end{aligned}$$

By setting $\eta \leq \min\{\frac{1}{\sqrt{d}}, \frac{1}{dK_1}\}$, we can get

$$\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] \leq \frac{2\Delta_f}{\eta T} + 4\sqrt{d} \cdot \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| \right] + \eta d (K_0 + K_1 + 2K_2). \quad (15)$$

For $\alpha = 1$, according to Lemma 1 and setting $\eta \leq \frac{1}{L_1 \sqrt{d}}$, we have

$$\begin{aligned}
& f(\mathbf{x}_{t+1}) \\
& \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 (L_0 + L_1 \|\nabla f(\mathbf{x}_t)\|) \exp(L_1 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|) \\
& \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{\exp(1)}{2} \eta^2 d (L_0 + L_1 \|\nabla f(\mathbf{x}_t)\|) \\
& \leq f(\mathbf{x}_t) + 2\eta\sqrt{d} \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| - \eta \|\nabla f(\mathbf{x}_t)\|_1 + \frac{3\eta^2 d}{2} (L_0 + L_1 \|\nabla f(\mathbf{x}_t)\|),
\end{aligned}$$

where the second inequality is due to $L_1 \|\mathbf{x}_{t+1} - \mathbf{x}_t\| \leq 1$, and others follow the previous proof.

Rearranging and summing up, we then have

$$\begin{aligned}
& \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] \\
& \leq \frac{\Delta_f}{\eta T} + 2\sqrt{d} \cdot \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| \right] + \frac{3\eta d L_0}{2} + \frac{3\eta d L_1}{2T} \mathbb{E} \left[\sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\| \right].
\end{aligned}$$

By setting $\eta \leq \frac{1}{3dL_1}$, we can get

$$\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] \leq \frac{2\Delta_f}{\eta T} + 4\sqrt{d} \cdot \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| \right] + 3\eta d L_0. \quad (16)$$

Then we begin to bound the term $\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| \right]$. According to the definition of \mathbf{v}_t , we have:

$$\begin{aligned} \mathbf{v}_t - \nabla f(\mathbf{x}_t) &= (1 - \beta)(\mathbf{v}_{t-1} - \nabla f(\mathbf{x}_{t-1})) + \beta \left(\frac{1}{B_1} \sum_{k=1}^{B_1} \nabla f(\mathbf{x}_t; \xi_t^k) - \nabla f(\mathbf{x}_t) \right) \\ &\quad + (1 - \beta) \left(\frac{1}{B_1} \sum_{k=1}^{B_1} (\nabla f(\mathbf{x}_t; \xi_t^k) - \nabla f(\mathbf{x}_{t-1}; \xi_t^k)) + \nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_t) \right). \end{aligned}$$

Denote that $G_t = \left(\frac{1}{B_1} \sum_{k=1}^{B_1} (\nabla f(\mathbf{x}_t; \xi_t^k) - \nabla f(\mathbf{x}_{t-1}; \xi_t^k)) + \nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_t) \right)$ and $\Delta_t = \frac{1}{B_1} \sum_{k=1}^{B_1} (\nabla f(\mathbf{x}_t; \xi_t^k) - \nabla f(\mathbf{x}_t))$. By summing up, we have

$$\begin{aligned} &\mathbf{v}_t - \nabla f(\mathbf{x}_t) \\ &= (1 - \beta)(\mathbf{v}_{t-1} - \nabla f(\mathbf{x}_{t-1})) + \beta \Delta_t + (1 - \beta) G_t \\ &= \dots \\ &= (1 - \beta)^{t-1} (\mathbf{v}_1 - \nabla f(\mathbf{x}_1)) + \beta \sum_{s=1}^t (1 - \beta)^{t-s} \Delta_s + (1 - \beta) \sum_{s=1}^t (1 - \beta)^{t-s} G_s. \end{aligned}$$

Thus, we can know that

$$\begin{aligned} \mathbb{E} [\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|] &\leq (1 - \beta)^{t-1} \mathbb{E} [\|\mathbf{v}_1 - \nabla f(\mathbf{x}_1)\|] \\ &\quad + \beta \mathbb{E} \left[\left\| \sum_{s=1}^t (1 - \beta)^{t-s} \Delta_s \right\| \right] + (1 - \beta) \mathbb{E} \left[\left\| \sum_{s=1}^t (1 - \beta)^{t-s} G_s \right\| \right]. \end{aligned}$$

Then we give the following two important lemmas, and their proofs can be found in Appendix E.1.1 and Appendix E.1.2, respectively.

Lemma 3

$$\begin{aligned} \mathbb{E} \left[\left\| \sum_{s=1}^t (1 - \beta)^{t-s} \Delta_s \right\| \right] &\leq \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-r} (1 - \beta)^{t-s} \Delta_s \right\|^2 + \frac{1}{B_1} \sum_{s=1}^r (1 - \beta)^{2s-2} \Lambda^2} \right] \\ &\quad + \sum_{s=t+1-r}^t \frac{\Gamma}{\sqrt{B_1}} (1 - \beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \end{aligned}$$

Lemma 4

$$\begin{aligned} \mathbb{E} \left[\left\| \sum_{s=1}^t (1 - \beta)^{t-s} G_s \right\| \right] &\leq \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-r} (1 - \beta)^{t-s} G_s \right\|^2 + \frac{1}{B_1} \sum_{s=1}^r 2(1 - \beta)^{2s-2} \eta^2 L_3^2 d} \right] \\ &\quad + \frac{\sqrt{2d}\eta L_4}{\sqrt{B_1}} \sum_{s=t+1-r}^t (1 - \beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \end{aligned}$$

Using these lemmas and setting $r = t$, we then have

$$\begin{aligned}
\mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} \Delta_s \right\| \right] &\leq \sqrt{\frac{\Lambda^2}{B_1} \sum_{s=1}^t (1-\beta)^{2s-2} + \sum_{s=1}^t \frac{\Gamma}{\sqrt{B_1}} (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|]} \\
&\leq \frac{\Lambda}{\sqrt{B_1} \beta} + \sum_{s=1}^t \frac{\Gamma}{\sqrt{B_1}} (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
\mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} G_s \right\| \right] &\leq \sqrt{\frac{2\eta^2 L_3^2 d}{B_1} \sum_{s=1}^t (1-\beta)^{2s-2} + \frac{\sqrt{2d}\eta L_4}{\sqrt{B_1}} \sum_{s=1}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|]} \\
&\leq \frac{\sqrt{2d}\eta L_3}{\sqrt{B_1} \beta} + \frac{\sqrt{2d}\eta L_4}{\sqrt{B_1}} \sum_{s=1}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|].
\end{aligned}$$

Combining above inequalities and setting $\beta = \eta\sqrt{d}$, we derive

$$\begin{aligned}
&\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| \right] \\
&\leq \frac{1}{T} \sum_{t=1}^T (1-\beta)^{t-1} \mathbb{E} [\|\mathbf{v}_1 - \nabla f(\mathbf{x}_1)\|] + \beta \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} \Delta_s \right\| \right] \\
&\quad + \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} G_s \right\| \right] \\
&\leq \frac{\sigma}{\sqrt{B_0} T} \sum_{t=1}^T (1-\beta)^{t-1} + \frac{\Lambda\sqrt{\beta}}{\sqrt{B_1}} + \frac{\sqrt{2d}\eta L_3}{\sqrt{B_1} \beta} + \left(\frac{\beta\Gamma}{\sqrt{B_1}} + \frac{\sqrt{2d}\eta L_4}{\sqrt{B_1}} \right) \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
&\leq \frac{\sigma}{\beta T \sqrt{B_0}} + \frac{\Lambda\sqrt{\beta}}{\sqrt{B_1}} + \frac{\sqrt{2d}\eta L_3}{\sqrt{B_1} \beta} + \left(\frac{\beta\Gamma}{\sqrt{B_1}} + \frac{\sqrt{2d}\eta L_4}{\sqrt{B_1}} \right) \left(\sum_{i=1}^T (1-\beta)^i \right) \frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|] \\
&\leq \frac{\sigma}{\beta T \sqrt{B_0}} + \frac{\Lambda\sqrt{\beta}}{\sqrt{B_1}} + \frac{\sqrt{2d}\eta L_3}{\sqrt{B_1} \beta} + \left(\frac{\Gamma}{\sqrt{B_1}} + \frac{\sqrt{2d}\eta L_4}{\sqrt{B_1} \beta} \right) \frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|] \\
&\leq \frac{\sigma}{\eta T \sqrt{B_0 d}} + \frac{(\Lambda + \sqrt{2}L_3)d^{1/4}\eta^{1/2}}{\sqrt{B_1}} + \left(\frac{\Gamma}{\sqrt{B_1}} + \frac{\sqrt{2}L_4}{\sqrt{B_1}} \right) \frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|].
\end{aligned}$$

For $\alpha \in (0, 1)$, by setting that

$$\begin{aligned}
B_0 &= 1, \\
B_1 &\geq \max\{256\Gamma^2, 512L_4^2\}d, \\
\beta &= \frac{d^{1/3}}{T^{2/3}}, \\
\eta &= \frac{1}{d^{1/6}T^{2/3}},
\end{aligned}$$

and suppose that iteration number

$$T \geq \mathcal{O}(d^2),$$

then we can guarantee

$$\begin{aligned}
& \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] \\
& \leq \frac{2\Delta_f}{\eta T} + \eta d(K_0 + K_1 + 2K_2) + \frac{4\sigma}{\eta T \sqrt{B_0}} + \frac{4(\Lambda + \sqrt{2}L_3)d^{3/4}\eta^{1/2}}{\sqrt{B_1}} \\
& \quad + \left(\frac{4\sqrt{d}\Gamma}{\sqrt{B_1}} + \frac{4\sqrt{2d}L_4}{\sqrt{B_1}} \right) \frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|] \\
& \leq \mathcal{O} \left(\frac{d^{1/6}}{T^{1/3}} \right) + \frac{1}{2} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\| \right] \\
& \leq \mathcal{O} \left(\frac{d^{1/6}}{T^{1/3}} \right) + \frac{1}{2} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right],
\end{aligned}$$

which indicates that

$$\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] \leq \mathcal{O} \left(\frac{d^{1/6}}{T^{1/3}} \right).$$

Note that the batch size for each iteration is $B_1 = \mathcal{O}(d)$, by assuming that $N = B_1 * T$, we know that the convergence concerning N is

$$\mathcal{O} \left(\frac{d^{1/6}}{T^{1/3}} \right) = \mathcal{O} \left(\frac{d^{1/2}}{(B_1 T)^{1/3}} \right) = \mathcal{O} \left(\frac{d^{1/2}}{N^{1/3}} \right).$$

Similar results can be easily obtained for $\alpha = 1$, i.e., we can also guarantee the following for $\alpha = 1$:

$$\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] \leq \mathcal{O} \left(\frac{d^{1/6}}{N^{1/3}} \right).$$

E.1.1 Proof of Lemma 3

We prove this lemma by mathematical induction.

1) When $r = 0$, we have the following:

$$\mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} \Delta_s \right\| \right] = \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^t (1-\beta)^{t-s} \Delta_s \right\|^2} \right],$$

which satisfies the above lemma.

2) Then, suppose the lemma holds for $r = k$. For $r = k + 1$, we have

$$\begin{aligned}
& \mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} \Delta_s \right\| \right] \\
& \leq \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k} (1-\beta)^{t-s} \Delta_s \right\|^2 + \frac{1}{B_1} \sum_{s=1}^k (1-\beta)^{2s-2} \Lambda^2} \right] + \sum_{s=t+1-k}^t \frac{\Gamma}{\sqrt{B_1}} (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
& = \mathbb{E} \left[\mathbb{E}_{\xi_{t-k}} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} \Delta_s + (1-\beta)^k \Delta_{t-k} \right\|^2 + \frac{1}{B_1} \sum_{s=1}^k (1-\beta)^{2s-2} \Lambda^2} \right] \right] \\
& \quad + \sum_{s=t+1-k}^t \frac{\Gamma}{\sqrt{B_1}} (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
& \leq \mathbb{E} \left[\sqrt{\mathbb{E}_{\xi_{t-k}} \left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} \Delta_s + (1-\beta)^k \Delta_{t-k} \right\|^2 + \frac{1}{B_1} \sum_{s=1}^k (1-\beta)^{2s-2} \Lambda^2} \right] \\
& \quad + \sum_{s=t+1-k}^t \frac{\Gamma}{\sqrt{B_1}} (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
& \leq \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} \Delta_s \right\|^2 + \mathbb{E}_{\xi_{t-k}} \left[(1-\beta)^{2k} \|\Delta_{t-k}\|^2 \right] + \frac{1}{B_1} \sum_{s=1}^k (1-\beta)^{2s-2} \Lambda^2} \right] \\
& \quad + \sum_{s=t+1-k}^t \frac{\Gamma}{\sqrt{B_1}} (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
& \leq \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} \Delta_s \right\|^2 + \frac{(1-\beta)^{2k}}{B_1} \left(\Lambda^2 + \Gamma^2 \|\nabla f(\mathbf{x}_{t-k})\|^2 \right) + \frac{1}{B_1} \sum_{s=1}^k (1-\beta)^{2s-2} \Lambda^2} \right] \\
& \quad + \sum_{s=t+1-k}^t \frac{\Gamma}{\sqrt{B_1}} (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
& \leq \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} \Delta_s \right\|^2 + \frac{1}{B_1} \sum_{s=1}^{k+1} (1-\beta)^{2s-2} \Lambda^2} \right] + \frac{(1-\beta)^k \Gamma}{\sqrt{B_1}} \mathbb{E} [\|\nabla f(\mathbf{x}_{t-k})\|] \\
& \quad + \sum_{s=t+1-k}^t \frac{\Gamma}{\sqrt{B_1}} (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
& = \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} \Delta_s \right\|^2 + \frac{1}{B_1} \sum_{s=1}^{k+1} (1-\beta)^{2s-2} \Lambda^2} \right] + \sum_{s=t-k}^t \frac{\Gamma}{\sqrt{B_1}} (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|],
\end{aligned}$$

where the second inequality is due to the Jensen Inequality.

E.1.2 Proof of Lemma 4

We prove this lemma by mathematical induction.

1) When $r = 0$, we can easily prove $\mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} G_s \right\| \right] = \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^t (1-\beta)^{t-s} G_s \right\|^2} \right]$

2) Suppose the inequality holds for $r = k$. Then, for $r = k + 1$, we derive

$$\begin{aligned}
& \mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} G_s \right\| \right] \\
\leq & \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k} (1-\beta)^{t-s} G_s \right\|^2 + \frac{1}{B_1} \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 L_3^2 d} \right] \\
& + \frac{\sqrt{2d}\eta L_4}{\sqrt{B_1}} \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
= & \mathbb{E} \left[\mathbb{E}_{\xi_{t-k}} \left[\sqrt{\left\| \sum_{s=1}^{t-k} (1-\beta)^{t-s} G_s \right\|^2 + \frac{1}{B_1} \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 L_3^2 d} \right] \right] \\
& + \frac{\sqrt{2d}\eta L_4}{\sqrt{B_1}} \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
= & \mathbb{E} \left[\mathbb{E}_{\xi_{t-k}} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} G_s + (1-\beta)^k G_{t-k} \right\|^2 + \frac{1}{B_1} \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 L_3^2 d} \right] \right] \\
& + \frac{\sqrt{2d}\eta L_4}{\sqrt{B_1}} \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
\leq & \mathbb{E} \left[\sqrt{\mathbb{E}_{\xi_{t-k}} \left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} G_s + (1-\beta)^k G_{t-k} \right\|^2 + \frac{1}{B_1} \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 L_3^2 d} \right] \\
& + \frac{\sqrt{2d}\eta L_4}{\sqrt{B_1}} \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
= & \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} G_s \right\|^2 + (1-\beta)^{2k} \mathbb{E}_{\xi_{t-k}} [\|G_{t-k}\|^2] + \frac{1}{B_1} \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 L_3^2 d} \right] \\
& + \frac{\sqrt{2d}\eta L_4}{\sqrt{B_1}} \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
\leq & \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} G_s \right\|^2 + \frac{2}{B_1} (1-\beta)^{2k} \eta^2 L_3^2 d + \frac{1}{B_1} \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 L_3^2 d} \right] \\
& + \frac{\sqrt{2d}\eta L_4 (1-\beta)^k}{\sqrt{B_1}} \mathbb{E} [\|\nabla f(\mathbf{x}_{t-k})\|] + \frac{\sqrt{2d}\eta L_4}{\sqrt{B_1}} \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
= & \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} G_s \right\|^2 + \frac{1}{B_1} \sum_{s=1}^{k+1} 2(1-\beta)^{2s-2} \eta^2 L_3^2 d} \right] \\
& + \frac{\sqrt{2d}\eta L_4}{\sqrt{B_1}} \sum_{s=t-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|],
\end{aligned}$$

where the third inequality is due to the following, for simplify we denote $\xi_{t-k} = \xi_{t-k}^1, \dots, \xi_{t-k}^{B_1}$:

$$\begin{aligned}
\mathbb{E}_{\xi_{t-k}} \left[\|G_{t-k}\|^2 \right] &= \mathbb{E}_{\xi_{t-k}} \left[\frac{1}{B_1^2} \sum_{j=1}^{B_1} \left\| \nabla f(\mathbf{x}_{t-k}; \xi_{t-k}^j) - \nabla f(\mathbf{x}_{t-k-1}; \xi_{t-k}^j) + \nabla f(\mathbf{x}_{t-k-1}) - \nabla f(\mathbf{x}_{t-k}) \right\|^2 \right] \\
&= \mathbb{E}_{\xi_{t-k}} \left[\frac{1}{B_1^2} \sum_{j=1}^{B_1} \left\| \nabla f(\mathbf{x}_{t-k}; \xi_{t-k}^j) - \nabla f(\mathbf{x}_{t-k-1}; \xi_{t-k}^j) \right\|^2 \right] + \frac{1}{B_1} \left[\|\nabla f(\mathbf{x}_{t-k-1}) - \nabla f(\mathbf{x}_{t-k})\|^2 \right] \\
&\quad - 2\mathbb{E}_{\xi_{t-k}} \left[\frac{1}{B_1^2} \sum_{j=1}^{B_1} \left\langle \nabla f(\mathbf{x}_{t-k}; \xi_{t-k}^j) - \nabla f(\mathbf{x}_{t-k-1}; \xi_{t-k}^j), \nabla f(\mathbf{x}_{t-k}) - \nabla f(\mathbf{x}_{t-k-1}) \right\rangle \right] \\
&= \mathbb{E}_{\xi_{t-k}} \left[\frac{1}{B_1^2} \sum_{j=1}^{B_1} \left\| \nabla f(\mathbf{x}_{t-k}; \xi_{t-k}^j) - \nabla f(\mathbf{x}_{t-k-1}; \xi_{t-k}^j) \right\|^2 \right] - \frac{1}{B_1} \left[\|\nabla f(\mathbf{x}_{t-k-1}) - \nabla f(\mathbf{x}_{t-k})\|^2 \right] \\
&\leq \mathbb{E}_{\xi_{t-k}} \left[\frac{1}{B_1^2} \sum_{j=1}^{B_1} \left\| \nabla f(\mathbf{x}_{t-k}; \xi_{t-k}^j) - \nabla f(\mathbf{x}_{t-k-1}; \xi_{t-k}^j) \right\|^2 \right]
\end{aligned}$$

For $\alpha \in (0, 1)$, denoting $L_3^2 = (\bar{K}_0 + \bar{K}_1 + \bar{K}_2)^2 + \bar{K}_1^2 \Lambda^2$, and $L_4^2 = \bar{K}_1^2 (1 + \Gamma^2)$, we have:

$$\begin{aligned}
&\mathbb{E}_{\xi_{t-k}} \left[\frac{1}{B_1^2} \sum_{j=1}^{B_1} \left\| \nabla f(\mathbf{x}_{t-k}; \xi_{t-k}^j) - \nabla f(\mathbf{x}_{t-k-1}; \xi_{t-k}^j) \right\|^2 \right] \\
&\leq \frac{1}{B_1^2} \sum_{j=1}^{B_1} \|\mathbf{x}_{t-k} - \mathbf{x}_{t-k-1}\|^2 \left(\bar{K}_0 + \bar{K}_1 \mathbb{E}_{\xi_{t-k}} \left[\left\| \nabla f(\mathbf{x}_{t-k}; \xi_{t-k}^j) \right\|^\alpha \right] + \bar{K}_2 \|\mathbf{x}_{t-k} - \mathbf{x}_{t-k-1}\|^{\frac{1-\alpha}{1-\alpha}} \right)^2 \\
&\leq \frac{\eta^2 d}{B_1^2} \sum_{j=1}^{B_1} \left(\bar{K}_0 + \bar{K}_1 + \bar{K}_2 + \bar{K}_1 \mathbb{E}_{\xi_{t-k}} \left[\left\| \nabla f(\mathbf{x}_{t-k}; \xi_{t-k}^j) \right\| \right] \right)^2 \\
&\leq \frac{2\eta^2 d}{B_1} (\bar{K}_0 + \bar{K}_1 + \bar{K}_2)^2 + \frac{2\eta^2 d}{B_1^2} \sum_{j=1}^{B_1} \bar{K}_1^2 \left(\mathbb{E}_{\xi_{t-k}} \left[\left\| \nabla f(\mathbf{x}_{t-k}; \xi_{t-k}^j) \right\| \right] \right)^2 \\
&\leq \frac{2\eta^2 d}{B_1} (\bar{K}_0 + \bar{K}_1 + \bar{K}_2)^2 + \frac{2\eta^2 d}{B_1} \bar{K}_1^2 \left((1 + \Gamma^2) \|\nabla f(\mathbf{x}_{t-k})\| + \Lambda^2 \right) \\
&\leq \frac{2\eta^2 d L_3^2}{B_1} + \frac{2\eta^2 d L_4^2}{B_1} \|\nabla f(\mathbf{x}_{t-k})\|^2,
\end{aligned}$$

where the second inequality holds by setting $\eta \leq d^{-1/2}$ such that $\|\mathbf{x}_{t-k} - \mathbf{x}_{t-k-1}\| \leq \eta\sqrt{d} \leq 1$.

For $\alpha = 1$, denoting $L_3^2 = 3(L_0^2 + 2L_1^2 \Lambda^2)$, and $L_4^2 = 6L_1^2 (1 + \Gamma^2)$, we have:

$$\begin{aligned}
&\mathbb{E}_{\xi_{t-k}} \left[\frac{1}{B_1^2} \sum_{j=1}^{B_1} \left\| \nabla f(\mathbf{x}_{t-k}; \xi_{t-k}^j) - \nabla f(\mathbf{x}_{t-k-1}; \xi_{t-k}^j) \right\|^2 \right] \\
&\leq \frac{2}{B_1^2} \sum_{j=1}^{B_1} \|\mathbf{x}_{t-k} - \mathbf{x}_{t-k-1}\|^2 \left(L_0^2 + 2L_1^2 \mathbb{E}_{\xi_{t-k}} \left[\left\| \nabla f(\mathbf{x}_{t-k}; \xi_{t-k}^j) \right\|^2 \right] \right) \exp \left(12L_1^2 \|\mathbf{x}_{t-k} - \mathbf{x}_{t-k-1}\|^2 \right) \\
&\leq \frac{6\eta^2 d}{B_1^2} \sum_{j=1}^{B_1} \left(L_0^2 + 2L_1^2 \mathbb{E}_{\xi_{t-k}} \left[\left\| \nabla f(\mathbf{x}_{t-k}; \xi_{t-k}^j) \right\|^2 \right] \right) \\
&\leq \frac{6\eta^2 d}{B_1} \left(L_0^2 + 2L_1^2 \left((1 + \Gamma^2) \|\nabla f(\mathbf{x}_{t-k})\|^2 + \Lambda^2 \right) \right) \\
&\leq \frac{2\eta^2 d L_3^2}{B_1} + \frac{2\eta^2 d L_4^2}{B_1} \|\nabla f(\mathbf{x}_{t-k})\|^2,
\end{aligned}$$

where the second inequality holds by setting $\eta \leq \frac{1}{\sqrt{12L_1^2 d}}$, such that $\|\mathbf{x}_{t-k} - \mathbf{x}_{t-k-1}\|^2 \leq \frac{1}{12L_1^2}$.

E.2 Proof of Theorem 6

Similar to Lemma 1 and 2, we can have the following lemma for generalized individual smoothness.

Lemma 5 For $\alpha \in (0, 1)$, generalized individual smoothness (Assumption 3') leads to

$$f_i(\mathbf{x}_{t+1}) \leq f_i(\mathbf{x}_t) + \langle \nabla f_i(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 (K_0 + K_1 \|\nabla f(\mathbf{x}_t)\|^\alpha + 2K_2 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^{\frac{\alpha}{1-\alpha}}),$$

as well as

$$\|\nabla f_i(\mathbf{x}_{t+1}) - \nabla f_i(\mathbf{x}_t)\| \leq \|\mathbf{x}_{t+1} - \mathbf{x}_t\| (K_0 + K_1 \|\nabla f(\mathbf{x}_t)\|^\alpha + K_2 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^{\frac{\alpha}{1-\alpha}}).$$

where $K_0 := L_0(2^{\frac{\alpha^2}{1-\alpha}} + 1)$, $K_1 := L_1 \cdot 2^{\frac{\alpha^2}{1-\alpha}} \cdot 3^\alpha$, $K_2 := L_1^{\frac{1}{1-\alpha}} \cdot 2^{\frac{\alpha^2}{1-\alpha}} \cdot 3^\alpha(1-\alpha)^{\frac{\alpha}{1-\alpha}}$.

Lemma 6 For $\alpha = 1$, generalized individual smoothness (Assumption 3') leads to

$$f_i(\mathbf{x}_{t+1}) \leq f_i(\mathbf{x}_t) + \langle \nabla f_i(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 (L_0 + L_1 \|\nabla f(\mathbf{x}_t)\|) \exp(L_1 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|),$$

as well as

$$\|\nabla f_i(\mathbf{x}_{t+1}) - \nabla f_i(\mathbf{x}_t)\| \leq \|\mathbf{x}_{t+1} - \mathbf{x}_t\| (L_0 + L_1 \|\nabla f(\mathbf{x}_t)\|) \exp(L_1 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|).$$

Then, we can begin our proof. For $\alpha \in (0, 1)$, according to Lemma 5, by setting $\eta \leq d^{-\frac{1}{2}}$, we have:

$$\begin{aligned} & f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \\ & \leq \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{x}_{t+1}) - \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{x}_t) \\ & \leq \frac{1}{m} \sum_{i=1}^m \langle \nabla f_i(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 (K_0 + K_1 \|\nabla f(\mathbf{x}_t)\|^\alpha + 2K_2 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^{\frac{\alpha}{1-\alpha}}) \\ & \leq \langle \nabla f(\mathbf{x}_t), -\eta \text{Sign}(\mathbf{v}_t) \rangle + \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 (K_0 + K_1 (1 + \|\nabla f(\mathbf{x}_t)\|) + 2K_2) \\ & \leq \eta \langle \nabla f(\mathbf{x}_t), \text{Sign}(\nabla f(\mathbf{x}_t)) - \text{Sign}(\mathbf{v}_t) \rangle - \eta \langle \nabla f(\mathbf{x}_t), \text{Sign}(\nabla f(\mathbf{x}_t)) \rangle \\ & \quad + \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 (K_0 + K_1 (1 + \|\nabla f(\mathbf{x}_t)\|) + 2K_2) \\ & = \eta \langle \nabla f(\mathbf{x}_t), \text{Sign}(\nabla f(\mathbf{x}_t)) - \text{Sign}(\mathbf{v}_t) \rangle - \eta \|\nabla f(\mathbf{x}_t)\|_1 \\ & \quad + \frac{1}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 ((K_0 + K_1 + 2K_2) + K_1 \|\nabla f(\mathbf{x}_t)\|) \\ & \leq 2\eta\sqrt{d} \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| - \eta \|\nabla f(\mathbf{x}_t)\|_1 + \frac{\eta^2 d}{2} ((K_0 + K_1 + 2K_2) + K_1 \|\nabla f(\mathbf{x}_t)\|), \end{aligned}$$

where the second inequality is due to the fact that $\alpha < 1$ and $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \leq \eta^2 d \leq 1$.

Rearranging and summing up, we then have

$$\begin{aligned} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] & \leq \frac{\Delta f}{\eta T} + 2\sqrt{d} \cdot \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| \right] \\ & \quad + \frac{\eta d(K_0 + K_1 + 2K_2)}{2} + \frac{\eta d K_1}{2T} \mathbb{E} \left[\sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\| \right]. \end{aligned}$$

By setting $\eta \leq \min\{\frac{1}{\sqrt{d}}, \frac{1}{dK_1}\}$, we can get

$$\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] \leq \frac{2\Delta f}{\eta T} + 4\sqrt{d} \cdot \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| \right] + \eta d(K_0 + K_1 + 2K_2). \quad (17)$$

For $\alpha = 1$, by setting $\eta \leq \frac{1}{3dL_1}$, we can apply the very similar analysis and obtain

$$\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] \leq \frac{2\Delta_f}{\eta T} + 4\sqrt{d} \cdot \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| \right] + 3\eta d L_0. \quad (18)$$

Then we bound the term $\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| \right]$. According to the definition of \mathbf{v}_t , we have:

$$\begin{aligned} \mathbf{v}_t - \nabla f(\mathbf{x}_t) &= (1 - \beta) (\mathbf{v}_{t-1} - \nabla f(\mathbf{x}_{t-1})) + \beta (\mathbf{h}_t - \nabla f(\mathbf{x}_t)) \\ &\quad + (1 - \beta) (\nabla f(\mathbf{x}_t; \xi_t^k) - \nabla f(\mathbf{x}_{t-1}; \xi_t^k) + \nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_t)) \end{aligned}$$

Denote that $G_t = (\nabla f(\mathbf{x}_t; \xi_t^k) - \nabla f(\mathbf{x}_{t-1}; \xi_t^k) + \nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_t))$ and $\Delta_t = \mathbf{h}_t - \nabla f(\mathbf{x}_t)$. By summing up, we have

$$\begin{aligned} &\mathbf{v}_t - \nabla f(\mathbf{x}_t) \\ &= (1 - \beta) (\mathbf{v}_{t-1} - \nabla f(\mathbf{x}_{t-1})) + \beta \Delta_t + (1 - \beta) G_t \\ &= (1 - \beta)^{t-1} (\mathbf{v}_1 - \nabla f(\mathbf{x}_1)) + \beta \sum_{s=1}^t (1 - \beta)^{t-s} \Delta_s + (1 - \beta) \sum_{s=1}^t (1 - \beta)^{t-s} G_s. \end{aligned}$$

Thus, we can know that

$$\begin{aligned} &\mathbb{E} [\|\mathbf{v}_t - \nabla f(\mathbf{x}_t)\|] \\ &\leq (1 - \beta)^{t-1} \mathbb{E} [\|\mathbf{v}_1 - \nabla f(\mathbf{x}_1)\|] + \beta \mathbb{E} \left[\left\| \sum_{s=1}^t (1 - \beta)^{t-s} \Delta_s \right\| \right] + (1 - \beta) \mathbb{E} \left[\left\| \sum_{s=1}^t (1 - \beta)^{t-s} G_s \right\| \right] \\ &\leq \beta \mathbb{E} \left[\left\| \sum_{s=1}^t (1 - \beta)^{t-s} \Delta_s \right\| \right] + (1 - \beta) \mathbb{E} \left[\left\| \sum_{s=1}^t (1 - \beta)^{t-s} G_s \right\| \right], \end{aligned}$$

where the first term vanishes since we use full batch in the first iteration.

Then we give the following two important lemmas, and their proofs can be found in Appendix E.2.1 and Appendix E.2.2, respectively.

Lemma 7

$$\begin{aligned} \mathbb{E} \left[\left\| \sum_{s=1}^t (1 - \beta)^{t-s} \Delta_s \right\| \right] &\leq \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-r} (1 - \beta)^{t-s} \Delta_s \right\|^2 + \sum_{s=1}^r 2(1 - \beta)^{2s-2} \eta^2 I^2 L_5^2 d} \right] \\ &\quad + \sqrt{2d} \eta L_6 \sum_{s=t+1-r}^t (1 - \beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \end{aligned}$$

Lemma 8

$$\begin{aligned} \mathbb{E} \left[\left\| \sum_{s=1}^t (1 - \beta)^{t-s} G_s \right\| \right] &\leq \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-r} (1 - \beta)^{t-s} G_s \right\|^2 + \sum_{s=1}^r 2(1 - \beta)^{2s-2} \eta^2 L_7^2 d} \right] \\ &\quad + \sqrt{2d} \eta L_8 \sum_{s=t+1-r}^t (1 - \beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \end{aligned}$$

Using these lemmas and setting $r = t$, we then have

$$\begin{aligned}
& \mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} \Delta_s \right\| \right] \\
& \leq \sqrt{2\eta^2 I^2 L_5^2 d \sum_{s=1}^t (1-\beta)^{2s-2} + \sqrt{2d\eta} I L_6 \sum_{s=1}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|]} \\
& \leq \frac{\sqrt{2d\eta} I L_5}{\sqrt{\beta}} + \sqrt{2d\eta} I L_6 \sum_{s=1}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|],
\end{aligned}$$

as well as

$$\begin{aligned}
& \mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} G_s \right\| \right] \\
& \leq \sqrt{2\eta^2 L_7^2 d \sum_{s=1}^t (1-\beta)^{2s-2} + \sqrt{2d\eta} L_8 \sum_{s=1}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|]} \\
& \leq \frac{\sqrt{2d\eta} L_7}{\sqrt{\beta}} + \sqrt{2d\eta} L_8 \sum_{s=1}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|].
\end{aligned}$$

Combining above inequalities and setting $\beta = 1/m$ and $I = m$, we derive

$$\begin{aligned}
& \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t) - \mathbf{v}_t\| \right] \\
& \leq \beta \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} \Delta_s \right\| \right] + \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} G_s \right\| \right] \\
& \leq \sqrt{2d\beta\eta} I L_5 + \frac{\sqrt{2d\eta} L_7}{\sqrt{\beta}} + \left(\sqrt{2d\eta\beta} I L_6 + \sqrt{2d\eta} L_8 \right) \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
& \leq \sqrt{2dm\eta} (L_5 + L_7) + \sqrt{2d\eta} (L_6 + L_8) \left(\sum_{i=1}^T (1-\beta)^i \right) \frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|] \\
& \leq \sqrt{2dm\eta} (L_5 + L_7) + \sqrt{2d\eta} m (L_6 + L_8) \frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|].
\end{aligned}$$

For $\alpha \in (0, 1)$, by setting $\eta = \min \left\{ \frac{1}{m^{1/4} d^{1/2} T^{1/2}}, \frac{1}{8\sqrt{2}md(L_6+L_8+1)} \right\}$, we can guarantee

$$\begin{aligned}
& \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] \\
& \leq \frac{2\Delta_f}{\eta T} + \eta d (K_0 + K_1 + 2K_2) + 4d\sqrt{2m\eta} (L_5 + L_7) + 4d\sqrt{2\eta} m (L_6 + L_8) \frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\nabla f(\mathbf{x}_t)\|] \\
& \leq \mathcal{O} \left(\frac{m^{1/4} d^{1/2}}{T^{1/2}} + \frac{md}{T} \right) + \frac{1}{2} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right],
\end{aligned}$$

which indicates that $\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\nabla f(\mathbf{x}_t)\|_1 \right] \leq \mathcal{O} \left(\frac{m^{1/4} d^{1/2}}{T^{1/2}} + \frac{md}{T} \right)$. Similar results can be easily obtained for $\alpha = 1$.

E.2.1 Proof of Lemma 7

We prove this lemma by mathematical induction.

1) When $r = 0$, we have the following:

$$\mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} \Delta_s \right\| \right] = \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^t (1-\beta)^{t-s} \Delta_s \right\|^2} \right],$$

which satisfies the above lemma.

2) Then, suppose the lemma holds for $r = k$. For $r = k + 1$, we have

$$\begin{aligned} & \mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} \Delta_s \right\| \right] \\ & \leq \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k} (1-\beta)^{t-s} \Delta_s \right\|^2 + \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 I^2 L_5^2 d} \right] \\ & \quad + \sqrt{2d} \eta I L_6 \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\ & = \mathbb{E} \left[\mathbb{E}_{i_{t-k}} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} \Delta_s + (1-\beta)^k \Delta_{t-k} \right\|^2 + \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 I^2 L_5^2 d} \right] \right] \\ & \quad + \sqrt{2d} \eta I L_6 \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\ & \leq \mathbb{E} \left[\sqrt{\mathbb{E}_{i_{t-k}} \left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} \Delta_s + (1-\beta)^k \Delta_{t-k} \right\|^2 + \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 I^2 L_5^2 d} \right] \\ & \quad + \sqrt{2d} \eta I L_6 \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\ & = \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} \Delta_s \right\|^2 + (1-\beta)^{2k} \mathbb{E}_{i_{t-k}} [\|\Delta_{t-k}\|^2] + \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 I^2 L_5^2 d} \right] \\ & \quad + \sqrt{2d} \eta I L_6 \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\ & \leq \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} \Delta_s \right\|^2 + 2(1-\beta)^{2k} \eta^2 I^2 L_5^2 d + \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 I^2 L_5^2 d} \right] \\ & \quad + \sqrt{2d} \eta I L_6 (1-\beta)^k \mathbb{E} [\|\nabla f(\mathbf{x}_{t-k})\|] + \sqrt{2d} \eta I L_6 \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\ & = \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} \Delta_s \right\|^2 + \sum_{s=1}^{k+1} 2(1-\beta)^{2s-2} \eta^2 I^2 L_5^2 d} \right] \\ & \quad + \sqrt{2d} \eta I L_6 \sum_{s=t-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|], \end{aligned}$$

where the third inequality is due to the following:

$$\begin{aligned}\mathbb{E}_{i_{t-k}} \left[\|\Delta_{t-k}\|^2 \right] &= \mathbb{E}_{i_{t-k}} \left[\|\nabla f_{i_{t-k}}(\mathbf{x}_{t-k}) - \nabla f_{i_{t-k}}(\mathbf{x}_\tau) + \nabla f(\mathbf{x}_\tau) - \nabla f(\mathbf{x}_{t-k})\|^2 \right] \\ &\leq \mathbb{E}_{i_{t-k}} \left[\|\nabla f_{i_{t-k}}(\mathbf{x}_{t-k}) - \nabla f_{i_{t-k}}(\mathbf{x}_\tau)\|^2 \right]\end{aligned}$$

For $\alpha \in (0, 1)$, denoting that

$$\begin{aligned}L_5^2 &= (K_0 + K_1 + K_2)^2, \\ L_6^2 &= K_1^2,\end{aligned}$$

we have:

$$\begin{aligned}&\mathbb{E}_{i_{t-k}} \left[\|\nabla f_{i_{t-k}}(\mathbf{x}_{t-k}) - \nabla f_{i_{t-k}}(\mathbf{x}_\tau)\|^2 \right] \\ &\leq \|\mathbf{x}_{t-k} - \mathbf{x}_\tau\|^\alpha (K_0 + K_1 \|\nabla f(\mathbf{x}_{t-k})\|^\alpha + K_2 \|\mathbf{x}_{t-k} - \mathbf{x}_\tau\|^{\frac{\alpha}{1-\alpha}})^2 \\ &\leq \eta^2 I^2 d (K_0 + K_1 + K_2 + K_1 \|\nabla f(\mathbf{x}_{t-k})\|)^2 \\ &\leq 2\eta^2 I^2 d (K_0 + K_1 + K_2)^2 + 2\eta^2 I^2 d K_1^2 \|\nabla f(\mathbf{x}_{t-k})\|^2 \\ &\leq 2\eta^2 d I^2 L_5^2 + 2\eta^2 I^2 d L_6^2 \|\nabla f(\mathbf{x}_{t-k})\|^2,\end{aligned}$$

where the second inequality holds by setting

$$\eta \leq I^{-1} d^{-1/2}$$

such that

$$\|\mathbf{x}_{t-k} - \mathbf{x}_\tau\| \leq \eta I \sqrt{d} \leq 1.$$

For $\alpha = 1$, denoting that

$$\begin{aligned}L_5^2 &= 9L_0^2 \\ L_6^2 &= 9L_1^2,\end{aligned}$$

we have:

$$\begin{aligned}&\mathbb{E}_{i_{t-k}} \left[\|\nabla f_{i_{t-k}}(\mathbf{x}_{t-k}) - \nabla f_{i_{t-k}}(\mathbf{x}_\tau)\|^2 \right] \\ &\leq 2 \|\mathbf{x}_{t-k} - \mathbf{x}_\tau\|^2 \left(L_0^2 + L_1^2 \|\nabla f(\mathbf{x}_{t-k})\|^2 \right) \left(\exp \left(L_1^2 \|\mathbf{x}_{t-k} - \mathbf{x}_\tau\|^2 \right) \right)^2 \\ &\leq 18\eta^2 d I^2 \left(L_0^2 + L_1^2 \|\nabla f(\mathbf{x}_{t-k})\|^2 \right) \\ &\leq 2\eta^2 I^2 d L_5^2 + 2\eta^2 I^2 d L_6^2 \|\nabla f(\mathbf{x}_{t-k})\|^2,\end{aligned}$$

where the second inequality holds by setting

$$\eta \leq \frac{1}{\sqrt{L_1^2 I^2 d}},$$

such that we have

$$\|\mathbf{x}_{t-k} - \mathbf{x}_{t-k-1}\|^2 \leq \eta^2 I^2 d \leq \frac{1}{L_1^2}.$$

E.2.2 Proof of Lemma 8

We prove this lemma by mathematical induction.

1) When $r = 0$, we can easily prove that

$$\mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} G_s \right\|^2 \right] = \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^t (1-\beta)^{t-s} G_s \right\|^2} \right].$$

2) Suppose the inequality holds for $r = k$. Then, for $r = k + 1$, we derive

$$\begin{aligned}
& \mathbb{E} \left[\left\| \sum_{s=1}^t (1-\beta)^{t-s} G_s \right\| \right] \\
\leq & \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k} (1-\beta)^{t-s} G_s \right\|^2 + \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 L_7^2 d} \right] \\
& + \sqrt{2d} \eta L_8 \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
\leq & \mathbb{E} \left[\mathbb{E}_{i_{t-k}} \left[\sqrt{\left\| \sum_{s=1}^{t-k} (1-\beta)^{t-s} G_s \right\|^2 + \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 L_7^2 d} \right] \right] \\
& + \sqrt{2d} \eta L_8 \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
= & \mathbb{E} \left[\mathbb{E}_{i_{t-k}} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} G_s + (1-\beta)^k G_{t-k} \right\|^2 + \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 L_7^2 d} \right] \right] \\
& + \sqrt{2d} \eta L_8 \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
\leq & \mathbb{E} \left[\sqrt{\mathbb{E}_{i_{t-k}} \left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} G_s + (1-\beta)^k G_{t-k} \right\|^2 + \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 L_7^2 d} \right] \\
& + \sqrt{2d} \eta L_8 \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
= & \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} G_s \right\|^2 + (1-\beta)^{2k} \mathbb{E}_{i_{t-k}} [\|G_{t-k}\|^2] + \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 L_7^2 d} \right] \\
& + \sqrt{2d} \eta L_8 \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
\leq & \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} G_s \right\|^2 + 2(1-\beta)^{2k} \eta^2 L_7^2 d + \sum_{s=1}^k 2(1-\beta)^{2s-2} \eta^2 L_7^2 d} \right] \\
& + \sqrt{2d} \eta L_8 (1-\beta)^k \mathbb{E} [\|\nabla f(\mathbf{x}_{t-k})\|] + \sqrt{2d} \eta L_8 \sum_{s=t+1-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|] \\
= & \mathbb{E} \left[\sqrt{\left\| \sum_{s=1}^{t-k-1} (1-\beta)^{t-s} G_s \right\|^2 + \sum_{s=1}^{k+1} 2(1-\beta)^{2s-2} \eta^2 L_7^2 d} \right] \\
& + \sqrt{2d} \eta L_8 \sum_{s=t-k}^t (1-\beta)^{t-s} \mathbb{E} [\|\nabla f(\mathbf{x}_s)\|],
\end{aligned}$$

where the third inequality is due to the following:

$$\begin{aligned}
& \mathbb{E}_{i_{t-k}} \left[\|G_{t-k}\|^2 \right] \\
&= \mathbb{E}_{i_{t-k}} \left[\left\| \nabla f_{i_{t-k}}(\mathbf{x}_{t-k}) - \nabla f_{i_{t-k}}(\mathbf{x}_{t-k-1}) + \nabla f(\mathbf{x}_{t-k-1}) - \nabla f(\mathbf{x}_{t-k}) \right\|^2 \right] \\
&= \mathbb{E}_{i_{t-k}} \left[\left\| \nabla f_{i_{t-k}}(\mathbf{x}_{t-k}) - \nabla f_{i_{t-k}}(\mathbf{x}_{t-k-1}) \right\|^2 \right] + \left[\left\| \nabla f(\mathbf{x}_{t-k-1}) - \nabla f(\mathbf{x}_{t-k}) \right\|^2 \right] \\
&\quad - 2\mathbb{E}_{i_{t-k}} \left[\left\langle \nabla f_{i_{t-k}}(\mathbf{x}_{t-k}) - \nabla f_{i_{t-k}}(\mathbf{x}_{t-k-1}), \nabla f(\mathbf{x}_{t-k}) - \nabla f(\mathbf{x}_{t-k-1}) \right\rangle \right] \\
&= \mathbb{E}_{i_{t-k}} \left[\left\| \nabla f_{i_{t-k}}(\mathbf{x}_{t-k}) - \nabla f_{i_{t-k}}(\mathbf{x}_{t-k-1}) \right\|^2 \right] - \left[\left\| \nabla f(\mathbf{x}_{t-k-1}) - \nabla f(\mathbf{x}_{t-k}) \right\|^2 \right] \\
&\leq \mathbb{E}_{i_{t-k}} \left[\left\| \nabla f_{i_{t-k}}(\mathbf{x}_{t-k}) - \nabla f_{i_{t-k}}(\mathbf{x}_{t-k-1}) \right\|^2 \right]
\end{aligned}$$

For $\alpha \in (0, 1)$, denoting $L_7^2 = (K_0 + K_1 + K_2)^2$, and $L_8^2 = K_1^2$, we have:

$$\begin{aligned}
& \mathbb{E}_{i_{t-k}} \left[\left\| \nabla f_{i_{t-k}}(\mathbf{x}_{t-k}) - \nabla f_{i_{t-k}}(\mathbf{x}_{t-k-1}) \right\|^2 \right] \\
&\leq \|\mathbf{x}_{t-k} - \mathbf{x}_{t-k-1}\|^2 (K_0 + K_1 \|\nabla f(\mathbf{x}_{t-k})\|^\alpha + K_2 \|\mathbf{x}_{t-k} - \mathbf{x}_{t-k-1}\|^{\frac{\alpha}{1-\alpha}})^2 \\
&\leq \eta^2 d (K_0 + K_1 + K_2 + K_1 \|\nabla f(\mathbf{x}_{t-k})\|)^2 \\
&\leq 2\eta^2 d (K_0 + K_1 + K_2)^2 + 2\eta^2 d K_1^2 \|\nabla f(\mathbf{x}_{t-k})\|^2 \\
&\leq 2\eta^2 d L_7^2 + 2\eta^2 d L_8^2 \|\nabla f(\mathbf{x}_{t-k})\|^2,
\end{aligned}$$

where the second inequality holds by setting $\eta \leq d^{-1/2}$ such that $\|\mathbf{x}_{t-k} - \mathbf{x}_{t-k-1}\| \leq \eta\sqrt{d} \leq 1$.

For $\alpha = 1$, denoting $L_7^2 = 9L_0^2$ and $L_4^2 = 9L_1^2$, we have:

$$\begin{aligned}
& \mathbb{E}_{i_{t-k}} \left[\left\| \nabla f_{i_{t-k}}(\mathbf{x}_{t-k}) - \nabla f_{i_{t-k}}(\mathbf{x}_{t-k-1}) \right\|^2 \right] \\
&\leq 2 \|\mathbf{x}_{t-k} - \mathbf{x}_{t-k-1}\|^2 \left(L_0^2 + L_1^2 \|\nabla f(\mathbf{x}_{t-k})\|^2 \right) \left(\exp \left(L_1^2 \|\mathbf{x}_{t-k} - \mathbf{x}_{t-k-1}\|^2 \right) \right)^2 \\
&\leq 18\eta^2 d \left(L_0^2 + 2L_1^2 \|\nabla f(\mathbf{x}_{t-k})\|^2 \right) \\
&\leq 2\eta^2 d L_7^2 + 2\eta^2 d L_8^2 \|\nabla f(\mathbf{x}_{t-k})\|^2,
\end{aligned}$$

where the second inequality holds by setting $\eta \leq \frac{1}{\sqrt{L_1^2 d}}$, such that we have $\|\mathbf{x}_{t-k} - \mathbf{x}_{t-k-1}\|^2 \leq \eta^2 d \leq \frac{1}{L_1^2}$.

F Additional experiments

In this section, we present additional experiments on the CIFAR-10 dataset to validate whether the proposed SSVR method is sensitive to hyper-parameters such as learning rate η , momentum parameter β , and batch size. Specifically, we fix the value of β as 0.5 and try different learning rates from the set $\{5e-3, 1e-3, 5e-4, 1e-4, 5e-5\}$. Then, we fix the learning rate as $1e-3$ and enumerate β from the set $\{0.3, 0.5, 0.7, 0.9, 0.99\}$. The results are reported in Figure 3 and Figure 4, respectively, which indicate that our method is insensitive to the choice of hyper-parameters within a certain range. Finally, we also try different batch sizes from the set $\{64, 128, 256, 512\}$, and the results are shown in Figure 5. It can be seen that our algorithm does not necessitate large batches for convergence and is not sensitive to variations in batch sizes.

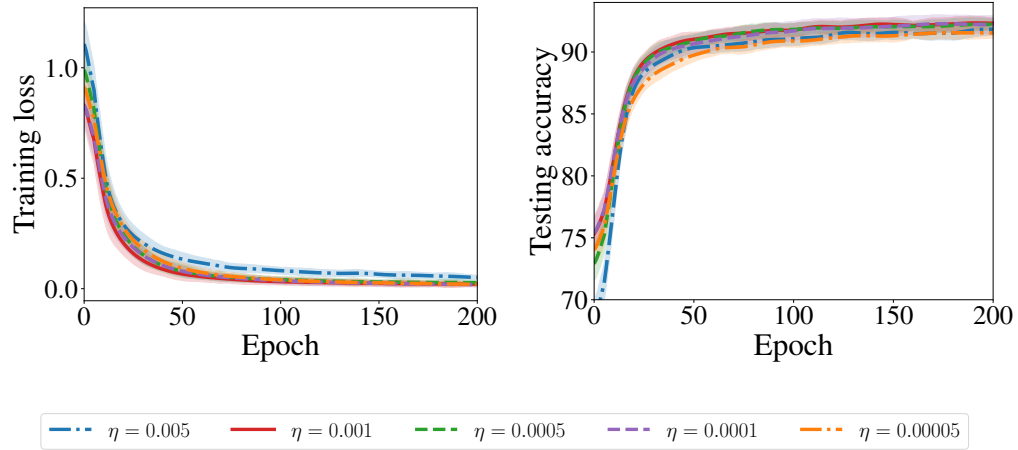


Figure 3: Results for CIFAR-10 dataset with different learning rates.

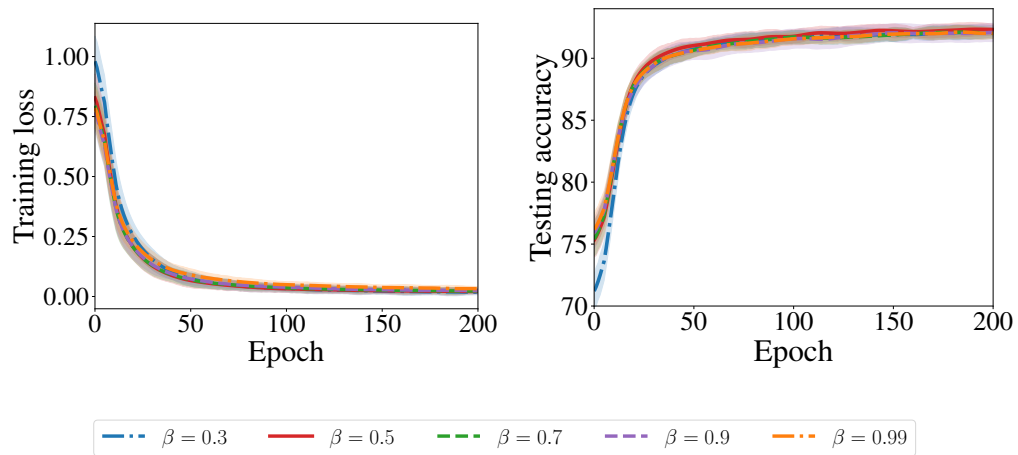


Figure 4: Results for CIFAR-10 dataset with different β .

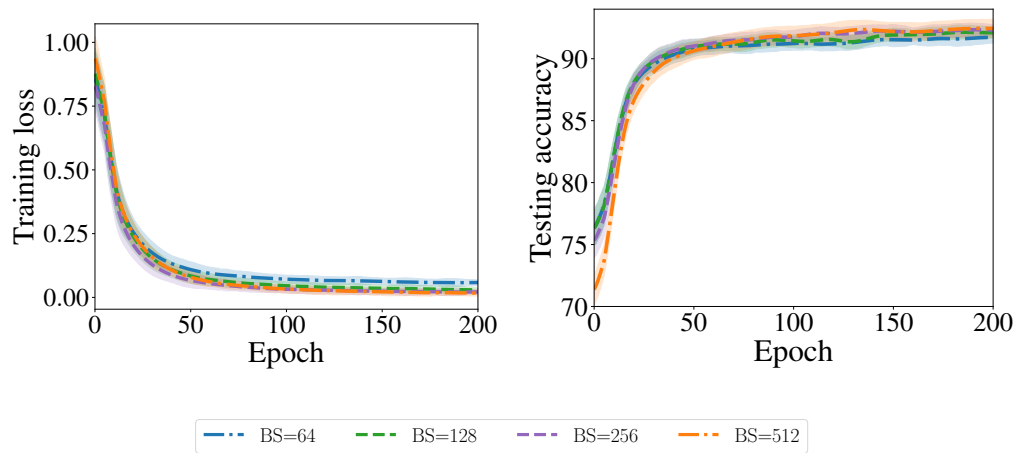


Figure 5: Results for CIFAR-10 dataset with different batch sizes.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The claims presented in the abstract and introduction accurately represent the contributions and scope of the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The theoretical results demonstrated in the paper rely on specific assumptions, which have been clearly stated in the main text.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper provides assumptions and proofs for each theoretical result.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper discloses the information necessary to reproduce the main experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Due to privacy concerns and ongoing research, we do not include the code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper describes the training and testing details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We have provided the relevant information.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics [https://neurips.cc/public/EthicsGuidelines?](https://neurips.cc/public/EthicsGuidelines)

Answer: [Yes]

Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This is primarily a theoretical paper with no potential negative social impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The creators or original owners of assets used in the paper are properly credited and the license and terms of use explicitly are properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.